# Evaluating Text-to-Speech (TTS) Synthesis for use in Computer-Assisted Language Learning (CALL)

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Humanities

**2005**

**Zöe L. Handley**

**School of Informatics**



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

# Table of contents

# List of figures

# List of tables

# List of abbreviations

| | |
|---|---|
| API | Application Interface |
| *Appeal* | *A Pleasant Personal Environment for Adaptive Learning* |
| ASR | Automatic Speech Recognition |
| AV | Audio-Visual |
| CALL | Computer-Assisted Language Learning |
| CEF | Common European Framework |
| CP | Conversational Partner |
| CT | Caretaker Talk |
| CTS | Concept-to-Speech |
| DV | Dependent Variable |
| EAGLES | Expert Advisory Group on Language Engineering Standards |
| EFL | English as a Foreign Language |
| ELSE | Evaluation in Language and Speech Engineering |
| ESL | English as a Second Language |
| F0 | fundamental frequency |
| FT | Foreigner Talk |
| HCI | Human-Computer Interaction |
| *HD TTS* | *High Density Text-to-Speech* |
| *HQ TTS* | *High Quality Text-to-Speech* |
| IPA | International Phonetic Alphabet |
| ISO | International Organization for Standardization |
| ITU-T | International Telecommunication Union Telecommunication Standardisation |
| IV | Independent Variable |
| JEIDA | Japanese Electronic Industry Development Association |
| KTH | Kungliga Tekniska Högskolan (Royal Institute of Technology) |
| L1 | first language |
| L2 | second language |
| LL&T | Language Learning and Teaching |
| LTS | Letter-to-Sound |
| *MOS-CALL* | *Mean Opinion Score for CALL* |

| | |
|---|---|
| *MOS-X* | *Mean Opinion Score Expanded* |
| MT | Mother Tongue |
| NLP | Natural Language Processing |
| phonetic PM | Pronunciation Model at the phonetic level |
| PM | Pronunciation model |
| POS | part of speech |
| prosodic PM | Pronunciation Model at the prosodic level |
| PTS | Phoneme-to-Speech |
| RM | Reading Machine |
| S1 | TTS synthesis system 1 |
| S2 | TTS synthesis system 2 |
| S3 | TTS synthesis system 3 |
| S4 | TTS synthesis system 4 |
| S5 | TTS synthesis system 5 |
| S6 | TTS synthesis system 6 |
| *SAFRAN* | *Système d'Apprentissage du FRANçais* |
| SALT | Speech And Language Technology |
| *SAM* | *Speech Assessment Methods* |
| SLA | Second Language Acquisition |
| SNOR | Standard Normalised Orthographic Representation |
| SR | Speech Rate |
| STM | Short Term Memory |
| TL | Target Language |
| TT | Teacher Talk |
| TTP | Text-to-Phoneme |
| TTS | Text-to-Speech |
| TWP | Talking Word Processor |
| USS | Unit Selection Synthesis |
| VLC | Virtual Learning Center |
| W3C | World Wide Web Consortium |
| WTC | Willingness To Communicate |

# Abstract

Despite the fact that Text-to-Speech (TTS) synthesis has the potential to bring a number of new possibilities to Computer-Assisted Language Learning (CALL), it has not yet made an impact on CALL. It is believed that this is because it has not been adequately evaluated for the purposes. With the aim of validating this claim, an infrastructure for the evaluation of CALL applications integrating TTS synthesis is put forward and evaluations conducted to date are assessed with respect to it. This analysis indicates that TTS synthesis has indeed not been adequately evaluated for the purposes of CALL, specifically, that an important stage in the process of evaluation has been omitted in all evaluations that have been conducted to date, namely requirements analysis. With the aim of developing a benchmark test for the evaluation of the adequacy of TTS synthesis systems for use in CALL applications, this thesis looks to SLA research for indications of what the requirements of TTS synthesis for use in CALL might be. This literature review suggests that CALL applications place demands on the quantity, quality and flexibility of the speech generated by TTS synthesis systems. Regarding the demands that it is suggested that CALL applications place on the quality of the speech generated, in order to validate these requirements, two investigations are carried out. The results of these investigations, which also attempt to determine whether the different roles that TTS synthesis systems may assume in CALL applications impose different requirements on the quality of the speech generated, suggest that, as suggested by the SLA literature, CALL applications do place demands on the comprehensibility, accuracy and naturalness of the speech generated by TTS synthesis, but that, in addition, they also place demands on intelligibility, choice of pronunciation, naturalness of voice, expressiveness and register and that the different roles do indeed place different demands on the quality of the speech generated by TTS synthesis systems, but that teachers and CALL researchers have difficulty differentiating between the different roles and their requirements. It is believed that the results of these investigations imply that evaluations of the adequacy of TTS synthesis systems for use in CALL applications ought to address all of the aspects of the quality of the speech generated by the TTS synthesis systems mentioned above. Regarding the different roles that TTS synthesis systems may assume within CALL applications, it is believed that they imply that, while the different roles do impose different demands, it will not be possible to ask participants to differentiate between these roles and their requirements. Rather, participants can only be asked to rate the quality of the speech generated by TTS synthesis systems for use in CALL applications in general.

## Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

## Copyright statement

# Acknowledgements

First, I would like to thank my first supervisor Dr. Marie-Josée Hamel for giving me the opportunity to work on this project. She was a real source of encouragement and motivation. I would also like to thank Prof. Harold Somers who took over from Marie-Josée when she left UMIST to work in Canada for carefully reading through the final versions of each of the chapters. In addition, I would also like to thank Dr. Blaise Nkwenti-Azeh and Prof. Paul Blenkhorn who served as supervisors whilst Marie-Josée was on maternity leave.

I am also indebted to all the French teachers and CALL researchers who kindly gave up their time to participate in my investigations and the Grundy Educational Trust and the former Centre for Computational Linguistics at UMIST who supported me financially throughout the completion of this study.

Finally, I would like to thank all my family and friends who encouraged and supported me through the ups and downs of the last four years. In particular, I would like to thank my mum, Rob, Lucy, Stepf and Alan.

# The Author

In 2001, Zöe Handley gained a first class honours degree from the Centre for Computational Linguistics, UMIST, Manchester in French Language Technology (Applied French) BSc. She began her research career during her year abroad in the Natural Language team at France Telecom R&D (formerly Centre National d'Etudes des Télécommunications (CNET)), Lannion, France. Her main responsibility whilst working in the Natural Language team at France Telecom R&D was to increase the coverage of the team's shallow parser for English. Before commencing the research reported in this thesis she assisted in the design of *ICT 3c: An Introduction to Speech Technology in Language Learning*, one of the modules of the on-line course *International Modules in Information and Communication Technology for Language Learning*.[1] Her contribution to this project was a review of the use of speech technology in CALL. In September 2001, she began working towards a PhD on the use of Text-to-Speech (TTS) synthesis in Computer-Assisted Language Learning (CALL) in the Centre for Computational Linguistics at UMIST. This work has been presented at a number of conferences as well as in a peer reviewed journal (see below). In addition, she worked as a research assistant on the EU project FreeText and was involved in editing the proceedings of the 1999 and 2000 Integrating Speech Technology in (Language) Learning (InSTIL) conferences for republication.

Publications in peer reviewed journals

Handley, Z and Hamel, M.-J. (2005). Establishing a Methodology for Benchmarking Speech Synthesis for Computer-Assisted Language Learning (CALL). *Language Learning and Technology Journal*. 9 (3): 99-119.

Conference presentations

Handley, Z and Hamel, M.-J. (2004). Investigating the Requirements of Speech Synthesis for CALL with a View to Developing a Benchmark. In *Procs. InSTIL/ICALL 2004* (pp. 71-74). Venice, Italy.

Handley, Z. and Hamel, M.-J. (2002). Text-to-Speech (TTS) Synthesis in CALL: Developing an Evaluation Methodology to Determine the Suitability of TTS Output for Integration in CALL Applications. In *Procs. EUROCALL 2002* (p. 66). Jvaskyla, Finland.

[1] http://www.ilo.uva.nl/Ontwikkeling/imictll/index.html

Note:
All URLs and e-mail addresses, unless stated otherwise, were still available on 28th September 2005.

# 1 Introduction

## 1.1 Motivation

> Writing must be taught, whereas spoken language is acquired automatically. All children naturally learn to speak the language of the community in which they are brought up. They acquire the basics of their native language before they enter school, and even if they never attend school they become fully competent speakers (Jannedy *et al.*, 1994: 6).

In other words, as Trask (1995: 2) notes: "For most people, most of the time, the ordinary medium of language is speech". The importance of speech as a medium of communication is emphasised by a number of other researchers including the phoneticians Sweet (1877) and Abercrombie (1967), and the Second Language Acquisition (SLA) researcher Brown (1977).

Learners who wish to acquire general proficiency, i.e. reading, writing, listening and speaking skills, in a second or foreign language would therefore, as Sweet (1877) recommended,[2] benefit from a programme of study which privileges the spoken medium, i.e. the development of listening and speaking skills.

Moreover, some teachers and researchers believe that the written word should not be introduced to second language (L2) learners until they have acquired a certain level of proficiency in speaking, and pronunciation in particular, because it is believed that exposure to the written word leads to the development of bad pronunciation habits (Dabène, 1974; Dansereau, 1995). Specifically, it has been observed that learners who are exposed to the written word too early have a tendency to acquire what Dansereau terms a 'reading' pronunciation, that is pronunciation characterised by mispronunciations introduced through the application of first language (L1) grapheme-to-phoneme (letter-to-sound) translation rules to the Target Language (TL), and the incorrect application of TL grapheme-to-phoneme (letter-to-sound) translation rules.

---

[2] Sweet (1877), in fact went so far as to say that phonetics, the scientific study of speech, should form the basis of all language study. We do not go this far here because it is possible to teach pronunciation without recourse to phonetics (Brown, 1992).

Regarding the development of written language skills, when the spoken language is taught before the written language, written language development, in particular spelling, is typically approached through the introduction of phoneme-grapheme (sound-letter) correspondences (Dabène, 1974). Awareness of phoneme-grapheme and grapheme-phoneme relationships is in fact important when learners engage in real world tasks. Real world tasks require the use of different language skills in integration (Warschauer and Healey, 1998) and hence the ability to switch between spoken and written language with ease (Groschel, 1979). In order to switch between spoken and written language with ease, learners need to master the phoneme-to-grapheme and grapheme-to-phoneme relationships of the TL (*ibid.*).

Consequently, given that it is known that "Learning increases with time spent on task" (Massaro *et al.*, 1999: 1291) and "spacing practice over a longer time leads to better learning than massing practice within a shorter time" (*loc. cit.*), there is a need for practice in the use of the spoken language and its relationship to written language outside the classroom. One way in which this can be provided is through Computer-Assisted Language Learning (CALL).

That TL input is essential for SLA is undisputed (Gass, 1997). Nevertheless, while it has long been possible to provide learners with spoken TL input in CALL applications this possibility has been under-utilised (Pennington and Esling, 1996; Sobkowiak, 1998).

One way in which the first computers could 'talk' was through a computer-controlled cassette-recorder (Ahmad *et al.*, 1985). Coupling the computer with cassette-recorder offered one advantage over the conventional audio-lingual language laboratory: it made it possible to produce branching programs (*ibid.*). The use of computer-controlled cassette-recorders, however, suffered from the same limitations as the conventional language laboratory, namely that "Cassette-recorders are a slow-speed device for storing and retrieving data" (*ibid.*: 131), and cassette tape wears out. Another way in which early computers could 'talk' was through special-purpose circuit boards which permitted the digital recording (i.e. as a long string of numbers) and subsequent playback of speech. Digitised speech overcame two of the limitations of cassette-recordings: retrieval times were shorter because speech samples could be randomly accessed (*ibid.*), and digitised speech is more durable than cassette-recordings. However, while the quality of digitised speech was satisfactory – "similar to that found on a cassette-recorder" (*ibid.*: 132), the amount of speech that could be stored and subsequently

retrieved was very limited: for example, it was only possible to store a few minutes of speech on a floppy disc (*ibid.*), the main storage medium used at the time. Since then, as predicted by Ahmad *et al.*, the storage capacity of the computer has increased considerably making the use of digitised speech and other media, such as (digitised) video and computer animation, in CALL more viable. It has also become possible to link various media: for example, a word or a picture might have a link to a sound file giving its pronunciation.[3] As a result,

> More and more of the computer-aided foreign-language instruction proceeds in the medium of speech ... Teaching techniques crucially dependent on listening, like dictation, repetition or sound-letter matching, are now available to teachers using CALL for EFL [English as a Foreign Language] (Sobkowiak, 1998: 27).

Another advance which has permitted the provision of more spoken TL input to learners is Computer Mediated Communication (CMC). Specifically, technologies including audio chat, and audio and video conferencing permits learners to communicate orally with native speakers of the language that they are learning who may be situated anywhere in the World.

Despite these advances, we believe that the provision of spoken TL input to learners in CALL is still limited.

Text-to-Speech (TTS) synthesis, the automatic generation of speech from text, is another technology which might be used to provide learners with spoken TL input in CALL applications. Yet despite the fact that some of its potential benefits were identified more than twenty years ago (Sherwood, 1981) and that it is believed to be more robust than Automatic Speech Recognition (ASR), "the process of converting an acoustic signal [i.e. speech], captured by a microphone, or telephone, to a set of words" (Zue *et al.*, 1996:4), which is now widely used in CALL, TTS synthesis has as yet not made an impact on CALL (Sobkowiak, 2005).

The most likely reasons for the neglect of TTS synthesis in CALL, we believe, is that the technology is not proven, i.e. has not been adequately evaluated for the purposes of CALL – since the failure of the heralded but unsubstantiated language laboratory (Pederson, 1987; Doughty, 1990), teachers have been sceptical about unproven, i.e. unevaluated, technologies

[3] Linked media of this type is commonly referred to as *hypermedia*.

(Dunkel, 1990). In this thesis, we therefore attempt to determine whether TTS synthesis has been adequately evaluated for the purposes of CALL.

## 1.2 Aims and objectives

As said, one of the aims of this thesis is to determine whether TTS synthesis has been adequately evaluated for the purposes of CALL. In order to achieve this aim our objectives are to establish:

(1) an infrastructure for the evaluation of TTS synthesis for CALL purposes; and,

(2) what stage in this infrastructure has been reached so far.

On the basis of our findings, our second aim is to propose an agenda for the further evaluation of TTS synthesis for CALL purposes.

One of the first and an essential stage in the process of evaluation is requirements analysis (EAGLES, 1999; ISO, 1999). As far as it was possible to establish from the literature the requirements that CALL applications impose on TTS synthesis systems have not been investigated. The third aim of this thesis is therefore to establish what requirements CALL applications impose on TTS synthesis systems. Specifically, our objectives are:

(1) to identify which aspects of TTS synthesis systems SLA models and best practice suggest that CALL applications place demands on;

(2) to validate those requirements; and,

(3) to determine whether those requirements differ across the different uses of TTS synthesis in CALL, specifically the different roles that it might assume within CALL applications.

As a by-product it was hoped that we would also be able to establish the readiness of TTS synthesis for use in CALL applications.

## 1.3 Context

The research presented here began within the EU project *FreeText* (Hamel, 2003b), the goal of which was the production of a web-based multimedia CALL system for French featuring Natural Language Processing (NLP) tools, one of which was a TTS synthesis system (the uses of TTS synthesis within *FreeText* are discussed in chapter 3), for a smart treatment of authentic documents and free production exercises. The investigations presented in this thesis

therefore look at the use of French TTS synthesis for teaching French as a foreign or second language (L2).

## 1.4 Structure of the thesis

The remainder of the thesis is structured as follows. In chapter 2, TTS synthesis is introduced. Specifically, before focusing on the different types of speech synthesis (see section 2.3) and TTS synthesis more specifically (see section 2.4), speech synthesis in general is considered in more detail (see section 2.2). Regarding TTS synthesis, the different designs, or architectures, that have been proposed for TTS synthesis systems are presented in 2.4. The implications of these designs on the quality and the flexibility of the speech generated by TTS synthesis systems are then presented in sections 2.5.1 and 2.5.2 respectively. Then, in section 2.5.3, the computational demands of the different designs of TTS synthesis systems are presented. Integration issues are discussed in section 2.5.4. And, finally, in section 2.5.4, current applications in TTS synthesis are presented with a view to permitting a comparison between suggested CALL applications and applications in which TTS synthesis is already used.

Chapter 3 looks at the use of TTS synthesis in CALL. First, in section 3.2, the suggested benefits of TTS synthesis are discussed. Then, in section 3.3, actual and suggested uses of TTS synthesis in CALL are presented. Finally in section 3.4, dimensions of CALL applications and the settings in which they are used which it is believed might have an effect on the requirements placed on TTS synthesis and hence implications for the evaluation of TTS synthesis for CALL purposes are presented and discussed.

The focus of chapter 4 is evaluation. Having considered very generally what evaluation is (see section 2.2) and why people conduct evaluations (see section 3.2), best practice in evaluation is discussed (see sections 4.4, 4.5 and 4.6). More specifically, in section 4.4, the different stages in designing and conducting an evaluation are presented, in section 4.5, the levels at which CALL applications integrating TTS synthesis ought to be evaluated are considered, and in section 4.6, the features of good methods of evaluation are discussed. Then, in section 4.7, the evaluations of CALL software integrating TTS synthesis that have been conducted to date are assessed with respect to best practice in evaluation. Section 4.8 looks at reasons why evaluation might be neglected in this context. And then, in section 4.9, a potential solution to the most likely of these reasons is put forward. Finally, on the basis of the findings of this chapter, an agenda for further evaluation is proposed (see section 4.10).

As already mentioned in section 1.2, requirements analysis is an essential stage in the process of evaluation which it would appear has not been conducted in this context. CALL draws on a number of different fields of research (see section 5.1) all of which may suggest requirements of TTS synthesis for use in CALL applications. Of these fields it is believed that SLA ought to be the primary consideration (also see section 5.1). In chapter 5, SLA research is therefore reviewed. Specifically, in section 5.2, we look at the goal of language learning, namely communicative competence, and its implications regarding the requirements that CALL applications impose on TTS synthesis systems. In section 5.3, we look at a model of SLA, specifically one that is believed to unify the main views on SLA, namely Gass's (1997) Interactionist Model, and its implications regarding the requirements that CALL applications impose on TTS synthesis systems. Finally, the results identified as a result of these reviews are summarised in section 5.4.

In chapter 6, first, two investigations (see sections 6.2 and 6.3) which look into some of the requirements identified in section 5 in more detail are presented. Then, on the basis of the results of these investigations, recommendations are made regarding: the evaluation of TTS synthesis for CALL purposes (see section 6.4.1) the use of TTS synthesis systems in CALL applications in their current state (see section 6.4.2).

Conclusions are drawn in chapter 7.

# 2 TTS synthesis

## 2.1 Overview

TTS synthesis systems are a specific type of speech synthesis system, which automatically generate speech from text. Contact with language teachers, CALL researchers and publishers of CALL software during the completion of this study revealed that speech synthesis is often confused with other technologies which enable the user to interact with the computer using spoken language, in particular ASR (see section 1.1). Before focusing on the different types of speech synthesis (see section 2.3) and TTS synthesis more specifically (see sections 2.3.1 and 2.4), it is therefore appropriate to consider what speech synthesis is in more detail (see section 2.2).

The suitability of TTS synthesis systems for use in different setups, or contexts of use (Sparck Jones and Galliers, 1996), such as CALL, depends on both the quality and flexibility of their output and the ease with which they can be integrated into those setups. These factors in turn depend on the design, or architecture, of the TTS synthesis system. Several different TTS architectures have been proposed, and these are presented in section 2.4. The implications of these designs on the quality and the flexibility of the output of TTS synthesisers are presented in sections 2.5.1 and 2.5.2 respectively. The computational demands of different designs of TTS synthesis systems are presented in section 2.5.3. Integration issues are discussed in section 2.5.4. Finally, in section 2.5.4, current applications in TTS synthesis are presented with a view to permitting a comparison between suggested CALL applications and applications in which TTS synthesis is already used.

## 2.2 What is speech synthesis?

As already mentioned in the overview, speech synthesis systems, or speech synthesisers, are computer programs which automatically generate speech, i.e. systems which enable the computer to 'talk' or 'speak' to the user.

Due to the fact that the output of speech synthesis systems is termed 'synthetic speech', some (Docherty and Shockey, 1988; van Bezooijen and van Heuven, 1997; Bickley *et al.*, 1999; Pitt and Edwards, 2003) consider the following technologies also to be examples of speech synthesis: the digital recording of speech, speech coding (i.e. the compression of digitally

recorded speech to reduce storage and transmission requirements), and speech editing (i.e. the manipulation of digitally recorded speech along one or more dimensions including, but not limited to, volume and pitch). The output of these processes is indeed synthetic speech in the sense that it is not natural, i.e. is artificial (see definitions of 'synthetic' and 'synthesis' below). These processes are, however, *not* considered to be speech synthesis in this thesis. In this thesis, rather, speech synthesis is restricted to:

> the generation of novel [oral] messages, either from scratch (i.e. entirely by rule) or by recombining shorter pre-stored units (van Bezooijen and van Heuven, 1997: 481).

In other words systems that produce speech which is synthetic in both senses of the term:

> *adj* **synthetic** (*-thet'*) or **synthet'ical** relating to, consisting in, or formed by, synthesis; artificially produced but of similar nature to, not a mere substitute for, the natural product (Chambers, 1998: 1679)
>
> **synthesis** *sin'thi-sis, n* building up; putting together; making a whole out of parts" (*loc. cit.*)

Benefits of the use of both speech coding and speech editing for CALL purposes have been proposed. Before moving on to consider speech synthesis in relation to the other speech technologies, these benefits are briefly presented along with the reasons for their exclusion from this study.

As previously stated, speech coding decreases the storage requirements of speech samples. Consequently, more examples can be stored for subsequent retrieval and presentation thus increasing the amount of input that can be provided to the learners. Exposure to large quantities of TL input is believed by many to be a prerequisite for SLA (Krashen, 1982; Gass, 1997; see also sections 5.3 and 5.3.1). The ability to provide learners with more input is merely an improvement on what is possible with other technologies. It does not bring any new possibilities, i.e. add value, to CALL. Moreover, as the storage capacity of computers is continually and rapidly increasing, the use of speech coding is likely to bring only short-term benefits. TTS synthesis, on the other hand, as we shall see in chapter 3, has more to offer to CALL than speech coding.

With respect to the use of speech editors in CALL, at least one speech editor has been developed specifically for the purposes of Language Learning and Teaching (LL&T), namely *WinPitch LTL* (Language Teaching and Learning) (Germain-Rutherford and Martin, 2000; 2001), and at least one other has been equipped with tools for the specific purposes of LL&T, namely *WinSnoori* (Bonneau *et al.*, 2000; 2004). The use of speech editors in LL&T is interesting because they provide opportunities to promote apperception, or noticing, of features of the TL, a process that is believed to be necessary for SLA to take place (Gass, 1997; see also sections 5.3 and 5.3.2). For example, they can be used to provide learners with a range of different types of modified input (see section 3.2) which may make features of the TL more salient and hence promote apperception. The types of modified input that they can produce include, but are not limited to: examples with globally or locally modified (in particular slowed) Speech Rate (SR), examples with globally or locally modified (in particular exaggerated) intonation, examples with enhanced phonemes (in particular stop bursts and fricatives) (Bonneau, *et al.*, 2000), 'synthetic humming', in other words examples from which the segmental content has been removed (Yoram and Hirose, 1996), and examples with "intonation, but no rhythm" (Keller and Zellner-Keller, 2000: 110; Keller, 2002: 7). In addition, speech editors can be used to provide learners with a particular type of feedback which may promote apperception. Specifically, speech editors can be used to manipulate learner's incorrect productions and subsequently provide them with feedback consisting in examples of themselves correctly producing the TL (Nagano and Ozawa, 1990; Yoram and Hirose, 1996; Bonneau *et al.*, 2000). It is believed that in such feedback, errors are more salient to the learner, and therefore more likely to be apperceived, because distractions caused by the particular characteristics of another person's voice are eliminated (Yoram and Hirose, 1996).

The use of speech editors, however, requires the ability to interpret and manipulate complex visual displays of speech, namely waveforms and spectrograms, an ability which in general only trained phoneticians (Pennington and Esling, 1996) and speech engineers possess. "teaching students and teachers what these displays mean might take longer than the pedagogical potential their use might warrant" (Komissarchick and Komissarchick, 2000: 86). It is therefore argued that speech editors are not currently suitable for CALL.[4] In this thesis,

[4] If visual displays were adapted to make them easier for learners, and teachers for that matter, to interpret as was done in the Spell project (Hiller *et al.*, 1994), it is acknowledged that they

speech editors will therefore only be considered when used in combination with TTS synthesis.

To sum up, in this thesis, speech synthesis refers only to systems that permit the generation of novel unrestricted oral messages.

## 2.3 A broad classification of speech synthesisers

Speech synthesis systems can be classified along a number of axes. Treating the systems as a 'black-box' they can be classified according to:

- the type of input that they support, and,
- the type of output that they produce.

### 2.3.1 A classification according to input type

With respect to the type of input that they support, two classes of speech synthesis system are distinguished, namely:

- TTS systems, and,
- Concept-to-Speech (CTS) systems.

TTS systems are driven by text input (Bhaskararao, 1994; van Bezooijen and van Heuven, 1997; d'Alessandro, 2001; Bailly, 2002a). In other words, they mimic the human process of reading in some sense (Tatham, 1993; Pfister and Traber, 1994).

CTS synthesis systems, also referred to as message-to-speech systems (Sproat, 1996; Theune, 2000), are driven by concepts (Rodman, 1999; Benoît *et al.*, 2000; Bailly, 2002a), in other words, a semantic representation (Benoît *et al.*, 2000) of the utterances to be pronounced. The following is a simplified example of what the input to a CTS synthesis system for the generation of the response to the query *How far is it from New York to Los Angeles?*:

> ASSERT(STATE(MEASURE(DISTANCE(NEW YORK, LOS ANGELES)
> (Rodman, 1999: 207).

---

would be a very useful tool in LL&T. In the Spell project, in order to enable learners to distinguish acceptable deviations in pitch from unacceptable ones, pitch contour displays were adapted to present acceptable pitch ranges through the use of pitch tunnels. Similarly, in order to enable learners to distinguish acceptable deviations from model vowels, plots of the values of the first and second formants were adapted to present acceptable deviations from models through the use of vowel targets.

From this input the CTS synthesis system may generate the following output: "*The distance between New York and Los Angeles is two thousand, four hundred and sixty one miles*" (*loc. cit.*). A CTS synthesis system must generate the utterances to be pronounced itself.

While CTS synthesis systems are, on the one hand, more complex than TTS synthesis systems because they must themselves generate the utterances that are to be pronounced, CTS systems should be able to generate higher quality output than TTS synthesis systems because, to a certain extent, they should 'know' and 'understand' what they are saying (Witten, 1982; Allen, 1992; Sproat, 1996; van Bezooijen and van Heuven, 1997; Huang *et al.*, 2001). For example, a CTS synthesis system should know the syntactic and semantic structures and the communicative purpose of the utterances to be generated (Allen, 1992; van Bezooijen and van Heuven, 1997; Rodman, 1999)[5]. Producing high quality output from text on the other hand is more difficult because this information must be derived from text which is an inadequate representation of spoken language.

While it has been suggested by some in the field of CALL that CTS synthesis systems might be more appropriate for use in CALL because they ought to generate higher quality speech than TTS synthesis, in particular speech that is less monotonous and has more human-like prosody (Thomas *et al.*, 2004), TTS synthesis is more frequently (re-)used for CALL purposes. TTS synthesis is hence the focus of this thesis.

An example of a CTS synthesis system developed for use in CALL is *VINCI* (*ibid.*). According to the developers, this system could be used within the CALL context for the presentation of dictation exercises, pronunciation and auditory discrimination exercises focusing on intonation, and as a tool for teaching phonetic transcription (*ibid.*). Uses of TTS synthesis for CALL purposes are presented in chapter 3.

[5] It should be noted, however, that many state-of-the-art CTS synthesis systems do not produce higher quality output than TTS synthesis systems. Many rely heavily on the use of templates, 'carrier sentences', such as the following, which contain slots which can be filled with real information: *Flight [flight_no] is scheduled to land at {sch_time]* (Huang *et al.*, 2001). Others convert the input to text which they then pass through a TTS synthesiser (Benoît *et al.*, 2000).

### 2.3.2 A classification according to output type

Although human speech is bimodal (Beskow, 1996; Huang *et al.*, 2001), consisting in both audio and visual cues (lip, tongue, jaw, movements (Schomaker *et al.*, 1995) eyebrow movements, nodding (Beskow *et al.*, 2000) etc.), most speech synthesis systems generate audio output only. Such systems are simply termed 'speech synthesisers'.

However, a smaller number of speech synthesis systems, through the use of animation techniques or through the concatenation of segments of videos of a speaker (Bailly *et al.*, 2003a), generate visual cues as well as audio cues. Such systems are referred to as Audio-Visual (AV) speech synthesis systems, or 'talking heads'. With respect to input, most AV synthesis systems are driven by text and are hence referred to as text-to-audio-visual speech synthesis systems.

Regarding the benefits of the use of AV synthesis, the auditory and visual modes of speech are complementary:

> For example, the difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually (Massaro and Cole, 2000: 157).

Moreover, the visual mode supplements the auditory mode:

> Both verbal and visual signals are used in conversation to signal feedback and turntaking. Verbal signals can complement syntax and interact with the prosodic (accentual and phrasal) structure of utterance. For example, a phrase-final intonation pattern can function as both a cue for prosodic grouping and as a verbal turngiving signal. Visual cues such as eyebrow movements and nodding for accentuation can function as parallel signals to intonation (i.e. as linguistic signals) as well as being used as conversational signals (e.g. raised eyebrows to signify an interested, listening agent, or nodding to provide encouragement).
>
> Fundamental frequency [f0] and duration are the primary auditory cues used for signaling prominence, grouping and feedback. Visual cues for emphasizing stress placement and phrasing include blinking and changes of gaze, eyebrow raising, frowning, head nodding and head turning (Beskow *et al.*, 2000: 138).

Consequently, when the audio and visual cues are coherent, i.e. in synchrony (Le Goff and Benoît, 1996; Bailly *et al.*, 2003a), AV synthesis is more robust, i.e. more intelligible and comprehensible in adverse, noisy conditions and to the hearing impaired, than simple audio

speech synthesis (Beskow, 1996; Le Goff and Benoît, 1996; Benoît and Le Goff, 1998; Massaro *et al.*, 1999; Beskow *et al.*, 2000; Huang *et al.*, 2001; Massaro and Cole, 2000). 'Intelligibility' is used here, as it is generally in speech synthesis circles, to refer to the ease with which a listener can recognise individual speech sounds and words (Francis and Nusbaum, 1999);[6] 'Comprehensibility' is used to refer to the ease with which a listener can understand a speaker's intended message (*ibid.*).

With respect to language learning, non-native speakers rely more heavily on the 'extra' information provided by the visual mode than native speakers (Beskow *et al.*, 2000). The use of AV synthesis in CALL is therefore particularly attractive.

Innovative functionalities also make the use of AV synthesis in CALL all the more attractive. Such functionalities include rendering the talking head see-through (*ibid.*), pivoting the head, and exaggerating the articulation of individual segments (Massaro and Cole, 2000). Rendering the head see-through enables learners to view articulatory behaviour that would otherwise be hidden behind the lips and/or cheeks (Beskow *et al.*, 2000). Pivoting the head enables learners to view articulatory behaviour from any orientation, in particular from behind:

> It is possible that this back-of-head view would be much more conducive to learning language production. The tongue in this view moves away from and towards the student in the same ways as the student's own tongue would move (Massaro and Cole, 2000: 159)

The exaggeration of the articulation of segments etc. should increase the saliency of their articulatory features and promote apperception, which, as mentioned above, is believed to be necessary for acquisition to take place (Gass, 1997; see sections 5.3 and 5.3.2).

While, as we have seen, AV synthesis has a number of advantages over simple audio synthesis, AV synthesis will not be considered further here. Rather, simple audio TTS synthesis will be the focus of this thesis, for it is more widely available and easier to use. Free demonstrations of TTS synthesisers are available on the Internet for use by teachers and learners. AV synthesis systems, on the other hand, typically need to be downloaded and can be complex to install.

---

[6] As we shall see in section 5.2, in LL&T the term intelligibility is used to refer to what we refer to as comprehensibility.

It is however believed that the research presented here will also have implications for the design and evaluation of AV synthesis for CALL purposes. Specifically, it is believed that AV synthesis and TTS synthesis for CALL purposes will have many common requirements or needs. Generally, AV synthesis has been evaluated in the same way as TTS synthesis (Le Goff and Benoît, 1996; Le Goff *et al.*, 1997; Benoît and Le Goff, 1998). It is therefore also believed that it will be possible to evaluate the satisfaction of many of the requirements of AV synthesis for CALL purposes in the same way that they are with TTS synthesis.

To sum up, this thesis focuses on speech synthesis systems that produce speech only from text.

TTS synthesis systems themselves are further classified according their design or internal architecture. This is the subject of the next section.

## 2.4 The architecture of TTS synthesisers

If writing was a faithful and adequate representation of spoken language, TTS synthesis systems would have a very simple architecture: they would consist in a simple set of rules for converting letters and symbols to speech sounds and for assigning prosodic specifications such as phrase boundaries, stress and intonation. Despite the fact, as we shall see in section 2.4.1, that text is not an adequate representation of speech even in languages whose writing system is more systematic than that of English or French (Bailly *et al.*, 2003b), the first TTS synthesisers carried out a very limited linguistic analysis of input texts (Ainsworth, 1973; Allen, 1973; Coker *et al.*, 1973).

Today, on the other hand, the importance of analysing input text to determine what the synthesiser is to say is recognised. As a result, today's synthesisers typically consist of two modules (Klatt, 1987; Pfister and Traber, 1994; Edgington *et al.*, 1996a; 1996b), namely:

- a Text-to-Phoneme (TTP) module, and
- a Phoneme-to-Speech (PTS) module.

The goal of the former, which is also referred to as the 'text analysis module' (Allen, 1992; Edgington *et al.*, 1996a; 1996b) or the 'NLP module' (Dutoit, 1997), is to determine what the synthesiser is to say (van Santen, 1993; Rodman, 1999), in other words to generate an unambiguous narrow phonetic transcription augmented with prosodic specifications (phrasing,

stress, rhythm, and intonation) of the utterances to be pronounced from the input text (Klatt, 1987; Allen, 1992; Dutoit, 1997).[7] The goal of the latter, which is referred to as the 'speech synthesis module' (Allen, 1992; Huang *et al.*, 2001), the 'speech generation module' (Edginton *et al.*, 1996a; 1996b), or the 'Digital Signal Processing' module (Dutoit, 1997), is to produce those phonemes with the appropriate prosody, i.e. to vocalise the utterances (van Santen, 1993; Rodman, 1999).

Several alternative designs are possible for these two modules. They are discussed in sections 2.5.1 and 2.4.2 respectively.

### 2.4.1 The TTP module

As stated, the goal of the TTP module is to determine what the TTS synthesiser is to say. In order to achieve this goal, a number of different levels of linguistic processing are carried out each by a dedicated sub-module. In the following sections, the roles of the sub-modules most frequently found in the TTP module are presented followed by a discussion of the various ways in which they could be implemented. Other modules, not presented here, that have been suggested include a mark-up interpreter (Huang *et al.*, 2001; Monaghan, 2002a), a module for the detection of document structure (headings, lists, etc.) (Huang *et al.*, 2001), a language identification module (Rodman, 1999), and a module dedicated to homograph disambiguation (Yarowsky, 1997).

#### 2.4.1.1 The Letter-to-Sound (LTS) conversion module

Of all the potential sub-modules, the LTS conversion module, also termed the 'grapheme-to-phoneme' module (Pfister and Traber, 1994; Dutoit, 1997), was traditionally the one essential component of the TTP module because it was the main module responsible for generating the phonetic transcription of input text. Traditional TTS systems only resorted to a lexicon for the provision of the phonetic transcription of tokens whose pronunciation could not be generated by this module (Liberman and Church, 1992; Edgington *et al.*, 1996a; Divay and Vitale, 1997; Huang *et al.*, 2001).[8] Today, on the other hand, phonetic transcriptions are provided in the

[7] In some systems, such as the TTS system developed at British Telecom (Edgington *et al.*, 1996a), prosodic specifications are not generated at this stage, rather they are generated by the PTS module.

[8] An exception is the MITalk system which relied on its lexicon for the provision of the phonetic transcription of the majority of words (Allen *et al.*, 1987; Allen, 1992; Liberman and Church, 1992)

lexicon, and TTS systems only resort to LTS rules to generate the phonetic transcription of out-of-vocabulary words (Edgington *et al.*, 1996a; Divay and Vitale, 1997; Huang *et al.*, 2001).

Typically, the LTS conversion module generates the phonetic transcription of words through the application of LTS conversion rules. In order to permit the generation of the phonetic transcription of unrestricted text, two sets of LTS rules are required. Specifically, in addition to a set of rules for the transcription of standard words, a set of letter-by-letter rules is required to provide the phonetic transcription of initials and 'initialisms'.[9] In addition, if a TTS system is intended to handle bilingual texts, the LTS conversion module should also incorporate a separate set of LTS rules for each language that the synthesiser is expected to handle (Rodman, 1999) because the grapheme-phoneme relationship differs across languages. Borrowed words are typically treated as exceptions by systems intended to handle monolingual texts and handled by the lexicon (Allen, 1992).

As mentioned in section 2.4.1, LTS rules alone are not sufficient to provide an accurate phonetic transcription of unrestricted text. In the following sections, further levels of analysis which may be carried out and the problems in TTP conversion that they are proposed to overcome are presented.

### 2.4.1.2 The text pre-processing module

The first level of analysis that is carried out before LTS rules are applied to a text is text pre-processing. Text pre-processing, also referred to as 'format analysis' (Allen, 1992), consists in three processes:

- segmentation,
- tokenisation, and,
- normalisation.

The primary aim of segmentation is to identify punctuation and split input text into paragraphs, sentences and phrases (*ibid.*; Edgington *et al.*, 1996a; Dutoit, 1997; Huang *et al.*, 2001). In general, punctuation marks coincide with the boundaries of intonational and

[9] Initialisms are short forms created from the initial letters of the words which constitute a longer expression which are read out letter by letter such as *UFO (Unidentified Flying Object)* (Cabré, 1998).

phonological phrases and hence pauses (Horne and Filipsson, 1997; Rodman, 1999). In addition, punctuation marks also encode the pragmatic function of utterances and hence the intonation of utterances because specific pragmatic functions are typically associated with specific intonation patterns.

Regarding paragraph boundaries,

> The pitch range of good readers or speakers in the first few clauses at the start of a new paragraph is typically substantially higher than that for mid-paragraph sentences, and it narrows further in the final few clauses, before resetting for the next paragraph (Huang *et al.*, 2001).

Segmentation therefore provides a lot of information that can be, and has been (see for example Klatt (1987) Dutoit (1997), Horne and Filipson (1997), Huang *et al.* (2001)), exploited to determine the prosodic specifications of utterances. Moreover, the information provided by segmentation is often crucial in determining the prosody of an utterance in French because syntax is often neutral, as in the following examples: *Vous ne dites rien. Vous ne dites rien ? Vous ne dites rien !* (Leon, 1992: 121).

The goal of tokenisation is to split the input text into tokens or lexical units (Sproat *et al.*, 2001). On the basis of the orthographic form of the tokens, specifically "character content (all alphabetic, numeric, or mixture), vowel/consonant content, casing (all upper, or lower, or mixed), and if it contains some specific punctuation marks (slash, dot or dash)" (*ibid.*: 307), the tokenisation module also attempts to determine whether each token identified is a standard word, an abbreviation, an acronym, an initialism, a number, an ordinal, a date, a time, a telephone number, a monetary expression, a Roman numeral, a symbol, an e-mail address, or a URL, etc. (Allen, 1992; Liberman and Church, 1992; Dutoit, 1997; Huang *et al.*, 2001; Sproat *et al.*, 2001). Tokenisation therefore provides clues as to which set of LTS conversion rules, standard or letter-by-letter, should be applied to tokens (see section 2.4.1.1). In addition, tokenisation also provides clues as to how tokens should be normalised.

Normalisation is the process of converting non-standard words into their full orthographic forms (Divay and Vitale, 1997; Huang *et al.*, 2001; Sproat *et al.*, 2001), also known as Standard Normalised Orthographic Representation (SNOR) (Huang *et al.*, 2001). Examples of a range of different types of token and their SNOR are presented in Table 1.

**Table 1Examples of non-standard words and their SNOR (examples taken from Rodman (1999: 205))**

| Token | SNOR |
|---|---|
| Mrs. | Missus |
| Dr. | Doctor, drive |
| Ph.D. | Pee aitch dee |
| St. | Street, saint, stanza |
| Ch. | Chapter, chaplain |
| 1,111 | One thousand, one hundred (and) eleven |
| 111th | On hundred (and) eleventh |
| 5/19/40 | May nineteenth, nineteen forty; the nineteenth of May, nineteen forty |
| 5:30 A.M. | Five thirty a em |
| $44.44 | Forty-four dollars and forty four cents |
| NASA (National Aeronautics and Space Administration) | Nassa |
| NCSU (North Carolina State University | En see ess you |
| DOD (Department of Defense) [*sic*] | Dee oh dee |

Once converted to SNOR, other levels of linguistic analysis necessary to determine the phonetic transcription of tokens can be applied.

### 2.4.1.3 The lexicon

The functions of the lexicon in a TTS synthesis system are to provide:

- pronunciations of words (or morphemes),[10]
- SNOR of abbreviations for exploitation in text normalisation (see section 2.4.1.2), and,
- part of speech (POS) and other information on lexical items required by other levels of linguistic analysis (Huang *et al.*, 2001).

Regarding the pronunciation of lexical items, as mentioned in section 2.4.1.1, traditionally the lexicon was only used to provide the phonetic transcription of irregular words, that is, words whose pronunciation cannot be accounted for by the LTS rules of the system (Liberman and Church, 1992; Edgington *et al.*, 1996a; Huang *et al.*, 2001). The number of words whose pronunciation cannot be accounted for by LTS rules varies from language to language. In languages like German, Finnish, Russian, Spanish Swahili and Turkish there is a near one-to-one relationship between graphemes and phonemes (Tench, 1992; Sproat, 1996; Divay and

[10] Lexicons are often morpheme-based because, while it is common for people to coin new words, "it is rare for a new morpheme ... to enter the language" (Allen, 1992: 754) and because morphemes are generative: "A lexicon of N morph[eme]s can readily generate between 5N and 10N words" (*loc. cit.*).

Vitale, 1997; Lemmetty, 1999; Rodman, 1999; Huang *et al.*, 2001), consequently few phonetic transcriptions need to be provided in the lexicon. In languages such as English and French, on the other hand, the relationship is less transparent (Liberman and Church, 1992; Sproat, 1996; Rodman, 1999). For example, the English grapheme sequence *ough* has 7 phonetic realisations: "*rough* [ʌf], *through* [u ˑ], *bough* [aʊ], *thought* [ɔ ˑ], *dough* [əʊ], *cough* [ɒf], and *hiccough* [ʌp]" (Divay and Vitale, 1997: 498), and the French grapheme *x* has 5 "[ks] in *axiome*, [gz] in *exemple*, [s] in *soixante*, [z] in *sixième*, [ ] (not pronounced) in *auxquels*" (*ibid.*: 499). In addition to the likely candidates proper nouns, including personal names and place names (Rodman, 1999), the pronunciation of frequent function words deviates from standard pronunciation, (see Table 2). Their phonetic transcriptions must therefore also be provided in the lexicon.

**Table 2 Examples of function words whose pronunciation deviates from standard pronunciation (Allen, 1992: 752)**

| Function words | Content words |
|---|---|
| of | fun, roof |
| the, this, that, ... | thesis, thimble, earth |
| have | shave, behave |
| was, is, has | atlas, canvas |

In section 2.4.1.2, it was mentioned that the form of a token can provide clues as to whether it is an acronym or an initialism, and hence whether a token should be pronounced as a word or spelled out letter-by-letter. Acronyms are short forms created from the initial letters of the words which constitute a longer expression which are pronounced as if they were a standard word, e.g. *UNESCO* (United Nations Educational, Scientific and Cultural Organization) is pronounced [junɛskəʊ] and *NATO* (North Atlantic Treaty Organization) [neɪtəʊ] (Cabré, 1998). Initialisms, however, are pronounced letter-by-letter, e.g. *UFO* (Unidentified Flying Object) [ju ˑ ɛf əʊ] and *AC* (Alternating Current) [eɪ si] (*ibid.*). Regarding the orthographic form of acronyms and initialisms, the initials which constitute an acronym are, in general, not separated by full stops, whereas the initials which make up an initialism are. Another factor that might be taken into account is whether the sequence of letters is speakable (Dutoit, 1997; Rodman, 1999; Huang *et al.*, 2001). In order to qualify as an acronym the sequence of letters must in general be pronounceable (Dutoit, 1997; Rodman, 1999; Huang *et al.*, 2001). There are, however, many exceptions to these rules: initialisms may be written without periods as in the examples above; *DOD* (Department of Defense) is pronounceable, however it is read out letter-by-letter (Rodman, 1999); and, the theoretically unpronounceable *RFRA* (Religious

Freedom Restoration Act) and *SCSI* (Small Computer Systems Interface) are pronounced as words, [rɪfra] and [sku ˈ zi] respectively (Liberman and Church, 1992; Rodman, 1999). Consequently, it is safer to include the phonetic transcription of common acronyms and initialisms in the lexicon (Liberman and Church, 1992).

As mentioned in section 2.4.1.1, today, the lexicon of most TTS synthesis systems provides a phonetic transcription for most words that the system is expected to come across and TTS synthesis systems rely on LTS rules only for the generation of the phonetic transcription of out-of-vocabulary words (Edgington *et al.*, 1996a; Divay and Vitale, 1997; Huang *et al.*, 2001).

Regarding the third function of the lexicon mentioned above, examples of the information that the lexicon might provide for exploitation by the other levels of linguistic analysis include: POS, morpheme-type (root, bound, free, prefix, suffix, infix), and semantic representation (Huang *et al.*, 2001).

### 2.4.1.4 The morphological analyser

The goal of the morphological analyser is to determine the morphological structure of words. When synthesising languages which have highly productive vocabularies, such as agglutinative languages like Basque, Finnish, and Turkish (Lopez de Ipina, 2002), and languages which rely heavily on derivational morphology such as Dutch, German, Swedish and Welsh (Dutoit, 1997), morphological analysis is essential because it is impossible to include every possible word in the lexicon. Specifically in combination with the lexicon, morphological analysis may be able to provide the phonetic transcription, POS, and semantic representation of out-of-vocabulary words: certain suffixes are typically associated with a particular POS, and the meaning of words is often derivable from the meaning of their constituent morphemes.

In English, LTS rules do not apply across morpheme boundaries (Allen, 1992; Divay and Vitale, 1997). Consider the pronunciation of "*uni* in *uniformed-uninformed*, *th* in *pothole-matthew*, [and] *ph* in *flophouse-sphere*" (Divay and Vitale, 1997: 498). The same is true in French:

> forms like *tournesol*, *entresol*, *télésiège* are formed from two morphemes, each of which retains its pronunciation. Usually, in French, *s* between two vowels is pronounced [z], otherwise [s]. The *s* in *tournesol*, *entresol*, *télésiège*, *contresens*, *antisocial* must be considered the beginning of a morpheme, and although it occurs between two vowels, is pronounced [s] (*ibid.*: 499f).

Morphological analysis is therefore necessary in order to generate accurate phonetic transcriptions of out-of-vocabulary words in these languages.

### 2.4.1.5 The syntax module

Syntactic analysis permits POS disambiguation and hence the disambiguation of heteronyms, words which have a single orthographic realisation, but more than one phonetic realisation, such as English *lives* [lɪvz] vs. [laɪvz] (Yarowsky, 1997), and French *couvent* [kuvɑ̃] 'convent' vs. [kuv] '[they] brood' (Mertens *et al.*, 2001).

The POS information provided by syntactic analysis is also exploited in lexical stress and sentence stress assignment. For example, the stress pattern of the English word 'import' depends on whether it is being used as a noun ('import) or a verb (im'port). Regarding sentence stress, content words tend to receive stress whereas function words tend not to receive stress (Linggard, 1985; Dutoit, 1997; Huang *et al.*, 2001).

Syntactic analysis also permits the identification of the tense of a word and hence the disambiguation of heteronyms like the English word *read* [rid] vs. [rɛd] (Sagisaka, 1990; Allen, 1992).

Number, gender and case agreement can also be resolved through syntactic analysis. Syntactic analysis hence permits a TTS synthesis system to determine the SNOR of numbers, abbreviations and symbols, which must agree in number and/or gender and/or case. For example:

- measure word abbreviations must agree in number with the noun that they qualify in English: "Measure word abbreviations are systematically ambiguous between singular and plural interpretations: *1 in.* is 'one inch' while *2 in.* is 'two inches'" (Liberman and Church, 1992: 805);
- the number *1* must agree in gender with the noun that it qualifies in both French and Spanish, and in gender and case with the noun that it qualifies in German; and,

- in Russian, when % takes on the form of an adjective, it must agree in number, gender and case with the noun that it modifies (Sproat *et al.*, 1998).

Finally, through syntactic analysis it is possible to determine the syntactic structure of utterances. "syntactic and prosodic phrases || tend to be aligned || in simple phrases" (Dutoit, 1997: 154). In addition, the intonation pattern of many utterances depends on their function. This is the case of the different question types in English: 'yes/no' questions are associated with a rising pitch contour, whereas, 'wh' questions are associated with a declining pitch contour (Allen, 1992; Dutoit, 1997). Syntactic analysis therefore also provides information that can be exploited by the prosody generation module to determine the prosodic specifications of utterances.

### 2.4.1.6 The semantics module

By building a picture of the semantic context of an utterance, semantic analysis permits the disambiguation of heteronyms (see section 2.4.1.5) whose phonetic realisation depends solely on semantic context, such as *bass* [bas] vs. [beɪs] in English (Yarowsky, 1997), and *fils* [fil] 'threads' vs. [fis] 'son(s)' in French (Mertens *et al.*, 2001)

Semantic analysis also permits topic tracking and hence the identification of given and new information and cases of antithesis, or contradiction. This information is important for determining the prosodic specification of utterances because new information and cases of antithesis typically receive stress, whereas given information is deaccented (Allen, 1992; Hiyakumoto *et al.*, 1997; Huang *et al.*, 2001; Pitt and Edwards, 2003).

### 2.4.1.7 The prosody generation module

The goal of the prosody generation module is to determine the prosodic specifications of utterances, i.e. the F0, amplitude and duration of the constituent segments.

As mentioned in the preceding sections:

- text pre-processing, in particular, segmentation provides clues to phrasing, and to the intonation pattern of utterances (see section 2.4.1.2);
- the lexicon and morphological analysis provide clues which could be exploited in stress assignment (see sections 2.4.1.3 and 2.4.1.4 respectively);

- syntactic analysis provides clues which could be exploited in accent assignment, phrasing, and intonation assignment (see section 2.4.1.5); and,
- semantic analysis provides clues which could be exploited in accent assignment (see section 2.4.1.6).

> [N]ot all prosodic phrases or even all clause boundaries happen to be delimited by punctuation marks. Major syntactic breaks indeed can appear as conjunctions, or simply not appear at all (as in *I found the book I wanted to read*) (Dutoit, 1997: 149).

The information provided by segmentation is therefore not adequate to determine the phrasing of utterances. Nor is it adequate to determine the intonation pattern of utterances: a single punctuation mark is used to encode a number of different question types each of which is associated with a different intonation pattern (Allen, 1992; Dutoit, 1997).

The information provided by the lexicon and morphological analysis is not sufficient for accurate stress assignment either: "When words are grouped into sentences, ... they do not necessarily keep their lexical stress" (*ibid.*: 160).

> The classic example is 'thirteen men'. In isolation, 'thirteen' has a w[eak] s[trong] syllabic stress pattern, but appears to undergo a reversal to s w when embedded in 'thirteen men' (Edgington *et al.*, 1996b: 87)

Regarding accent assignment, a system which assigns accent on the basis of POS may

> work[ ] adequately for many short, isolated sentences, such as "*The **cat sat** on the **mat***", where the words selected for accentuation appear in bold face. For more complex sentences, appearing in document of dialog context, such an algorithm will sometimes fail (Huang *et al.*, 2001: 751)

In particular, if one were to accent all content words in more complex sentences, the speech would often sound *overaccented* (Dutoit, 1997).

Regarding phrasing, syntactic and prosodic phrases may not be aligned in complex sentences such as those containing embedded phrases (Edgington *et al.*, 1996b; Dutoit, 1997).

There is therefore more to prosody generation than integrating the information provided by the other levels of linguistic analysis and converting it to acoustic specifications, i.e. F0,

amplitude and duration. Further processing must be carried out (see Edgington *et al.*, 1996b; Dutoit, 1997).

### 2.4.1.8 Techniques

Regarding the techniques employed in TTP conversion, at almost every level of linguistic analysis, the TTS synthesis developer has a choice between the use of: simple hand-derived rules, theoretically motivated grammar-based systems, and rules derived from corpora by statistical analysis.

## 2.4.2 The PTS module

As stated, the goal of the PTS module is to generate speech output from the phonetic transcription that it is provided by the TTP module. Currently, four main approaches are employed.[11] These are:

- articulatory synthesis,
- formant synthesis,
- concatenative synthesis, and,
- Unit Selection Synthesis (USS).

In the following sections, these approaches are presented in turn.

### 2.4.2.1 Articulatory synthesis

Articulatory synthesis is based on the simulation of the physical processes involved in human speech production (Docherty and Shockey, 1988; Denes and Pinson, 1993; d'Alessandro and Liénard, 1996; Edgington *et al.*, 1996b; Huang *et al.*, 2001). Early articulatory synthesisers, such as the one proposed by Wolfgang von Kempelen, consisted in direct physical simulations of the human speech organs (lungs, vocal cords, tongue, lips, etc.) which were operated much like a musical instrument (Denes and Pinson, 1993; Olive *et al.*, 1998; Rodman, 1999; Sondhi, 2002).

Today's articulatory synthesisers, on the other hand, consist in indirect computational simulations of human speech production. Taking neuromotor commands, articulator positions,

[11] A number of hybrid techniques have recently been suggested in the literature (Bickley and Bruckert, 2002; Hertz, 2002; Prudon *et al.*, 2002; Sondhi, 2002; Stevens, 2002). However, as far as it is possible to establish, these are not yet employed in commercial systems.

or vocal tract shapes as input (O'Shaughnessy, 1987) systems based on articulatory synthesis calculate the movement of the vocal cords and the shapes and volumes of the different cavities of the vocal tract as it changes over time (Linggard, 1985; Klatt, 1987; Docherty and Shockey, 1988; Gabioud, 1994; Huang et al, 2001; Shadle and Damper, 2002). Waveforms are then generated by simulating the flow of air through the vocal tract (Allen *et al.*, 1987; Klatt, 1987; Styger and Keller, 1994).

### 2.4.2.2 Formant synthesis

Formant synthesis is a specific type of parametric synthesis, a technique which is based on the assumption that it is possible to model human speech satisfactorily by simulating the perceptually relevant characteristics of the acoustic signal (Pfister and Traber, 1994). Of the many parameters that constitute the acoustic signal, the frequencies and patterning of stationary and transitional formants, have been shown to be important perceptual cues in speech perception (Denes and Pinson, 1993; Borden *et al.*, 1994; Dutoit, 1997; Strange, 1999a; 1999b). Most parametric synthesisers therefore simulate the amplitude and frequencies of formants. Such synthesisers are referred to as formant synthesisers. Formant synthesisers are based on the source-filter model of speech production (Klatt, 1987; Stevens, 1992; Styger and Keller, 1994; Olive *et al.*, 1998; Bickely *et al.*, 1999; Lemmetty, 1999; Rodman, 1999; Stevens, 2002):

> To prepare a synthetic vowel sound, one begins with an electronic tone [the source] at the frequency of vibrating vocal cords, and all its harmonics, as ingredients. Filters modify the harmonics to accentuate the formant frequencies, and de-emphasize other frequencies, in the proportions specific to the vowel being synthesized
>
> Recipes for consonants require additional ingredients and modifications. Sound must be cut off [*sic*] for stops and affricates, resonances added for nasals, hissing and buzzing noises tossed in [*sic*] for sibilants like /s/ and /z/, and white noise (aperiodic sounds) blended into the mixture in just the right proportion for voiceless consonants.
>
> Cooking is time-dependent. "Roast 20 minutes per pound," "stir for five minutes," or "baste every half hour." Speech synthesis is highly time-dependent, and the proportion of ingredients must be adjusted frequently. Indeed, the dynamic character of speech sounds requires that parameters be updated hundreds of times per second, typically, once every five milliseconds (Rodman, 1999: 179).

The values of the acoustic parameters, including but not limited to the F0 of the source, the frequency and the amplitude of the formants, and duration, are derived by rule from the phonetic transcription provided by the TTP module (Klatt, 1987; O'Shaughnessy, 1987;

Docherty and Shockey, 1988). Formant synthesis is therefore also known as 'rule(-based) synthesis' (Docherty and Shockey, 1988; Dutoit, 1997; Olive *et al.*, 1998; Huang *et al.*, 2001).

### 2.4.2.3 Concatenative synthesis

Concatenative synthesis is based on the assumption that it is possible to generate novel utterances satisfactorily by concatenating segments of pre-recorded natural human speech (Moulines, 1992; Bhaskararao, 1994; Edgington *et al.*, 1996b; Olive *et al.*, 1998; Rodman, 1999; Huang *et al.*, 2001; Rank, 2002). Such systems therefore consist in a database of segments which have been extracted from a corpus of recordings of human speakers (Dutoit, 1997; Rodman, 1999; Huang *et al.*, 2001). Concatenative synthesis may be based on the concatenation of phonemes, diphones,[12] triphones or tetraphones,[13] syllables, disyllables,[14] demi-syllables,[15] words or whole phrases. According to Huang *et al.* (2001), the optimal segment for concatenative synthesis:

- leads to low distortion at concatenation points,[16]
- is generalisable, i.e. permits the synthesis of unrestricted text; and,
- is trainable, i.e. can be automatically extracted from recordings of human speakers.

In addition, it is also suggested that the optimal segment ought to capture interallophonic effects such as coarticulation (Bhaskararao, 1994; Dutoit, 1997), and be computationally manageable, i.e. the number of units required for the generation of unrestricted text is not excessive (Bhaskararao, 1994). Most systems are diphone-based (Olive *et al.*, 1998; Huang *et al.*, 2001). Diphones, are generalisable (Huang *et al.*, 2001) and computationally manageable – a database of around 1,200 diphones is needed for the synthesis of unrestricted French (Dutoit, 1997). Diphones capture most inter-allophonic transitions (O'Shaughnessy, 1987; Moulines,

---

[12] Diphones, or dyads, consist in the second half of one allophone and the first half of the following allophone (Dutoit, 1997; Rodman, 1999; Huang *et al.*, 2001).

[13] Triphones, also referred to as context-dependent phonemes (Huang *et al.*, 2001), consist in the second half of an allophone, the following allophone and the first half of the next allophone (O'Shaughnessy, 1992; Dutoit, 1997; Huang *et al.*, 2001). Tetraphones, by extension, cover two and two half allophones.

[14] Disyllables consist in the second half of one syllable and the first half of the following syllable (Dutoit, 1997)

[15] Demi-syllables consist in either the first half or the second half of a syllable. That is they consist in either the syllable onset and the first half of the syllable nucleus, or the first half of the syllable nucleus and the syllable coda (Allen *et al.*, 1987; Lemmetty, 1999; Rodman, 1999).

[16] When segments of pre-recorded human speech are concatenated distortions occur due to the fact that the amplitude and frequency of the formants and/or the fundamental frequency of the concatenated segments do not match (Huang *et al.*, 2001).

1992; Dutoit, 1997; Huang *et al.*, 2001), and lead to low distortion at concatenation points (Lemmetty, 1999; Rodman, 1999). Diphones are, however, not optimal because they do not capture all coarticulatory effects – coarticulatory effects may span several phonemes (O'Shaughnessy, 1987; Moulines, 1992; Dutoit, 1997; Portele *et al.*, 1997; Olive *et al.*, 1998) – and they give rise to a high density of concatenation points (Dutoit, 1997). Similarly, none of the other segments, mentioned above meets all of these criteria (see Bhaskararao (1994), Edgington *et al.* (1996b), Dutoit (1997), Rodman (1999), Huang *et al.* (2001), and Henton (2002) for discussions of the properties of the different segments). Consequently, the database of many systems comprises a combination of the different types of segments (Dutoit, 1997; Portele *et al.*, 1997; Olive *et al.*, 1998).

Regarding synthesis, concatenative synthesis systems take as input a phonetic transcription of the utterance to be synthesised. First the system searches through its database for all strings of segments that can be concatenated to generate the desired utterance. If several strings are retrieved, a network of segments is generated and the system searches for the best path through the network, i.e. the string of segments that best accounts for coarticulation (Takeda *et al.*, 1992; Olive *et al.*, 1998) and gives rise to the least distortion at concatenation points (Takeda *et al.*, 1992; Dutoit, 1997; Olive *et al.*, 1998; Huang *et al.*, 2001). The selected strings of segments are then concatenated. As stated, distortions occur at segment boundaries, so during concatenation the selected segments are manipulated in order to smooth the boundaries (see Edgington *et al.* (1996b), Dutoit (1997), Rodman (1999), and Huang *et al.* (2001) for a discussion of the algorithms employed). Due to the fact that typically only one instance of each segment is stored in the database, segments generally do not fit the prosody of the utterances to be generated. During concatenation, segments must therefore also be manipulated in order to fit the prosody (Edgington *et al.*, 1996b). This is achieved using the same algorithms as are used to smooth segment boundaries.

#### 2.4.2.4 USS

Like concatenative synthesis, USS is based on the assumption that it is possible to generate novel utterances satisfactorily by concatenating segments of pre-recorded natural human speech. In USS systems, these segments may either be stored in a database (Campbell and Black, 1997; Huang *et al.*, 2001; Schroeter, 2001) or extracted directly from a corpus during synthesis (Black and Taylor, 1994; Conkie, 1999; Prudon *et al.*, 2002). Also like concatenative synthesis, USS may be based on the concatenation of phonemes, diphones, triphones or

tetraphones, syllables, disyllables, demi-syllables, words or whole phrases (see section 2.4.2.3). The features of the optimal segment for USS are the same as those for concatenative synthesis. Due to the fact that no one segment meets all those criteria, USS, like concatenative synthesis, is also often based on a combination of the aforementioned segment types (Portele *et al.*, 1997; Huang *et al.*, 2001; Schroeter *et al.*, 2001). When based on a combination of different segment types, USS, is referred to as non-uniform USS (Schroeter, 2001; Henton, 2002). The difference between USS and concatenative synthesis is that in USS an attempt is made to capture prosody in the segments (Olive *et al.*, 1998; van Santen *et al.*, 2002). This is achieved by including in the system's database or corpus several instances of each segment each in a different context (Campbell and Black, 1997). Thus, in USS the segments need only be manipulated for the purposes of smoothing segment boundaries (Conkie, 1999). This is kept to a minimum by selecting the string of segments that leads to the least distortion at segment boundaries (*ibid.*).

Perhaps the most well-known system based on USS is the CHATR system from ATR (Black and Taylor; 1994; Conkie, 1999). This system, which is phoneme-based, takes as input a phonetic transcription augmented with phoneme duration and target F0. The system then searches through its corpus for all examples of each phoneme in the string in turn. The corpus is marked up with indications of the prosodic features of each phoneme. The prosodic features of the phonemes provided in the corpus are then compared with the prosodic specifications in the input and a score of similarity to the target, referred to as 'target cost', is assigned to each phoneme. Next, adjacent phonemes are compared and a score of similarity at segment boundaries, referred to as 'concatenation', 'join', or 'transition cost', is assigned to each pair. The phonemes are then organised into a network and the system searches for the string of phonemes which gives rise to the lowest cost. Next, the selected phonemes are concatenated. Although, the string of segments is selected so as to lead to the least distortion at segment boundaries, a small amount of distortion may remain. In order to reduce this distortion, during concatenation the selected segments are manipulated in order to smooth any such distortions (the algorithms employed, which are the same as those employed in concatenative synthesis, are discussed in Edgington *et al.* (1996b), Dutoit (1997), Rodman (1999), and Huang *et al.* (2001)).

## 2.5 Using TTS synthesis

As we shall see in section 2.5.5, the suitability of a TTS synthesis system for use in a particular application depends on the quality and flexibility of the speech generated and the computational demands of the system. These aspects of TTS synthesis systems are therefore reviewed in sections 2.5.1, 2.5.2 and 2.5.3 respectively. Integration is another important issue with respect to the use of TTS synthesis systems. Integration issues are therefore considered in section 2.5.4. Current applications in TTS synthesis are then presented in section 2.5.5. Finally, the benefits of using TTS synthesis over other media are presented in section 2.5.6.

### 2.5.1 Quality of the speech output

The quality of the speech generated by TTS synthesis systems depends on the techniques employed in both TTP and PTS conversion. These effects are considered in turn in the sections that follow.

#### 2.5.1.1 TTP conversion

The quality of the speech generated by TTS synthesis systems is affected by the following aspects of TTP conversion:

- whether LTS conversion is predominantly rule- or lexicon-based,
- whether other levels of linguistic analysis are carried out, and,
- the type of technique employed in the different levels of linguistic analysis.

Regarding LTS conversion, lexicon-based approaches are more accurate than rule-based approaches (Liberman and Church, 1992) and hence produce higher quality output.

As the examples presented in sections 2.4.1.2, 2.4.1.4, 2.4.1.5, 2.4.1.6, and 2.4.1.7 demonstrate, LTS rules or dictionary look up alone are not adequate to generate high quality speech from text. Due to the discrepancies between text and speech, other levels of linguistic analysis are necessary to generate high quality speech output. The quality of the final output however depends on the techniques employed at each level of linguistic analysis. As mentioned in section 2.4.1.8, at almost every level of linguistic analysis, the TTS synthesis developer has a choice between the use of:

- simple hand-derived rules,
- theoretically motivated grammar-based systems, and,

- rules derived from corpora by statistical analysis.

The effects of these different techniques on the quality of the speech generated by TTS synthesis systems are presented in the following paragraphs.

Simple hand-derived rules are robust (Liberman and Church, 1992), i.e. are "able to cope with errors and mishaps during program execution without producing wrong results or stopping" (BCS, 1995: 2000). Simple hand-derived rules can also be hand-tuned to produce optimal quality output. While many simple hand-derived rules perform remarkably well in isolation, "the combined degradation contributed by several imperfect processes is [however] likely to impair the speech quality very seriously" (Witten, 1982: 234).

Like simple heuristics, systems based on linguistic theories can also be hand-tuned to produce optimal quality output. They, however, do not guarantee better output than simple hand-derived rules. In particular, they are less robust than simple rules and rules derived from corpora through statistical analysis (Dutoit, 1997). The main reason for this is that "Real sentences are often much more complex than those found in linguistic texts" (*ibid.*: 152).

Rules derived from aligned corpora through statistical analysis should in theory have a broader more uniform coverage than both simple hand-written rules and theoretically motivated grammar-based systems (Edgington *et al.*, 1996a) and hence be more robust. The rules derived from corpora are, however, often not readable by humans and therefore do not lend themselves to hand tuning (Dutoit, 1997).

### 2.5.1.2 Articulatory synthesis

Of all the techniques of PTS conversion described here, it is believed that articulatory synthesis has the potential to produce the best quality speech (Edwards, 1991; Huang *et al.*, 2001; Shadle and Damper, 2002). Firstly, articulatory synthesis should reproduce "all the perceptually relevant effects that occur in real speech" (Linggard, 1985: 39); in contrast, as mentioned in section 2.4.2.2, formant synthesis only reproduces those features of speech which it is *believed* are necessary for perception (Pfister and Traber, 1994). Secondly, because articulatory synthesis takes into account the constraints of the vocal tract (Linggard, 1985; O'Shaughnessy, 1987; Stevens, 1992; Dutoit, 1997; Shadle and Damper, 2002), the generated speech should sound natural. In particular, it should only consist in possible sounds of the

synthesised language (Bailly, 2002b). Thirdly, interallophonic effects characteristic of natural speech, such as coarticulation, should "occur naturally – as they do in real speech" (Linggard, 1985: 38).

Although it was demonstrated in the early 1960s that articulatory synthesis can be manually tuned to produce highly intelligible speech (Klatt, 1987), the quality of automatically generated output is poor in comparison with that generated by formant, concatenative, and unit selection synthesis. This is mainly due to the difficulty of obtaining the articulatory parameters required to drive it. Traditionally X-rays were used for these purposes. The use of X-rays, however, had several limitations: X-ray data is 2D while the vocal tract is 3D (O'Shaughnessy, 1987; Lemmetty, 1999); "X-ray data do not characterise the masses of degrees of freedom of the articulators" (Klatt, 1987; Lemmetty, 1999); and, little data is available due to the dangers of exposure to X-rays (O'Shaughnessy, 1987; Lemmetty, 1999). Several other techniques have been proposed for the acquisition of articulatory parameters, including: magnetic resonance imaging, articulography, electropalatography, and ultrasound (Huang *et al.*, 2001; Shadle and Damper, 2002; Sondhi, 2002). While the results of these techniques can be combined to obtain a full view of articulation (Shadle and Damper, 2002), data must be obtained under artificial conditions (Sondhi, 2002).

Another reason for the poor quality of articulatory synthesis is that it has been necessary to simplify articulatory models in order to keep the computational costs down. The problem is that it is not known where simplifications can be made without causing losses to quality (Klatt, 1987). Most frequently, articulatory synthesis is simplified through the direct specification of the shape of the vocal tract. As a consequence of the use of this approach one of the most frequently cited benefits of articulatory synthesis is lost; coarticulation rules must be stated explicitly (O'Shaughnessy, 1987).

### 2.5.1.3 Formant synthesis

It is proposed that it should be possible to generate high quality speech output using formant synthesis because formant synthesis permits fine manipulation of the acoustic parameters of speech (Henton, 2002), and acoustic cues which result from articulatory behaviour which is difficult to measure can be copied directly from the acoustic signal (Linggard, 1985).

Just as it has been demonstrated that it is possible to obtain high quality speech output by manually tuning articulatory synthesis systems (Klatt, 1987), it has been demonstrated that it is possible to generate speech output which is practically indistinguishable from real speech by manually tuning formant synthesis systems (Denes and Pinson, 1993; Edgington *et al.*, 1996b; Bickely *et al.*, 1999).

Regarding the quality of automatically generated speech, performing uniformly across different utterances (Huang *et al.*, 2001), formant synthesisers have produced highly intelligible speech without recourse to manual tuning for many years (Allen *et al.*, 1987; Edgington *et al.*, 1996b). In fact, in a recent experiment which compared the intelligibility of a number of commercial TTS synthesis systems, the output of the formant synthesiser, *IBM ViaVoice*, was found to be only slightly less intelligible than the output of the concatenative synthesisers, *AT&T Next-Gen TTS* and *Festival* from the University of Edinburgh, and more intelligible than the hybrid system, *FlexVoice* (Venkatagiri, 2003). Regarding the differences in intelligibility between the formant and the concatenative synthesisers, in the aforementioned experiment, it was found that while the formant synthesiser made more consonant errors than the concatenative synthesisers, the formant synthesiser made fewer vowel errors than the concatenative synthesisers (*ibid.*). Regarding the specific types of errors made by formant synthesisers at the phonetic level, Rodman (1999) notes that parametric synthesis is poor at pronouncing consonant clusters and modelling vowel durations in stressed syllables.

In addition to being highly intelligible, the speech generated by formant synthesis is by design smooth (Olive *et al.*, 1998; Bickley *et al.*, 1999; Henton, 2002). It does not, however, sound very natural (Allen *et al.*, 1987; Edgington *et al.*, 1996b; Rodman, 1999; Huang *et al.*, 2001; Henton, 2002). In particular, it sounds "robotic" (Edgington *et al.*, 1996b; Henton, 2002), and monotonous (Henton, 2002).

There are several reasons for this lack of naturalness. Firstly, formant synthesisers model only a small number of acoustic cues (Docherty and Shockey, 1988). Secondly, they do not take into account the articulatory constraints of the vocal tract (Stevens, 1992). Thirdly, regarding naturalness at the phonetic level, formant synthesisers "make use of too few allophones" (Rodman, 1999: 183). Finally, regarding naturalness at the prosodic level, the speech generated by formant synthesis lacks naturalness at this level not because it is not possible to

produce natural sounding prosody, rather because the rules for determining the prosody of utterances are inadequate (Styger and Keller, 1994; Dutoit, 1997; Rodman, 1999)

### 2.5.1.4 Concatenative synthesis

It is believed that the speech generated by concatenative synthesis systems should be as intelligible and natural as human speech "within the limits of digitization" (Rodman, 1999: 185) (see van Santen *et al.* (2002) for a discussion of the effects of digitisation): concatenative synthesis should only produce "possible human speech sounds" (Keller, 2002: 5) and for those speech sounds it should reproduce all perceptually relevant acoustic cues (Olive *et al.*, 1998) because it is based on the concatenation of segments of recordings of natural human speech.

The actual quality of the speech generated by concatenative synthesis varies significantly from system to system depending on the type(s) of segment(s) used (Dutoit, 1997; Edgington, 1997), the features and quality of the corpus from which those segments were extracted (Dutoit, 1997), the quality of segmentation (*ibid.*), and the algorithm used for digitising the segments, smoothing segment boundaries, and manipulating the prosody (Dutoit, 1997; Olive *et al.*, 1998; Lemmetty, 1999; Huang *et al.*, 2001; van Santen *et al.*, 2002). Moreover, the quality of the output generated by a single concatenative synthesis system may vary from very high to very low quality depending on the utterance to be synthesised (Huang *et al.*, 2001).

In general, the speech generated by concatenative synthesis is, however, highly intelligible. As mentioned in the previous section 2.5.1.3, in an experiment which compared the intelligibility of a number of commercial TTS synthesis systems, the output of the concatenative synthesisers, *AT&T Next-Gen TTS* and *Festival* from the University of Edinburgh, was found to be more intelligible than the output of both the formant synthesiser, *IBM ViaVoice*, and the hybrid system, *FlexVoice* (Venkatagiri, 2003). Regarding the differences in intelligibility between the concatenative synthesisers and the formant synthesiser, as also mentioned in section 2.5.1.3, in the aforementioned experiment it was found that while the concatenative synthesisers made fewer consonant errors than the formant synthesiser, they made more vowel errors (*ibid.*). More specifically, Rodman (1999) observed that concatenative synthesisers are better at modelling consonant clusters and vowel durations in stressed syllables than formant synthesisers.

Beyond the phonetic level, the speech generated by concatenative synthesis, while it sounds more natural than that generated by formant synthesis (Edgington, 1997; Rank, 2002), does not sound as natural as human speech. There are several reasons for this lack of naturalness. Firstly, distortions at concatenation points make the speech less smooth than natural human speech (Huang *et al.*, 2001). Secondly, due to the fact that the algorithms used to manipulate the prosodic features of segments cannot make the changes to segmental quality which accompany changes in prosody in natural human speech, the individual segments may sound hyper-articulated (Campbell and Black, 1997). Thirdly, during manipulation segments may become distorted (Takeda *et al.*, 1992; Campbell and Black, 1997). Finally, due to the inadequacy of methods for determining the prosodic specification of utterances (Dutoit, 1997; Rodman, 1999; Henton, 2002), and the fact that segments are typically extracted from corpora of prosodically neutral speech (Campbell and Black, 1997), the prosody of speech generated by concatenative synthesis, like that of formant synthesis, tends to be monotonous (Henton, 2002).

#### 2.5.1.5 USS

The attraction of USS is that the speech generated should be highly intelligible and natural at both the phonetic and prosodic levels (Conkie, 1999). There are several reasons why USS is expected to generate high quality speech. Firstly, like concatenative synthesis, USS should only produce "possible human speech sounds" (Keller, 2002: 5) and for those speech sounds it should reproduce all perceptually relevant acoustic cues (Olive *et al.*, 1998) because it is based on the concatenation of segments of recordings of natural human speech. Secondly, the prosody should sound natural because segments with the appropriate prosody are extracted from recordings of human speakers (Conkie, 1999). Thirdly, less distortion should be present in the generated speech because little segment manipulation is carried out (*ibid.*).

According to Conkie, the speech generated by USS is indeed highly intelligible and natural both at the phonetic and prosodic levels. In fact, Schroeter *et al.* (2002) believe that short utterances, such as voice prompts, generated by USS sound so natural that they might pass the Turing test. Like the speech generated by concatenative synthesis, the speech generated by USS is, however, not of consistent quality (Conkie, 1999).

### 2.5.2 Flexibility of the output

Like the quality of the output generated by TTS synthesis systems, the flexibility of the speech generated by TTS synthesis systems depends on the techniques employed in both TTP and PTS conversion. The effects of the different TTP and PTS techniques are considered in turn in the sections that follow.

#### 2.5.2.1 TTP conversion

Choices in the design of TTS synthesis systems at the level of TTP conversion affect their ability to provide options over accent and speaking style.

In order to provide options over accent and style in a system in which LTS conversion is rule-based, the system developer would need to develop an alternative set of LTS rules for each accent and style of speech to be generated. In a lexicon-based system, on the other hand, the system developer would need to identify all words affected by changes in accent and style and provide alternative transcriptions for them. It is believed that the latter is more difficult and time-consuming than the former.

Regarding the precise degree of flexibility of the rule-based approaches to LTS conversion, this depends on the technique employed. Hand-written rules permit human intervention and therefore permit hand tuning. As mentioned in section 2.5.1.1, rules derived from corpora through statistical analysis are often not readable by humans and therefore cannot be adapted by hand (Dutoit, 1997). Rather, a whole new corpus for each accent and style would need to be collected and analysed in order to obtain the rules. This is believed to be more time consuming than hand-tuning hand-written rules.

#### 2.5.2.2 Articulatory synthesis

Articulatory synthesis is highly flexible because it provides direct control over the speech organs (Docherty and Shockey, 1988). Specifically, "The parameter values can be manipulated to produce a range of vocal tract shapes and glottal configurations" (*ibid.*: 145) and hence their acoustic and perceptual correlates (*ibid.*). In other words, articulatory synthesis provides control over segmental quality, voice quality, pitch, volume and duration. Articulatory synthesisers should therefore be able to provide options over Speech Rate (SR), volume, prosody, emotion, and voice. Regarding language options, little adaptation of

articulatory synthesisers is necessary to permit the synthesis of most languages: "Languages with sounds (e.g. clicks) other than pulmonic egressives [such as Zulu and Xhosa (O'Connor, 1973)] require some simple modifications of the synthesiser architecture" (O'Shaughnessy, 1987: 110).

Regarding the control of the parameters mentioned above, it is believed that articulatory synthesisers should be easier to control than formant synthesisers (Linggard, 1985; Styger and Keller, 1994). Indeed, through the integration of *HLsyn*, a system that maps articulatory to acoustic parameters, the number of parameters needed to control the *DECtalk* formant synthesiser was reduced from over 40 to 13 (Stevens, 2002).

### 2.5.2.3 Formant synthesis

Formant synthesis is highly flexible because it provides direct control over the acoustic parameters of speech. Specifically, it permits the independent manipulation of the voice source and the formants (Dutoit, 1997). More specifically, the F0, amplitude and shape of the source and the duration, amplitude and frequency of the formants can be manipulated. In other words, like articulatory synthesis, formant synthesis provides control over segmental quality, voice quality, pitch, volume and duration (Docherty and Shockey, 1988). Formant synthesisers should therefore be able to provide options over SR, volume, prosody, emotion, and voice. SR, in particular, is highly flexible (Bickley *et al.*, 1999; Rodman, 1999): formant synthesisers can generate the high speaking rates required in reading machines (RMs) for the blind without introducing distortions to the speech output (Bickely *et al.*, 1999). Regarding options over voice, while formant synthesisers can theoretically generate different voices (Dutoit, 1997; Huang *et al.*, 2001), "female and child voices (derived from male parameters and head sizes) sound neither feminine nor human" (Henton, 2002: 121). It is also theoretically possible for formant synthesisers to provide options over language (Henton, 2002). Due to differences in the place of articulation of phonemes across languages (O'Shaughnessy, 1987), derived languages tend to sound even more unnatural than the language that the TTS synthesis system was originally programmed to generate, however:

> Multiple languages produced by a parametric system sound as though they are being spoken by a close relative of the original artificial speaker, who carries all the infelicities of voice quality that are generated from a less-than-perfectly modelled glottal source, and suffers from the same monotonic or repetitious prosodic

deficiencies. In popular literature this is referred to as the 'drunken Swede' syndrome (Henton, 2002: 127).

### 2.5.2.4 Concatenative synthesis

The flexibility of the output of concatenative synthesis depends on the size of the segment used (Henton, 2002) and the algorithm for boundary smoothing and prosody manipulation used (Dutoit, 1997; Olive *et al.*, 1998; Lemmetty, 1999; Huang *et al.*, 2001). In general, concatenative synthesis is significantly less flexible than formant and articulatory synthesis (Huang *et al.*, 2001). In particular, concatenative synthesis does not provide direct control over voice quality (Edgington, 1997). Duration, amplitude and pitch can only be manipulated within a narrow range (Hertz *et al.*, 2000). Concatenative synthesisers can therefore provide options over SR, volume and prosody – control over voice quality is necessary to provide options over emotion (Edgington, 1997; Bailly *et al.*, 2003b). The amount of control that can be provided over these parameters is however limited. Regarding the provision of options over speaking style and voice, a new database of segments must be collected for each style and voice that is to be generated (Lemmetty, 1999; Keller and Zellner-Keller, 2000; Keller, 2002; van Santen *et al.*, 2002). Database collection is labour intensive and hence costly (see Henton (2002) for a discussion of the problems that might be encountered during database collection). Consequently, most concatenative synthesisers only 'speak' in one style, reading (Keller and Zellner-Keller, 2000; Keller, 2002). Regarding the quality of the speech generated by TTS synthesis systems based on concatenative synthesis which do provide options over voice and style, as presented in section 2.5.1.4, the quality of the speech generated by systems based on concatenative synthesis depends on the quality of the corpus from which the database on which it is based is drawn. Due to the fact that each voice and style is based on a different database, the speech generated by systems based on concatenative synthesis which do provide options over voice and style can therefore differ significantly across voices and styles.

### 2.5.2.5 USS

Given the similarities between USS and concatenative synthesis, it is expected that speech generated by USS will demonstrate a similar degree of flexibility to that of speech generated by concatenative synthesis. Specifically, it is expected that systems based on USS will be able to 'speak' in a male or a female voice in reading style at a range of different SRs, volumes and pitches.

### 2.5.3 Computational demands

The computational demands of TTS synthesis systems depend for the most part on the method of PTS conversion employed. In the following paragraphs, the computational demands of articulatory synthesis, formant synthesis, concatenative synthesis and USS are considered in turn.

In order to permit the generation of high quality speech output, articulatory synthesisers must model an endless number of articulatory parameters (Docherty and Shockey, 1988). Articulatory synthesis is therefore extremely computationally expensive (Allen *et al.*, 1987; Klatt, 1987; Docherty and Shockey, 1988) in terms of both storage and processing.

Formant synthesisers, on the other hand, "can generate intelligible speech with relatively few parameters (about 40)" (Huang *et al.*, 2001: 802). Formant synthesisers therefore have lower storage requirements than articulatory synthesisers. The demands placed on processing however remain high (Allen *et al.*, 1987) because formant synthesisers must process a new set of parameters every 2-10ms (Keller, 2002).

Consisting in a database of pre-recorded segments of natural human speech, concatenative synthesis has high storage requirements (O'Shaughnessy, 1992; Olive *et al.*, 1998). The demands placed on processing are, on the other hand, low.

Consisting in a database or corpus of pre-recorded natural human speech, USS, like concatenative synthesis, has high storage requirements. The demands of USS are, however, greater than those of concatenative synthesis (Prudon *et al.*, 2002): having searched the database or corpus for all possible strings of segments, complex calculations must be carried out in order to determine which string best matches the phonetic and prosodic specifications of the utterance to be generated and leads to the least distortion at segment boundaries (see section 2.4.2.4).

### 2.5.4 Integration

Integration is achieved by means of an Application Programming Interface (API) (Huang *et al.*, 2001). An API is a set of commands which can be used by a program to call up the different functions of the TTS synthesis system. APIs are thus mark-up languages. Most TTS

synthesis vendors provide their own APIs (DISC, 1999). In order to facilitate the integration of TTS synthesis into different applications, a number of speech synthesis mark-up standards have been proposed. These include: Sun Microsystems Java Synthesis Markup Language (JSML), Microsoft Speech API (SAPI), Apple Speech Manager, Speech Synthesis Markup Language (SSML), SABLE, Voice Extensible Markup Language (VoiceXML), Java Speech API (JSAPI), and the World Wide Web Consortium's (W3C's) SSML (Huang *et al.*, 2001; Monaghan, 2002a). The goal of the W3C[17] project is to arrive at a universal standard for speech synthesis mark-up (*ibid.*).

According to Monaghan, the ideal speech synthesis mark-up language would be device-independent, that is it would have the same effect on any speech synthesiser, and would permit the application developer to specify exactly how an utterance should be pronounced, i.e. to control "pronunciation, emphasis, pitch contour, duration, amplitude, voice quality, articulatory effort, etc." (*ibid.*: 313), in an intuitive and meaningful way. While it should be possible to develop an intuitive and meaningful mark-up language that permits applications developers to specify exactly how an utterance should be pronounced, it is unlikely that the mark-up will have the same effect on all TTS synthesis systems, due to the fact that, as presented in section 2.4, they are implemented in many different ways (*ibid.*). To get the maximum control over TTS synthesis systems, use of the APIs supplied by TTS vendors is therefore recommended.

### 2.5.5 Applications

Generally, people are not prepared to listen to poor quality speech on a regular basis for long periods of time (Henton, 2002; Pitt and Edwards, 2003). Consequently, for a long time, TTS synthesis systems were only used where they presented significant advantages over other methods of generating speech output (*ibid.*), i.e. added value. In particular, they only tended to be found in applications for the blind, such as RMs, Talking Word Processors (TWPs), talking clocks, and screen readers, "software used to convert visual displays into speech" (*ibid.*: 10) in order to "make standard software accessible to blind users" (Edwards, 1991: 6), communication aids for the speech-impaired, and to deliver warning messages to pilots (Pitt and Edwards, 2003).

[17] http://www.w3.org/TR/speech-synthesis/

As a result of improvements in the quality of the speech generated by TTS synthesis systems, TTS synthesis is now acceptable for use in a much wider variety of applications. Most frequently, it is used to provide information services over the telephone, including directory assistance, travel information, movie guides, weather reports, customised news, real estate listings and stock quotes (Denes and Pinson, 1993; Rodman, 1999; Henton, 2002; Pitt and Edwards, 2003; Schroeter *et al.*, 2002). Permitting the generation of speech from text input on demand, TTS synthesis is particularly suited to the provision of information such as the weather, the news and stock quotes which are continually changing (Pitt and Edwards, 2003). Information services generally require that information is presented objectively (*ibid.*). The neutral intonation (see section 2.5.1.4) of TTS synthesis is therefore a benefit in these applications.

TTS synthesis is also now found in educational applications. In particular, it is already used in the teaching of basic skills and adult literacy and in speech therapy (Henton, 2002). The benefits of the use of TTS synthesis are that it permits unsupervised practice (Klatt, 1987), and is perceived as non-judgemental by learners because it is *not* human (Keller and Zellner-Keller, 2000; Keller, 2002). With respect to the latter point, formant synthesis is particularly suited to the creation of "robotic and other non-human sounding voices" (Henton, 2002: 124). Another area of education in which TTS synthesis has already been used is in foreign language learning. As mentioned in the introduction, this is the focus of this thesis. Applications of TTS synthesis in foreign language learning are discussed in detail in chapter 3.

These are but a few of the many commercial applications of TTS synthesis. TTS synthesis has also been used in the remote monitoring of patients (Klatt, 1987), to provide access to email over the telephone (Rodman, 1999), in proofreading applications (Henton, 2002), and in voicemail systems (Pitt and Edwards, 2003). Other suggested applications include 24-hour advice lines (*ibid.*) and talking home appliances (Schroeter *et al.*, 2002) (see references for more details on these applications).

TTS synthesis has also been used in research. Specifically, the flexibility of both articulatory and formant synthesis have been exploited to explore speech production and speech perception (Linggard, 1985; Docherty and Shockey, 1988; Bickley *et al.*, 1999; Henton, 2002; Keller, 2002).

### 2.5.6 Advantages of TTS synthesis

As mentioned in section 2.5.5, because people are not prepared to listen to poor quality speech, TTS synthesis should only be used where it presents significant advantages over other methods of providing speech output (Pitt and Edwards, 2003). The main alternative to the use of TTS synthesis is the use of digitised speech. Through the presentation of the applications of TTS synthesis in the preceding section, section 2.5.5, a number of the general advantages of TTS synthesis over digitised speech were presented, namely:

- ability to generate speech on demand,
- high degree of flexibility, and,
- low storage requirements.

In addition, it is cheaper to use TTS synthesis than to hire voice talent to make digital recordings (Henton, 2002). The fact that there is no need to hire voice talent also means that applications can be developed more quickly (Schroeter *et al.*, 2002).

As presented in section 2.5.5, TTS synthesis has specific advantages in certain applications. As presented, the specific advantage of using TTS synthesis in information services is that one can guarantee that TTS synthesis, unlike voice talent, will remain objective (Pitt and Edwards, 2003). The specific advantage of using TTS synthesis in educational applications is that it is not human and therefore perceived as non-judgemental (Keller, 2002).

## *2.6 Summary*

In this chapter, it was established that TTS synthesis systems are systems which permit the generation of novel oral utterances from unrestricted text. Regarding the architecture of such systems, it was established that there are two main modules, a TTP module and a PTS module. The goal of the former, the TTP module, is to generate a narrow phonetic transcription of the input text augmented with prosodic specifications. As established in sections 2.4.1.2 to 2.4.1.7, LTS rewrite rules are not sufficient for this purpose: many words are not written in full orthographic form, many words have more than one possible phonetic realisation, and punctuation provides few reliable clues to the prosody of utterances. The generation of a narrow phonetic transcription of input text augmented with prosodic specifications is therefore a complex task which involves many levels of linguistic analysis. At each of these levels of analysis the developer has the choice between using simple hand-written rules, fully-fledged linguistically-based grammars, or rules derived from corpora through statistical analysis. None

of these techniques is entirely robust. Errors may therefore occur at both the phonetic and prosodic levels. Regarding the evaluation of TTS synthesis systems for use in CALL applications, or any other type of application for that matter, this implies that both the pronunciation and the prosody of the speech generated by TTS synthesis systems ought to be evaluated. Regarding any errors found in the speech generated by TTS synthesis systems through evaluation, the cause of these errors may be difficult for the evaluator to diagnose because of the number of interacting levels of processing involved. Turning now to the second module, the PTS module, as established in section 2.4, the goal of this module is to convert the narrow phonetic transcription provided by the TTP module to speech output. Four main approaches to PTS conversion have been employed in TTS synthesis systems, namely: articulatory synthesis, formant synthesis, concatenative synthesis, and USS. In section 2.5.1, it was established that the choice of approach has a significant effect on the quality of the speech generated by TTS synthesis systems. Regarding the evaluation of TTS synthesis systems for use in CALL, or any other application area, this implies that every TTS synthesis that is intended for use in a CALL application ought to be evaluated. Moreover, regarding concatenative synthesis and USS, it was established that the quality of different voices and styles generated by the same TTS synthesis system may differ significantly in quality. Every voice, accent and speech style to be used in a given application may therefore need to be evaluated when using concatenative or USS. Regarding the use of TTS synthesis in applications, it was established that applications in general impose demands on the quality of the output, flexibility of the output and computational demands. Like the quality of the speech generated, it was established that the flexibility of the speech generated and the computational demands of the TTS synthesis system depended on the approach to PTS employed. In section 2.5.5, current applications in TTS synthesis were presented and their requirements with respect to the aforementioned qualities of TTS synthesis system were discussed. Whether CALL applications place the same demands on TTS synthesis systems is discussed in chapter 5. According to Sparck Jones and Galliers (1996), the function which a technology adopts in an application may have implications for evaluation. As can be seen from section 2.5.5, in most applications to date TTS synthesis is used as an RM. The functions in which TTS synthesis is used in CALL applications are discussed in section 3.4. Finally, in section 2.5.6, the benefits of TTS synthesis over other media were considered. The benefits of the use of TTS synthesis in CALL applications will be discussed in section 3.2.

# 3 TTS synthesis in CALL

## 3.1 Overview

Having established in the previous chapter what TTS synthesis is and what its advantages and limitations for use in applications in general are, we now turn to its use in CALL. More specifically, the focus of this chapter is a review of the use of TTS synthesis in CALL (see section 3.3). Regarding the evaluation of CALL applications in TTS synthesis, according to Sparck Jones and Galliers (1996), "NLP systems [, such as TTS synthesis,]... cannot in general be effectively or usefully evaluated in isolation" (*ibid.*: 11). Rather, they believe that, in addition to the system, it is necessary to consider the setting, or operational context, in which the system is used, that is the users of the system and any other apparatus that they may have available to them (*ibid.*). Moreover, where the NLP system is a component of a larger system, as is the case of TTS synthesis in CALL, they believe that it is necessary to consider the function of the NLP system within the larger system (*ibid.*). This is because they believe that the requirements imposed on NLP technologies differ according to the 'setup', that is the "system plus operational context" (*ibid.*: xiv), as in this example provided by Francis and Nusbaum (1999):

> in some situations listening to a machine may be considered more acceptable than in others [i.e. the demands placed on naturalness may not be as stringent]. For example, when requesting information or assistance, such as when asking for directions or computer support, users may prefer to listen to a human-sounding voice, under the assumption that more mechanical sounding voices are less likely to "know" how to solve the problem at hand. In contrast ... when listening to sensitive personal information such as a bank balance, users may prefer to hear a less natural voice because they want to be sure that the information remains private (*ibid.*: 79).

The dimensions of CALL setups in which TTS synthesis is being and has been suggested for use are therefore considered in section 3.3.1.5.

First, however, in section 3.2, we believe that it is appropriate to consider the motivation for the research presented in this thesis in more detail, specifically the benefits that TTS synthesis is proposed to bring to CALL.

## 3.2 Benefits of the use of TTS synthesis in CALL applications

As presented in the introduction, Sherwood (1981) was the first to identify potential benefits and uses of TTS synthesis in CALL. Regarding the benefits of its use in CALL more specifically, Sherwood observed that it is easier to create, edit and navigate through speech models created through the use of TTS synthesis than those created through the use of a cassette-recorder, the dominant speech output medium of the time, and TTS synthesis, unlike the cassette-recorder, permits, as presented in section 2.5.6, the generation of speech models on demand.

Since Sherwood, both TTS synthesis specialists and CALL developers have seen further benefits in the use of TTS synthesis in CALL. According to Ahmad *et al.*:

> It is helpful to divide the advantages of the computer into three types: those which are part of its inherent nature, those which benefit the teacher, and those which benefit the learner (*ibid.*: 4).

Within this framework, the benefits identified by Sherwood (1981) are examples of benefits which result from the inherent nature of TTS synthesis. Further benefits which result from the inherent nature of TTS synthesis have been identified by others. Firstly, it has been suggested that learners may perceive it to be non-judgemental because it is *not* human (Keller and Zellner-Keller, 2000). Secondly, it has been suggested that the inherent nature of TTS synthesis could be exploited in order to generate speech models which may promote SLA. In particular, it has been suggested that the prosodic quality of speech generated by TTS synthesis may be particularly suitable as a model for teaching the intonation of the TL:

> The advantage of synthesized intonation contours compared to those realized by the speaker is that their melodic pattern is more formalized. Thus the generalized contours free from additional emotional colouring are used as models (Skrelin and Volskaya, 1998: 24).

Specifically, Skrelin and Volskaya appear to be suggesting that the lack of variability characteristic of the speech generated by TTS synthesis allows learners to focus more easily on the aspect of the intonation of the TL being taught.

Also regarding the generation of speech models, it has been suggested that the inherent manipulability of TTS synthesis could be exploited to generate different types of modified

input which may facilitate apperception, a process which is believed to promote acquisition Gass (1997), by making aspects of the TL more salient, or noticeable (see sections 5.3 and 5.3.2). These include: slower speech (Bonneau *et al.*, 2000; Keller and Zellner-Keller, 2000), enhanced speech (Bonneau *et al.*, 2000), exaggerated speech (*ibid.*; Keller and Zellner-Keller, 2000), "speech with intonation, but no rhythm" (*ibid.*: 110), 'synthetic humming' Yoram and Hirose (1996), speech without content, i.e. segmental information,[18] and contrastive examples (Germain Rutherford, 2001).

In addition, it has been suggested that the manipulability of TTS synthesis could be exploited in order to generate artificial phonological models which might be useful for language teaching, such as the model for international English proposed by Jenkins (2000) (Hincks, 2002):

> a 'lingua franca core' in which features from RP [Received Pronunciation], GA [General American] and L2 varieties of English have been selected for their practicality in functioning as features that can easily be taught and learnt, perceived and produced (*ibid.*: 4).

Other benefits resulting from the inherent nature of TTS synthesis which have been put forward include: the low storage requirements of TTS synthesis in comparison with digitised speech (Pennington and Esling, 1996), the fact that TTS synthesis frees up space on screen for the presentation of other media, (Skrelin and Volskaya, 1998), and the fact that TTS synthesis permits the simultaneous and synchronised presentation of text and speech (Pennington and Esling, 1996).

The benefits of being able to produce of new types of speech models and present text and speech simultaneously and in synchronisation are also teacher-oriented benefits because these features provide teachers with the means to produce new types of CALL activities and exercises. Similarly, some of the benefits identified by Sherwood (1981), presented at the beginning of this section, namely the ease with which speech models can be created, edited and navigated using TTS synthesis, are also teacher-oriented benefits because they facilitate the authoring of CALL activities and exercises.

[18] "Synthetic humming is created by summing sinusoid waveforms, using a given pitch and intensity specification. For prosodic training, this can help remove some of the irrelevant details, by removing segmental data" (Yoram and Hirose, 1996: 1452).

Similarly, a number of the benefits of the inherent nature of TTS synthesis are also learner-oriented benefits. These include the fact that TTS synthesis is perceived as non-judgemental by learners, its low storage requirements and its ability to generate speech models on demand. Regarding the low storage requirements of TTS synthesis, these mean that both more speech output and other media, such as graphics and video clips, can be stored on the computer and delivered to the learner in CALL applications (Hart *et al.*, 1988; Esling, 1992; Skrelin and Volskaya, 1998). This is advantageous to the learner because exposure to large quantities of TL input is believed to promote acquisition (Gass, 1997; see sections 5.3 and 5.3.1). The ability to generate speech models on demand, identified by Sherwood (1981), is also of benefit to the learner. Specifically, it makes the provision of individualised aural feedback on demand in CALL applications possible.

In addition to those resulting from the inherent nature of TTS synthesis, a number of other learner-oriented benefits of the use of TTS synthesis in CALL have been identified. Firstly, it has been suggested that TTS synthesis will make it possible to make language learning more accessible to the blind (Gray, 1984). Secondly, it has been suggested that TTS synthesis will make it possible to provide learners who are not literate in their L1 explanations in their L1 to support their learning (*ibid.*). And, finally, it has been suggested that TTS synthesis has "a novelty value and may stimulate play, chat, or discussion around the computer" (Pennington and Esling, 1996: 157).

According to Stevens (1989), computers:

> are best exploited in the ways that take advantage of their particular characteristics rather than when they are used to try to "improve" deliveries in the media they seem to be replacing (*ibid.*: 33).

Of the aforementioned benefits of the use of TTS synthesis, the ability to generate the following types of speech models is particular to TTS synthesis: enhanced speech (Bonneau *et al.*, 2000), "speech with intonation, but no rhythm" (Keller and Zellner-Keller, 2000: 110), 'synthetic humming' Yoram and Hirose (1996), and speech without content, i.e. segmental information. Another benefit particular to TTS synthesis is the ability to generate speech models on demand.

Regarding the other suggested advantages of the use of speech synthesis in CALL applications, other technologies already exploited in CALL applications permit the synchronised presentation of text and speech:

> The computer makes possible presentation of speech coordinated with written text of with other visuals on a computer screen. The simplest system is one which synchronizes the recording and / or the playing of audiotaped speech with computer text. Such systems have been used in instructional contexts for training or testing purposes to coordinate spoken productions of lexical items, sentences, or discourses with their written transcriptions appearing on the computer screen. Common uses of this capability are for listening cloze and audio reinforcement of vocabulary or illustrations of grammatical structures (Pennington and Esling, 1996: 157).

And, the remaining suggested benefits merely *improve* what is possible with existing technologies:

- the fact that speech synthesis is non-human and therefore perceived as non-judgemental *reduces* learner anxiety;
- the fact that speech can be produced from text input makes the construction, and editing of examples *easier* for the teacher;
- the low storage requirements of speech synthesis *increase* the number of exercises and the range of media that it is possible to provide; and,
- the novelty value of speech synthesis *increases* learner interest and motivation.

Moreover, the low storage requirements and novelty value are only likely to be beneficial in the short term; the storage capacity of computers is constantly increasing and the novelty value will no doubt wear off.

The next sections look at how the benefits of TTS synthesis presented in this section are being and could be exploited in CALL through a review of the uses of TTS synthesis in CALL. In particular, one of the aims of this section is to identify the best applications of TTS synthesis in CALL, that is, the applications that exploit the benefits which are particular to TTS synthesis identified in this section. In addition, this review will also serve to expose some of the dimensions of CALL setups integrating TTS synthesis.

## 3.3 Uses of TTS synthesis in CALL

As presented in the overview of this chapter, the setup in which an NLP system is used is believed to have implications for its evaluation. It therefore seems appropriate to present the

actual and suggested uses of TTS synthesis in CALL according to dimensions of their setups. One dimension of NLP setups in general, as presented in section 2.1, is the type of system into which the NLP systems are embedded. CALL systems in general have been classified in a number of different ways. Most commonly they are classified according to "the relationship between and roles assumed by learner and computer" (Wyatt, 1988: 85) (see Davies and Higgins (1985), Jones and Fortescue (1987), Hardisty and Windeat (1989), de Quincey (1986), Warschauer (1996), and Colpaert and Decoo (1999), and Bax (2003) for other possible classifications). Several classifications of this type, referred to as 'relational' classifications (Wyatt, 1988), have been proposed (see Taylor (1980), Phillips (1987), Higgins (1983), Wyatt (1988), and Levy (1997)). As Levy (1997) highlights, the classes 'tutor' and 'tool' are common to most of these classifications, though they go by different names and have slightly different scopes.[19] Tutors, according to Taylor (1980), who first used the term in CALL, "*evaluates* the learner, and then proceeds on this basis" (Levy, 1997: 180). In other words, tutors are an attempt to "emulate or replace the teacher" (*ibid.*: 184). Tools, on the other hand, do not evaluate the learner (*ibid.*). Rather, their function is to "enhance or improve the [quality and] efficiency of the work of the teacher or student" (*ibid.*: 184; author's addition). TTS synthesis has been suggested for use in both CALL tutors and tools. These systems will be presented in sections 3.3.1 and 3.3.3, respectively. In addition, TTS synthesis is being used in systems which do not clearly fit into these two classifications. These systems, presented in section 3.3.2, we believe, more clearly fit the class of 'stimulus' (Taylor and Perez, 1989).

### 3.3.1 Tutors integrating TTS synthesis

Another dimension of CALL setups is the aspect of language proficiency on which the systems focus:

> The prevailing view held that language proficiency could be divided into unrelated skills (listening, speaking, reading and writing) and knowledge of language components (vocabulary, phonology and grammar) (Larsen-Freeman and Long, 1991: 38)

It has been suggested that TTS synthesis could be used in tutors focusing on each of the skills that comprise language proficiency. Those focusing on reading are presented in section 3.3.1.1, those focusing on writing in section 3.3.1.2, those focusing on listening in section

[19] Phillips' (1987) 'expert system', Higgins' (1988) 'magister', and Wyatt's (1988) 'instructor' are very similar to Taylor's (1980) 'tutor' and Phillips' 'prosthetic model', Higgins' 'pedagogue', and Wyatt's 'facilitator' are very similar to Taylor's 'tool'.

3.3.1.3, and those focusing on speaking in section 3.3.1.4. Of the language components that comprise language proficiency, TTS synthesis has been used in tutors focusing on grammar. These applications are presented in section 3.3.1.5.

In addition, it has been suggested that TTS synthesis could be used in CALL tutors designed to teach learners phonetic transcription, a skill which it is believed can be helpful in learning foreign languages (Underhill, 1985; Tench, 1992). These tutors are presented in section 3.3.1.5.

### 3.3.1.1 TTS synthesis for teaching reading

The ability to match letters, or graphemes, to the speech sounds that they represent is believed by some, specifically those who subscribe to the 'phonics' approach, to be a prerequisite to being able to read (Wallace, 1992). The CALL tutor *Système d'Apprentissage du FRANçais (SAFRAN)* (Hamel, 2003a) uses the TTS synthesis system *FIPSvox* (Guadinat and Wehrli, 1997) to provide exercises dedicated to the development of this skill. Specifically, it exploits TTS synthesis for the provision of feedback in the following activities designed to reinforce the grapheme-to-phoneme relationship in French:

- *charivari* (anagrams),
- *virelanguse* (tongue twisters), and,
- *karaoké* (karaoke).

In the case of the anagrams, the learner's task is to find the solution, record their pronunciation of it, and then compare their production with the model provided by the TTS synthesis system. In the case of the tongue twisters, the learners were asked to record their attempts to pronounce them and to compare their productions with the models provided by the TTS synthesis system. And finally, in the case of the karaoke, the learners were asked to record their attempts to sing them and to compare their productions with the models provided by the TTS synthesis system. A limitation of the use of *FIPSvox* for the provision of feedback in the karaoke activities was that it could not sing, and still cannot (Hamel, 2003a). Most TTS synthesis systems in fact cannot sing. One, and perhaps the most famous, example of a speech synthesis system that sings is the *LYRICOS* system developed at the Georgia Institute of Technology (Macon *et al.*, 1997).

Regarding the question of whether TTS synthesis adds value to this CALL application, in our view it does not because the speech models to be generated are known in advance and could therefore be provided through the use of digital recordings of native speakers.

### 3.3.1.2 TTS synthesis for teaching writing

The ability to match speech sounds to letters is believed to be a prerequisite to being able to write, in particular if learners were taught to speak and listen in the TL without recourse to the written language as is recommended by some (Dabène, 1974). The CALL system *SAFRAN* (Hamel, 2003a) also uses TTS synthesis to provide exercises dedicated to the development of this skill. Specifically, it exploits TTS synthesis for the provision of oral stimuli in the following activities:

- *blancs de mémoire* (gap-filling (or cloze) exercises),
- *pendu* (hangman),
- *mot-mystère* (word search),
- *homophones* (homophones), and,
- *dictées* (dictation).

The gap-filling exercises are a variation on traditional gap-filling exercises in which learners are presented with a text with a number of words (or letters) blanked out, and their task is to reconstitute the original text. The gap-filling exercises proposed to the learner in *SAFRAN* differ from traditional gap-filling exercises in that all the gaps correspond to the same phoneme but different graphemes. The TTS synthesis system is used to present the phonemes to the learner, whose task is to use their knowledge of the phoneme-to-grapheme correspondences of the TL in order to fill in the graphemes.

The hangman exercises proposed to the learner in *SAFRAN* differ from traditional hangman exercises in that learners are given a clue to the identity of the word, more specifically they are told that it contains a particular phoneme. This phoneme is presented to the learner both graphically via phonetic transcription and auditorily through the use of the TTS synthesis system. Learners are expected to use their knowledge of the phoneme-to-grapheme correspondences in the target language in order to determine the identity of the word.

The word searches are a variation on traditional word searches in which the learner is presented with a grid of letters and a list of words which can be found in that grid. The hidden

word exercises proposed by *SAFRAN* differ from traditional word searches in that the learner is given a list of phonemes. The learner's task is to discover which word remains when they have crossed off all possible orthographic correspondences of the list of phonemes that they were given. In order to help them determine the identity of the hidden word, they are told that it contains a certain phoneme. This phoneme and the list of phonemes whose orthographic correspondences are to be found in the grid are presented to the learner both auditorily using the TTS synthesis system and graphically using phonetic transcription.

In the homophones exercises, the TTS synthesis system is used to present homophones to the learner. The learner's task is to write down all the possible orthographic forms of the homophones.

Finally, in the dictation exercises, the TTS synthesis system was used to read aloud the texts that were to be transcribed by the learners.

Of the exercises presented above, dictation is quite a popular application of TTS synthesis in CALL. Other examples of dictation tutors developed for use in LL&T can be found in Sherwood (1981) and Mercier *et al.* (2000) both of which are in our opinion particularly interesting because they both exploit the unique capabilities of TTS synthesis. Sherwood exploited the capacity of TTS synthesis to generate speech models on demand to provide learners with individualised feedback in single word dictation exercises. For example, if a learner had been presented the stimulus 'sonis' and transcribed it as 'sanus', they received the following feedback 'not *sanus*, *sonis*' (*ibid.*: 179) – Esperanto was the object of instruction in Sherwood's system. Mercier *et al.* (2000), on the other hand, exploited the manipulability of TTS synthesis in order to dynamically adapt the speed of delivery of the dictation to the rate at which the learners typed their responses, in other words to the ability of the individual learner. Exploiting the unique features of TTS synthesis systems, TTS synthesis is believed to add value to these applications.

### 3.3.1.3 TTS synthesis for teaching listening

Traditionally, listening was taught by presenting learners with a passage to listen to and then asking them to answer questions and complete activities based on that passage (Anderson and Lynch, 1998). TTS synthesis is being used to present activities of this type in software

developed by the Virtual Learning Center (VLC), Hong Kong.[20] The activities which VLC proposes, *Storyboard*, *Cloze* and *Jumbler* games, are variants on the following exercises found in traditional text-based CALL software, respectively: storyboard exercises, gap-filling exercises (or cloze tests), and text-unscrambling exercises. In traditional storyboard exercises, learners are given a text to read and then presented with the same text with all the words blanked out. In traditional gap-filling exercises, learners are given a text to read and then presented with the same text with a number of words blanked out. In traditional text unscrambling exercises, learners are given a text to read and are then presented with sections (words, sentences, or paragraphs) of the text in randomised order. In the VLC versions of these exercises the texts are presented orally to the learner using a TTS synthesis system, as opposed to visually. As in the traditional versions of the exercises, the learners' task in all of the activities is to reconstruct the original texts.

While many attempt to teach listening in this way, for many others this is not teaching listening it is testing listening (Anderson and Lynch, 1998). Regarding the actual teaching of listening, several different approaches have evolved from the different models of speech perception, listening and SLA (Rost, 2001). One of these approaches contends that listening tasks consist in a hierarchy of goals (see Table 3) and that in order to achieve the first-order goals, learners must automatise 'lower-level' processing so that they have enough Short Term Memory (STM) to be able to devote attention to higher-order goals (*ibid.*).

[20] http://www.edict.com.hk/vlc/default.htm

**Table 3 Hierarchical structure of listening (adapted from Rost, 2001: 110)**

| | |
|---|---|
| **First-order goal:** | respond to relevant aspects of what is heard |
| **Second-order goal:** | establish appropriate connection with speaker or content<br>activate relevant knowledge to understand speakers and topic<br>understand social meaning of input (including speaker's intentions) |
| **Third-order goals:** | understand gist of input<br>understand cohesion between utterances<br>understand words and structures<br>understand pragmatic conventions |
| **Lower-order goals:** | understand sounds speaker uses |

"regular targeted practice" (*ibid.*: 111) is believed to be necessary in order to gain automatic control over lower-level processing, which in the case of listening consists in the ability to identify and discriminate phonemes, stress patterns, intonation patterns, etc. As far as it is possible to establish TTS synthesis has been used in two CALL tutors which provide such practice at the phonetic level.

In addition to exercises focusing on the grapheme-phoneme and phoneme-grapheme relationships in French, *SAFRAN* (Hamel, 2003a) also uses TTS synthesis for the provision of stimuli in exercises designed to develop learners' proficiency in phoneme identification and discrimination. More specifically, *SAFexo*, *SAFRAN's* phonetic training system, uses TTS synthesis to deliver stimuli in four types auditory discrimination and identification exercises, namely: *Identique ou Différent ?* (Same or different?), *Présent ou Absent ?* (Present or absent?), *Où ?* (Where?), and *Combien ?* (How many?). In the first of these four exercise types, *Identique ou Différent ?*, learners are presented two stimuli and asked to identify whether they were different or not. In the second, *Présent ou absent?*, learners are presented a stimulus and asked to identify whether it contained a specific phoneme or not. In the third, *Où?*, learners are presented a stimulus and asked to identify which syllable of the stimulus (the 1st, 2nd, 3rd , etc) a particular phoneme appears in. And finally, in the fourth type, *Combien?*, learners are presented a stimulus and asked to identify how many instances of a particular phoneme it contains.

TTS synthesis was also used by Sherwood (1981) for the provision of auditory discrimination exercises in Esperanto. Specifically, it was used to present both stimuli and feedback in

minimal pair exercises, exercises in which learners are presented two or more words whose pronunciations differ in only one phoneme and are asked to indicate which one they heard.

Regarding the question of whether the use of TTS synthesis adds value to these applications, one might be tempted to consider that Sherwood's use of TTS synthesis for the provision of feedback is a smart use of TTS synthesis. In our opinion, on the other hand, it is not. The speech models to be generated are known in advance and therefore could be provided through the use of digital recordings of native speakers. The same, we believe, is true of the stimuli in the other activities presented in this section.

### 3.3.1.4 TTS synthesis for teaching speaking

Like listening, it has been suggested that speaking consists in a hierarchy of goals. The first of these goals, according to Levelt (1989) is to conceptualise the utterance to be spoken, that is to decide on the purpose or communicative intention of the utterance and select and structure the information that must be conveyed in order to express that intention. The second goal is to formulate an articulatory plan for the utterance, that is to retrieve the necessary lexical items, structure them and assign them grammatical and phonological features. Finally, the third goal is to execute the articulatory plan, i.e. to articulate the utterance. According to Levelt, native speakers have automatic control over articulation and the processes involved in forming the articulatory plan. Regarding automatic control, it is generally believed that automatic control over TL forms can be achieved through practice (McLaughlin, 1987). Regarding the acquisition of automatic control over articulation, TTS synthesis has also been used in one CALL tutor for the provision of practice in articulation. In this tutor, *SAFRAN* (Hamel, 2003a), TTS synthesis is used to present models for imitation in the following types of exercise focusing on the pronunciation of individual phonemes: *Échauffement* (warm-up), *Entraînement* (training). In the both the *Échauffement* and *Entraînement* exercises, TTS synthesis was used as a model to imitate. In the *Échauffement* exercises, a phoneme was presented "first on its own, then in mono and pluri-syllabic contexts and finally in logatoms [*sic*]"[21] (*ibid.*: 243). In the *Entraînement* exercises, which followed the *Échauffement* exercises, "the same phoneme was opposed to another one in minimal pairs and pluri-syllabic contexts" (*loc. cit.*). In a final type of exercise, *Épreuve* (test), in which learners were presented the same phoneme in a range of contexts and asked to replace it with a contrasting

[21] 'Logatom' is the technical term for nonsense word.

phoneme, TTS synthesis was used to provide the learner feedback as well as a pronunciation model.

Regarding the question of whether TTS synthesis adds value to this application, the models and feedback to be generated are known in advance. We therefore do not think that TTS synthesis adds value in this application.

Also regarding the value of this application, in all exercises the learners are expected to diagnose their own errors by comparing their attempts with the pronunciation models and feedback provided by the TTS synthesis system. While some believe that self-diagnosis is effective and that learners prefer it over other types of feedback, in particular certain types of visual displays (Reeser, 2002), it is argued that self-diagnosis is limited by the learner's perception (Pennington and Esling, 1996; Sioufi, 2000). For example, on the use of this same approach in audiolingual language laboratories Pennington and Esling (1996) observe:

> If students do not perceive the difference between their performance and that of the model on tape, they will simply reproduce the same output in response to the tape on every repetition. Thus, although they might practice for hours on end, they never improve because they fail to perceive the gap between their production and that of the native speaker voice on tape (*ibid.*: 170).

TTS synthesis was also used for this purpose by Stratil *et al.* (1987a) in a language laboratory setting. An evaluation of learners' reactions to the use of TTS synthesis for this purpose in a language laboratory setting is presented in section 4.7.2.3.

### 3.3.1.5 TTS synthesis for teaching grammar

Two examples of the use of TTS synthesis for the presentation of grammar exercises were found in the literature. In the first example discussed in Stratil *et al.* (1987a; 1987b), TTS synthesis was used to present Spanish grammar exercises from the learners' course books in a language laboratory setting. Optical character recognition was used to generate the exercises. Evaluations of this CALL application that were carried out by Stratil *et al.* (1987a) and Stratil *et al.* (1987b) are presented in sections 4.7.2.3 and 4.7.1, respectively. In the second example, the *Appeal (A Pleasant Personal Environment for Adaptive Learning)* system for teaching Dutch as a foreign language (de Pijper, 1997), TTS synthesis was used to present individualised grammar exercises which were "created on the fly, according to predefined models" (*ibid.*: 581).

Such systems would be of particular benefit to auditory learners, learners who have a preference for auditory input such as lectures and audiotapes (Brown, 1994), but also to learners in general; according to Coleman (1991), learners in general retain more if they both hear and read the material they are intended to learn:

> One formula states that trainees retain 25% of what they hear, 45% of what they see, and 70% of what they see, hear and do. Another source suggests we remember 10% of what we read, 20% of what we hear, 30% of what we see, 50% of what we see and hear, 80% of what we say, and 90% of what we say and do at the same time (*ibid.*: 101).

While the use of TTS synthesis in the first example does lighten the load on the teacher by automating the creation of the exercises, in our opinion the use of TTS synthesis does not, however, add value. The speech models to be created were known in advance, consequently digital recordings of native speakers could have been made. The second example, on the other hand exploits the unique capacity of TTS synthesis to generate speech models on demand. We therefore believe that the use of TTS synthesis adds value in this application.

### 3.3.1.6 TTS synthesis for teaching phonetic transcription

Knowledge of phonetic transcription, according to Underhill (1985), is useful to language learners, in particular those learning languages such as French and English where "there is a huge discrepancy between pronunciation and spelling" (Tench, 1992: 97), because it provides them with:

> 1 the ability to find the pronunciation and stress of any word in the dictionary;
> 2 the ability to record in their own handwriting the pronunciation and stress of new words, phrases etc.; and
> 3 the ability to objectify the string of sounds contained in a word and to study the sequences and clusters
> (*loc. cit.*).

Moreover, if learners are familiar with phonetic transcription, teachers can use it in pronunciation training. The benefit of using phonetic transcription in pronunciation training, according to Tench, is that it can draw learners' attention to aspects of pronunciation which they might not otherwise notice. This, however, is not sufficient to improve learners' pronunciation (*ibid.*). In order to improve pronunciation, "A good spoken model is required" (*ibid.*: 100). This is where TTS synthesis comes in. According to Knowles (1986), this model

could be provided through the use of a good quality TTS synthesis system. Moreover, if the system accepted phonetic symbols as output, learners could hear their own attempts at phonetic transcription read aloud, i.e. receive feedback on their attempts. This according to Knowles, would be particularly useful for making learners aware of the difference between phonemic, or broad, transcription, and phonetic, or narrow, transcription: "The reality of the distinction should quickly become clear: a word synthesised with the wrong allophone is likely to sound odd or unintelligible" (*ibid.*: 146). Moreover, TTS synthesis, according to Knowles permits the presentation of more examples of contextually conditioned allophonic variation than traditional methods:

> By conventional methods, it is difficult to give more than a few cursory examples or this, e.g. the distribution of clear and dark /l/, or the aspiration of /p,t,k/, and the point of making the distinction probably eludes most beginners. As part of a speech synthesis program, beginners can see as many examples as they want (*ibid.*: 146).

### 3.3.2 TTS as a stimulus

Regarding the use of TTS synthesis as a stimulus, a couple of researchers, one from the field of CALL (Hincks, 2002) and one from the field of language teaching more generally (Kenworthy, 1987), have suggested that TTS synthesis could be used as a stimulus to draw learner's attention to features of the pronunciation of the TL, a condition which is believed to promote SLA (Gass, 1997; see sections 5.3 and 5.3.2). Specifically, Kenworthy (1987) suggests dividing learners into groups, giving each a few utterances in the TL generated by a TTS synthesis system accompanied by descriptions of the contexts in which they were uttered, and asking them to go away and discuss any problems that they notice in the output and how they might modify the speech in order to eliminate them and then come back and present their findings to the whole class. As a follow up exercise, she suggests that learners could be asked to: "(1) mark a written version to show the points they would change, ... (2) record the sentence themselves using a more natural pronunciation" (*ibid.*: 81), or (3) incorporate the modifications that they have suggested to the utterances generated by the TTS synthesis system using speech editing software.

This, we believe, is a smart use of TTS synthesis because it exploits unique features of TTS synthesis. It, however, relies on the fact that the speech generated is imperfect. As presented in section 2.5.1.4, pronunciation at the segmental level is to a certain extent given in state-of-the-art TTS synthesis systems, i.e. those based on concatenative and unit selection synthesis, and

the prosody has significantly improved in particular in systems based on USS (see section 2.5.1.5). It is therefore questionable whether today's state-of-the-art TTS synthesis systems would be suitable for use in such an activity. Systems based on articulatory and formant synthesis (see sections 2.5.1.2 and 2.5.1.3), on the other hand, generate less perfect speech. They might, therefore, still be suitable for use in activities of this type.

Hincks (2002), for her part, suggested that TTS synthesis in conjunction with a speech editor could be used as a stimulus to draw learners' attention to "pitch and duration and how they differ between cognates" (*ibid.*: 153) in Swedish and English. Specifically, she suggested asking learners to synthesise a Swedish noun which has a cognate in English, such as *parameter*; asking them to think about how it differs from its English cognate; and then, guiding them to make the modifications necessary to make the cognate sound English using the speech editor.

While this is a very innovative CALL activity, we do not believe that TTS synthesis adds value to it; learners could, in our opinion, make a digital recording of themselves producing the Swedish noun. Moreover, if learners were to use a digital recording of themselves pronouncing the cognate as the starting point, they would be able to hear themselves pronouncing the TL word correctly, a type of feedback, which as presented in section 2.2, is believed to promote apperception, or noticing, of features of the TL. An evaluation of the effectiveness of this activity is presented in section 4.7.2.1.

It has long been recognised that conversation plays an important role in SLA, not only as a "medium for practice" (Gass, 1997: 104), but also as a "means by which learning takes place" (*loc. cit.*). CALL applications therefore ought to be able to provide learners with opportunities to engage in conversation in the TL. According to Egan and LaRocca (2000), the capacity of TTS synthesis to generate any utterance on demand could be harnessed to generate stimuli for interactive conversations. More recently a system that exploits TTS synthesis for this purpose has been developed, namely *SCILL* (Spoken Conversational Interaction for Language Learning) (Seneff *et al.* 2004). This system for English and Mandarin proposes learners a conversation-based activity that consists in three stages. In the first stage, the aim of which is to familiarise learners with the vocabulary and syntax of a given topic, learners examine a simulated conversation on that topic. In the second stage learners are given the opportunity to

practice the TL forms that they have been introduced to by engaging in an interactive conversation with the system on that topic. In the final stage learners examine the conversation that they had with the system. This, we believe, is another smart use of TTS synthesis in CALL as it exploits one of the unique features of TTS synthesis to provide an activity that could not be provided through the use of other media; interactive conversations could not be provided through the use of digital recordings of native speakers because of the unpredictability of conversations. The conversations in which learners can engage in this system are, however, restricted to a number of topics, namely: booking a hotel, booking a flight, inquiring about the weather, making an acquaintance, getting around a city, and asking for an alarm call (*ibid.*).

### 3.3.3 Tools integrating TTS synthesis

Regarding tools, it has been suggested that TTS synthesis could be exploited to provide learners with three main tools, namely talking dictionaries and talking texts. The benefits and potential uses of these tools in LL&T are presented in sections 3.3.3.1 and 3.3.3.2 respectively.

#### 3.3.3.1 Talking dictionaries

'Talking dictionaries' are electronic dictionaries which integrate TTS synthesis for the presentation of dictionary entries. They are the most popular application of speech synthesis in CALL. One possible reason for this is that they are one of the easiest applications to develop – if the aim is only to provide the pronunciation models of headwords, the TTP module (see section 2.4.1) can be by-passed and phonetic transcriptions used as input to the TTS synthesis system, if it accepts them, as in Mercier et al's (2000) Breton talking dictionary. Another possible reason for this is that there is more funding available for the development of such CALL applications, as they have a wider market, that is, they are also a useful tool for native speakers (Levy, 1999b). The benefit of talking dictionaries for learners and native speakers, is that they overcome the problem of trying to interpret phonetic transcriptions (L'Haire 2000); even if users are familiar with phonetic transcription, they may encounter problems because there are several different sets of phonetic symbols in use in addition to the International Phonetic Alphabet (IPA) (Knowles, 1986). Moreover, there is no standard set of IPA symbols for transcribing a given language (*ibid.*):

> Contrary to the widespread belief, these is no such thing as an 'IPA' transcription for RP. Scholars are free to choose their own symbols from the International Phonetic Alphabet: and different scholars make different choices (*ibid.*: 136).

It was in fact already possible to produce talking dictionaries through the integration of digital recordings of human speakers (see Myers (2000), for example). However, typically only one pronunciation model per entry, typically the headword, is provided in such systems. The generative power of TTS synthesis, on the other hand, makes it possible to provide learners with pronunciation models for all of the forms of a word's paradigm, its derivations, synonyms and antonyms as well as to present its definition and examples of its usage orally (L'Haire, 2000: 44). According to L'Haire, the possibility to provide learners with pronunciation for all the forms of a word's paradigm could also be exploited to provide learners with talking conjugators. As far as it is possible to establish no examples of such tools have been developed. Talking dictionaries, on the other hand, have been developed for a wide range of different languages. *Ectaco*[22] alone has developed talking dictionaries, which it also refers to as 'talking phrase books', for numerous languages including both popular languages, such as English, French, German, Spanish and Italian, and less widely spoken languages, such as Albanian, Armenian, Azeri (or Azerbaijani), Bulgarian, Croatian, Hungarian, Romanian, and Serbian.

Regarding the question of whether TTS synthesis adds value to talking dictionaries, few of the many talking dictionaries on the market exploit the unique capacities of TTS synthesis. Two examples which do are Mercier *et al.*'s (2000) talking Breton dictionary, and the *Oxford Hachette French Dictionary on CD-ROM* (Oxford Hachette, 2003). Mercier *et al.*'s dictionary exploits both the capacity to generate pronunciation models for all of the forms of a word's paradigm and to read aloud examples of usage. The former capacity is particularly useful because the several dialects of Breton. The *Oxford Hachette French Dictionary on CD-ROM*, for its part, exploits the capacity of TTS synthesis to generate speech models on demand to read aloud both definitions and examples of usage in addition to the headwords. This talking dictionary also includes a talking text facility. Talking texts, as said, are the subject of the next section.

[22] http://www.ectaco.co.uk/

### 3.3.3.2 Talking texts

The term 'talking text' is used here to refer to a tool, which will read aloud any section of text (a single word, a sentence, a paragraph, etc.) typed or copied into it from either the CALL application or an external source such as a Web page (Hamel 2003a; 2003b). Talking texts like talking dictionaries are quite a popular CALL application. We believe that this is because they are relatively easy to develop – the talking text facility provided in the *Oxford-Hachette French Dictionary on CD-ROM* hardly differs from the demonstrations of TTS synthesis systems that you find on-line – and like talking dictionaries, they are also useful to native speakers and hence funding is also available for their development. Further examples of CALL applications that integrate talking text facilities are: *Spanish for Business Professionals* produced by L. Kirk Hagen of the University of Houston-Downtown (Andersen, 2000), *FreeText* (Hamel, 2003b), and *Filoglossia+* produced by the Institute for Language and Speech Processing, Athens, Greece.[23]

*Spanish for Business Professionals* consists in a number of units of exercises each centred around a dialogue. It integrates *Macintosh's* Spanish-to-Speech and English-to-Speech programs, i.e. Spanish and English TTS synthesis programs (Anderson, 2000). The learner can call up the English TTS synthesis program to listen to the English translations of the dialogues which are spoken by native speakers (*ibid.*). And, the Spanish TTS program can be called up to listen to other participants' messages in a chat room and sentences that the learner has produced in free production exercises (*ibid.*).

Talking Word Processors (TWPs), an extension of the talking text paradigm, are another tool which are being used in LL&T. One example which has been used in LL&T is *Composition* produced by the Learning Group at the Centre Mondial Resources Humaines (Global Centre for Human Resources) in Paris (Cohen, 1993). This TWP allowed learners to create pictures "by choosing words from a list offered by the program ... or by typing them" (*ibid.*: 26) and then to write stories about the pictures that they had created. Regarding speech output, learners could hear every letter and every word as well as the whole text that they typed read aloud by the TTS synthesis system (*Ferma F 5000*) that was used in the software. An evaluation of the use of *Composition* by young learners of French as an L2 is presented in section 4.7.2.2.

[23] http://www.ilsp.gr/filoglossia_plus_eng.html

In addition, it is possible to turn standard word processors such as *Microsoft Word* into TWPs by installing an add-in. An example of such an add-in which has been marketed for use by language learners is *WordPilot 2002* from Compulang.[24]

Regarding the potential uses of these technologies in LL&T, it has been suggested that talking texts may be useful for on-line reading (Moisa and Ontanu, 1999). In this context, it is suggested that talking texts might reduce the distraction caused by the many links and colours typical of hypermedia environments (*ibid.*). This benefit, we believe, applies generally to any CALL application presented on-line. Regarding the use of talking texts to support writing, a number of benefits were identified through the evaluation of *Composition* mentioned earlier in this section. This evaluation, as said, is reported in section 4.7.2.2.

## 3.4 Dimensions of CALL setups integrating TTS synthesis

As presented in section 2.1, there are two dimensions to setups in general, the system and the setting, or operational context (Sparck Jones and Galliers, 1996). So far two dimensions of CALL systems which may have implications for evaluation have been presented namely the type of system, tutor, stimulus or tool, and the aspect of language proficiency on which they focus. A further dimension of systems in general is the function or role that the NLP system assumes within the system (*ibid.*). TTS synthesis systems are, in our opinion, being used in three main functions in CALL applications: in additions to being used as an *RM* to read out texts for learners to listen to, as it is in applications in TTS synthesis in general (see section 2.6), it is being used as a *pronunciation model* (PM) for learners to imitate, and a *conversational partner* (CP). In Table 4 the applications presented in section 3.3 are classified according to the role that TTS synthesis is believed to assume within them. A final dimension is the mode of oral communication that is to be simulated in the application.

[24] http://www.compulang.com

**Table 4 Classification of CALL applications integrating TTS synthesis according to the role that it assumes within them**

| RM | PM | CP |
|---|---|---|
| Grapheme-to-phoneme exercises<br>Phoneme-to-grapheme exercises<br>Dictation exercises<br>Listening comprehension exercises<br>Auditory discrimination exercises<br>Grammar exercises<br>Talking dictionaries<br>Talking conjugators<br>Talking texts<br>TWPs | Pronunciation drills<br>Phonetics training<br>Pronunciation stimuli | Conversation systems |

All of these different dimensions, it is believed, may have an effect on the requirements imposed on TTS synthesis and hence implications for the evaluation of CALL applications integrating TTS synthesis. Regarding the type of system, it is believed that use in tutors and tools will place higher demands on TTS synthesis than use as a stimulus: tutors are intended to improve learners' command of the TL; tools are intended to help learners improve the quality of the documents etc. that they create in the TL; when used as a stimulus, on the other hand, TTS synthesis is just intended to get learners thinking about a particular aspect of the TL or get learners talking in the TL (see section 3.3.2).

Regarding the aspect of language proficiency on which the application focuses, the goal of reading and writing tutors is to improve learners' knowledge of grapheme-phoneme and phoneme-grapheme correspondences in the TL, respectively. It is essential, we believe, for TTS synthesis used in such applications to get grapheme-to-phoneme conversion right. As presented in section 2.4.1, this can be complex and it is therefore not given that TTS synthesis systems will achieve this. Speaking, on the other hand, will, we believe, place high demands on the quality of the speech generated, as these applications aim to improve learners' ability to pronounce the TL.

Regarding the function of TTS synthesis within the CALL system, it is believed that use as a pronunciation model will place higher demands on the quality of the speech generated than use as a RM, which will in turn place higher demands on the quality of the speech generated than use as a CP: as a PM, learners are expected to imitate the speech generated by the TTS

synthesis system; as a RM, learners are only expected to listen to it; and, as a CP, it is merely meant to stimulate learners to talk. This hypothesis is investigated in chapter 6.

Regarding the mode of oral communication that is to be simulated in CALL applications, a main distinction is made between "spontaneous speech [which] involves a small number of participants [and] non-spontaneous speech [which] usually [involves] a passive audience" (Offord, 1990: 110). Spontaneous speech is further divided into conversing which involves more than one active participant who are typically non-specialists in a relaxed informal environment and monologuing which involves only one speaker who is typically a specialist addressing a generally passive audience in a typically formal environment (*ibid.*). A typical example of conversing is therefore chatting with a friend. Examples of monologuing include "domestic reports[,] ... classroom teaching, television and radio commentaries, court-room and parlimentary proceedings" (*ibid.*: 112), etc. Non-spontaneous speech is further divided into reciting and the speaking of what is written. And, the speaking of what is written is further sub-divided into two further modes one in which the speaker(s) attempt(s) to conceal the fact that the written origin of their speech and one in which they do not (*ibid.*). Examples of the former include plays and films. Examples of the latter include news broadcasts (*ibid.*). Depending on the speaker an attempt may or may not be made to conceal the written origin of a lecture or a seminar (*ibid.*). Each of these models of oral communication are characterised by different differences at the levels of pronunciation, morpho-syntax and vocabulary (Battye and Hintze, 1992). Of these only pronunciation is determined by how the TTS synthesis system handles text input – morpho-syntax and vocabulary are determined by the author of the CALL materials. We will therefore only consider differences at the level of pronunciation here. According to Battye and Hintze (1992), at this level, the modes of speaking of what is written and monologuing in formal environments in French are characterised by: avoidance of regionalisms, careful delivery, monotonous intonation, careful syllable-timing, slow tempo, maintenance of vowel and length contrasts and the distinction between future and conditional endings which are unstable and in the process of disappearing from the French language such as [œ- ] and [ɛ- ], *mettre* (put) [mɛtR] and *maître* (master) [mɛ:tR], and *je partirai* (I will leave) [e] and *je partirais* (I would leave) [ɛ] respectively, resistance to vowel assimilation and vowel harmony, retention of a high percentage of *e-muet* or schwa, realisation of a high number of optional liaisons, and occurrences of *liaisons sans enchaînement*, i.e. "pronunciation of the liaison consonant without resyllabification; in other words,

[proncunation of] ... the liaison consonant ... despite a following pause: *(ceci) est intolerable* [et|ɛ- . to. le. Rabl], as opposed to [e. tɛ- . to. le. Rabl] with no pause or [e|ɛ- . to. le. Rabl]" (*ibid.*: 337). In French, monologuing in more informal situations such as when reporting an incident verbally and in classrooms and conversing in more formal environments are characterised by more rapid and spontaneous delivery than *français standard*, the term used to describe the variety of pronunciation found in speaking of what is written and monologuing in formal environments, assimilation, simplification of consonant clusters, reduction of hiatus, i.e. "the loss of a vowel when occurring before another: *qui était* (who was) [kete]" (*ibid.*: 342), deletion of [y] prevocalically, disappearance of unstressed syllables and unstressed *e-muets*, and use of emphatic stress (*ibid.*). And, conversing in more informal situations in French, according to Battye and Hintze, is characterised at the level of pronunciation by "presence of an identifiable or regional accent" (*ibid.*: 350), frequent assimilations, reduction of consonant clusters, non-deletion of *e-muet* prevocalically, umotivated insertions of *e-muet*, non-realisation of optional liaisons, and deletion of unstressed vowels and unstressed syllables. It is believe that different CALL applications require the simulation of different modes of communication. Specifically, it is believed that: listening comprehension activities require the simulation of monologuing, speaking as if not written and speaking as if written – television and radio commentaries, plays and news broadcasts are all used in listening activities; dictation activities require the simulation of speaking as if written; interactive speaking practice requires the simulation of conversing; and, talking dictionaries, talking texts and TWPs require the simulation of speaking as if written. It is therefore believed that different CALL applications will place different demands on the register – linguistic variation according to purpose or social setting – of the speech generated by TTS synthesis systems. The demands that CALL applications integrating TTS synthesis place on the register of the speech generated are also investigated in the main investigation presented in chapter 6. The investigation of the dimension of text type and the first two dimensions of CALL systems mentioned above are subjects for further research.

In addition, there are, we believe, a number of dimensions to the settings in which CALL applications integrating TTS synthesis are used. These include, but are not limited to:

- whether the software is being used by an individual learning the language autonomously or whether it is being used by a learner attending classes,
- the approach to language teaching adopted by the class teacher, and,

- the level of proficiency in the TL of the learners using the software.

Dimensions of the setting may, in our opinion, also have an effect on the requirements imposed on TTS synthesis and hence an effect on evaluation. If the software is being used by entirely autonomous learners, then, we believe, the speech generated by the TTS synthesis systems used in CALL applications will need to be of higher quality than for those attending classes because such learners, unlike those attending classes, will, in general, not have access to a proficient speaker of the TL. Regarding the use of CALL applications integrating TTS synthesis in conjunction with traditional language classes, whether in class hours or outside class hours, we believe that the teacher's approach to language teaching will impose different demands on the quality of the TTS synthesis used in the CALL application because different approaches are associated with different goals: while the goals of the Audiolingual approach, the Silent Way (Gategno, 1972; 1976), and other early approaches to language learning are nativelike pronunciation, the goal of the Communicative Approach is effective communication and hence a 'threshold level' of pronunciation:

> there is a threshold level of pronunciation for non-native speakers of English; if they fall below this threshold level, they will have oral communication problems no matter how excellent and extensive their control of English grammar and vocabulary might be (Celce-Murcia *et al.*, 1996: 7).

The quality of TTS synthesis used in CALL applications ought, in our opinion, to meet the learning goals of the class in which it is used.

Regarding the proficiency level of the learners using the applications, with respect to the teaching of pronunciation, it has been suggested that teachers (MacCarthy, 1975) need only be more proficient than the learners they teach.

> Now a person teaching his own language must be assumed to be a perfect model (barring any marked regionalisms and aside from any purely personal 'speech defects'), so here it will simply be a matter of saying: listen to me, then getting the learner to imitate as accurately as possible and insisting on something pretty close to one's own speech. For the teacher of a language not his own, the position in relation to his class is not really very different. Many a teacher who could not for one moment pass as a native speaker in everyday conversation can provide a thoroughly adequate, even admirable, demonstration model at the level of the word and probably the sentence. Only an occasional lapse in articulation during rapid speech or, in the higher reaches, some intonation pattern that could not be accepted as authentic, need reveal

> the non-native, and this is of negligible importance in the general run of teaching situations. It is true that the learner cannot be expected to acquire in class a *better* pronunciation than that of his teacher, but there is no reason why he should not strive for one as good; and if, in advance level work, an exceptionally gifted individual ever reaches the point where his teacher can help him no further in this particular respect, he (the teacher) can feel well satisfied that he has prepared his pupil for taking advantage of everything the foreign residence will be able to add by way of a final polish to an outstanding performance (*ibid.*: 12f).

If this is true, then higher proficiency learners will place higher demands on the quality of TTS synthesis. Regarding the use of TTS synthesis for teaching listening comprehension, on the other hand, less advanced learners will, in our opinion, place higher demands on the quality of the speech generated by TTS synthesis systems than more advanced learners because of the very fact that their proficiency in listening is less developed.

The effects of dimensions of CALL settings on the requirements imposed on TTS synthesis and implications for evaluation are subjects for further research.

## 3.5 Summary

With the goal of providing further motivation for this study, in this chapter we began by looking at the advantages which it is believed that TTS synthesis would bring to CALL (section 3.2). Of the benefits suggested, we identified, that the following have the potential to add value to CALL because they are unique to TTS synthesis:

- the ability to generate enhanced speech (Bonneau *et al.*, 2000), "speech with intonation, but no rhythm" (Keller and Zellner-Keller, 2000: 110), 'synthetic humming' Yoram and Hirose (1996), and,
- the ability to generate speech models in general on demand.

CALL applications integrating TTS synthesis were then discussed. This discussion highlighted that only around half of the applications actually exploited the unique capabilities of TTS synthesis to provide activities and exercises which it is not possible to provide through the use of other media such as digital recordings of native speakers, i.e. were smart applications of TTS synthesis in CALL. Another aim of this discussion was to identify dimensions of the CALL setups that might have implications with respect to evaluation. As presented in section 2.1, different setups may impose different requirements on NLP systems and hence require

different types of evaluation. Regarding CALL applications integrating TTS synthesis, it is believed that the following dimensions may have implications for evaluation:

- type of CALL system (tutor, tool or stimulus),
- the aspect of language proficiency on which the application focuses (reading, writing, listening or speaking),
- the function of TTS synthesis within the CALL system (RM, PM or CP),
- whether the software is being used by an individual learning the language autonomously or whether it is being used by a learner attending classes,
- the approach to language teaching adopted by the class teacher, and,
- the level of proficiency in the TL of the learners using the software.

As said, the implications of the function that TTS synthesis plays within CALL applications for evaluation are investigated in chapter 6.

# 4 Evaluation

## 4.1 Overview

As mentioned in the introduction, one of the potential reasons why there is not much CALL software integrating TTS synthesis on the market, despite its potential benefits (see section 3.2) and applications (see section 3.3), is that it has not been evaluated sufficiently for the purpose. In this chapter, this question is examined.

In order to permit the assessment of the evaluations of TTS synthesis in CALL that have been conducted to date, having considered very generally what evaluation is (see section 4.2) and why people conduct evaluations (see section 4.3), best practice in evaluation is discussed (see section 4.4). Standards organisations perhaps have the greatest experience of evaluation. This discussion is therefore based on the relevant standards found in the literature, namely:

- ISO (International Organization for Standardization) (1999) *Information Technology – Software Product Evaluation*, and,
- EAGLES (1999) Expert Advisory Group on Language Engineering Standards (EAGLES) *Evaluation Working Group Final Report.*

The sections that follow expand on several of the issues raised by these standards, namely:

- the levels of evaluation that ought to be conducted (see section 4.5), and,
- the features of good methods of evaluation (see section 4.6).

Then, in section 4.7, the evaluations of CALL software integrating TTS synthesis that have been conducted to date are assessed with respect to these findings. Section 4.8 looks at reasons why evaluation might be neglected in this context. And then, in section 4.9, a potential solution to the most likely of these reasons is put forward. Finally, on the basis of the findings of this chapter, an agenda for further evaluation is proposed (see section 4.10).

## 4.2 What is evaluation?

In the Chambers dictionary, the following definition of evaluation, or more precisely the verb evaluate, is provided:

> evaluate i- or e-val'u-at, vt to determine or estimate the value of. – n evaluation. – adj eval'uative tending or serving to evaluate, or functioning as an evaluation. [Fr évaluer] (Chambers 1998: 559).

Given the diverse areas in which evaluation is used – for example, organisations, institutions, policies, strategies, programs, projects, products, services, systems, processes, performance, job candidates, jobs, and proposals – evaluation has many different purposes and takes many different forms. The dictionary definition is therefore necessarily vague.

While it may be possible to be more specific when defining evaluation within a specific context, evaluation remains difficult to define: even within a specific context, there are several different stakeholders each of whom have different motivations for conducting evaluations (see section 3.2), and thus conduct different types (see section 4.4) and employ different techniques of evaluation.

Consequently, while some, such as Fourcin (1992) (see below), attempt to distinguish evaluation from similar concepts such as assessment, others make no attempt at a definition of the concept of evaluation and proceed directly to a presentation of the purposes, types, and techniques of evaluation.

> Whilst assessment gives an overall operational rating [of a system], evaluation is the process which, based on assessment, give and an insight into the essential fundamental factors which contribute to a system's performance (*ibid.*: 432).

The aim here is not to provide a precise definition of evaluation. The dictionary definition is satisfactory for the present purposes because it is supported, in the sections that follow, with a presentation of the purposes (see section 3.2), process (see section 4.4), types (see section 4.5) and principles (see section 4.5.2.2) of evaluation.

## *4.3 Why do people conduct evaluations?*

As said, the type of evaluation to be conducted and the most appropriate techniques of evaluation to be used depend on the purpose(s) of evaluation. Before proceeding to discuss the types and techniques of evaluation, it is therefore appropriate to consider the motives of the different stakeholders for conducting evaluations.

Three main groups of individuals are interested in the results of evaluations of software products in general, namely:

- developers,
- end-users, and,
- funding agencies.

Evaluation enables developers to:

- identify potential applications of their technologies,
- establish whether their technologies fulfil the requirements of their end-users,
- identify and diagnose any weaknesses in their technologies for subsequent improvement,
- "Decide on the completion of a process and when to send products to the next process" (EAGLES, 1999: 1.2.4),
- "Predict or estimate end product quality" (*loc. cit.*),
- "Compare the product with competitive products" (*loc. cit.*), and,
- "Decide when to release a product" (*loc. cit.*).

Regarding the evaluation of Speech And Language Technologies (SALTs), more specifically, White (2003) introduces a specific type of developer, namely productisers. Productisers, of which CALL researchers are an example, according to White use evaluation to determine whether:

- a marketable product can be produced, and,
- it will meet real needs.

Moving on to end-users, in common with researchers and developers, end-users, in general, are interested in:

- identifying potential applications of technologies, that is potential solutions to the problems that they face (ELSE, 1999),
- establishing whether technologies meet their requirements, and,
- comparing competing technologies in order to establish which is most suitable to their needs.

Regarding CALL software, more specifically, there are two distinct groups of end-user, teachers and learners who are interested in the results of evaluations of CALL software. Regarding teachers' interest in evaluation, evaluations enable teachers to:

- make a case for the use of CALL in their classes (Chapelle, 2003), specifically, to determine "whether to use CALL and for what purpose" (*ibid.*: 76),
- identify "successful strategies for using software" (*ibid.*: 82), and,
- identify "the best ways to structure learning tasks" (*loc. cit.*).

The identification of successful strategies for using CALL software is also of interest to learners (*ibid.*).

From the point of view of funding agencies, evaluation is a tool which enables them to find out whether money invested in a particular technology has been used effectively, that it has led to technological progress, and to identify areas where further investment would be of benefit (ELSE, 1999).

Regarding the evaluation of SALTs, more specifically, on the evaluation of machine translation systems, White (2003) adds to this list managers and vendors. Similarly, regarding the evaluation of CALL software, Chapelle (2003) adds to this list administrators and publishers. Evaluations enable administrators, like managers, to make a case for the use of technology (*ibid.*; White, 2003), specifically to decide whether it is worth funding (Chapelle, 2003; White, 2003). Publishers, according to Chapelle (2003), use evaluation to generate marketing data to make a case for the use of their CALL applications.

In general, positive evaluation benefits all stakeholders. It leads to the acceptance of technologies which in turn leads to further investment in further R&D and consequently improvements in the quality of technologies made available to end-users.

## 4.4 How are evaluations conducted?

Regarding the evaluation process, guidelines for planning and conducting evaluations of software products in general are provided in the standard *ISO 14598 Information Technology – Software Product Evaluation* (ISO, 1999). The EAGLES standard (EAGLES, 1999) is an expansion of the ISO guidelines for the specific purposes of evaluating speech and language processing systems. Neither standard presents a single, universal evaluation technique which

can be applied to any system. Rather, they provide a standard methodology for evaluation design and implementation, intended to be applicable to the evaluation of any software product at any point in its life cycle.

In the EAGLES standard, the evaluation process is broken down into four stages:

- establish the evaluation requirements,
- specify the evaluation,
- design the evaluation, and,
- execute the evaluation.

These stages will be presented in more detail in the following subsections.

#### 4.4.1.1 Establish the evaluation requirements

The first of these stages, establish the evaluation requirements, is itself broken down into three stages:

- establish the purpose of the evaluation,
- identify the types of products to be evaluated, and,
- specify the quality model.

For examples of the potential purposes of evaluation, the reader is referred back to section 4.3 where the motives of the different stakeholders for conducting evaluations are presented.

The objective of the second stage is to identify what is to be evaluated: "a research proposal, a system component, a partially developed system, a complete system and so on" (EAGLES, 1999: 1.1.1). In other words, 'types of products' refers to the level which the technology has reached in the software product life cycle. This will determine the level of evaluation that is to be conducted. The different levels of evaluation that might be conducted are discussed in section 4.5.

Regarding the third stage, the specification of the quality model, the goal of this stage is to identify attributes of the system which can be measured in order to determine whether a system fulfils user requirements. This is achieved through the construction of a quality model, a definition of features of a product that determine its ability to satisfy the needs of the end-user (EAGLES, 1999). The construction of a quality model typically begins with the

identification of general criteria such as functionality, reliability, usability, efficiency, maintainability, portability (*ibid.*; ISO, 1999). These are then successively broken down until measurable attributes of the system are identified. One or more attributes may need to be measured in order to get at a specific criterion. For example the usability, ease of use (Spencer, 1985), of TTS synthesis in a CALL context may depend on among other things the overall quality of the speech generated by the TTS synthesis system, which itself is determined by a number of different measurable attributes including: intelligibility, comprehensibility, naturalness, etc.

Regarding what determines user requirements, the setup, that is operational context into which the system is embedded and the function of the system within that operational context (Sparck Jones and Galliers, 1996), is an important consideration (see section 3.1). The different operational contexts in which TTS synthesis might be embedded for CALL purposes and the functions, or roles, that TTS synthesis might play in those operational contexts were discussed in section 3.4. Whether these different CALL setups place different demands on TTS synthesis remains to be seen. The investigations presented in chapter 6 are a first attempt to answer this question.

#### 4.4.1.2 Specify the evaluation

This stage is also broken down into three stages:

- select metrics,
- establish rating levels for metrics, and,
- establish criteria for assessment.

The objective of the first sub-stage is to select (or develop if necessary) metrics, methods and scales for the measurement of the attributes (ISO, 1999) identified in the previous stage. For example, a common metric used for the evaluation of the intelligibility of the output of TTS systems, the articulation test, involves presenting human listeners a list of words produced by the speech synthesiser and measuring the number of words correctly transcribed by the human participants (van Bezooijen and van Heuven, 1997).

In order to ensure that the results of evaluations are meaningful, metrics and measures should meet a number of requirements. These are presented in section 4.5.2.2.

Once metrics have been selected, rating levels should be established. That is the relationship between scores and the degree of satisfaction of user requirements should be established.

In the third sub-stage, the procedure for summarizing the results of the evaluation should be specified, that is the procedure for obtaining a score for a system for a given attribute or criterion.

Regarding the aforementioned metric for the evaluation of the intelligibility of the output of TTS systems, the average number of words correctly transcribed across participants might be reported as the system score. Alternatively, if the test had been presented in closed response format, that is participants were asked to select from a pre-defined list of words the word they heard, the number of words correctly identified might be adjusted through the application of a formula, in order to take into account the probability of correctly guessing the identity of a word, before averaging the measures across participants in order to arrive at the system score. The following is the example of a formula which could be used to adjust the scores of such a test to account for correct guessing:

> This can be calculated using the formula:
>
> $$R_A = R - W(n - 1)$$
>
> In which $R_A$ represents the adjusted percent score (the measure of intelligibility), R represents the number of items recognized correctly, W is the number of items identified incorrectly, and n is the number of possible answers in each response set (Francis and Nusbaum, 1999: 85).

It is important that procedures such as this for summarizing results be established prior to the execution of an evaluation in order to avoid any temptation to influence the results.

### 4.4.1.3 Design the evaluation

This third stage involves the production of an evaluation plan. This should describe the evaluation methods and provide a schedule of the evaluation including the following specifications:

- who will produce any required test materials,
- when and by whom the test(s) will be conducted, and,
- how the results of the test(s) will be reported and communicated.

#### 4.4.1.4 Execute the evaluation

The final stage, as its name suggests involves the execution of the evaluation plan, that is applying the metrics, taking measures, comparing the measures obtained through the use of defined procedures, and assessing, summarizing, reporting and communicating the results.

## *4.5 At what levels should evaluation be conducted?*

Regarding the levels at which CALL software integrating TTS synthesis ought to be evaluated, as far as it is possible to establish, no recommendations exist. The different infrastructures proposed for the evaluation of SALTs and CALL software are therefore considered in sections 4.5.1 and 4.5.2, respectively. Then, in section 4.5.2.2, on the basis of these infrastructures, an infrastructure for the evaluation of CALL software integrating TTS synthesis is proposed.

### 4.5.1 At what levels should SALTs be evaluated?

Regarding the evaluation of SALTs, such as TTS, several different infrastructures for evaluation have been proposed (Hirschman and Thompson, 1996; Sparck Jones and Galliers, 1996; ELSE, 1999; White, 2003). The most comprehensive of these infrastructures is the ELSE (Evaluation in Language and Speech Engineering) infrastructure (ELSE, 1999), which consists in five stages of evaluation:

- 'basic research evaluation', the objective of which is to determine whether a new technology, or an improvement on an existing technology, is worth pursuing, that is whether it is viable and whether it will bring significant improvement on existing solutions.
- 'technology evaluation', the goal of which is to determine whether a system meets its objectives. This is typically achieved through measurement of the performance of the system in a control task.
- 'usage evaluation', the goal of which is to determine whether a system fulfils its function in a given operational context, that is to determine whether end-users find the system useful and easy to use as well as whether the system meets its objectives.
- 'impact evaluation', the objective of which is to assess the effects of the system beyond its primary function, such as the socio-economic effects of its use.
- 'programme evaluation', the goal of which is to determine whether a funding programme was worthwhile, that is whether the investment resulted in progress.

Of these levels of evaluation (near) equivalents to basic research evaluation, technology evaluation, usage evaluation, and impact evaluation are found in the other infrastructures.[25] Program evaluation is unique to the ELSE infrastructure. Two further levels of evaluation mentioned in the other infrastructures are 'adequacy evaluation' and 'formative evaluation'. The goal of adequacy evaluation is to determine whether a system meets user requirements (Hirschman and Thompson, 1996). And, the goal of formative evaluation is to guide system design through the identification of where a system needs improvement in order to meet user requirements, i.e. where a system fails to meet user requirements (*ibid.*). Adequacy evaluation and formative evaluation are therefore near equivalents. Both are achieved through a combination of 'diagnostic evaluation', the identification of the successes and limitations of a system with respect to a taxonimisation of possible inputs, that is the production of a profile of a system's performance, and technology evaluation (*ibid.*).[26]

### 4.5.2 At what levels should CALL applications be evaluated?

As presented in the previous chapter, in addition to its use in off-the-shelf CALL software, it is implied that TTS synthesis could be used by teachers to create their own CALL activities and exercises, i.e. TTS synthesis could be used in CALL authoring tools. Ready-to-use CALL software and CALL authoring tools require different levels of evaluation. The infrastructures proposed for their evaluation are presented in sections 4.5.2.1 and 4.5.2.2 respectively.

#### 4.5.2.1 At what levels should CALL software be evaluated?

Regarding the levels at which CALL ought to be evaluated, two infrastructures were found in the literature. The first, referred to in Swartz *et al.* (1990), and Levy (1999a), consists in two levels of evaluation: 'formative' and 'summative'.

> In formative evaluations, one wants to find out the effects of the program as they relate to specified goals and to uncover any unanticipated results. Additionally, one wants to

[25] What ELSE refer to as 'basic research evaluation' is also known as 'feasibility testing' (White, 2003); what ELSE refer to as 'technology evaluation' is also known as 'intrinsic evaluation' (Sparck Jones and Galliers, 1996), 'performance evaluation' (Hirschman and Thompson, 1996), 'summative evaluation' (*ibid.*), and 'declarative evaluation' (White, 2003); what ELSE refer to as 'usage evaluation' is also known as 'extrinsic evaluation' (Sparck Jones and Galliers, 1996), and 'usability evaluation' (White, 2003); and what ELSE refer to as 'impact evaluation' is also known as 'operational evaluation' (*ibid.*).

[26] 'Diagnostic evaluation' is also referred to as 'internal evaluation' (White, 2003).

> know system operation, the adequacy of presentation, format, and so on (Swartz *et al.*, 1990: 55)
>
> Formative evaluation provides feedback that is cycled back into the CALL design in order to improve it. Summative evaluation is the final analysis of the results and effects of the CALL (*ibid.*: 54f)

In other words formative and summative evaluation are used in the same sense in CALL circles as they are among language engineers (see section 4.5.1, and footnote 25).

The second infrastructure was proposed by Chapelle (2001). According to Chapelle, CALL software, including both tutors and tools (see section 3.3), would benefit from three stages of evaluation. In a first stage, she recommends the judgmental evaluation of the CALL application for its potential to provide conditions that promote SLA. Then, in a second stage, she recommends that a similar evaluation of the activities that teachers plan around the CALL software be carried out – as Jones (1986) demonstrates, teachers may use a piece of software in a number of different ways in addition to the way in which it was originally intended for use. And, finally, in a third stage of evaluation, she recommends that learners' performance in those activities be empirically evaluated – the fact that a particular piece of CALL software provides learners opportunities to engage in certain types of interaction (some of the types of interaction which are believed to promote SLA are presented in section 5.3) does not guarantee that learners will actually engage in them and acquire language through them.

Chapelle's (2001) infrastructure is adopted here, as it is believed to be more comprehensive.

#### 4.5.2.2 At what levels should authoring tools be evaluated?

Regarding the levels at which authoring tools ought to be evaluated, according to Bickerton *et al.* (2001), there are two approaches to their evaluation:

- 'taxonomic evaluation', and,
- 'implementational evaluation'.

Taxonomic evaluation, according to Bickerton *et al.*:

> establishes lists of possible or desirable features, and checks whether tools possess them. The result provides a score, and this converts into a rating

In other words, the goal of taxonomic evaluation is to determine whether an authoring tool has the potential to be able to create the types of activities and exercises that the author wishes to create.

Implementational evaluation, according to Bickerton *et al.*, involves "lesson or task benchmarking" (*ibid.*), that is, evaluation of the performance of authors in creating exercises. Variables typically measured include time and cost of creation of activities and exercises (*ibid.*). In other words, implementational evaluation is an evaluation of the usage of the authoring tool (see section 4.5.1).

### 4.5.3 At what levels should CALL applications integrating TTS synthesis be evaluated?

Just as ready-to-use CALL software and CALL authoring tools require different levels of evaluation, it is proposed that ready-to-use CALL software integrating TTS synthesis and CALL authoring tools integrating TTS synthesis will require different levels of evaluation. On the basis of the infrastructures proposed for the evaluation of SALTs and CALL software, an infrastructure for the evaluation of CALL software integrating TTS synthesis is proposed in section 4.5.2.1. Then, in section 4.5.2.2, on the basis of the infrastructures proposed for the evaluation of SALTs and CALL authoring tools, an infrastructure for the evaluation of CALL authoring tools integrating TTS synthesis is proposed.

#### 4.5.3.1 At what levels should CALL software integrating TTS synthesis be evaluated?

Regarding the evaluation of CALL software integrating TTS synthesis systems, it would seem appropriate to combine the most comprehensive infrastructures from each field. In other words, when evaluating CALL software integrating TTS synthesis, it is suggested that Chapelle's (2001) infrastructure for the evaluation of CALL software be extended to include two further stages of evaluation, namely basic research evaluation and adequacy evaluation of TTS synthesis for use in CALL, used here to refer to the combination of adequacy diagnostic and technology evaluation. Generally, when a funding programme has been involved in the development of the software, program evaluation should also be conducted. Regarding impact evaluation, positive impact is one of the ideal conditions for SLA identified by Chapelle, which should be addressed at each stage of evaluation. A separate stage of impact evaluation is therefore superfluous. The result is an infrastructure consisting in six levels of evaluation. The objects of these levels of evaluation are presented in Table 5.

**Table 5 An infrastructure for the evaluation of CALL software integrating TTS synthesis**

| | |
|---|---|
| Level 1 | Viability and potential benefits of the use of TTS synthesis in CALL. |
| Level 2 | Adequacy of TTS synthesis for use in CALL. |
| Level 3 | Potential of the CALL program to provide conditions which promote SLA. |
| Level 4 | Potential of the teacher-planned CALL-based activity to provide conditions which promote SLA. |
| Level 5 | Learner's performance in the CALL activity. |
| Level 6 | Success of the funding programme. |

Regarding the different stakeholders in the evaluation of CALL software integrating TTS synthesis, it is believed that levels 1 and 2 will be of interest to TTS synthesis researchers, levels 1 to 5 will be of interest to CALL researchers, administrators, teachers, and learners, and publishers of CALL software, and level 6 will be of interest to funding agencies.

#### 4.5.3.2 At what levels should authoring tools integrating TTS synthesis be evaluated?

Regarding the evaluation of authoring tools integrating TTS synthesis, we believe that the approaches to the evaluation of authoring tools presented in Bickerton *et al.* (2001) are complementary and hence that both ought to be conducted: first taxonomic evaluation, and then implementational evaluation. Implementational evaluation is necessary because even if an option is available it might not fulfil the intended function and it might not be easy to use. Taxonomic evaluation is recommended because implementational evaluation is costly and implementational evaluation is pointless if the desired options are not available. But, first before integrating TTS synthesis into authoring tools, it is believed that basic research evaluation and adequacy evaluation (see section 4.5.1) of TTS synthesis for use in CALL ought to be conducted. Also, if a funding programme has been involved in the development of the authoring tool, program evaluation (see section 4.5.1) should also be conducted once the project has been completed. The resulting infrastructure is presented in Table 6.

**Table 6 An infrastructure for the evaluation of CALL authoring tools integrating TTS synthesis**

| | |
|---|---|
| Level 1 | Viability and potential benefits of the use of TTS synthesis in CALL. |
| Level 2 | Adequacy of TTS synthesis for use in CALL. |
| Level 3 | Potential of the authoring tool to permit the creation of CALL activities and exercises. |
| Level 4 | Ease and cost of the creation of CALL activities and exercises using the |

| | |
|---|---|
| | authoring tool. |
| Level 5 | Success of the funding programme. |

As regards the activities and exercises that authors create within these environments, it is believed that Chapelle's (2001) infrastructure ought to be applied to them (see section 4.5.2.1).

Regarding the different stakeholders in the evaluation of CALL authoring tools integrating TTS synthesis, it is believed that levels 1 and 2 will be of interest to TTS synthesis researchers, levels 1 to 4 will be of interest to CALL researchers, publishers of CALL authoring tools, administrators and teachers, and level 5 will be of interest to funding agencies.

## *4.6 Features of good methods of evaluation*

As said, in order for the results of evaluations to be meaningful, metrics and measures should meet a number of requirements, namely they must be:

- valid, and,
- reliable.

These concepts are discussed individually in the following sections.

### 4.6.1.1 Validity

Regarding validity, measures must have both:

- internal validity, and,
- external validity.

#### *4.6.1.1.1 Internal validity*

An internally valid measure is one that measures what it set out to measure (Chapelle and Jamieson, 1990; Lindgaard, 1994; Sparck Jones and Galliers, 1996). That is to say results are due to the attribute measured and not to some other variable or chance. The effects of other variables can be controlled through the use of a number of techniques including but not limited to randomisation, experimental controls, and matching (McGrath, 1995). Statistical tests can be applied in order to determine the likelihood that the results could have been obtained by chance (*ibid.*).

Regarding the evaluation of different aspects of the quality of the speech generated by TTS synthesis systems, spoken utterances are not the result of the simple concatenation of speech segments: there are no spaces between the individual words; adjacent (and non-adjacent for that matter) segments influence one another's pronunciation; and, other layers of information are superimposed onto the segmental layer, namely prosody and voice quality (Abercrombie, 1967; van Bezooijen and van Heuven, 1997). Consequently, it is difficult to isolate and evaluated them separately. For example:

> Since prosody is highly redundant given the segmental information (with the exception of the signalling of sentence type and emotion/attitude), it can be functionally tested only if measures are taken to reduce its redundancy. The first course of action, then, has been to concentrate on atypical, rather contrived materials in which prosody is non-redundant. That is the materials consist of segmental structures that would be ambiguous without the prosody, and listeners are asked to solve the ambiguity. ... The second route is to make prosody less redundant by degrading the segmental quality, such that without prosody (i.e. in the baseline conditions identified above) the intelligibility of the speech output would be poor (*ibid.*: 534)

Internal validity is also difficult to attain when evaluating the intelligibility and comprehensibility (see section 2.3.2 for definitions of these terms) of the speech generated by TTS synthesis systems because, in addition to the cues provided in the acoustic signal, listeners rely on their knowledge of the phonotactics and possible lexical items of the language, the syntactic and semantic context, and the real world (van Bezooijen and van Heuven, 1997).

The TTS synthesis evaluations considered so far are typically conducted in laboratories. Controlling for extraneous variables is even more difficult in field settings (Möller *et al.*,

2001), such as CALL settings. Regarding evaluations of CALL applications more specifically, attaining internal validity is difficult because evaluators typically work with intact classes of learners (Chapelle and Jamieson, 1990; Motteram, 1999). Consequently, they are unable to control (Chapelle and Jamieson, 1990) the many variables which may affect the results of CALL evaluations which include, but are not limited to:

> behaviour/personality of the instructor, the type of methods used, students' reasons for studying the language (their objectives of language study), the amount of time students used CALL, and whether CALL is an assigned or an optional activity (*ibid.*: 50).

In response to this problem, in order to permit users of the results of evaluations to assess for themselves the validity of evaluations, it has been recommended that evaluators of CALL applications "identify and explain the factors that may have influenced their ... results" (*ibid.*: 43). This has also been recommended in the literature on the evaluation of TTS synthesis (JEIDA, 1995), though not for the purposes of assessing the internal validity of evaluations, rather for the purposes of assessing the external validity of evaluations (see section 4.6.1.1.2).

#### *4.6.1.1.2 External validity*

An externally valid measure is one that has predictive validity, that is which is generalisable to other setups (Sparck Jones and Galliers, 1996), than the one specifically tested. To ensure the generalisability of results from evaluations conducted under laboratory conditions to field conditions, i.e. actual setups, the conditions under which an evaluation is conducted should be as representative of real-world conditions as possible. For example, the environment in which an evaluation is conducted should be representative of the real-world environment, the materials used should be representative of actual materials used in the real world, and the attitudes and motivations of participants should be representative of those of the real-world end-users (van Bezooijen and van Heuven, 1997). To ensure the generalisability of results from one field evaluation to other setups, characteristics of the setups must be as similar as possible. The external validity of a measure can be tested by replication of the evaluation in other setups.

Regarding the evaluation of TTS synthesis, evaluations may be conducted in the laboratory or in the field in which TTS synthesis is intended for use. The results of laboratory evaluations may not generalise well to actual setups due to differences between the laboratory setup and the field setup:

> Speech output systems typically form an element of a larger human-human interface in an application with a specific, dedicated task. In practice this means that, quite probably, the vocabulary and types of information exchanges are restricted and domain-specific, so that situational redundancy is likely to make up for poor intelligibility. On the other hand, speech output systems will often be used in complex information processing tasks, so that the listener has only limited resources available for attending to the speech input (van Bezooijen and van Heuven, 1997: 487).

Another reason why the results of laboratory evaluations may not generalise well to actual setups is that actual users often have different attitudes (*ibid.*; Möller *et al.*, 2001) to and motivations (van Bezooijen and van Heuven, 1997) for the use of TTS synthesis than the participants in laboratory evaluations. For example:

> If people have a choice between human and synthetic speech, the synthetic speech will have to be good if it wants to have a chance of being accepted. However, if people do not have a choice, e.g., the visually handicapped who without synthesis (or braille) [*sic*] will not have access to a daily newspaper, synthesis will be accepted more easily (*ibid.*: 179f).

Regarding the generalisation of evaluation results from one setup to another, according to Schmidt-Nielsen (1995), "Field tests are ... highly specific to a particular situation and do not generalise well to other situations" (*ibid.*: 197), a view which is shared by van Bezooijen and van Heuven (1997): "The use of *field tests* will be limited to one system in one specific application; results of a field test cannot, as a rule, be generalised to other systems and/or other applications" (*ibid.*: 488).

It is believed that this also applies to generalisations of results from one CALL setup to another because learners and instructional settings can vary with respect to so many different variables. These variables are the same as those that might affect internal validity (Chapelle and Jamieson, 1990).

In order to permit the users of the results of evaluations (the different stakeholders in evaluation are presented in section 3.2) to make an informed assessment of whether the results would generalise to the setup in which they intend to use TTS synthesis, the Japanese Electronic Industry Development Association (JEIDA) recommend that evaluators report all variables that may have affected the results of the evaluation (JEIDA, 1995). This approach

has also been recommended when evaluating CALL applications (Chapelle and Jamieson, 1995).

### 4.6.1.2 Reliability

A reliable measure is one which yields the same results when repeated in identical setups (Lindgaard, 1994). The reliability of a metric can be tested by repeating the task on different occasions and/or with different samples of participants (test-retest reliability), or by splitting the test into two halves and asking participants to perform the same task at different stages in the evaluation (split half reliability), for example (*ibid.*). Attaining reliable results in evaluations of TTS synthesis can be difficult due to the fact that not only do participants possessing different levels of experience with TTS synthesis and listening analytically to speech perform differently in evaluations of TTS synthesis, but even when these variables are controlled differences among participants persist (van Bezooijen and van Heuven, 1997). In CALL evaluations, attaining reliable results is believed to be impossible, because "there is no way to replicate the conditions of the experiment exactly" (Pederson, 1987: 106): as presented in section 4.6.1.1.1, CALL researchers must work with intact classes of students (Chapelle and Jamieson, 1990; Motteram, 1999), and, as presented in the preceding section, no two classes are the same.

## *4.7 Evaluations of CALL software integrating TTS synthesis*

To date, as far as it is possible to establish, only four 'formal' evaluations of CALL applications integrating TTS synthesis have been conducted (Stratil *et al.*, 1987a; Stratil *et al.*, 1987b; Cohen, 1993; Hincks, 2002). A few others mention the results of 'informal' evaluations (Sherwood, 1981; Mercier *et al.*, 2000). They, however, give no indications of the methods of evaluation that they employed. In the sections that follow, the formal evaluations are presented and assessed with respect to the evaluation theory presented in the preceding sections.

### 4.7.1 Evaluations of the adequacy of TTS synthesis for use in CALL applications

One evaluation was found in the literature which appears to address the adequacy of TTS synthesis for use in CALL. This evaluation, conducted by Stratil *et al.*, (1987b), addressed the adequacy of the quality of the speech generated by the *SSI 263 Spanish TTS chip* for use for the presentation of grammar exercises to learners of Spanish in a multifacet language laboratory which also integrated video disc, video tape, ASR, and an audio system.

Specifically, the ability of non-Spanish speaking, i.e. beginner, and Spanish speaking, i.e. more advanced, learners of Spanish to repeat and transcribe Spanish sentences presented via the TTS synthesis chip was compared with their ability to repeat and transcribe the same sentences spoken by a native speaker. The learners who participated in the evaluation were split into two groups. Each group were presented an audiotape of alternate synthesised and spoken sentences. One group heard the synthesised sentences first, and the other group heard the spoken sentences first. Each sentence was presented twice. On the first hearing, the learners' task was to repeat the sentence that they heard. Recordings of these repetitions were made. On the second hearing, the beginners were again asked to repeat the sentence that they heard and a recording was made, while the more advanced learners were asked to transcribe the sentence that they heard. The recordings and transcriptions of the sentences were assessed by a Spanish lecturer. This assessment revealed that while the beginners performed equally well in the two presentation conditions, the more advanced learners performed better when the sentences were spoken than when they were synthesised Analysis of the data using the Man-Whitney test revealed the differences in performance for the advanced learners across the presentation conditions were statistically significant. Regarding differences in performance from the first to the second hearing, neither the performance of the beginners nor the more advanced learners improved from the first to the second presentation for the synthetic speech. Improvements from the first to the second presentation were, on the other hand, observed for the spoken sentences, in particular for the more advanced learners.

As an evaluation of the adequacy of TTS synthesis for use for the presentation of grammar exercises in an audio-lingual language laboratory setting, it is believed that this evaluation is an attempt to assess the intelligibility (see section 2.3.2) of the speech generated by the TTS synthesis chip for use in that CALL setup. As such, it has a number of weaknesses. First, it is unclear whether the presentation of grammar exercises in CALL imposes requirements on the intelligibility of the speech generated by TTS synthesis and whether intelligibility is the only aspect of the quality of the speech generated by the TTS synthesis upon which it imposes requirements because, as far as it is possible to establish, the requirements that this CALL setup imposes on TTS synthesis have not been investigated. Evaluations of TTS synthesis for use in applications, in general, have addressed a number of other aspects of the quality of speech in addition to intelligibility. These include, but are not limited to, overall quality, comprehensibility, naturalness, prosodic form and prosodic function, and voice quality (van

Bezooijen and van Heuven, 1997). Establishment of a quality model, i.e. requirements analysis, as presented in section 4.4.1.1, is an essential stage in planning and conducting an evaluation. Another important stage in the evaluation process is to define acceptance levels (see section 4.4.1.2). While the authors appear to have defined acceptance levels, namely human speech, the appropriateness and validity of the reference condition used to define the acceptance levels is questionable. Regarding the appropriateness of the reference condition for defining acceptance levels, it is unclear whether it is necessary for the speech to be as intelligible as human speech. It has been suggested that teachers (MacCarthy, 1975; see section 3.4) and TTS synthesis (Keller and Zellner-Keller, 2000; see section 5.2) need only be more proficient than the learners they teach.

Secondly, it is unclear from the paper whether the reference condition is provided by the learners' regular Spanish lecturer or another speaker of the TL. If the reference condition is provided by the learners' regular Spanish lecturer, the internal validity of the evaluation is brought into question because this is not a fair comparison. Like anyone, a learner is more likely to perform well if they are familiar with the speaker.

Regarding the internal validity of the method of evaluation, while asking the participants to imitate the sentences controls for the effects of knowledge of the phoneme-grapheme relationship in Spanish, something which the learners may not yet have mastered, there are a few other variables which may affect the results that have not been controlled for. Firstly, the following comment suggests that semantic knowledge has not been controlled for (see section 4.6.1.1.1):

> Students with more knowledge of Spanish tended to reconstruct the sentences and sometimes gave a valid sentence that was different from the one on the tape (Stratil *et al.*, 1987b: 118).

Secondly, the effects of the participants' ability to pronounce Spanish are not controlled for:

> it is not clear that discrimination between L1 and L2 phones necessarily implies a correct pronunciation of the latter. It is well known for example that, although most anglophones [*sic*] can discriminate between North American /r/ and French /r/, only a few are successful at producing the latter accurately. Even when they have internalized a detailed representation ("sound image") of French /r/, they often resort to their L1 articulatory habits. In general, it appears that "if people can hear a difference between a

> pair of sounds, then they *can* make the difference – but do not necessarily do so in their ordinary speech (Ladefoged, 1967, p169) (Rochet, 1990: 120).

The evaluation does, however, have its strengths. The use of real end-users, and the fact that the evaluation was conducted in the CALL setup mean that the results of the evaluation are more likely to generalise to actual use than typical evaluations of TTS synthesis which are often conducted using with participants who are not representative of real users (van Bezooijen and van Heuven, 1997) under strictly controlled laboratory conditions (*ibid.*). As presented in section 4.6.1.1.2, no two classes are the same. The results of CALL evaluations cannot, however, therefore be expected to perfectly generalise to another class etc. On this point, as presented in section 4.6.1.1.2, it has been suggested that all the relevant features of the participants be reported, in order to permit users of evaluation results to make informed assessments as to whether the results are likely to generalise to the groups of learners with whom they are working. It is believed that the details provided in this evaluation are insufficient for these purposes: the age, sex, and linguistic background of the participants are not reported among other possible variables.

In summary, not only is it likely that Stratil *et al.*'s (1987b) is only partial, it is also questionable whether the method of evaluation used is valid.

### 4.7.2 Evaluations of learners' performance in teacher-planned CALL activities integrating TTS synthesis

According the Chapelle (2001) both product-oriented and process-oriented evaluations of learners' performance in teacher-planned CALL activities integrating TTS synthesis are conducted. Product-oriented evaluations on their own are not very illuminative (Scholfield and Ypsiladis, 1992; Chapelle, 2001) because:

> The real question is not whether the *provision for* interactional modifications [or any other conditions which are believed to promote SLA] increases acquisition, but whether the *use of* interactional modifications [etc.] increases acquisition of those forms for which interactional modifications [etc.] are used (*ibid.*: 79; author's addition).

Process-oriented evaluations on their own are not adequate – that a learner has engaged in a process that is believed to promote acquisition does not guarantee that they have acquired the intended TL forms (Howatt, 1969; Chapelle, 2001). One example of each of these two types of evaluation has been found in the literature. These evaluations conducted by Hincks (2002) and

Cohen (1993) are presented in sections 4.7.2.1 and 4.7.2.2 respectively. In addition, Chapelle (2001) recommended that at all levels of her infrastructure for the evaluation of CALL application, their impact, that is "The ... effects of the CALL activity on those who participate in it" (*ibid.*: 55) ought to be evaluated. One example of an evaluation of the impact of a CALL application integrating TTS synthesis was also found in the literature (Stratil *et al.*, 1987a). This evaluation is presented in section 4.7.2.3.

### 4.7.2.1 Product-oriented evaluations

As said, one product-oriented evaluation of learners' performance in a teacher-planned task using a CALL application integrating TTS synthesis was found in the literature. This evaluation conducted by Hincks (2002) looked at the effectiveness of the use of the male voice of the Infovox diphone-based concatenative Swedish TTS synthesis system in conjunction with the speech editor WaveSurfer (Beskow *et al.*, 2000) for the teaching of English lexical stress to Swedophones. Specifically, the evaluation focused on the acquisition of the pronunciation of the English words *component* and *parameter*. These two words are typically mispronounced by Swedish learners of English because they have Swedish cognates, namely *komponent* and *parameter*, which differ from their English counterparts with respect to the placement of lexical stress.

The participants were 13 students of Technical English enrolled on engineering courses at the Kungliga Tekniska Högskolan (KTH) (Royal Institute of Technology) College of Engineering, Stockholm. 8 were male, and 5 were female. All spoke Swedish as either their Mother Tongue (MT) or an L2. Specifically, eleven spoke Swedish as their MT, one spoke Serbo-Croatian, and the other "had grown up in a bilingual Swahili/English environment" (Hincks, 2002: 154). Regarding their proficiency in English,

The students' scores on a 100-pt placement test ranged from 32-71, with a mean of 54, in line with the average score of KTH students studying English at the intermediate level (*loc. cit.*).

The corpus used in the evaluation consisted in a short text from the *New Scientist* containing three instances of each of the two words. Before participating in the training, which was carried out over a period of 4 weeks, the participants were asked to read aloud this text. Digital audio tape recordings of their productions were made. Another recording was made in the last week of the course of training. As presented in section 3.3.2, the training consisted in guiding

learners to produce an English sounding version of the cognates from the version generated by the Swedish TTS synthesis system using the speech editor:

Comparison of pre-test and post-test results revealed improvements in pronunciation for both words. From this Hincks (2002) concluded that "The exercise helped students achieve long-term acquisition of correct lexical stress for the particular words" (*ibid.*: 153).

According to Chapelle (2001):

> The most convincing way to demonstrate the language learning potential of CALL activity is through the study of learning outcomes. In other words, if learners were to have acquired particular grammatical forms of vocabulary through a CALL task, then results of an assessment after learners have completed the task can provide some evidence for the language learning potential of the task. The evidence is much stronger, of course, if pretest data indicate that the learners did not know the target forms before beginning to work with CALL. Still stronger evidence is obtained if a contrasting group that did not use the CALL task or used the CALL task in another form failed to make similar gains. Any of these designs is strengthened if learners are shown to have retained what was learned at a later time (*ibid.*: 74).

As presented, Hincks' (2002) evaluation included both a pre-test as well as a post-test. If we assume that the methods of evaluation that Hincks employed were both valid and reliable, this is quite a convincing product-oriented evaluation of learners' performance. It, however, lacks a control study[27] and a delayed post-test. The absence of the latter, in our opinion, means that it is not possible to draw the conclusion made by Hincks that the training led to long term acquisition of the pronunciation of *parameter* and *component* from this evaluation. Moreover, Higgins (1995) cautions that:

> we can prove less with [the results of evaluations such as these] ... than we want to, since the stimuli given in the test question are so different from the needs of real-life communication, and test performance is at best an indirect measure of competence (*ibid.*: 74).

Regarding the internal validity of the evaluation, as presented in section 4.7.2, it is believed that process-oriented evaluations ought to be conducted in addition to product-oriented evaluations in order to verify that learners do indeed engage in the processes that are believed

[27] Hincks (2002) indicates that one was planned. We were, however, unable to find a report of it in the literature.

to promote SLA that are being investigated. No such evaluation is reported. It is therefore impossible to tell whether the learners followed the teacher's instructions and hence to attribute the performance gains to the method of instruction. Moreover, other instruction, if any, that the learners were receiving is not reported. It is therefore impossible to say whether the performance gains were a result of the training or the other instruction that the learners were receiving. Another source often cited is exposure and practice outside the classroom (Chapelle and Jamieson, 1990). In this case, practice outside the classroom is unlikely to have caused their improvement because, according to Hincks (2002), L2 speakers of English at KTH who the learners would have come into contact with regularly pronounce *parameter* and *component* incorrectly. The planned control study (*ibid.*) would have, in our opinion, permitted the control of these extraneous variables.

Regarding the external validity of this evaluation, as with any evaluation of CALL software, the results are unlikely to generalise perfectly to another class and instructional setting because no two classes are the same (see section 4.6.1.1.2). Anyone wishing to use Hincks technique for the teaching of lexical stress with their students should be aware that the participants in this evaluation were engineers. As presented in section 2.2, the use of speech editors, requires the ability to interpret and manipulate complex visual displays of speech, namely waveforms and spectrograms, an ability which in general only trained phoneticians (Pennington and Esling, 1996) and perhaps speech engineers possess.

In summary, while quite convincing, the internal validity of this evaluation is questionable. But, the problems could, as said, be easily overcome by conducting a control study.

#### 4.7.2.2 Process-oriented evaluations

One process-oriented evaluation of learners' performance using CALL applications that integrate TTS synthesis was found in the literature. This evaluation, which was conducted by Cohen (1993), looked at the processes that young learners of French as an L2 engaged in when using the writing tool *Composition* described in section 3.3.3.2.

The participants in the evaluation were non-Francophone children from non-Francophone working class families attending pre-school in a suburb of Paris. Qualitative methods were used to evaluate the processes that the learners engaged in when using the software which was permanently available to the learners in a self-service situation, namely:

- observation,
- case studies,
- analyses of learner productions, and,
- interviews with the class teacher.

Regarding the processes in which the learners engaged while using the software, the data gathered revealed:

- "rich verbal exchanges" (Cohen, 1993: 27),
- active participation from all learners,
- learners intuitively repeated after the computer as it spelled out the words that they typed, and later pronounced the letters before the computer pronounced them in game-like fashion, and,
- learners' stories became increasingly elaborate over time.

In other words, they revealed that learners engaged in the following flow of activities presented in Figure 1:

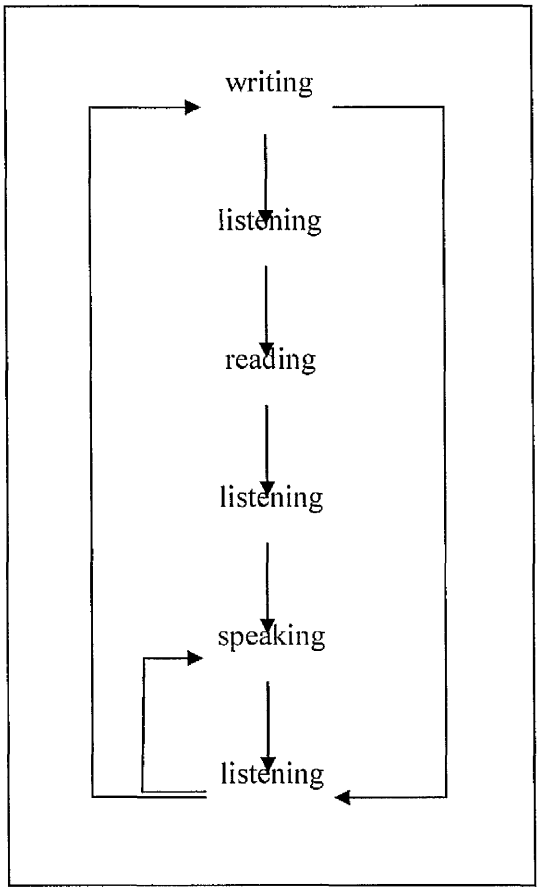



**Figure 1 Naturally occurring flow of activities observed among the young using the voice synthesiser for second-language acquisition (Cohen, 1993: 28)**

From this data, Cohen concluded that, contrary to popular ideas about language acquisition, "writing and reading can develop simultaneously with oral language" (Cohen, 1993: 28).

In addition to data regarding the processes that the learners engaged in, the evaluation also provided the evaluators with data regarding the learners' reactions to the software and their general behaviour. Regarding the learners' reactions to the software, the evaluators observed that the learners were extremely enthusiastic about the use of the software – such was their enthusiasm that it became necessary to control access to the software, and that "The children loved hearing their own stories over and over again, especially when they had signed it" (*ibid.*: 27). Regarding the learners' general behaviour, the data gathered suggested that, as a result of

the use of the software, the learners became more autonomous, gained self-esteem, and their attention span increased.

Regarding the validity of this evaluation, the internal validity is questionable for a number of reasons. First, it is impossible to definitively attribute the processes and strategies in which the learners engaged to the CALL software because the instruction that the learners received from their class teacher was not described. The learners may have developed these processes and strategies from the instruction that they received from their teacher (see Chapelle and Jamieson, 1990). Second, it is impossible to determine whether the reactions reported by the evaluators are really reactions to the software, or rather reactions to the teacher, their classmates, or the general instructional setting:

The internal validity of research based on students' reports depends on the truthfulness and accuracy of the students' reports. When students indicate their attitudes toward CALL, how do researchers know whether they are indicating their sentiments toward CALL, or whether they are indicating their more general attitudes toward the overall L2 learning environment? If, for example students are pleased with their English class and like their teacher, the books, and their peers, the learning environment could generate a "halo error" (Thorndike & Hagen, 1977) in the attitude scores (*ibid.*: 45).

Regarding the external validity of the evaluation, as we have already mentioned numerous times, it is unlikely that the results of any evaluation of any CALL software will generalise perfectly to another class and instructional setting, because no two classes are the same (see section 4.6.1.1.2). As mentioned in section 4.6.1.1.2, teachers and researchers can make an informed assessment of how well such an evaluation would generalise to the learners and the instructional setting in which they are working if sufficient details of the learners and the instructional setting in which the evaluation was conducted. In this respect, while Cohen (1993) provides details regarding the learners, this evaluation is limited because she does not provide details of the instruction that was provided to the learners by their teacher.

In summary, like the other evaluations discussed in previous section, the internal validity of this evaluation is questionable.

### 4.7.2.3 Impact evaluations

As said, one evaluation was found that addressed the impact of the use of CALL applications integrating TTS synthesis. This evaluation, conducted by Stratil *et al.* (1987a), looked at learners' reactions to the use of the aforementioned multifacet language laboratory (see section 4.7.1), which as said integrates TTS synthesis, for the presentation of Spanish grammar explanations and exercises and the rules of Spanish pronunciation.

The participants in the evaluation were 90 pre-O-level (i.e. pre-GCSE) learners of Spanish as a foreign language. Ranging from children to mature students, they were learning Spanish in a variety of different instructional settings:

> a group of first-year university students (with their lecturer), a group of fifth- and sixth-form students from a direct grant school (with their teacher), a group of third- and fourth- form pupils from a convent school (with their teacher), and three eight year olds (children of friends) (*ibid.*: 311).

Their reactions to a demonstration of the software which consisted in 2 sections from a book of Spanish grammar and exercises on those grammar topics proposed in a book of Spanish exercises explanations of the rules of Spanish pronunciation were probed using the following methods of evaluation:

- logs of learners' use of the software,
- questionnaires, and,
- observation.

More specifically, the routes that learners took through the software and the time that they spent using the software were logged. The former were thought to reflect how easy or difficult the learners found the materials, specifically repetition of a grammar point was believed to reflect an inadequate explanation. The latter were thought to reflect learners' enthusiasm for the software; the longer the learners spent using the software, the more enthusiastic they were assumed to be. Regarding the time that learners spent using the software, while one of the learners, a mature student, gave up using the software after only half an hour, many of the learners continued using the software long after the class had ended and expressed a desire to use the software again.

There were four sections to the questionnaire: the first section probed learners' reactions to the physical location of the laboratory and the equipment; the second probed their reactions to the courseware; the third probed their reactions to the presentation of the course materials; and, in the final section, learners were invited to make any further comments they had regarding the software. Two questions in the first section specifically probed learners' reactions to the TTS synthesis. These questions read:

- "As you can see, it is possible to link the computer to an audio system; would you find this sort of package useful in a classroom situation?" (Stratil *et al.*, 1987a.: 315); and,
- "Did you have difficulty with ... the quality of the sound?" (*loc. cit.*).

The questionnaire revealed that learners' reactions to the software in general were on the whole positive. Regarding the use of TTS synthesis specifically:

> Over 68 per cent found the sound linked to the computer a useful classroom aid and, in spite of its lack of natural intonation, 56 per cent had no difficulty with the speech (*ibid.*: 312).

Regarding observation, two researchers attended each of the software demonstrations. They made notes on any difficulties that the learners encountered, learners' reactions to the software as well as their teachers' reactions. Very generally, it was observed that younger learners more readily accepted the software than older learners and that the older learners' reluctance to accept the software was initially shared by the teachers – though seeing the learners use the software changed their attitudes. Learners' initial reactions to the TTS synthesis, specifically, were "to dissolve into giggles" (*ibid.*: 313). Despite this initial reaction, some of the learners proceeded to "repeat aloud the spoken words after the computer, using the language laboratory technique without being prompted to do so and taking to the sound teaching facility quite naturally" (*loc. cit.*). One of the learners did, however, initially question the point of using TTS synthesis. And some of the more advanced learners noticed errors in its pronunciation. Specifically, they "noticed the American pronunciation of the 'R'" (*loc. cit.*). These mispronunciations were believed to result from the fact that "the speech chip used to produce the sounds was originally produced for the US market" (*loc. cit.*).

Regarding the validity and reliability of the evaluation, in our opinion, there are several threats to the internal validity of the evaluation. Firstly, regarding the use of logs, Higgins (1995) and Goodfellow (1999) issue the following warnings:

> Usage can be distorted, with teachers' recommendations having some of the same effect as advertisers' hype. As with sales figures, the best evidence is provided when predictions are falsified, when 'entertaining' material is rejected in voluntary sessions in favour of 'boring' drill, for instance (Higgins, 1995: 73);
>
> however thorough and detailed a computer's record may be, it does not describe what the learners think about what they are doing (Goodfellow, 1999: 112).

In order to overcome the latter problem, it is recommended that more than one researcher look at the data (Chapelle and Jamieson, 1990). If they come up with the same interpretation then that interpretation can be assumed to be valid (*ibid.*). In addition, in order for others to make an informed assessment of the findings of an evaluation, it is recommended that evaluators provide justifications for the inferences that they make (*ibid.*).

The validity of the use of questionnaires is also debatable:

> I suspect that learners are very inaccurate reporters of what they have enjoyed, tending often to report what they think they ought to have enjoyed, or not quite knowing what enjoyment consists of in the context of a learning activity (Higgins, 1995: 73);
>
> students feel much less inhibited and tend to be more 'honest' when asked to voice an opinion, rather than having to put it in writing (L'Huillier, 1990: 84).

Regarding the internal validity of the data obtained through observation of the learners, it is well known that the presence of an observer may affect the results of an evaluation (the observers' paradox). On the other hand, it has been observed that the presence of the evaluator can be positive:

> The participants' comments suggest they approached the language learning situation with a fair degree of anxiety. These people were asked to do several stressful and potentially threatening things. One thing that seemed helpful in alleviating the stress was my being there, engaging in chitchat before and after sessions, getting to know the participants as people; i.e. what one researcher has referred to as 'participatory consciousness', a 'being with' in 'an attitude of profound openness and receptivity' as opposed to 'being there' as an observer (Heshusius, 1994, p. 16) (Murray, 1999: 188).

> During this time, I attempted to build a rapport based on respect, which some researchers claim strengthens the trustworthiness of self-report and introspective data (Grotjahn, 1991; Oxford, 1995) (Murray, 1999: 190).

Like the reactions observed in the evaluation discussed in the preceding section, the reactions observed through the three different methods of evaluation may have been reactions to the teacher, their peers or the instructional setting, and not reactions to the CALL software and the TTS synthesis facility specifically. In addition, as the technology was new to the learners, the results of the evaluation may have been subject to the 'wow' factor (Murray and Barnes, 1998).

Regarding the external validity of this evaluation, in our opinion, it is quite good, because it looks at a range of different ages of learners from a range of different instructional settings. It could, however, be improved by distinguishing the scores and reactions of the different subgroups of participants.

In summary, the internal validity of this evaluation like that of the other evaluations that have been conducted to date is questionable.

## 4.8 Potential reasons for the neglect of evaluation of CALL applications integrating TTS synthesis

There are several reasons why the evaluation of CALL applications integrating TTS synthesis may have been neglected:

- "There is a tendency to assume that if a program is technologically at the cutting edge, it is also effective as a language learning tool" (Fox, 1997: 444);
- evaluations are often not publishable (Pederson, 1987):

> because evaluative research is software-specific, its results should not be published as research articles in professional journals. This recommendation does not diminish the importance of formative and summative evaluative research; it simply acknowledges the fact that most software evaluations become obsolete. CALL research that should be reported are those basic research projects that offer new theoretical information about the psycholinguistic nature of language learning and the way in which a specific computer capability of delivering instruction – its coding elements – has been shown to affect or interact with the learning process. Such studies will stand the test of time (*ibid.*: 109);

and,

- evaluation is costly in terms of time and resources (*ibid.*; Higgins, 1995; Hirschman and Thompson, 1996) which are particularly limited in CALL (Pederson, 1987; Higgins, 1995; Holland, 1995; Levy 1999b) and for which it competes with further development (Hirschman and Thompson, 1996); and,
- "systems are not finished enough to support evaluations" (Holland, 1995: vii).

Of these reasons, it is believed that cost is the biggest reason why evaluations are not conducted: not all teachers are prepared to jump on the latest technological bandwagon, after the failure of the language laboratory many teachers are sceptical about the introduction of unproven technologies (Dunkel, 1990); it is possible to get evaluations published if they

> focus[ ] on the nature of the language acquisition process in terms of second language acquisition theory [or any other theory that CALL draws upon (see section 5.1), my comment] rather than in terms of superficial manipulation of some pedagogical variables (Garrett, 1995: 355);

and, as presented in section 4.5, evaluation ought to be conducted at a number of different stages throughout the lifecycle of a product.

In order to overcome the cost of evaluation, it is necessary to make evaluations as efficient as possible. One way in which evaluations can be made more efficient is presented in the following section.

## 4.9 Benchmarking as a solution to the limitations of evaluation

In order to render evaluation more efficient, and hence less costly, benchmarking, is commonly used in software evaluation in general (Lindgaard, 1994; Grace, 1996; Ralston *et al.*, 2000), and the evaluation of SALTs (Sparck Jones and Galliers, 1996; van Bezooijen and van Heuven, 1997) more specifically.

Originating in the field of surveying where "A benchmark is a surveyor's mark, used as a reference for determining further heights and distances" (Codling, 1998: 7), benchmarking has a long history and today is used in a wide range of fields including surveying, management, economics, education, computing, and so on. Regarding the evaluation of computer systems of

which CALL systems are an example, benchmarking was first used for the comparative evaluation of the adequacy of the processing speed of computers. In this context, a benchmark was a computer program used to measure processing speed which typically produced a single numerical score for the system being tested. Using these programs, developers would compare the scores obtained by their system with those obtained by their competitors, and potential end-users would compare the scores obtained by systems that they were considering acquiring (Grace, 1996). Since then, benchmarking has been applied to other features of the performance of computer hardware including the access time of memory systems, I/O bus traffic, bandwidth, etc., as well as to the performance (Cai *et al.*,1998) and usability (ease of use) of computer software (Lindgaard, 1994). Benchmarking the usability of computer software typically involves measuring the performance of end-users in the completion of a number of 'typical' tasks with the aid of the software.

Today, benchmarking is also used to refer to an activity in which a group of organisations get together to identify the 'best-in-class', through the use of a common test, with the goal of identifying what it is possible to achieve, areas for improvement, and realistic targets (Hetzel, 1993). Through the sharing of information about best solutions and increased communication within the research community, this approach often leads to rapid technological progress (Sim *et al.*, 1998). While TTS systems for use in CALL would benefit from this form of benchmarking, this is not the type of benchmarking that we are proposing here.

CALL developers often find themselves working in the situation that we found ourselves working in the *FreeText* project (Hamel 2003b), that is working in collaboration with a developer of SALTs towards the development of CALL applications. In this situation, what the CALL developer wants to know is whether the TTS system, or any other SALT for that matter, offered by their partner organisation is ready for use in the CALL applications that they wish to develop. They therefore need a test or set of tests that can tell them whether the system meets user requirements. In other words, they need a benchmark of the type described by van Bezooijen and van Heuven (1997):

> By a *benchmark test* we mean an efficient, easily administered test, or set of tests, that can be used to express the performance of a speech output system (or some module thereof) in numerical terms. The *benchmark* itself is the value that characterizes some reference system, against which a newly developed system is (implicitly) set off. The benchmark is preferably chosen such that it guarantees user satisfaction. Consequently,

> if the performance of a new product exceeds the benchmark, its designer or prospective buyer is assured of at least a satisfactory product, and probably even better. (*ibid.*: 497).

As pointed out by van Bezooijen and van Heuven, benchmarking "is more efficient than pairwise or multiple testing of competing products" (*ibid.*: 497) and therefore more cost-effective. Consequently, it overcomes one of the major limitations of evaluation. A further advantage of benchmarking is the ease of interpretation of the results: In most cases, benchmark scores are expressed as single numbers (Grace, 1996; Ralston *et al.*, 2000). In addition, other advantages are brought about by the fact that benchmarking involves evaluation in a common task. Specifically, evaluation in a common task leads to comparability and consistency of results across evaluations (Sparck Jones and Galliers, 1996). Benchmarking could therefore provide a good solution to the neglect of the evaluation of applications of TTS synthesis in CALL.

It would be even better if a generic benchmark test were to be developed for the evaluation of the adequacy of TTS synthesis for use in applications in general and an independent organisation was to apply that benchmark to all systems available on the market and publish the results. CALL developers who find themselves in a situation where they have an idea for an application that would integrate TTS synthesis, or any other SALT for that matter, and want to go out and buy a system that meets their requirements so that they can build that application would then simply need to obtain that publication, compare the scores achieved by the different systems with the acceptance levels for use in the type(s) of CALL application(s) that they are intending to build.

## 4.10 Summary

In this chapter, the principles of SALT and CALL evaluation were reviewed. On the basis of this review, an infrastructure for the evaluation of CALL applications integrating TTS synthesis was put forward and the evaluations of CALL applications integrating TTS synthesis that have been conducted to date were assessed with respect to this infrastructure and the aforementioned principles of evaluation. It was found that very few 'formal' evaluations of CALL applications integrating TTS synthesis have been conducted. Regarding the formal evaluations that had been conducted, as far as it was possible to establish, only one evaluation of the adequacy of TTS synthesis for use in CALL has been conducted (see section 4.7.1), only one product-oriented and one process-oriented evaluation of learners' performance using

a CALL application integrating TTS synthesis has been conducted (see sections 4.7.2.1 and 4.7.2.2), and only one evaluation of the impact of a CALL application integrating TTS synthesis has been conducted (see section 4.7.2.3) – identification of the potential benefits speech synthesis could bring to CALL is considered to fulfil the function of basic research evaluation. Yet, as said, every TTS synthesis system to be used in CALL should be evaluated (Huang *et al.*, 2001) in not only every CALL application that it is intended to be used, but also in every activity planned around those applications (Chapelle, 2001). Moreover, as we have demonstrated, the validity of the evaluations that have been conducted is questionable.

Even though CALL applications integrating TTS synthesis are already available on the market, we believe that it is important to go back and conduct adequacy evaluation because if it is omitted, considerable time and resources may be wasted integrating the technology into applications for which it is not suitable. Moreover, it is believed that "It is not possible to design an effective project *a priori*" (Nelson and Oliver, 1999: 112) and "working systems are hard to build and once built are hard to change, unless carefully designed for flexibility" (Hamburger, 1990: 19). Adequacy evaluation is also important because it enables us to identify the successes and limitation of a technology which is useful because technologies should be exploited for what they are best at (Stevens, 1989).

We believe that the main reason for this is that evaluation is costly. This, as presented in section 4.8, could be overcome through the use of benchmarking. Before benchmarks, generic or specifically for CALL, can be developed, however, the requirements that CALL applications impose on TTS synthesis must be established. Analysis of the requirements that CALL applications impose on TTS synthesis is the subject of the following chapters.

# 5 Requirements of TTS synthesis for CALL: Literature Review

## 5.1 Overview

As mentioned in the previous chapter, it appears that the requirements of CALL applications integrating TTS synthesis for CALL have not been investigated. This is an essential stage in the evaluation process (see section 4.4.1.1). In this chapter, the literature relevant to CALL is analysed for indications of the potential requirements that should be evaluated at the level of adequacy evaluation with a view to developing a benchmark test for the evaluation of the adequacy of TTS synthesis for use in CALL applications. Regarding the literature relevant to CALL, (Chapelle, 1997; Levy, 1997; Chapelle, 1998; Hamel, 2003a), CALL draws on a diverse range of fields including, but not limited to, (second) language acquisition, linguistics, psychology, instructional design and Human-Computer Interaction (HCI) (Chapelle, 1997; Levy, 1997). According to Chapelle (1998), criteria for the design and evaluation of CALL applications should therefore be based on the collective findings of these fields (MacWhinney, 1995; Chapelle, 1998). Similarly, it is suggested here that criteria for the evaluation of TTS synthesis for use in CALL setups should be based on such findings. However, to analyse the findings of all the fields that might provide insights into the optimal TTS synthesis for use in CALL setups would be a huge undertaking, one which time constraints do not permit most CALL developers to engage in (Hamel, 2003a) and one which this study did not permit. CALL refers to "any process in which a learner uses a computer and, as a result, improves his or her language" (Beatty, 2003: 7). It has therefore been argued that the findings of SLA research should be the primary consideration in the design and evaluation of CALL applications (Pederson, 1987; Chapelle, 1997, 1998, 2001). Despite the fact that there is as yet no fully articulated model of SLA, this view is subscribed to here. Like Doughty (1987), Pederson (1987), MacWhinney (1995), and Chapelle (1997; 1998), we believe that CALL provides the ideal conditions to operationalise hypotheses about SLA and further knowledge in the field.

Regarding the selection of a model of SLA for use in CALL, Doughty (1987) recommends the use of a unified model of SLA, that is, one that brings together the main perspectives on SLA. This approach is adopted here. Chapelle (1998; 2001) based her criteria for the design and evaluation of CALL tasks on such a theory, namely Gass's (1997) Interactionist Model. This

model is adopted here. The suitability of this model of SLA as a basis for the articulation of criteria for the evaluation of TTS synthesis for use in CALL setups is assessed in section 5.3. Then, in sections 5.3.1 to 5.3.6, the implications of each of the different stages of the model with respect to the evaluation of TTS synthesis for use in CALL setups are considered in turn.

Further to models of SLA, like Oxford *et al.* (1993), we believe that the goals of language learning, i.e. communicative competence, ought to be taken into consideration when designing and evaluating CALL applications. Before looking at Gass's (1997) Interactionist Model, communicative competence is considered and its implications with respect to the evaluation of TTS synthesis for use in CALL setups are discussed (see section 5.2).

Finally, in section 5.4, the requirements identified as a result of the review of communicative competence and the Interactionist Model (*ibid.*) of SLA are summarised and compared with the findings of chapter 2 regarding the state of the art of TTS synthesis.

## 5.2 Communicative competence

Defined very generally as "that aspect of our competence that enables us to convey and interpret messages and to negotiate meanings interpersonally within specific contexts" (Brown, 1994; 227), opinions are divided with respect to the skills and competences that communicative competence comprises. According to Canale and Swain (1980) and Canale (1983), communicative competence consists in: grammatical competence, sociolinguistic competence, discourse competence, and, strategic competence.

More specifically, grammatical competence, according to Canale and Swain (1980), consists in:

> knowledge of lexical items and of rules of morphology, syntax, sentence-grammar, and phonology[28] (*ibid.*: 29).

For Canale and Swain, sociolinguistic competence consists in knowledge of sociocultural rules, i.e. how to use language appropriately in different sociocultural contexts.

[28] The authors of the article are American. In the USA, 'phonology' has a broader sense than in the UK. While in the UK, 'phonetics', subsumes phonology, in the USA, 'phonology' subsumes phonetics (Brown, 1992). In this instance, 'phonology' refers to phonetics, phonology and prosody, therefore.

> The primary focus of [sociocultural] rules is on the extent to which propositions and communicative functions are appropriate within a given sociocultural context depending on cultural factors such as topic, role of participants, setting, and norms of interaction. A secondary concern of such roles is the extent to which appropriate attitude and register or style are conveyed by a particular grammatical form within a given sociocultural context (*ibid.*: 30).

Discourse competence, according to Canale and Swain (1980) and Canale (1983), consists in knowledge of rules of discourse, i.e. "how to combine grammatical forms and meanings to achieve a unified spoken or written text in different genres" (*ibid.*: 9). Regarding rules of discourse:

> Unity of text is achieved through *cohesion* in form and *coherence* of meaning. Cohesion deals with how utterances are linked structurally and facilitates interpretation of a text. ... [and] Coherence refers to the relationship among the different meanings in a text (*loc. cit.*; emphasis original).

And, finally, strategic competence consists in knowledge of how to:

> handle breakdowns in communication: for example, how to deal with false starts, hesitations, and other performance factors, how to avoid grammatical forms that have not been mastered fully, how to address strangers when unsure of their social status – in short, how to cope in an authentic communicative situation and how to keep the communicative channel open (Canale and Swain, 1980: 25).

Another popular model of communicative competence is the one proposed by Bachman (1990). In this model communicative competence, or 'communicative language ability' as Bachman refers to it, comprises:

- language competence, "a set of knowledge components that are utilized in communication via language" (*ibid.*: 84),
- strategic competence (see above), and,
- psychophysiological mechanisms, the neurological and physical processes necessary for the execution of language (*loc. cit.*).

Language competence, according to Bachman, is further broken down into: organisational competence and pragmatic competence. "Organizational competence comprises those abilities involved in controlling the formal structure of language" (*ibid.*: 87), namely grammatical competence (see page 124 of this thesis) and textual competence (Canale's (1983) 'discourse

competence'). And, pragmatic competence, knowledge of "the relationships between ... signs and referents on the one hand, and ... language *users* and the *context* of communication, on the other" (*idem.*; emphasis original), comprises, according to Bachman, illocutionary competence and sociolinguistic competence, where illocutionary competence refers to knowledge of how to perform different speech acts (see Searle, 1976) and language functions (see Halliday, 1973), while sociolinguistic competence, as in Canale and Swain's (1980) model, refers to knowledge of what language is appropriate for use in different sociocultural contexts (a more detailed description of sociolinguistic competence can be found on page 124 of this thesis).

The *Common European Framework* (CEF) (Council of Europe, 2001) is different again. In this model, like Bachman's (1990), strategic competence does not form part of communicative language competence, rather it is believed to be part of general competence, i.e. competences not specific to language. As regards communicative language competence, for the Council of Europe (2001), communicative language competence comprises: linguistic competence, sociolinguistic competence, and, pragmatic competence. Linguistic competence is the equivalent to grammatical competence, as presented earlier in this section; sociolinguistic competence, as for Canale and Swain (1980) and Bachman (1990), refers to knowledge about what language is appropriate for use in different sociocultural contexts; and, pragmatic competence refers to a combination of Bachman's illocutionary competence, and discourse competence.

Regarding the evaluation of the adequacy of TTS synthesis for use in CALL setups, the following aspects of grammatical competence are determined by the material that the author decides to present to the learner via TTS synthesis: morphology, syntax and sentence-grammar. They are therefore not believed to place demands on TTS synthesis. The choice of pronunciation, phonetics, phonology and prosody on the other hand, is to a certain extent determined by how the TTS synthesis system handles the input. Phonological competence, the ability "to produce a correct pronunciation from a written form" (Council of Europe, 2001: 116) therefore merits further consideration.

Phonological competence is discussed in more detail in the CEF (*ibid.*). According to the CEF, phonological competence comprises:

knowledge of, and skill in the perception and production of:

- the sound-units (*phonemes*) of the language and their realisation in particular contexts (*allophones*);
- the phonetic features which distinguish phonemes (*distinctive features*, e.g. voicing, rounding, nasality, plosion);
- the phonetic composition of words (*syllable structure*, the sequence of phonemes, word stress, word tones);
- sentence phonetics (*prosody*)
  - sentence stress and rhythm
  - intonation;
- phonetic reduction
  - vowel reduction
  - strong and weak forms
  - assimilation
  - elision.

(*ibid.*: 116f; emphasis original)

Regarding the level of phonological competence that learners are expected to attain, the CEF (*ibid.*) distinguishes three levels of user: (A) Basic User, (B) Independent User, and, (C) Proficient User.

Two levels of each type of user are further distinguished. These levels are presented and described in Appendix 1, Table 100. The level of phonological competence at these different levels, according to the CEF framework, are described in Table 7.

**Table 7 Levels of phonological competence in the CEF (Council of Europe, 2001: 117)**

| | **PHONOLOGICAL CONTROL** |
|---|---|
| **C2** | *As C1* |
| **C1** | *Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.* |
| **B2** | *Has acquired a clear, natural pronunciation and intonation.* |
| **B1** | *Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.* |
| **A2** | *Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, by conversational partners will need to ask for repetition from time to time.* |
| **A1** | *Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group.* |

According to Kenworthy (1987: 3), the goal of most learners is to be "comfortably intelligible", i.e. able to be "understood by a listener at a given time in a given situation" (*ibid.*: 13) without them having to put in too much effort, a view which is echoed by many in

the field (see Munro and Derwing, 1999). Intelligibility, as used here, therefore corresponds to the notion of comprehensibility as used by those concerned with the evaluation of TTS synthesis (see section 2.3.2). Regarding accent, Kenworthy (1987: 3) believes that "The great majority of learners will have a very practical purpose for learning English and will derive no particular benefit from acquiring native-like pronunciation". In other words, for Kenworthy accuracy appears to be more important than naturalness. Regarding accuracy, he seems to suggest that intelligibility is determined by the accuracy at the phonetic level alone:

> 'The more words a listener is able to identify accurately when said by a particular speaker, the more intelligible that speaker is.' Since words are made up of sounds, it seems that what we are talking about is the issue of equivalence of sounds. If the foreign speaker substitutes one sound feature of pronunciation for another, and the result is that the listener hears a different word or phrase from the one the speaker was aiming to say, we say that the foreigner's speech is unintelligible. Likewise, if the foreign speaker substitutes a sound in a particular word, but that word is nonetheless understood, then we say that the speech is intelligible (Kenworthy, 1987: 13).

Contrary to Kenworthy, many believe that accuracy at the prosodic level is more important in determining intelligibility than accuracy at the phonetic level:

> Sentence stress is like a backbone. Without it, the utterance is vague and shapeless. Children communicate via sentence stress before learning to say the words properly.
>
> Of all pronunciation skills to give to learners sentence stress is probably the most valuable, as it quickly helps them communicate effectively, even with very little English. Without it, students merely line up words. With it, they make sense (Haycroft, 1992: 57);
>
> Many teaching programmes begin by examining the vowel and consonant sounds, then going to contractions and weak forms, and then, time permitting, to stress and rhythm. To me, this seems back-to-front, since it is the sentence stress that determines which words are unstressed and the speaking speed that dictates just how they are contracted, the vowel and consonant sounds changing accordingly. How is the foreign learner to know which the unstressed words are, without establishing the stresses first? (*ibid.*: 71);
>
> Segmental consonant and word pronunciation relates only to the level of words in this categorisation, which stress rhythm, intonation and other suprasegmental features help to convey the tone. Segmentals may therefore have very limited importance in interactional terms. Smith and Nelson (1983) ... claim that problems of miscommunication arise more often in terms of comprehensibility and interpretability,

> rather than intelligibility.[29] Suprasegmentals relate more to the former levels than segmentals (Brown, 1992: 11).

By 1996, according to Celce-Murcia *et al.* (1996), phonetic and prosodic aspects were beginning to receive equal attention from teachers:

> Today's pronunciation curriculum ... seeks to identify the most important aspects of both the segmentals and suprasegmentals, and integrate them appropriately in courses that meet the needs of any given group of learners (*ibid.*: 10).

In summary, according to Kenworthy (1987), the goal of most language learners with respect to pronunciation is to reach the CEF's level B1, *Threshold Level.* Although there is some evidence to support Kenworthy's view that a slight foreign accent does not adversely affect intelligibility (Munro and Derwing, 1999), many, in particular employers, are intolerant of foreign accents (Sato, 1991). MacCarthy (1975) and Dalton and Seidlhofer (1994) also argue the case for the acquisition of native-like pronunciation, i.e. natural accent-free pronunciation. In their view, accentedness of pronunciation has an effect on how the learner is perceived by his/her interlocutors. Specifically, MacCarthy believes that native speakers will be more friendly towards those learners that speak their language "in a really acceptable manner" (MacCarthy, 1975: 3).

This all suggests that CALL setups place demands on the accuracy and naturalness of the speech generated by TTS synthesis systems at both the phonetic and prosodic levels. Further research is, however, necessary in order to determine the relative importance of accuracy and naturalness at the phonetic and prosodic aspects of the speech generated by TTS synthesis systems.

Regarding acceptability levels, the TTS synthesis in CALL specialists Keller and Zellner-Keller (2000) have suggested that "When the language competence of the system begins to

---

[29] 'Comprehensibility' refers to "the understanding of the meaning of words or utterances (the locutionary force, in speech act terms)" (Brown, 1992: 4). In other words the term is used as it is in speech synthesis circles (see section 2.3.2) and corresponds to Kenworthy's (1987) notion of intelligibility.
'Interpretability' "refers to the understanding of the meaning *behind* words or utterances (the illocutionary force)" (Brown, 1992: 4; my emphasis).
'Intelligibility' is used here as it is in speech synthesis circles (see section 2.3.2). That is it is "restricted to the low-level oral/aural recognition of words and utterances" (*loc. cit.*).

outstrip that of some of the better language users, such systems become useful new adjunct tools" (*ibid.*: 111; Keller, 2002: 7). Evidence to support this view is, however, not provided. Further research is therefore also necessary in order to establish acceptability levels for the aforementioned criteria.

Regarding sociolinguistic competence, one aspect is being able to produce the register appropriate to the sociocultural context (see page 124). Register, as presented in section 3.4, has an effect on phonetic and phonological aspects of speech. This suggests to us that TTS synthesis systems for use in CALL ought to be able to generate the register of speech appropriate to the sociocultural context of the teaching materials, in other words to provide options over register.

Regarding discourse competence, intonation is one of the linguistic mechanisms used to link utterances and meanings together (Tench, 1996). In addition, regarding pragmatic and illocutionary competence, intonation is also exploited to realise communicative functions (*idem.*). All this appears to suggest that the appropriateness of the intonation of the speech generated by TTS synthesis systems ought to be taken into consideration at this level of evaluation.

## 5.3 Gass's (1997) Interactionist Model

The Interactionist Model proposed by Gass (1997), is an expansion of the model of SLA proposed by Krashen (1982). A brief presentation of Krashen's model is therefore in order.

On the basis of a review of the SLA literature, Krashen made five hypotheses about SLA, namely:

(1) the acquisition-learning distinction,
(2) the natural order hypothesis,
(3) the Monitor Hypothesis,
(4) the input hypothesis, and,
(5) the Affective Filter hypothesis.

In his first hypothesis, Krashen proposes that there are two independent processes by which learners can develop proficiency in an L2, namely: (1) by *acquisition*, i.e. by subconsciously 'picking up' the language, (2) by *learning*, i.e. by developing conscious knowledge of TL

rules. The processes are independent in the sense that learnt knowledge of the TL cannot be turned into acquired knowledge (Lightbown and Spada, 1993).

In his second hypothesis, the natural order hypothesis, Krashen proposes that learners acquire the features of the TL in a predictable order,[30] an order which tends to be the same regardless of their L1.[31] Krashen presents studies of morpheme acquisition which demonstrate that learners tend to acquire morphemes in a certain order to support this hypothesis.

Krashen's third hypothesis, the Monitor hypothesis concerns how learners use acquired and learnt knowledge of the TL. According to Krashen, our performance in the TL is primarily accounted for by our acquired knowledge of it. Our learnt knowledge of the TL functions as a Monitor, or editor, which makes changes to the utterances, before or after (self-correction) we speak or write, which have been produced on the basis of our acquired knowledge of the TL. In other words, for Krashen "formal rules, or conscious learning, play only a limited role in second language performance" (Krashen, 1982: 16). Moreover, Krashen posits that learners only have access to the Monitor, i.e. learnt knowledge, when they are not under time pressure, when they are focusing on form and when they know the rule. Evidence for the Monitor Hypothesis is drawn from studies of the morphosyntax of learners' productions under different conditions: in casual conversation, written composition, and grammar tests. The results of use of the Monitor, namely 'unnatural' orders, only appeared in grammar tests, i.e. when all of the aforementioned conditions were met.

According to Krashen's fourth hypothesis, the input hypothesis, languages are acquired through focus on meaning, not through focus on form. More specifically, he proposes that for acquisition to take place TL input must be 'comprehensible', that is it must contain a few, but not too many, segments which are beyond the learner's current level: if input contains no unknown segments, there is nothing new for the learner to acquire, and, if there are too many unknown segments, the learner will not be able to understand the utterances and deduce the (meaning and) function of, and hence acquire, the new segments. Krashen presents simplified

[30] Krashen (1982) notes that this 'natural order' only appears under conditions of acquisition. Use of learnt knowledge of the TL, i.e. use of the Monitor, gives rise to other orders.

[31] On the other hand, Krashen notes that "the order of acquisition for second language is not the same as the order of acquisition for first language, but there are some similarities" (Krashen, 1982: 13).

codes, specifically Caretaker Talk (CT), the register of speech that parents, or other caregivers use to address infants, Foreigner Talk (FT), the register of speech that natives use to address non-native speakers of their L1, and Teacher Talk (TT), the register of speech that teachers use to address learners, and the silent period as evidence for this claim. Specifically, Krashen observes that caretakers, native speakers, and teachers 'rough-tune' the input that they provide to learners to their competency in the TL in order to permit comprehension and subsequently acquisition to take place. The silent period, Krashen claims, "indicates that the child is listening to and comprehending speech addresses to him or her, prior to beginning to produce" (Larsen-Feeman and Long, 1990: 140).

In his fifth and final hypothesis, Krashen proposes that affective variables, such as motivation, self-confidence, and anxiety, affect not only whether learners seek out opportunities to use the TL, but also how much of the TL input in their environment gets through to them, and how deeply they analyse the language that they are exposed to. In particular, he proposes that affective variables act as a filter on input. A learner who has negative attitudes to learning will have a strong or closed filter which will screen out input and make it unavailable for acquisition (Lightbown and Spada, 1993; Johnson, 2001). On the other hand a learner who has positive attitudes to learning will have a weak or open filter which will allow input to get through and hence make it available for acquisition (*idem.*). Attitudes to learning differ across learners. The affective filter therefore accounts for individual differences in SLA (Gass, 1997). Krashen cites studies of the following affective variables as support for this hypothesis: motivation, self-confidence, and anxiety. Specifically, he observes that: "Performers with high motivation generally do better in second language acquisition"; "Performers with self-confidence and a good self-image tend to do better in second language acquisition"; and, "Low anxiety appears to be conducive to second language acquisition, whether measured as personal or classroom anxiety" (Krashen, 1982: 31).

Returning to the Interactionist Model of SLA proposed by Gass (1997), just as models of SLA in general contend that learners require exposure to large quantities of TL input for acquisition to take place, so too does Gass's model. More specifically, like Krashen, Gass proposes that for acquisition to be possible, that input must be 'comprehensible'. However, unlike Krashen, she recognises, as proponents of interactionist models of SLA do in general, that the learner does not necessarily need to be provided with comprehensible TL input in the first place. TL

input may also be made comprehensible, and hence acquisition may be facilitated, through interaction with native speakers by means of negotiation of meaning, the clarification of meaning through conversational adjustment or modified interaction (Larsen-Freeman and Long, 1991; Lightbown and Spada, 1993). The initial evidence for this hypothesis was drawn from analyses of CT and FT (Long 1985). CT and FT, however, only provide partial support for this hypothesis, namely that caretakers and native speakers make conversational adjustments when interacting with learners. Whether negotiation of meaning brings about comprehension and subsequently acquisition is not clear from studies of CT and FT. While the initial evidence upon which the interactionist hypothesis was based was inadequate, it is corroborated by the findings of empirical studies. In support of the hypothesis that negotiation of meaning serves to render input comprehensible for learners, Gass presents a number of studies in which the type of input provided to learners was manipulated. Corroborating the results of an earlier study by Pica *et al.* (1987) which compared the effects of unmodified and negotiated input on learners' ability to follow a set of instructions, in these studies, negotiated input was found to have a positive effect on comprehension (Loschky, 1994; Gass and Varonis, 1994; Ellis *et al.*, 1994).[32] Regarding the question of whether negotiation of meaning brings about acquisition, Gass draws support for this hypothesis from the aforementioned study by Ellis *et al.* (1994), and a further study by Mackey (1995) In the study conducted by Ellis *et al.* (1994) learners who engaged in negotiation acquired more word order than those who were presented pre-modified input. Mackey's (1995) study focused on the acquisition of question formation by learners of English as a Second Language (ESL). She found that learners who engaged in negotiation of meaning proceeded through the different stages in the acquisition of English question formation quicker that those who did not. These results were replicated by Mackey (1999). Further evidence for this hypothesis is to be found in a study by (Sato, 1985: 185f) in which the phonological features of the speech of a young Vietnamese ESL learner in three contexts "(1) free conversation, (2) oral reading of continuous text, and (3) elicited imitation of words and short phrases" were investigated. In this study it was observed that new forms appeared first in communicative, i.e. interactional contexts.

A further difference between Krashen's model and the Interactionist Model proposed by Gass is that in the Interactionist Model exposure to 'comprehensible' input is not considered to be

[32] The results of Gass and Varonis' (1994) study suggested that pre-modified input might also facilitate comprehension, as did a previous study by Long (1985).

sufficient for acquisition to take place. In order for acquisition to be possible, the Interactionist Model contends that learners must *apperceive,* i.e. 'notice' (either consciously or unconsciously), and attend to, unknown segments in the input, i.e. gaps in his/her knowledge of the TL. The latter process is referred to as *focus on form* (Long, 1985). Evidence for the role of apperception and focus on form is drawn from Schmidt and Frota's (1986) study of Schmidt's own experience of learning Brazilian Portuguese. Analysis of the journal which Schmidt kept of her experience of learning Portuguese and the monthly recordings she had made of her conversations in Portuguese, showed that the presence of a particular form in the TL input did not guarantee her producing it in her own output, rather it was only once she had noticed the form and thought about it that she began to produce it (Schmidt, 1990; Gass, 1997).

Another difference regards *comprehension.* While Krashen contends that comprehension is a sufficient condition for acquisition, Gass does not. According to Gass, input must also be comprehended at the systemic or functional level (i.e. syntactic, morphological, or phonological levels) for acquisition to be possible. Input which has been understood at both the semantic and the systemic level is referred to as 'intake'. Gass draws support for this hypothesis from studies by VanPatten and Cadierno (1993) and VanPatten and Santz (1995) of grammar instruction. In these studies the researchers attempted to influence the way in which learners process input by manipulating the mode of instruction. Specifically, the effects of "traditional grammar instruction in which information is presented to the learner and then practiced" (Gass, 1997: 135) on the acquisition of direct object pronouns in Spanish were compared with those of instruction in which practice focused on processing mechanisms. Participants in the group that received processing instruction were found to be better able to perceive and produce the forms taught than those who received traditional instruction.

Moreover, for Gass *intake*, comprehension of TL input at the semantic and systemic levels, is not a sufficient condition for acquisition; rather, for acquisition to be complete intake must contribute to the development of the learner's model of the TL, i.e. intake must be integrated into the learner's existing model of the TL. *Integration*, the process of comparing intake with the learner's existing model of the TL and identifying opportunities for the integration of intake into that model, has two main outcomes: development of the learner's model of the TL and storage. Regarding input which has been put into storage, "Integration is not a one-time

affair" (Gass, 1997: 25), further evidence, i.e. input, may simply be needed in order to integrate it into the TL model.

A final difference between Krashen's model and the Interactionist Model proposed by Gass concerns the role of *output*. For Krashen output is simply the outcome of SLA and plays no role in the process of acquisition: "we acquire spoken fluency not by talking but by understanding the input, by listening and reading" (Krashen, 1982: 60). For Gass, on the other hand, output is more than the "overt manifestation" (Gass, 1997: 7) of the process, producing output plays a crucial role in SLA. According to Gass the production of output:

- enables learners to notice further gaps in their knowledge of the TL,
- pushes learners to make further hypotheses about the TL,
- enables learners to test further hypotheses about the TL, and,
- enables learners to automatise further aspects of their knowledge of the TL.

More specifically, according to Gass, gaps in TL knowledge may be noticed by the learner him/herself when s/he experiences difficulties producing utterances in the TL (Gass, 1997), or alternatively they may be drawn to the attention of the learner through feedback from interlocutors on their productions (Swain, 1985). Regarding the second role of output in SLA, producing output is believed to promote hypothesis generation because it "may force the learner to move from semantic to syntactic processing" (*ibid.*: 249). Regarding the third role of output, producing output and using it in his/her interaction with other speakers is one way in which a learner can test hypotheses that s/he has generated about the target language. In this context, feedback provides the learner with evidence as to whether s/he should accept, revise or reject hypotheses that s/he has made about the TL. Regarding the fourth role of output production, Gass proposes that repeated production of TL output by the learner may provide the practice necessary to move from controlled processing of TL forms to automatic processing. The SLA process is therefore not necessarily linear. Regarding whether learners actually engage in producing TL output, just as affect is believed to play a role in determining whether learners attend to input, it is believed to play a role in determining whether learners produce output in the TL. This is referred to as 'Willingness to Communicate' (WTC) (Chapelle, 2001). Evidence for the hypothesis that the production of output plays a crucial role in SLA is drawn from Swain's (1985) studies of children learning French in an immersion setting. Observing that even after years of study the proficiency of these learners did not improve as expected Swain compared a class of children learning French in immersion and a

class of native French-speaking children. What she noticed was a lack of opportunity for productive use of French in the French immersion setting. Since the publication of Gass (1997), a number of researchers have attempted to establish a direct link between output production and acquisition (Mitchell and Myles, 2004). In two studies of vocabulary acquisition by learners of ESL (Ellis and He, 1999; de la Fuente, 2002), one study of grammar acquisition by learners of ESL (Nobuyoshi and Ellis, 1993), and one study of the acquisition of Japanese honorifics (Nagata, 1998) learners who were encouraged to produce output in the TL outperformed those who were not given the opportunity to produce TL output in post-tests, in two other studies of second language grammar acquisition no differences were found between the experimental and control groups (Izumi *et al.*, 1999; Izumi and Bigelow, 2000). Regarding WTC, according to Gass evidence for this hypothesis can be found in numerous studies, she does not, however, provide any examples.

In the following sections, the path from input to acquisition according to the Interactionist Model is presented and an attempt is made to identify the demands that CALL setups place on TTS synthesis.

### 5.3.1 Input

*Input* is the starting point in the Interactionist Model (see 5.3). As presented in the previous section, just as models of SLA in general contend that a learner's success in acquiring an L2 depends on the quantity of input that they are exposed to, so too does the Interactionist Model. Exposure to large quantities of TL input is, however, not considered to be a sufficient condition for SLA. According to Gass, for input to constitute a candidate for acquisition three conditions must be met. First, there must be something to be learnt in the TL input, i.e. the input must contain some TL forms that the learner has not yet acquired – Chapelle (2001) refers to this as the *utility condition*. Second, the TL input must be 'comprehensible'. And, third, the learner must be 'open' to receiving and attending to the input, in other words their *affective filter* must be weak.

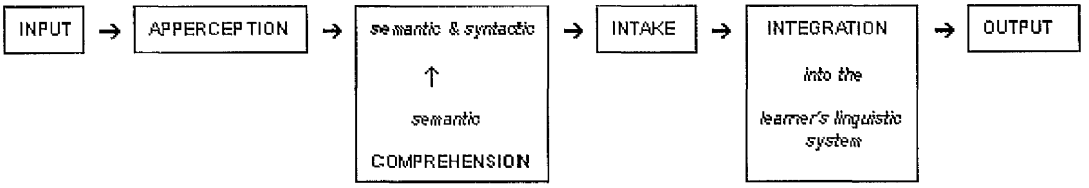



**Figure 2 Basic components in the SLA process in interactionist research (Chapelle, 1998: 23).**

TTS synthesis offers a means of providing input to learners in CALL setups. Regarding its evaluation, of the conditions mentioned above, quantity, comprehensibility and affect are believed to impose demands on TTS synthesis for use in CALL setups. These conditions are therefore discussed in more detail in the sections that follow.

Regarding the utility condition (Chapelle, 2001), the utility of the input provided to learners by means of TTS synthesis in CALL setups will be determined by the systemic (syntactic, morphological, phonological, etc.) features of the material that the author of an application decides to present to the learner by means of TTS synthesis. The utility condition is therefore not believed to place demands on TTS synthesis and will therefore not be considered further here.

### 5.3.1.1 Quantity of input

Regarding the quantity of input which learners need to be exposed to in order for SLA to take place, as said, like behaviourist and cognitivist models of SLA, Gass's Interactionist Model of SLA contends that learners require exposure to large amounts of TL input. The literature on best practice in LL&T corroborates this view (Kenworthy, 1987; Laroy, 1995; Pennington, 1996, 1999). More specifically, these authors believe that exposure to lots of different voices and accents is beneficial to learners for the development of both perception and production skills. That learners require exposure to a variety of different voices in order to develop L2 perception skills is corroborated in a study by Lively *et al.* (1993). Following on from a study by Logan *et al.* (1991) which demonstrated that high variability training brought about improvements in a group of Japanese learners' ability to discriminate English /r/ and /l/, Lively *et al.* (1993) conducted an investigation designed to isolate/determine the effects of the

two axes along which the stimuli presented to the learners in Logan's study varied, namely phonetic environment and talker. Specifically, in the Lively *et al.* study, one group of learners received training in which both the phonetic environment and talker were varied and another group of learners received training in which only phonetic environment was varied, i.e. all stimuli were produced by the same talker. While increases in performance were observed between pre-test and post-test for stimuli presented during the training sessions, those learners who received the high variability training in which both phonetic environment and talker were varied generalised to novel words produced by both a familiar and an unfamiliar talker, whereas those who received the training in which stimuli varied in phonetic environment only did not. That high variability training leads to improvements in auditory discrimination which generalise to novel stimuli and novel talkers is corroborated by a follow-up study by Lively *et al.* (1994) in which the durability of the effects of the training procedure were also investigated. Learners' performance did not decrease significantly from the initial post-test to a post-test presented 3 months after the training sessions and 6 out of the 8 learners still performed better in a post-test presented 6 months after the training sessions than in the pre-test. Regarding production, in two further studies the effects of the high variability training method on Japanese learners' production of /r/ and /l/ phonemes was investigated. In both studies learners productions of /r/ and /l/ in a post-test were more accurately identified by native speakers than those produced by the learners prior to completing the training (Akahane-Yamada, 1996; Bradlow *et al.*, 1997). That access to input produced by lots of different voices is desirable for the development of production skills is also corroborated by a study by Probst *et al.* (2002). They studied the effects of imitating (1) a learner-selected voice, (2) a similar voice (in terms of F0, SR, and gender), and (3) a dissimilar voice on learners of ESL pronunciation. Comparison of the first three utterances and the last three utterances imitated by the learners during the training sessions revealed that those who listened to a speaker with voice features similar to their own improved most and those who chose the speaker that they imitated improved least. A more fine-grained analysis across all the learners, revealed that:

- learners who imitated a speaker of the same gender (and hence pitch) improved less than those who imitated a speaker of the opposite gender,
- learners who imitated a speaker who spoke at a similar rate of articulation to them improved more than those who imitated a speaker who spoke at a dissimilar rate of articulation, and,

- learners who imitated a speaker whose phones were of a similar duration to theirs improved more than those who imitated a speakers whose phones were of a dissimilar duration.

These are but a few of numerous characteristics which distinguish individual speakers. Other characteristics which remain to be investigated include whether it is better to imitate a slower voice, a hyperarticulated voice, or the voice of someone of the same age (Probst *et al.*, 2002).

It is therefore suggested that TTS synthesis for use in CALL setups must be able to generate large quantities of TL input and provide options over voice, accent, SR and duration. Further research on SLA may reveal that further options are required.

### 5.3.1.2 Comprehensible input

As presented in section 5.3, input may be comprehensible *per se*, or it may be made comprehensible through negotiation of meaning, the clarification of meaning through conversational adjustment or modified interaction (Larsen-Freeman and Long, 1991; Lightbown and Spada, 1993). Input that is comprehensible *per se* contains a few but not too many TL forms which the learner has not yet acquired (see section 5.3). As presented in section 5.3.1, the TL forms that input contains are determined by the author. This condition is therefore is not believed to place demands on TTS synthesis.

Regarding negotiation of meaning, in studies of CT, FT, and TT input modifications have been observed at the following levels:

- phonetics, phonology and prosody,
- morphology,
- syntax,
- vocabulary,
- discourse structure, and,
- topic (Chaudron, 1988; Bingham Wesche, 1994; Ellis, 1994).

Like the TL forms that input presented by TTS synthesis in CALL applications contains, the morphology, syntax, vocabulary, discourse structure and topic of the input provided to learners via TTS synthesis in CALL setups will be determined by the authors of CALL applications. Modifications at these levels are also therefore not believed to place demands on TTS synthesis systems. Modifications of input at these levels are therefore not discussed any

further here. The phonetic features, phonology, and prosody of the input presented to learners by means of TTS synthesis in CALL setups, on the other hand, will be determined by how the TTS synthesis system handles the materials that authors wish to be presented via TTS synthesis. Modifications at the phonetic, phonological, and prosodic levels therefore merit further consideration.

Regarding modifications at the phonetic, phonological and prosodic levels, it has been observed that CT, FT, and TT are characterised by: slower SR, more frequent and longer pauses, clearer articulation, avoidance of contractions, clearer marking of word boundaries, and exaggerated intonation (Chaudron, 1988; Bingham Wesche, 1994; Ellis, 1994)

Regarding the role of these modifications in rendering input comprehensible, increased frequency and duration of pauses in CT, FT and TT are believed to aid comprehension by giving learners more time to process TL input (Chaudron, 1988; Bingham Wesche, 1994). Use of less contracted forms etc. is believed to aid comprehension by making utterances more transparent and giving learners more time to process the TL (Ellis, 1994). Exaggerations of intonation is believed to aid comprehension by directing learners attention to content words and making constituent boundaries more salient, i.e. easy to perceive (Bingham Wesche, 1994).

Regarding the degree of modification, very generally it has been observed that input is adjusted to the proficiency level of the learner (Krashen, 1982; Bingham Wesche, 1994; Ellis, 1985, 1994). This phenomenon has also been observed at the phonetic (Henzl, 1979) and prosodic levels (Sachs, 1977; Henzl, 1973; Henzl, 1979; Hakanson, 1986), more specifically. Regarding modifications at the phonetic level (Henzl, 1979) observed that "with low-level students, the teachers used more accurate, standard pronunciation" (Ellis, 1985: 145). Regarding modifications at the prosodic level, both Henzl (1973, 1979) and Hakanson (1986) observed that teachers adjusted their SR according to the learner's proficiency in the TL. Also regarding modifications at the prosodic level, "Sachs (1977) found that mothers tune the pitch, intonation, and rhythm of their speech to the perceptive sensitivity of their children" (Ellis, 1994: 248).

Regarding the evaluation of the adequacy of TTS synthesis for use in CALL setups, the findings presented above would seem to suggest that TTS synthesis for use in CALL need to

provide options over: SR, pause frequency and duration, clarity of articulation, phonology, and intonation.

The fact that caretakers, native speakers, and teachers make modifications to input does not mean that those modifications lead to comprehension and subsequently acquisition (Bingham Wesche, 1994; Gass, 1997). As presented in section 5.3, there is some evidence which suggests that negotiation of meaning in general leads to comprehension and acquisition. Regarding the effects of input modifications at the phonetic, phonological and prosodic levels on comprehension and acquisition, as far as it is possible to establish, the only variables that have been investigated are SR and pause frequency. Regarding the link between SR and comprehension, the results of studies by Grosjean (1972), Kelch (1985), and Griffiths (1990) suggest that decreasing SR has a positive affect on comprehension. The results of Grosjean's study also suggested that increasing pause frequency has a positive effect on comprehension.

As far as it is possible to establish, none of the studies have investigated the link between these variables and acquisition. It is therefore not possible to say definitively whether options over SR, pause frequency and duration, clarity of articulation, phonology, and intonation are necessary for SLA. As presented in section 5.3, CALL provides the ideal environment to operationalise and test such hypotheses about SLA (Doughty, 1987; MacWhinney, 1995). The provision of options over the aforementioned variables are therefore desirable at the current time for this purpose.

#### 5.3.1.3 The affective filter

Regarding the affective filter, it has been demonstrated that TTS synthesis manifests personality, in particular introversion and extroversion (Nass and Min Lee, 2001). Studies in HCI have shown that users react to and treat computers as if they were human beings (Weizenbaum, 1976; Reeves and Nass, 1996). According to Beebe (1985: 404), learners have "input preferences", which depend on, like many aspects of the SLA process, "Such concepts as solidarity, ethnocentricism, feelings of identification, loyalty, social identity, and intergroup dynamics" (Gass, 1997: 94). It is therefore expected that whether learners attend to the input presented to them via TTS synthesis in CALL setups will depend on the perceived personality of the voice of the TTS synthesis system. This is particularly expected to be the case when TTS synthesis is being used as a pronunciation model because "pronunciation is a key element

of one's self-image and the image that one projects to other" (Pennington, 1999: 433).[33] The perception of the personality of a TTS synthesis system is therefore believed to be an important consideration at this level of evaluation. As said, affective variables differ across learners and are responsible for individual differences in learning. The present research is limited to the identification of criteria which determine whether TTS synthesis is ready for use in CALL with learners in general. Affective variables and the specific demands that they place on TTS synthesis in the CALL context will therefore not be discussed further here.

### 5.3.2 Apperception

Noticing that there is something in the TL input to be acquired, i.e. *apperception*, and subsequently attending to what the TL form that has been noticed, i.e. *focus on form* (Chapelle, 2001), is the first stage in input utilisation (Gass, 1997). According to Schmidt (1990), whether or not a learner apperceives a given form in TL input depends on the:

- learner's expectations,
- frequency of the TL form,
- perceptual salience of the TL form,
- learner's knowledge of the TL, and
- task demands.

Regarding expectations, on the basis of research in psychology, Schmidt suggests that unexpected forms are more likely to attract learners' attention than expected forms. On the other hand, regarding the frequency of TL forms, forms which are very frequent in TL input are also very likely to be noticed by learners. Evidence for this hypothesis is drawn from Larsen-Freeman's (1976) study of the order of acquisition of English morphemes by L2 learners, in which it was observed that those morphemes which are most frequent in TL input are acquired first. The presence of specific TL forms in TL input is determined by the author. These conditions are therefore not believed to impose any demands on TTS synthesis systems.

Regarding Schmidt's (1990) third condition, "Salience refers to the ease with which learners are able to perceive grammatical features in input" (Ellis, 1985: 67). "as a result of their phonological form or their position in utterances" (*ibid.*: 120) some forms are inherently more salient than others. For example, research suggests that free morphemes are more salient than

[33] The link between pronunciation and identity is also highlighted by Guiora (1972), Kenworthy (1987), Dalton and Seidlhofer (1994), and Laroy (1995).

bound morphemes (Wode, 1981) and suffixes are more salient than prefixes (Slobin, 1973). In addition, it has been suggested that the saliency of TL forms may be increased through negotiation of meaning (Ellis, 1985, 1994; Chaudron, 1988; Gass, 1997; Chapelle, 2003). The view that input modification may increase the saliency of TL forms is also expressed in the literature on best practice in LL&T (Callamand, 1981; LeBel, 1990; Haycroft, 1992; Jenner, 1992; Wessels and Lawrence, 1992; Dalton and Seidlhofer, 1994; Laroy, 1995; Pennington, 1996). Studies of negotiation of meaning therefore merit further consideration. As presented in section 5.3.1.2, in studies of CT, FT, and TT input modifications have been observed at the following levels: phonetics, phonology and prosody, morphology, syntax, vocabulary, discourse structure, and topic (Chaudron, 1988; Bingham Wesche, 1994; Ellis, 1994). Regarding the use of TTS synthesis in CALL, it was suggested that modifications at the phonetic, phonological and prosodic levels ought to be considered in more detail because the phonetic features, phonology, and prosody of the input presented to learners by means of TTS synthesis in CALL setups, will be determined by how the TTS synthesis system handles the materials that the author or teacher wishes to present via TTS synthesis (see section 5.3.1.2). Regarding modifications at the phonetic, phonological and prosodic levels, as presented in section 5.3.1.2, CT, FT, and TT are characterised by: slower SR, more frequent and longer pauses, clearer articulation, avoidance of contractions, less phonological variation, clearer marking of word boundaries, and, exaggerated intonation (Chaudron, 1988; Bingham Wesche, 1994; Ellis, 1994). Regarding 'input enhancement', the modification of input to increase the perceptual saliency of TL forms (Sharwood Smith, 1991), in the literature on best practice in LL&T it is suggested that reducing SR might make the following phonetic features more salient: graveness (low pitch), darkness, closeness, backness, labialisation, nasalisation, aspiration, absence of voicing, affrication, retroflexion and velarisation (LeBel, 1990). In addition, at the phonological level, it is proposed that reducing SR might make vowel reduction more salient (*idem.*). Regarding the prosodic level, it is suggested that manipulating (both reducing and increasing) SR might make the rhythm of input more salient (*idem.*). Manipulating the frequency and distribution of pauses, it is suggested may make the rhythm (*idem.*) and intonation (Léon and Léon, 1969) of the TL more salient. Regarding the avoidance of contractions, Chaudron (1988: 71) suggests that avoidance of contractions, including consonant cluster and vowel length reduction, and phonological variation may render morphophonemic forms more salient as their full form will be "evident on the surface". As regards the exaggeration of intonation, it is believed that exaggeration (both of the depth and

height) of pitch makes both rhythm (LeBel, 1990) and intonation patterns (Pennington, 1996) more salient (Todaka, 1990). Regarding the exaggeration of the depth of pitch, it is suggested that this may render the following features of phonemes more salient: graveness and darkness (LeBel, 1990). Exaggeration of the height of pitch, on the other hand, is believed to render the acuteness and clearness of phonemes more salient (*ibid.*). Also regarding intonation, it is believed that manipulating stress (both increasing and decreasing) may render rhythm (*ibid.*) and stress more salient (Haycroft, 1992; Wessels and Lawrence, 1992; Bonneau *et al.*, 2000). Increasing stress, more specifically, is believed to increase the saliency of lexical phrases (Bingham Wesche, 1994; Chapelle, 2003) and the following phonetic features: acuteness and clearness (LeBel, 1990). Decreasing stress, on the other hand, is believed to increase the saliency of the graveness and darkness of phonemes (*ibid.*).

In addition to the modifications that are observed in CT, FT and TT, a number of other types of modified input are suggested in the literature on best practice in LL&T. In particular, it is suggested that:

- increasing SR might make the following phonetic features more salient: acuteness, clearness, openness, fronting, palatalisation, voicing, and frication (LeBel, 1990);
- modifying of volume might make rhythm more salient (*ibid.*);
- increasing the duration of voiced consonants might make voicing more salient (Léon and Léon, 1969);
- increasing the duration of stressed syllables might make stress more salient (*ibid.*);
- exaggerating of the articulation of phonemes might make them more salient (*ibid.*; Pennington, 1996);
- modifying rhythm, in particular rapping, might make stress and intonation patterns more salient (Wessels and Lawrence, 1992), as is whispering (Haycroft, 1992; LeBel, 1990);
- syllabifying speech might make rhythm more salient (Léon and Léon, 1969);
- whispering might make stress (Haycroft, 1992), intonation and the graveness of phonemes (LeBel, 1990) more salient;
- humming might render intonation more salient (Dalton and Seidlhofer, 1994; Laroy, 1995; Yoram and Hirose, 1996); and,

- reiterant speech, speech "in which a natural utterance is mimicked in a series of repetitions of a single syllable such as "ma"" (Cutler *et al.*, 1997: 145), might render rhythm more salient (LeBel, 1990).

The literature presented above would seem to suggest that TTS synthesis for use in CALL setups ought to provide options over: SR, pause frequency and pause duration, volume, articulation, phonology, intonation (including pitch, stress and rhythm), and voicing. Regarding voicing, the literature would seem to suggest that, in addition to normal voicing, TTS synthesis systems for use in CALL setups ought to be able to simulate whispering. Other options it would seem to suggest that TTS synthesis systems for use in CALL ought to provide are humming and reiterant speech. The ability of state-of-the-art TTS synthesis systems to meet these demands is discussed at the end of this section.

Just as the fact that caretakers, native speakers and teachers make these modifications to the input that they provide learners does not mean that the input will be comprehensible, it also does not mean that it makes TL forms salient, learners apperceive them and acquire them. As far as it is possible to establish the effects of input modifications at the phonetic, phonological, and prosodic levels have yet to be investigated. Nevertheless the provision of options over the aforementioned parameters is desirable in order to permit the research into their effects on apperception and focus on form, and ultimately acquisition. As said, CALL provides the ideal environment for the operationalisation of SLA models (Doughty, 1987; MacWhinney, 1995).

Regarding the learner's knowledge of the TL, on the basis of research on the role of memory in learning, Schmidt (1990) suggests if a learner has not developed automatic control over the other TL forms in the input, it is unlikely that s/he will notice the new TL forms in the input because all their STM will have been taken up processing the other TL forms. This condition depends on the learner. It is therefore not believed to have any implications as far as the evaluation of TTS synthesis for use in CALL setups is concerned.

Regarding task demands, Schmidt suggests that learners are more likely to notice TL forms if comprehension of those forms is required to complete a task. This condition is not believed to impose any demands on TTS synthesis systems and will not be considered any further here. It should, however, be taken into consideration at subsequent levels of evaluation.

Further to the conditions identified by Schmidt, Gass (1997) suggests that time pressure also has an affect on whether learners notice a given target form in TL input. Specifically, she suggests that the more time that the learner has at their disposal, the more likely that the learner is to be able to notice forms in the input. TTS synthesis, like digital recordings, by its very nature, reduces time pressure because it permits the repetition of utterances as many times as the learner requires (see section 3.2). Of course, whether or not a learner is able to listen to input as many times as they require will depend on the design of the task on which they are working. This condition is therefore an important consideration at subsequent levels of evaluation.

Regarding focus on form, Skehan (1998) identifies six conditions which may affect whether a learner attends to a form that they have apperceived in the TL input, namely: time pressure, modality, support, surprise, control, and stakes (*ibid.*; Chapelle, 2001). Of these conditions which are presented in Table 8, surprise is believed to place demands on TTS synthesis: as presented in section 2.5.1.4, the speech generated by TTS synthesis systems based on concatentative synthesis, the currently dominant approach to TTS synthesis, may contain distortions which may distract the learner when working on CALL tasks. We therefore believe that the smoothness of the speech generated by TTS synthesis systems ought to be addressed at this level of evaluation. The demands of time pressure, as presented above, are met by the very nature of TTS synthesis. Regarding modality TTS synthesis by its very nature overcomes some of the limitations of the spoken modality (see above). The other conditions are determined by the design of the task that the learner is working on and are not believed to place any demands on TTS synthesis.

**Table 8 Conditions that may affect focus on form during L2 tasks (adapted from Chapelle, 2001: 49)**

| Attention affected by … | Definition | Reason |
|---|---|---|
| **Time pressure** | An urgency in achieving communication caused by one's own anxiousness of external factors. | When no time pressure exists, attention to form is more likely. |
| **Modality** | Whether the language is spoken or written. | Written communication typically affords more opportunity for attention to form, whereas spoken language often occurs under time pressure to achieve fluency. |
| **Support** | Cues or information available to help in constructing meaning during task completion. | When some learners have help with some aspects of the language, their attentional resources are more free to be devoted to form. |
| **Surprise** | Introduction of an unexpected element during task completion. | The surprise element might be expected to decrease attention to form because of the interruption of plans and need to focus on the surprise, but these hypotheses would depend on the nature of the surprise. |
| **Control** | Who makes decisions about the direction that the task is to take. | Control of various aspects of the task by the teacher or the learner may help to prompt focus on form, but research is needed to investigate questions about control |
| **Stakes** | Learners' perception of the importance of accurate performance. | Tasks perceived as high stakes are likely to prompt more attention to form. |

In summary, the literature reviewed in this section suggests that TTS synthesis for use in CALL must be smooth, i.e. free from distortions that might distract the learner. In addition, the provision of options over the following parameters are desirable in order to permit the investigation of their contribution to SLA: SR, pause frequency and pause duration, volume, articulation, phonology, intonation (including pitch, stress and rhythm), and voicing. Regarding voicing, the literature would seem to suggest that, in addition to normal voicing, the possibility to generate humming and reiterant speech are also desirable for the same reason in

order to permit the investigation of their effects on SLA. As presented in section 2.5.2.4, state-of-the-art TTS synthesis systems do provide options over SR, volume and intonation, however, SR and intonation can only be manipulated within a narrow range (Hertz *et al.*, 2000). As presented in section 5.3.1.2, adequacy evaluation ought to investigate whether this is sufficient for the purposes of providing comprehensible input in CALL setups. Options over pause frequency and duration, clarity of pronunciation and phonology and the ability to generate whisper, humming and reiterant speech do not appear to be available.

### 5.3.3 Comprehension

As presented in section 5.3, once learners have identified that there is something to be learnt in the TL input, according to the Interactionist Model (Gass, 1997), *comprehension* of the input at the semantic level must be achieved. According to Gass, whether a learner can understand a given TL utterance at the semantic level depends on their prior knowledge of language universals, their MT, the TL, and other languages. These are learner-internal factors and are therefore not believed to place demands on TTS synthesis for use in CALL setups.

### 5.3.4 Intake

Next, it is proposed that comprehended input must be converted to *intake*, i.e. comprehended at the systemic or functional level (syntactic, phonetic, phonological, etc.). Some of the factors which affect apperception (see section 5.3.2) are also believed to affect the conversion of comprehended input to intake (Kamaravadivelu, 1994; Gass, 1997), namely the:

- individual characteristics of the learner (e.g. age, gender, etc.),
- preferred learning and communication strategies of the learner,
- the affective state of the learner, and,
- the learners' prior linguistic knowledge.

The implications of these factors with respect to the evaluation of TTS synthesis for CALL purposes have already been discussed in sections 5.3.1.3 and 5.3.2.

### 5.3.5 Integration

Finally, for acquisition to be complete, intake must be integrated into the learner's current model of the TL. According to Gass, the conditions which affect integration are similar to those that affect apperception. The implications of these factors with respect to the evaluation of TTS synthesis for CALL purposes are discussed in section 5.3.2.

### 5.3.6 Output

Once a new segment or rule has been integrated into the learner's model of the TL, it will then be possible for the learner to produce output on the basis of that new knowledge. As said, the production of output is the "overt manifestation" (Gass, 1997: 7) of the process. The SLA process does not necessarily finish at the stage of integration, however. Producing output might re-initiate the process. Through the production of output, learners might: notice further gaps in their knowledge of the TL, generate further hypotheses about the TL, test further hypotheses about the TL; and/or, automate further aspects of the TL. In addition, feedback on their TL productions from native speakers might lead learners to notice gaps in their knowledge of the TL. Due to their affective state, learners may, however, not be willing to communicate in the TL. For these learners, the SLA process will therefore stop at the stage of integration.

Regarding the demands that CALL places on TTS synthesis for CALL purposes, TTS synthesis offers a means of providing feedback to learners in CALL setups. Feedback is therefore considered in more detail in section 5.3.6.1.

Affect, as already discussed in section 5.3.1.3, may place demands on the personality of TTS synthesis systems for use in CALL setups and is an important subject for further research if the needs of individual learners are to be met.

#### 5.3.6.1 Feedback

Feedback (also referred to as negative evidence), information provided to learners on the correctness of their TL productions, according to Gass (1997) can be divided into two main types:

> Indirect feedback, as when an interlocutor says, "What?", "I didn't understand you," or "Could you say that again?" and direct feedback, as is often found in classrooms: "No, that isn't right, it should be ..." or " What would the correct form be?" (*ibid.*: 113).

Of these two types of feedback, the latter, direct feedback is believed to place demands on TTS synthesis. Regarding direct feedback, it has been observed that it can take three different forms:

> *metalinguistic feedback* (involving either a metalinguistic question designed to elicit a correct response or rule from a learner or provision of a metalinguistic rule), *repetition of incorrect production* (where the teacher repeat an incorrect learner sentence with a change in intonation to signal that it is incorrect); and *focus on form* (where 'the teacher draws attention to the error by using stress, snapping fingers, gasping, or stating outright that the production is incorrect' ([Spada and Lightbown, 1993]: 224) (Ellis, 1997: 79-80; emphasis original).

In order to be able to provide the latter two types of explicit feedback in CALL setups, it is suggested that TTS synthesis ought to provide options over intonation and stress. While there is evidence that explicit feedback in general has a positive effect on SLA (see Ellis, 1985), whether the use of intonation to indicate that a learner production is incorrect or the use of stress to draw attention to an incorrect form have a positive effect on SLA have not been investigated yet. It is therefore not possible to say whether the provision of options over intonation and stress are essential. As mentioned many times previously in this chapter, CALL provides the ideal environment to operationalise hypotheses about SLA (Doughty, 1987; MacWhinney, 1995). Options over intonation are therefore desirable for the purposes of further SLA research.

## 5.4 Summary: Requirements of TTS synthesis for CALL

In this chapter, the goals of SLA and models of SLA were discussed and analysed for indications of the requirements of TTS synthesis for use in CALL setups. This analysis would appear to suggest that CALL setups impose requirements on the following aspects of the speech generated by TTS synthesis systems: quantity, quality, and flexibility.

More specifically, regarding quantity, the literature suggests that TTS synthesis for use in CALL setups ought to be able to generate large quantities of speech. The generative nature of TTS synthesis (see section 3.2) permits it to meet the requirements placed on quantity.

Regarding quality, the literature suggests that CALL imposes requirements on the comprehensibility and naturalness of the speech generated by TTS synthesis systems. It suggests that comprehensibility is determined by the accuracy of the speech at both the phonetic and prosodic levels. The investigations presented in the following chapter attempt to determine whether CALL applications do indeed place demands on these aspects of the quality of the speech generated by TTS synthesis systems.

Regarding the question of whether state-of-the-art TTS synthesis systems might meet the demands placed on these aspects of the quality of the speech generated by TTS synthesis systems, if indeed CALL applications do place demands on them, comprehensibility is highly correlated with intelligibility (Manous *et al.*, 1985), high levels of which, as presented in section 2.5.1, it has been possible to attain for many (Allen *et al.*, 1987; Edgington *et al.*, 1996b). Naturalness at the phonetic level is to a certain extent given by the dominant approaches to TTS synthesis, namely concatenative synthesis and USS, because in these approaches segments of recordings of natural human speech are concatenated (see sections 2.4.2.3 and 2.4.2.4). While naturalness at the prosodic level is higher for systems based on USS than for TTS synthesis systems based on concatenative synthesis because in USS segments with the appropriate prosody are extracted from recordings of natural human speakers (Conkie, 1999), as mentioned in sections 2.5.1.4 and 2.4.2.4, the accuracy and naturalness at the prosodic level of the speech generated by TTS synthesis systems based on both USS and concatenative synthesis are, however, limited by the inadequacy of methods for determining the prosodic specification of utterances (Dutoit, 1997; Rodman, 1999; Henton, 2002). Acceptance levels for these requirements are not clear from the literature. It is therefore not possible to say whether they have been met. The investigations presented in the following chapter also provide an insight into whether state-of-the-art TTS synthesis systems have met acceptance levels for these aspects of the quality of the speech generated. The identification of acceptance levels is, however, a subject for further research.

Also regarding the requirements placed on the quality of the speech generated by TTS synthesis systems, in addition to comprehensibility, naturalness and accuracy, it is suggested that CALL applications place demands on smoothness. Regarding smoothness, as presented in section 2.4.2.3, the speech generated by TTS synthesis systems based on concatentative synthesis and USS, the currently dominant approaches to TTS synthesis, may contain distortions and discontinuities. As presented in section 2.4.2.4, the amount of distortion should, however, be less for USS because little segment manipulation is carried out (Conkie, 1999). Validation of this requirement and the identification of acceptance levels for it are subjects for further research.

Regarding flexibility, the literature suggests that TTS synthesis systems for use in CALL setups ought to provide options over: voice (gender, age, etc.), accent, register, SR, and, duration.

More specifically, the literature suggests that learners' reactions to the personality of the voice ought to be investigated and that it ought to be possible to match the SR and duration of the speech generated by the TTS synthesis systems to those of the speech of the learner.

In addition, in order to be able to test some of the hypotheses of and further develop current models of SLA, flexibility over the following are desirable: pause frequency and duration, clarity of articulation, phonology and intonation.

As presented in sections 2.5.2.4 and 2.5.2.5, the dominant technologies, concatentive synthesis and USS, permit the manipulation of duration, amplitude and pitch (Hertz *et al.*, 2000) and hence the provision of options over SR, duration and intonation (*ibid.*). Duration, volume and pitch, and hence SR, duration and intonation, can, however, only be manipulated within a narrow range (*ibid.*). Further research is necessary in order to determine what degree of flexibility over these options is required.

Also regarding flexibility, for the same purposes, the literature suggests that TTS synthesis ought to be able to provide options which permit the generation of the following types of speech: whisper, humming, and, generate reiterant speech (see section 5.3.2).

Options over pause frequency and duration, clarity of pronunciation and phonology and the ability to generate whisper, humming and reiterant speech, on the other hand, do not appear to be available.

In summary, if indeed CALL applications do place demands on the aspects of the quality and flexibility of the speech generated by TTS synthesis systems suggested by the research on SLA, the literature on TTS synthesis suggests that state-of-the-art TTS synthesis systems may already meet some of these requirements and hence be ready for use in some of the suggested applications. Further research is, however, necessary in order to determine whether CALL applications do actually place demands on the aspects of the quality and flexibility of the

speech generated by TTS synthesis systems suggested by the literature on SLA and to determine acceptance levels for those requirements. In the following chapter, two investigations designed to validate the findings of this literature review regarding the requirements that CALL setups impose on the quality of the speech generated by TTS synthesis systems are presented. As said, these investigations also give an insight into acceptance levels for these requirements and whether they have been met.

# 6 Requirements analysis: investigation

## 6.1 Overview

In the previous chapter, it was established that CALL setups place demands on the quantity, quality and flexibility of the speech generated by TTS synthesis systems. Of these aspects it was established that the demands placed on quality and flexibility required further investigation. Demands placed on the flexibility of the speech generated by TTS synthesis systems are evaluated in different ways to those placed on the quality: quality is in general[34] evaluated in perceptual experiments (Schmidt-Nielsen, 1995; van Bezooijen and van Heuven, 1997); flexibility is evaluated through the use of checklists (Bailly, 2001). It will therefore be necessary to consider them separately. As presented in sections 2.5.1.4 and 2.5.2.4, manipulating the speech generated by TTS systems may alter its quality (Beutnagel *et al.*, 1999; Bickley *et al.*, 1999; Hertz *et al.*, 2000). We therefore believe that the demands placed on the quality of the output of TTS synthesis systems for use in CALL should be addressed first, before those placed on the flexibility. The demands placed on the quality of the output of TTS systems for use in CALL are therefore the subject of the investigations presented here.

With respect to the quality of the speech generated, as stated, the literature suggests that comprehensibility, which is determined by accuracy, is a central requirement of TTS synthesis for CALL, that naturalness is an additional requirement, and that the smoothness of the speech ought to be evaluated as well. Further investigation is, as said, needed to validate the requirements suggested in the literature. This was the goal of a preliminary exploratory investigation. As presented in section 3.1, in general, requirements differ according to the setup in which an NLP system is used. TTS synthesis, as presented in section 3.4, is used in a variety of CALL setups. More specifically, the preliminary exploratory investigation, reported in detail in Handley and Hamel (2004; 2005) and presented briefly in section 6.2, looks at the requirements of the different roles that TTS synthesis systems may assume within CALL applications namely, (1) RM, (2) PM, and (4) CP (see section 3.4). While informative, the results of this experiment were inconclusive. A further investigation addressing this question was therefore conducted. This investigation is reported in detail in section 6.3.

[34] Attempts have also been made to automate the evaluation of the quality of the output of TTS synthesis systems (Quackenbush and Thomas, 1988). These methods have, however, not been validated yet (van Bezooijen and van Heuven, 1997).

On the basis of the results of these two investigations, recommendations for the evaluation of TTS synthesis for CALL purposes are presented in section 6.4.1, and recommendations for the use of TTS synthesis in CALL are presented in section 6.4.2.

## 6.2 Preliminary exploratory investigation

As stated, it is necessary to validate the requirements that the literature suggests that CALL setups impose on the quality of the speech generated by TTS synthesis systems, namely comprehensibility, accuracy, naturalness and smoothness. In order to achieve this goal, and investigate the hypothesis, presented in section 3.4, that the different roles TTS synthesis may assume within CALL applications, namely RM, PM, and CP, impose different demands on the quality of the speech generated by TTS synthesis systems, 12 teachers of French were asked to rate the speech generated by the TTS synthesis system *FIPSvox* (Gaudinat and Wehrli, 1997), a research system based on diphone concatenation (see section 2.4.2.3), for use in CALL applications in the three roles mentioned above. More specifically, 60 utterances, 20 representative of each of the 3 roles selected from existing CALL applications and one LL&T manual were synthesised and presented to participants blocked by role. For the RM corpus, 20 sentences from the text *Les voleurs d'écriture* (Begag, 1990) exploited in *FreeText* (Hamel, 2003b) were used. For the PM corpus, 20 utterances, each focusing on a different aspect of the pronunciation of French, were selected from *La Portée des Sons* (Garant-Viau, 1994). And, for the CP corpus, 20 consecutive turns were selected from one of the simulated dialogues in the ASR-based CALL program *Talk to Me Français: The Conversation Method* from Auralog (Auralog, 2002). The experimental procedure was a modified version of van Santen's (1993) word pointing paradigm. In a first pass, participants were presented the speech generated by the TTS synthesis system one utterance at a time and asked to rate its acceptability with respect to its use in the role indicated and its comprehensibility. As said, the utterances were blocked by role. At the end of each block, participants were also asked to rate the appropriateness of the speech generated by the TTS synthesis system for use in the role indicated. Acceptability was intended to get at the minimal level of acceptance, and appropriateness was intended to get at the top level of acceptance. Comprehensibility was used as it is in TTS synthesis circles (see section 2.3.2). Then, in a second pass participants were presented the speech generated by the TTS synthesis system accompanied by its orthographic transcription one utterance at a time and asked to highlight any errors in the speech which they believed affected its suitability for use in the role indicated. At the end of each block,

participants were also asked to indicate the types, frequency and seriousness of those errors.[35] It was believed that if comprehensibility was indeed an essential requirement of TTS synthesis for use in each of the three roles in CALL applications, there would be a strong correlation between the ratings of the comprehensibility of the speech generated by the TTS synthesis system and the ratings of the appropriateness and acceptability of the speech generated by the TTS synthesis system for use in the different roles. Regarding the other requirements imposed on the quality of the speech generated by TTS synthesis systems by CALL, it was believed that these would be reflected by the categories of errors identified in the speech.

Ratings of the appropriateness and acceptability of the speech generated by the TTS synthesis system for use in the three different roles were found to differ, providing support for the hypothesis that the different roles place different demands on the quality of the output of TTS synthesis systems. Regarding ratings of the comprehensibility of the speech generated by the TTS synthesis system, these were also found to differ across the three roles. A positive correlation was observed between ratings of the acceptability of the speech generated by the TTS synthesis system for use as an RM and ratings of the comprehensibility of the speech generated by the TTS synthesis system, but not between ratings of the acceptability of the speech generated by the TTS synthesis system for use as a PM and ratings of the comprehensibility of the speech generated by the TTS synthesis system, or between ratings of the acceptability of the speech generated by the TTS synthesis system for use as a CP and ratings of the comprehensibility of the speech generated by the TTS synthesis system. Contrary to the hypothesis that comprehensibility is an essential requirement of all CALL setups, these findings bring into question whether comprehensibility has a role to play in determining the acceptability of TTS synthesis for use as a PM and as a CP, but at the same time they provide support for the hypothesis that the different roles place different demands placed on the quality of the speech generated by TTS synthesis systems. Regarding other demands placed on the quality of the speech generated by TTS synthesis systems by the three

[35] In van Santen's (1993) original word pointing paradigm, participants were presented the text and the output generated by a TTS synthesis system of a corpus of utterances which they were asked to listen to paying particular attention to a predefined list of categories of error. These categories of error included 'outright mispronunciation', 'bad letters', 'missing letters or words', 'wrong stress', 'wrong word emphasised', 'overall voice quality', 'choppiness', and 'bad rhythm'. If they noticed one of these categories of error, they were asked to highlight the text corresponding to where in the corpus the error was made, to indicate the category of error made, and then to indicate its seriousness on a three-point scale.

different roles, the following classes of error highlighted by the participants suggest that accuracy is indeed a requirement: bad segments, bad words, bad phrasing, inappropriate intonation, bad sentence stress, and inappropriate rhythm. Cases of exaggerated intonation were also highlighted by participants, suggesting that naturalness is also a requirement. Classes of error relating to register, namely inappropriate dropping and retention of e-muet (schwa) and inappropriate omission and insertion of optional liaisons, were also highlighted, suggesting that register is another additional requirement. The rated frequency and seriousness of the classes of error highlighted differed across the three roles. This finding provides further support for the hypothesis that the different roles place different demands on the quality of the speech generated by TTS synthesis systems. Participants also commented that the speech lacked emotion, suggesting that TTS synthesis for use in CALL also needs to be expressive. Distortions and discontinuities were not mentioned by any of the participants. This suggests that either smoothness is not a requirement or that the levels of smoothness required are met.

Firm conclusions about the demands placed on the quality of the output of TTS synthesis systems in the 3 roles could, however, not be made due to the small sample size and further investigation was recommended. As said, the results of the experiment suggested that the three roles place demands on the register and the expressiveness of the output of TTS synthesis systems. It was therefore recommended that further experiments address the register and the expressiveness of the TTS output in addition to comprehensibility, accuracy, naturalness and smoothness. Such an experiment is reported in the next section, section 6.3. Regarding the method of investigation, one of the participants commented that it was difficult to judge whether the output was appropriate and acceptable for use in the different roles because examples of applications typical of the different roles were not provided. This suggested that further experiments should provide more contextualisation. Regarding contextualisation, the purpose of adequacy evaluation is to avoid wasting time and resources integrating a technology into an application for which it is not suitable. One way in which more contextualisation could be provided without recourse to the creation of actual applications, i.e. integration, is by presenting participants mock screen shots of potential applications.

## 6.3 Main investigation

The investigation presented here is a further attempt to validate the requirements of TTS synthesis for CALL identified in the literature review and investigate the hypothesis that the different roles that TTS synthesis systems may assume within CALL applications place

different demands on the quality of the speech generated by TTS synthesis systems. This investigation does not, however, address smoothness for it was decided that smoothness, or rather the effects of smoothness, that is whether distortions or discontinuities, if there are any, distract learners, ought to be evaluated by learners and not by teachers. On the other hand, in addition to the other aspects of the quality of the speech generated by TTS synthesis systems upon which the literature suggested that CALL applications place demands, namely comprehensibility, accuracy and naturalness, this investigation also addresses the additional aspects upon which the preliminary exploratory investigation suggested that CALL applications place demands, namely register and expressiveness.

Regarding the hypothesis that the different roles that TTS synthesis systems may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, in addition to the roles already considered, we suggest that the use of TTS synthesis systems for the provision of pronunciation models at the suprasegmental, or prosodic level, may place different demands on the quality of the speech generated by TTS synthesis systems to the use of TTS synthesis for the provision of pronunciation models at the segmental, or phonetic level, because different aspects of the quality of the speech are the focus of instruction and because the models typically differ in length and syntactic complexity. In the experiment presented here, this hypothesis was also investigated. The experiment presented here thus investigated the demands placed on the quality of TTS synthesis systems when used in the following roles in CALL applications: (1) as an RM, (2) as a PM at the phonetic level (phonetic PM), (3), as a PM at the prosodic level (prosodic PM), and (4) as a CP.

Regarding the method of evaluation employed in this investigation, there are two main approaches to 'acoustic evaluation', the evaluation of the speech generated by TTS synthesis systems (van Bezooijen and van Heuven, 1997), namely subjective and objective. In subjective approaches human participants evaluate the speech generated, whereas objective evaluation are automated evaluations. Automated evaluations, as presented in Footnote 34 have, as far as we are aware, not been validated yet (*ibid.*). Objective evaluations will therefore not be considered further here. Regarding subjective evaluations, two approaches are distinguished: functional and judgemental evaluations. In judgemental evaluations, human participants are asked to rate how well they *think* a system performs with respect to a

particular criterion (*ibid.*). In functional evaluations, on the other hand, human participants are asked to complete a task, their success in which is believed to indicate how well a system *actually* performs with respect to a particular criterion (*ibid.*). We believe that functional evaluations are more likely to be valid than judgemental evaluations because they evaluate actual system performance as opposed to participants' impressions of its performance. Ideally, therefore functional tests would have been used in this investigation. Functional tests are, however, time consuming. Regarding functional evaluations of the intelligibility of the speech generated by TTS synthesis systems, the *Diagnostic Rhyme Test (DRT)* (Peckels and Rossi, 1973; Voiers, 1983) takes 15 minutes per synthesiser, the *Modified Rhyme Test (MRT)* (House *et al.*, 1965) takes 25 minutes, the *Speech Assessment Methods (SAM) Standard Segmental Test* (Howard-Jones, 1992) takes 30 minutes, the *Bellcore Test* (Spiegel *et al.*, 1990) takes 40 minutes, and the *CLuster IDentification (CLID)* test (Jekosch, 1992) takes 2 hours (van Heuven and van Bezooijen, 1997). The use of functional tests was therefore not feasible in this investigation. It was therefore necessary to adopt a judgemental paradigm. Of all the judgemental paradigms that have been proposed for the evaluation of the quality of the speech generated by TTS synthesis systems, the *International Telecommunication Union Telecommunication Standardisation (ITU-T) Overall Quality Test* (Schmidt-Nielson, 1995; van Bezooijen and van Heuven, 1997; Polkosky and Lewis, 2003), which comprises 8 categorical estimation scales[36] which address acceptance, overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate, and voice pleasantness (see page 184 of this thesis), is the most efficient (van Bezooijen and van Heuven, 1997; Polkosky and Lewis, 2003):

> With 4 test items per system, testing 4 synthesis systems with 3 reference conditions (i.e. 7 different sources) takes about one hour for one group of subjects, including instructions to subjects and training (van Bezooijen and van Heuven, 1997: 562).

In addition, the validity, reliability, and sensitivity of *Mean Opinion Score-Expanded (MOS-X)* (Polkosky and Lewis, 2003), the latest version of the *ITU-T Overall Quality Test*, are proven

[36] Categorical estimation is used to refer to scales which are numbered and anchored. An example of such a scale would be "a 10-point scale which runs from 1: extremely incomprehensible to 10: extremely comprehensible" (van Bezooijen and van Heuven, 1997: 506). The alternatives to categorical estimation are "*Paired comparison*, where subjects indicate which of two synthesisers sounds more comprehensible" (*loc. cit.*) and "*Magnitude estimation*, where subjects assign a value expressing, or draw line of a length which is equal to the magnitude of, their impression of comprehensibility" (*loc. cit.*).

(*ibid.*). A modified version of *MOS-X* was therefore used in this investigation which compared the levels of adequacy (suitability for use when other options such as digitised speech are available), acceptability (suitability for use when other options are not available), and quality (see section 6.3.1.3.3) of the speech generated by one voice of each of 4 different French TTS synthesis systems and 2 voices of 1 other TTS synthesis system when used in the 4 different roles presented above, namely as (1) an RM, (2) a phonetic PM, (3) a prosodic PM, and (4) a CP. The systems included in the investigation were based on concatenative synthesis and USS. The purpose of the inclusion of two different voices generated by the same TTS synthesis system was to obtain preliminary data regarding whether different voices generated by such TTS synthesis systems ought to be considered as different systems from the point of view of evaluation. – as presented in section 2.5.1.4, the quality of the speech generated by two different voices provided by a single TTS synthesis system based on either concatenative synthesis or USS may differ significantly.

Although not the goal of this study, the ratings of the adequacy and acceptability of the speech generated by the different TTS synthesis systems provide an insight into the readiness of the state-of-the-art TTS synthesis systems for use in CALL.

Regarding the experimental hypotheses, the literature presented in chapters 2, 3, and 5 leads to 6 main predictions:

- CALL setups in general place demands on the following aspects of the quality of TTS synthesis systems: comprehensibility, intelligibility, accuracy, naturalness, expressiveness and register;
- the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems;
- the speech generated by different TTS synthesis systems differs in quality;
- different TTS synthesis systems will be suitable for use in different roles in CALL applications;
- the speech generated by different voices of the same TTS synthesis system differs in quality;
- different voices will be suitable for use in different roles in CALL applications.

With respect to the first hypothesis, as presented in section 5.2, it was expected that all CALL setups would place demands on the comprehensibility, accuracy and naturalness of the speech generated by the TTS synthesis systems because these are the goals of language learners. It was believed that if indeed CALL applications do place demands on these aspects of the quality of the speech generated by TTS synthesis systems, if the TTS synthesis systems evaluated in this investigation did not meet the demands placed on those aspects of the quality of the speech generated, then they would not receive top ratings, i.e. 6 or 7, for those aspects of the quality of the speech – it was not believed necessary to identify requirements which have already been met because in our opinion developers of TTS synthesis systems will not release a new version of a system that performs worse than the previous version of the system.

Regarding the second hypothesis, as presented in section 3.4 and earlier in this section, it was expected that the different CALL setups would place different demands on the quality of the output because the output is being used for different purposes, different aspects of the speech are the focus of instruction, and the utterances to be synthesised are of different lengths and complexity. More specifically, the role of the CP is to stimulate talk. It was therefore believed that comprehensibility is sufficient in this role and that high levels of accuracy and naturalness are not essential. Regarding the aspect of speech which is the focus of instruction, the use of TTS synthesis as a phonetic PM focuses on the phonetic level. It was therefore expected that the use of TTS synthesis as a phonetic PM would place greatest demands on accuracy and naturalness at the phonetic level. The use of TTS synthesis as a prosodic PM, on the other hand, focuses on the prosodic level. It was therefore expected that the use of TTS synthesis as a prosodic PM would place greatest demands on the prosodic level. It was believed that, if indeed the different roles investigated do place different demands on the quality of the speech generated by TTS synthesis systems, the ratings of the adequacy and the acceptability and some of the different aspects of the quality of the speech generated by the different TTS synthesis systems would differ across the different roles.

Regarding the third hypothesis, the TTS synthesis systems evaluated in this investigation differed primarily with respect to the PTS techniques that they employed. Regarding the PTS techniques employed by the systems, the systems evaluated in this investigation were based on either concatenative synthesis or USS because these are the techniques most commonly employed in state-of-the-art TTS synthesis systems. The quality of the speech generated by

such systems depends on the type of segments on which synthesis is based, the quality of the database, the quality of segmentation, the algorithm used for digitising segments, smoothing boundaries, and manipulating the prosody (see sections 2.4.2.3, 2.4.2.4, 2.5.1.4 and 2.5.1.5). The systems evaluated in this investigation differ in many of these respects (see section 6.3.1.3.1). It was therefore expected that the speech generated by the TTS synthesis systems evaluated in this investigation would differ in quality. In both concatenative synthesis and USS, the individual segments are extracted from a corpus of real human speech. It was therefore expected that the speech generated by all the synthesisers evaluated in this investigation would be highly comprehensible and highly accurate and natural at the phonetic level. In unit selections synthesis, unlike concatenative synthesis, an attempt is made to select and concatenate segments with the appropriate prosodic parameters. The speech generated by TTS synthesis systems based on USS is therefore expected to be more accurate and natural at the prosodic level and hence also more comprehensible – providing cues to segmentation and the structuring of utterances, prosody has an important role in determining comprehensibility (Cutler *et al.*, 1997) – than the speech generated by systems based on concatenative synthesis. It was believed that if indeed the quality of the speech generated differs across TTS synthesis systems, then ratings of the different aspects of the quality of the speech generated would differ across TTS synthesis systems.

The fourth hypothesis leads from the second and third hypotheses. If the different roles place different demands on the quality of the speech generated by TTS synthesis systems, and if different TTS synthesis systems produce speech of different quality, then it is expected that different TTS synthesis systems will be adequate and acceptable for use in different roles in CALL applications. More specifically, it was predicted that systems based on USS would be more adequate for use in all roles than systems based on concatenative synthesis because they are expected to generate higher quality speech than systems based on concatenative synthesis in general (see section 2.5.1.5). In particular, it was predicted that systems based on USS would be more adequate and acceptable for use as pronunciation model at the prosodic level than those based on concatenative synthesis: as presented earlier in this section, it is expected that the use of TTS synthesis as a prosodic PM will place high demands on the accuracy and naturalness of the speech generated at the prosodic level and systems based on USS are expected to generate speech which is more accurate and natural at the prosodic level than systems based on concatenative synthesis. It was believed that if indeed the systems based on

USS were more adequate and acceptable for use in all roles than systems based on concatenative synthesis, then ratings of the adequacy and acceptability of the systems based on USS would be higher than those of the systems based on concatenative synthesis for all roles.

Regarding the fifth hypothesis, as presented in section 2.5.1.4, due to the fact that the different voices provided by concatenative and unit selection synthesis systems are based on different databases, it was expected that the speech generated by two different voices provided by the same TTS synthesis system would differ in quality. Moreover, the voices in this study differed with respect to sex, one was male and one was female. In a study conducted by Bradlow *et al.* (1996), female speakers were found to have larger f0 ranges, larger vowel spaces, use less reduced forms, and articulate more precisely than male speakers. It was therefore expected that the speech generated in the female voice would be more comprehensible and accurate at the phonetic level than that generated in the male voice. It was believed that if indeed the quality of the speech generated in different voices by the same TTS synthesis system differs, then ratings of the different aspects of the quality of the speech would differ across the voices. More specifically, if the female voice was indeed more intelligible and accurate at the phonetic level than the male voice, it was expected that ratings of the comprehensibility and the precision of phonemes of the speech generated in the female voice would be higher than those of the speech generated in the male voice.

The sixth hypothesis leads from the second and fifth hypotheses. If the different roles indeed place different demands on the quality of the speech generated by TTS synthesis systems, and if different voices indeed produce speech of different quality, then it is expected that different voices will be adequate and acceptable for use in different roles. More specifically, as said, a male and a female voice were evaluated. If indeed the female voice is more intelligible and more accurate at the phonetic level, then it is believed that it will be more suitable for use as a phonetic PM than the male voice, because phonetic aspects of speech are the focus of instruction in this role.

### 6.3.1 Method

In the sections that proceed, the following aspects of the method are presented in turn: design, participants, apparatus and materials, and procedure.

### 6.3.1.1 Design

One voice offered by each of four different French TTS synthesis systems, and two voices offered by another French TTS synthesis system were evaluated in this investigation with respect to their use in each of the 4 roles identified in section 6.3. The TTS synthesis systems and voices evaluated are presented in Table 9.

**Table 9French TTS synthesis systems and voices evaluated**

| | TTS synthesis system | Proprietor | Voice |
|---|---|---|---|
| **S1** | *AT&T Next-Gen* | AT&T | Alain |
| **S2** | *Nuance Vocalizer* | Nuance | Julie Deschamps |
| **S3** | *eLite* | Multitel | Vincent |
| **S4** | *Elan Tempo*[37] | Elan | Cathy |
| **S5** | *Elan Tempo* | Elan | Robert |
| **S6** | *BrightSpeech*[38] | Babel Technologies | Julie |

The criteria upon which the TTS synthesis systems evaluated in this investigation were selected are discussed further in section 6.3.1.3.1, where the main features of the systems are also presented.

The 4 roles identified in section 6.3 were: (1) RM, (2) phonetic PM, (3) prosodic PM, and (4) CP.

The investigation had a related design. The Dependent Variables (DVs) were the quality of the speech generated by the different TTS synthesis systems with respect to their use in each of the four roles, specifically the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, appropriateness of register, and pleasantness of voice of the speech generated by the TTS synthesis systems, the adequacy and acceptability of the speech generated by the TTS synthesis systems for use in the 4 roles, and the readiness of TTS synthesis in general for use in the 4 roles. Regarding the quality of the speech generated by the TTS synthesis systems, while comprehensibility, intelligibility, choice of pronunciation, naturalness of voice, expressiveness, appropriateness of register, and pleasantness of voice were rated directly, ratings of accuracy were based on ratings of precision of phonemes and appropriateness of prosody and ratings of naturalness were based

[37] In January 2005, Elan merged with Babel Technologies and Babel Infovox to form Acapela Group who now market *Elan Tempo* as *High Density TTS (HD TTS)*.
[38] In January 2005, Babel Technologies merged with Elan and Babel Infovox to from Acapela Group. Acapela Group now market *BrightSpeech* as *High Quality TTS (HQ TTS)*.

on ratings of ratings of naturalness of the phonemes and naturalness of the prosody. Specifically, the mean was taken in each case. The Independent Variables (IVs) were the different TTS synthesis systems and the different roles that TTS synthesis systems may assume within CALL applications.

All measures were obtained using metrics which consisted in a question and a 7-point MOS scale. The questions used to probe the different DVs and the endpoints of the scales on which participants were asked to record their ratings of these variables are presented in Table 10 (see section 6.3.1.3.3 for a more detailed description of the design of the questionnaire).

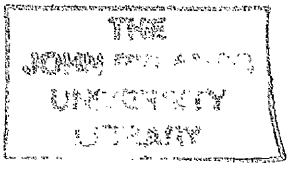


**Table 10 Metrics used in the investigation**

| Measure | Question | Endpoints |
|---|---|---|
| **Adequacy** | Is the speech synthesis adequate for use as a reading machine[39] (in comparison with other media)? | 1 (not at all adequate) to 7 (extremely adequate) |
| **Acceptability** | Is the speech synthesis acceptable for use as a reading machine[40] (when it is not possible to use other media)? | 1 (very unacceptable) to 7 (very acceptable) |
| **Readiness** | How ready do you think that TTS synthesis in general is for use as a reading machine?[41] | 1 (No, not at all ready) to 7 (Yes, entirely ready) |
| **Comprehensibility** | Is the message easy to understand? | 1 (very difficult) to 7 (very easy) |
| **Intelligibility** | Are the individual phoneme/sounds and words easy to recognise (and distinguish one from another)? | 1 (very difficult) to 7 (very easy) |
| **Choice of pronunciation** | Is the pronunciation correct? | 1 (incorrect) to 7 (correct) |
| **Precision of phonemes** | Is the articulation of individual phonemes/sounds precise? | 1 (very imprecise) to 7 (very precise) |
| **Prosody** | Is the prosody (music) of the phrase appropriate? | 1 (very inappropriate) to 7 (very appropriate) |
| **Naturalness of phonemes/sounds** | Do the phonemes/sounds sound natural/human? | 1 (not at all natural/human) to 7 (very natural/human) |
| **Naturalness of prosody** | Does the prosody (music) of the phrase sound natural/human? | 1 (not at all natural/human) to 7 (very natural/human) |
| **Naturalness of voice** | Does the voice sound natural/human? | 1 (not at all natural/human) to 7 (very natural/human) |
| **Expressiveness** | Are emotions expressed well? | 1 (very badly expressed) to 7 (very well expressed) |
| **Appropriateness of register** | Is the register appropriate? | 1 (very inappropriate) to 7 (very appropriate) |
| **Pleasantness of voice** | Is the voice pleasant to listen to? | 1 (very unpleasant) to 7 (very pleasant) |

[39] 'reading machine', was substituted for 'pronunciation model at the segmental (phoneme/word) level', 'pronunciation model at the suprasegmental (prosodic) level' or 'conversational partner' depending on the CALL setup with respect to which participants were being asked to rate the TTS synthesis systems.
[40] See Footnote 39.
[41] See Footnote 39.

### 6.3.1.2 Participants

Participants, as presented in section 4.6.1.1.2, should be representative of the end-users (van Bezooijen and van Heuven, 1997). There are three groups of end-user of TTS synthesis systems within the CALL context, namely learners, teachers and CALL developers. French teachers and CALL researchers fluent in French were to be recruited because both are end-users of TTS in this context and expert speakers of the TL. The recruitment of French teachers was particularly important because the success of CALL applications integrating TTS synthesis is dependent on their acceptance by teachers because teachers are the first people to whom learners turn when they want to find out about materials that could support their language learning. Moreover, as evaluations that have been conducted to date have shown, they are the end users that are the most sceptical about the use of TTS synthesis in CALL (Stratil *et al.*, 1987a).

Regarding the number of participants to be recruited, it was hoped that if 25 could be recruited the data would meet the assumptions of the more powerful of the two types of statistical test which could be used in the analysis of the data, namely parametric tests (see Dancey and Reidy, 2002).

Participants were a convenience sample recruited from French departments at universities and colleges across Manchester and the UK, the French department at Dalhousie University, Halifax, Nova Scotia, Canada, and the following mailing lists: *EDTECH* (Educational Technology),[42] *FLTEACH* (Foreign Language Teaching),[43] *Linguist List*,[44] and *LLTI* (Language Learning and Technology International).[45] Although, 63 French teachers and CALL researchers volunteered to participate in the investigation, in the end only 22 participants submitted their responses to the questionnaires. A few could not participate due to technical problems which included problems viewing the French characters, and problems submitting their responses. With regard to the second problem, this was solved by providing a downloadable version of the questionnaires on the website which could either be filled in electronically and submitted by e-mail, or filled in by hand and submitted by traditional mail.

[42] listserv@h-net.msu.edu
[43] listserv@listserv.buffalo.edu
[44] listserv@listserv.linguistlist.org
[45] listserv@dartmouth.edu

A few others found that ultimately their workload was too heavy to permit them to participate in the investigation.

During the period of data collection two of the TTS synthesis demonstrations used in the experiment, synthesisers 4 and 5, were taken off line, and the other, synthesiser 6, was not working for a short period of time. Consequently, 7 of the participants evaluated all 6 TTS synthesis systems, 10 participants evaluated only 4 of the systems, and 3 participants only evaluated 3 of the systems. The remaining participants did not complete the investigation: one, did not evaluate S1, only partially evaluated S2, and did not answer the questionnaire on her experience of speech synthesis, the use of CALL, linguistic background and demographics; and the other, did not evaluate S5 and S5, and only partially evaluated S4. The 2 participants that did not complete the experiment were eliminated. Those who only evaluated 3 synthesisers were also eliminated because it was felt that this considerably limited the range of synthesisers. The data analysed here is based on that provided by the remaining 17 participants. Regarding S4 and S5, only 7 of the participants evaluated these two TTS synthesis systems. It was therefore not possible to compare the data obtained for these two systems with that obtained for the other 4 systems. The main analysis presented here is therefore based on S1, S2, S3, and S6 only. As said, S4 and S5 were in fact two different voices, one female and one male respectively, produced by the same TTS synthesis system. With respect to the evaluation of TTS synthesis for CALL purposes, as said, it is interesting to see to what extent the adequacy and acceptability, and quality of the speech generated by different voices produced by a single TTS synthesis system vary: for, if the different voices produced by synthesisers were to differ considerably, it would be necessary to evaluate each voice intended for use in CALL separately. The results obtained for S4 and S5 will therefore be analysed separately at a later date.

In the following paragraphs, the critical features of the sample of participants that participated in the investigation are presented. In addition to sex, age, and occupation, which are typically reported in psychological studies (Harris, 2002), it is argued that it is necessary to consider the following features of the sample of participants and the conditions under which they completed the investigation: presence of the investigator, teaching experience, linguistic background, and experience of speech synthesis. Regarding presence of the investigator, it was decided that the experiment would be presented online so that participants could complete

the experiment whenever they had time in their own offices. Due to the low response rate reported earlier in this section, during the period of data collection, it was decided that it might be better to offer to be present while the participants completed the investigation. Consequently, some of the participants completed the experiment in the presence of the investigator, whilst others did not. Although, as can be seen in section 6.3.1.4, every possible effort was made by the investigator not to influence the results of the investigation, as in any experiment, it is possible that she may have influenced the way in which the participants responded to the questionnaires (Harris, 2002). Regarding teaching experience, learners from different linguistic backgrounds have different difficulties in acquiring French (Léon and Léon, 1965; Charliac and Motoron, 1998), or any other language for that matter (Lado, 1957). Learners and their teachers from different linguistic backgrounds will therefore focus on different aspects of speech. It is therefore expected that teachers of learners of different linguistic backgrounds may rate the speech generated by the TTS synthesis systems differently because they believe different aspects of the speech to be important. Regarding the linguistic background of the participants themselves, participants from different linguistic backgrounds may have different expectations with regard to pronunciation and so on and hence rate the speech differently with respect to its use in the different roles because they themselves speak differently. Non-native speakers do not perform as well as native speakers in intelligibility tests (Golstein, 1995; Reynolds *et al.*, 1996; Bradlow and Pisoni, 1999). It can therefore be expected that non-native speakers may rate the speech differently, in particular with respect to comprehensibility and intelligibility. Regarding experience of speech synthesis, previous research suggests that speakers who have had greater exposure to speech synthesis score higher in intelligibility tests than those who have had less exposure to speech synthesis (Nye and Gaitenby, 1974; Pisoni, 1978-9; McNaughton, *et al.*, 1994; Venkatagiri, 1994; van Bezooijen and van Heuven, 1997; Delogu *et al.*, 1998; Reynolds *et al.*, 2000). It can therefore be expected that participants who have had greater exposure to speech synthesis may rate the speech generated by the TTS synthesis systems more favourably than those who have had less or no exposure to speech synthesis.

Presence of investigator

9 of the 17 participants completed the experiment in the absence of the investigator. The remaining 8 participants completed the experiment in the presence of the investigator.

Sex and age

Of the 17 participants that were retained, 6 were male and 11 were female. Their ages ranged from under 25 to over 55 years. The precise distribution of their ages is presented in Figure 3

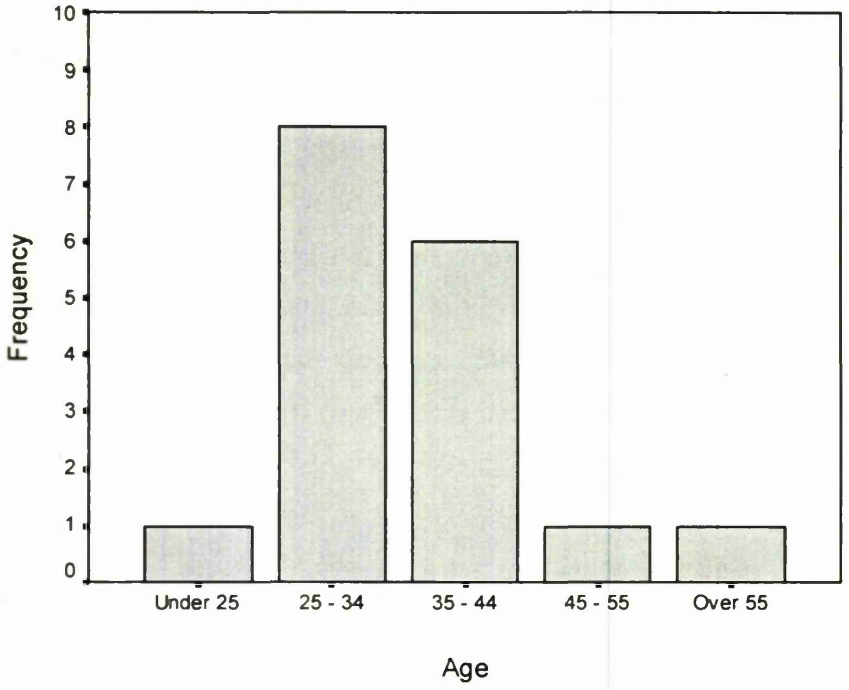



**Figure 3 Distribution of the age of the participants**

Occupation

10 out of the 17 participants were French teachers; 1 was a teacher-researcher specialising in French CALL; 1 was a retired French teacher; 1 was a teacher of French literature; 2 were French assistants; and, 2 were CALL researchers.

Teaching experience

Of the 11 participants that currently taught French, 10 taught French to Anglophones. 8 of these only taught French to learners whose first language was English. One of these also taught learners who spoke Chinese (Mandarin) and Japanese as their mother tongue. Another of these also taught learners who spoke Chinese (Mandarin), German, Greek, Italian, Japanese, or Spanish as their mother tongue. The remaining participant only taught French to learners who spoke Japanese as their mother tongue.

Regarding the level to which they taught French, 4 of the 11 participants that currently taught French taught it at all levels; 1 taught beginners only; 3 taught beginners, false beginners and intermediates; 1 taught beginners, intermediates, advanced and very advanced learners; 1 taught intermediate and advanced learners; and, 1 taught advanced learners only.

Regarding the teachers' use of CALL, only 2 of the 10 participants that responded to the questions relating to this,[46] had used CALL in the classroom. These 2 participants and another 2 had recommended the use of CALL outside the classroom to their students.

Linguistic background

13 of the 17 participants spoke French as their mother tongue. Regarding the 4 remaining participants, their mother tongues were Chinese, Croatian, English and German respectively. They all, however, had near native competency in French. As regards the dialect of French that they spoke, 7 were from metropolitan France, 7 were from Canada, 2 were from Cameroon and 1 was from Germany.

Experience of speech synthesis

7 of the 17 participants claimed to understand what speech synthesis is prior to participating in the experiment. All participants believed that having completed the experiment they better understood what speech synthesis is.

14 of the participants had used speech synthesis prior to participating in the experiment. Of these 14 participants, the applications that had most frequently been used by participants were vocal servers (10 participants), talking games (5 participants), and automated directories (3 participants). Other applications integrating speech synthesis that some participants had used included: talking books, talking games, talking home appliances, electronic phonetic inventories, and electronic dictionaries

These 14 participants varied in their experience of speech synthesis: 4 of these participants had used 3 of the above applications; 5 had used 2; and, 3 had only used one. Regarding the

[46] One of the French teachers who participated in the experiment failed to respond to the questions relating to their use of CALL inside and outside the classroom.

remaining participants, one had come into contact with speech synthesis through a colloquium, and another had only come into contact with it through the preliminary exploratory investigation that was carried out prior to this experiment.

Their frequency of use of applications integrating speech synthesis is presented in the following figure, Figure 4.

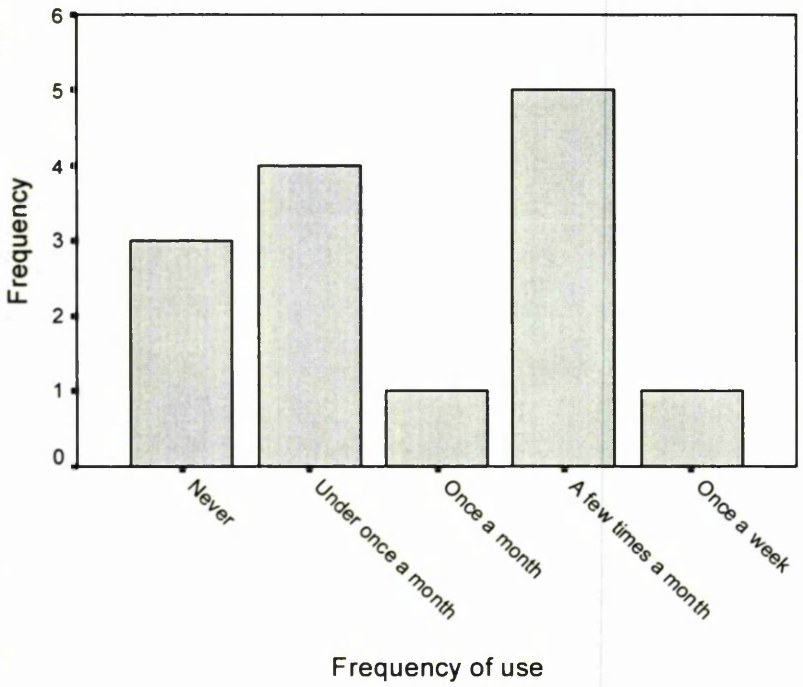



**Figure 4 Frequency of use of applications integrating speech synthesis**

In addition to the 1 participant that had only come into contact with speech synthesis through the pilot study carried out prior to this experiment, 3 other participants had also participated in this pilot study, and another 2 had participated in other speech synthesis experiments outside this study.

Regarding the use of CALL programs integrating speech synthesis, one of the teachers had used CALL programs integrating TTS synthesis in her classes. This was the only teacher to have heard of CALL programs integrating speech synthesis. The program that she had heard

of was *FreeText* (Hamel, 2003b). One of the researchers had also heard of CALL programs integrating speech synthesis, namely web dictionaries.

In previous evaluations of TTS synthesis systems reported in the literature, it has been observed that participants' ability to recognise the speech generated by TTS synthesis systems improves after only short periods of exposure to it (Pisoni, 1978-9; Francis and Nusbaum, 1999):

> substantial learning effects occur with synthetic speech. Even after an initial period of exposure, recognition performance continues to improve. Comparisons of word recognition performance in the first and second half of each of the tests indicated the presence of reliable learning effects (Pisoni, 1978-9: 159).

> As little as four hours training with synthetic speech results in significantly increased rate of recognition ..., and some improvement is noticeable after considerably less training (Francis and Nusbaum, 1999: 80).

In order to control for this effect and other order effects such as familiarity with the test materials and fatigue which may have an effect on the results of evaluations of TTS synthesis systems, the intention was that participants would be split into two groups, and one group would be presented the synthesisers in the order S1, S2, S3, S4, S5, S6, and the roles in the order RM, phonetic PM, prosodic PM, CP, and the other group would be presented the synthesisers and roles in the reverse order. Due to the aforementioned drop out rate and problems with the on-line demonstrations, the order of presentation of the synthesisers and roles was not balanced as intended: 3 participants were presented the synthesisers in the order S1, S2, S3, S4, S5, S6 and the roles in the order RM, phonetic PM, prosodic PM, CP; 4 participants were presented the synthesisers in the order S6, S5, S4, S3, S2, S1 and the roles in the order CP, prosodic PM, phonetic PM, RM; 5 participants were presented the synthesisers in the order S1, S2, S3, S6 and the roles in the order RM, phonetic PM, prosodic PM, CP; and, 5 participants were presented the synthesisers in the order S6, S3, S2, S1 and the roles in the order CP, prosodic PM, phonetic PM, RM.

### 6.3.1.3 Apparatus and materials

The investigation was presented on-line. In order to take part in the experiment, all participants required a PC equipped with a sound card and headphones or speakers and access to the Internet.

As said, one voice offered by each of four different French TTS synthesis systems, and two voices offered by another French TTS synthesis system were evaluated in this investigation.[47] Specifically, participants were presented one example of the use of each TTS synthesis system in each of the 4 roles, blocked by TTS synthesis system. Regarding how listeners recognise the speech of different speakers, according to Strange (1999a), theories of vowel recognition across speakers can be divided into two types, those that claim that listeners rely solely on the acoustic signal and those that claim that listeners rely on the "preceding (and perhaps even following) speech patterns" (Strange, 1999a: 157). If the claims of the latter, which are referred to as 'extrinsic models', are correct, then the participants' perception and hence ratings of the first example that they listen to for each TTS synthesis system will be lower than for the examples that follow. Prior to the presentation of examples of the use of the TTS synthesis systems in each of the 4 roles investigated, participants were therefore presented a brief passage to familiarise themselves with the voice of the synthesiser. Participants' ratings were collected either using on-line forms or using printed questionnaires depending on which response mode was the most convenient for the participants.

The familiarisation passage and examples of the 4 roles are presented in section 6.3.1.3.1. The features of the French TTS synthesis systems evaluated, the motivations for their selection, and how they were integrated is presented in section 6.3.1.3.2. The design of the questionnaire that was used to collect participants' ratings is presented in section 6.3.1.3.3.

#### *6.3.1.3.1 Corpora*

The familiarisation passage, which was taken from *Le Petit Prince* (Saint-Exupéry, 1999), was the same for all synthesisers. "individual vowels are perceived relative to the patterns of formant frequencies at a speaker's entire vowel inventory" (Strange, 1999a: 159). The passage was therefore selected so as to contain at least one instance of each of the vowels of Metropolitan French (the selected passage is presented in A2.1).

[47] From this point forward, for the purposes of presentation the two voices generated by referred to as if they were different TTS synthesis systems.

The examples of the use of TTS synthesis in each of the 4 roles were also the same for all of the TTS synthesis systems. As presented in section 4.6.1.1.2, in order ensure external validity the materials used in evaluations should be representative of materials used in the real world (van Bezooijen and van Heuven, 1997). The examples, which consisted in a corpus of approximately 50 words, were therefore selected from actual CALL applications. More specifically, the RM corpus consisted in 50 contiguous words from the text *Le Vieux Lit* proposed to learners in *FreeText* (Hamel, 2003b). The remaining corpora were all selected from *Talk to Me: The Conversation Method (French)* (Auralog, 2002), a pronunciation and conversation training program which uses ASR in combination with audio and video clips to provide learners pronunciation exercises at both the phonetic and prosodic levels as well as simulated conversations. More specifically, the phonetic PM corpus consisted in 10 lists of 5 words selected at random from the 'Phonetic Exercises', the prosodic PM corpus consisted in 10 sentences selected at random from the 'Sentence Pronunciation' exercises, and the CP corpus consisted in 50 contiguous words selected from one of the 'Dialogues: Comprehensibility'. These corpora are presented in Appendix 2.

In order to provide the extra contextualisation which it is believed is necessary to permit participants to make more reliable judgements of the adequacy, acceptability and quality of the speech generated by TTS synthesis systems for use in the different roles in CALL applications, these examples, which were the same for all the TTS synthesis systems, were accompanied by a screenshot of an application typical of each role. With the exception of the screenshots of the use of TTS synthesis as an RM, the screenshots were taken from the same sources as the examples. The screenshot of the use of TTS synthesis as an RM was taken from the following talking dictionary: *Oxford Hachette French Dictionary* (Oxford, 2003; see section 3.3.3.1).

#### *6.3.1.3.2 TTS synthesis systems*

Due to financial limitations, it was necessary to use on-line interactive TTS synthesis demonstrations in this investigation. The systems that were used were: *AT&T Next-Gen*, *Bright Speech* from Babel Technologies, *Elan Tempo*, *eLite* from Multitel, and *Nuance Vocalizer*. The main features of these systems are presented in the following paragraphs.

*AT&T Next-Gen*

*AT&T Next-Gen*[48] (Beutnagel *et al.*, 1999) combines the TTP techniques employed in *AT&T Flextalk* with a modified version of the approach to PTS conversion employed in the *CHATR* system (Black, 1996; Conkie, 1999). Regarding PTS conversion, *CHATR* is based on unit selection synthesis. The approach to PTS employed in *AT&T Next-Gen* differs from that employed in *CHATR* with respect to the minimal unit of synthesis. While the *CHATR* system employs phones, the AT&T system employs half phones in order to permit the system to imitate synthesis based on "diphones or phonemes or more complicated structures" (*ibid.*: 2). Two voices are provided for Parisian French, one male (Alain) and one female (Julie). The female voice was used in the experiment reported here. *AT&T Next-Gen* also supports the following languages: UK and US English, German, and Latin American Spanish. Control is provided over the way in which the synthesiser reads acronyms and abbreviations, SR and volume. It was not possible to obtain further details of the TTP techniques employed in this system.

*BrightSpeech* from Babel Technologies

*BrightSpeech* from Babel Technologies, now owned by Acapela Group and marketed as *HQ TTS*,[49] is based on non-uniform USS (Babel Technologies, 2003). Two female voices are provided for French, Claire and Julie. Julie was used in the experiment reported here. Other languages supported by the system include: modern Arabic, Dutch, US English, German, Spanish and Swedish.

*Elan Tempo*

*Elan Tempo* (Boula de Mareuil, *et al.*, 2001) is also now owned by Acapela Group who market it as *HD TTS*.[50] It consists in two main modules: a TTP module and a PTS module. Following text normalisation, five levels of linguistic analysis are carried out by the TTP module: morphological analysis, morpho-syntactic analysis, syntactic analysis, grapheme-to-phoneme conversion, and prosody generation. The techniques used to achieve these different levels of linguistic analysis depend on the language synthesised. The following techniques are used for the synthesis of French: n-grams for morpho-syntactic analysis (tagging), a handwritten dependency grammar for syntactic analysis, and handwritten rewrite rules for

[48] http://www.research.att.com/projects/tts/demo.html (Site no longer available)
[49] The on-line interactive demonstration of the TTS synthesis system is now available at: http://demo.acapela-group.com/
[50] http://tempo.elan.fr/

grapheme-to-phoneme conversion. Regarding the generation of speech output, concatenative synthesis is employed in the PTS module. The primary unit of concatenation is the diphone. The inventory is augmented with polyphones to provide a better synthesis of sonorants (glides and liquids). Two voices for French are available, one male (Robert) and one female (Cathy). Both voices were used in the experiment reported here. Other languages supported by the system include: Dutch, UK and US English, German, Italian, Polish, Brazilian Portuguese, Spanish, and Latin American Spanish. Control over pitch and speech rate is provided.

*eLite* from Multitel

*eLite*[51] (Enhanced Linguistically-based Text-to-speech synthesis) (Multitel, 2005) from Multitel consists in two modules, a TTP module and a PTS module. The TTP module, *Eliot* (Electronic processing linguistically-oriented texts), carries out the following levels of analysis: text pre-processing, morphological analysis, syntactic analysis, grapheme-to-phoneme transcription, and prosody generation. Text pre-processing is carried out using regular expressions. Tri-gram and n-gram language models are employed in syntactic analysis. Regarding grapheme-to-phoneme conversion, the phonetic transcription of most words is provided in the lexicon, *MBRDICO* (Multi-Band Re-synthesis Dictionary). Post-processing is carried out in order to account for interword effects including liaison and deletion. Prosody generation is achieved through the use of Classification And Regression Trees (CARTs), a corpus-based stochastic technique. Diphone-based concatenative synthesis, specifically *MBROLA* (Multi-Band Re-synthesis OverLap-Add), is employed by the PTS module. Four voices are provided for French, two male (Vincent and Thierry), and two female (Anne-Carole and Céline). Vincent was used in this experiment. *eLite* also supports English.

*Nuance Vocalizer*

*Nuance Vocalizer*[52] is based on concatenative synthesis (TMA Associates, 2003). One female voice is available for French (Julie Deschamps). Other languages supported by the system include: UK, US/Canadian, and Australian English, Brazilian Portuguese, and Latin American Spanish.

[51] http://www.multitel.be/TTS/layout.php?page=eLite_demo
[52] http://www.nuance.com/prodserv/demo_vocalizer.html

The features of the TTS synthesis systems evaluated in this investigation are summarised in Table 11.

**Table 11 Summary of the features of the TTS synthesis systems used in the experiment**

| | Synthesiser | Sex | Voice | Variety | Method of PTS conversion |
|---|---|---|---|---|---|
| **S1** | *AT&T Next-Gen* | M | Alain | Parisian French | Non-uniform USS |
| **S2** | *Nuance Vocalizer* | F | Julie Deschamps | Canadian French | Concatenative synthesis |
| **S3** | *eLite* | M | Vincent | French | Diphone-based concatenative synthesis |
| **S4** | *Elan Tempo* | F | Cathy | French | Diphone-based concatenative synthesis |
| **S5** | *Elan Tempo* | M | Robert | French | Diphone-based concatenative synthesis |
| **S6** | *BrightSpeech* | F | Julie | French | Non-uniform USS |

The TTS synthesis systems were selected in order to cover a range of different synthesis techniques, and hence qualities of output. In particular, in order to test the discriminative power of the method of evaluation, both systems that it was believed would produce output of very different quality, e.g. *eLite* and *BrightSpeech*, and systems that it was believed would produce speech of similar quality were included in the experiment, namely the two voices produced by *Elan Tempo*. As mentioned in section 6.3, including two voices produced by the same TTS synthesis system also made it possible to test whether different voices produced by a single TTS synthesis system differ significantly in quality and therefore require separate evaluation. The systems were also selected so as to include a range of varieties of French and a balance of male and female voices.

Regarding the presentation of the speech generated by the TTS synthesis systems in this investigation, the intention was to synthesise the examples with each of the TTS synthesis systems and collect the output for presentation to participants as pre-recorded utterances, which participants could listen to by clicking on the corresponding utterances. It was not possible to download the speech generated by all of the TTS synthesis systems that were to be investigated. The alternative approach was to provide hyperlinks to the TTS synthesis systems and ask participants to synthesise the examples themselves. This was thought to be difficult and time consuming for participants. Where possible, namely for synthesisers 1, 2 and 3, the

speech generated by the TTS synthesis was presented as pre-recorded utterances.[53] Given the added difficulty of rating synthesisers 4, 5, and 6, it was thought that participants might rate them less favourably than synthesisers 1 to 3 out of frustration. This, as can be seen in section 6.3.2.2, was not the case; S6 received the highest ratings overall.

#### *6.3.1.3.3 Questionnaire*

As presented in section 6.3, the questionnaire employed in this investigation was based on *MOS-X* (Polkosky and Lewis, 2003), the latest version of the *ITU-T Overall Quality Test* (Schmidt-Nielson, 1995; van Bezooijen and van Heuven, 1997; Polkosky and Lewis, 2003).

*MOS-X*, as can be seen on Page 183 of this thesis, consists in fifteen 7-point bi-polar categorical estimation scales which fall into four factors, namely intelligibility, naturalness, prosody, and social impression. Listening effort, comprehensibility problems, speech sound articulation, and precision formed the intelligibility factor; voice pleasantness, voice naturalness, humanlike voice, and voice quality formed the naturalness factor; emphasis, rhythm, and intonation formed the prosody factor; and, trust, confidence, enthusiasm, and persuasiveness formed the social impression factor. A measure of overall intelligibility could be obtained by combining ratings across all fifteen scales.

As presented in section 6.3, in addition to the ratings of the quality of the speech generated by the TTS synthesis systems, it was believed that the ratings of the adequacy and acceptability, i.e. the acceptance, of the speech generated by the TTS synthesis systems would differ across the different roles that TTS synthesis systems may assume within CALL applications if the roles did indeed place different demands on the quality of the speech generated. *MOS-X*, unlike the original *ITU-T Overall Quality Test* (see Page 184 of this thesis), does not include scales for the evaluation of the acceptance of the speech generated by TTS synthesis systems. It was therefore necessary to add scales for this purpose. As can be seen on Page 184 of this thesis, in the *ITU-T Overall Quality Test*, acceptance is measured on a 2-point scale. Such scales are limited to the extent that they do not give an idea of the degree of acceptance (Hubbard, 1987). The decision was therefore taken to use a larger scale. Traditionally, the *MOS* test is based on 5-point scales. In their investigation of the psychometric properties of *MOS*, Polkosky and Lewis (2003) found that 7-point scales produced more reliable results

[53] The wav files were downloaded from the AT&T demonstration on 18/10/04, from the Nuance demonstration on 18/10/04, and from the Multitel demonstration on 2/11/04.

than 5-point scales. 7-point scales were therefore used for the evaluation of acceptance in this investigation (see Page 185 of this thesis).

Regarding the quality of the speech generated by TTS synthesis systems, as said, the results of the literature review and the preliminary exploratory investigation suggest that CALL places demands on the following aspects of it: comprehensibility, accuracy, naturalness, expressiveness and the register.

Comprehensibility is used here as it is generally used in TTS synthesis circles (see section 2.3.2). Despite the fact that *MOS-X* is intended for use for the evaluation of TTS synthesis systems, its comprehensibility scale, scale 2, did not correspond with this sense of the term. This scale was adapted to fit comprehensibility as used here (see Page 185 of this thesis).

Regarding accuracy, it is believed that the different roles which TTS synthesis may assume within CALL applications place demands on accuracy at different levels, specifically that the use of TTS synthesis as a phonetic PM places demands on accuracy at the phonetic level, and the use of TTS synthesis as a prosodic PM places demands on accuracy at the prosodic level. Regarding the accuracy of the articulation of individual phonemes, scale 4, precision, appears to address this variable. Regarding accuracy at the prosodic level, one of the *MOS-X* scales appeared to address accuracy at the prosodic level, namely scale 9, emphasis. Emphasis, or sentence stress, is only one of a number of aspects of prosody. Judgemental evaluations, of which *MOS-X* is an example, already place high cognitive demands on participants (Goldstein, 1995). It was therefore decided that only one scale should be used to evaluate accuracy at the prosodic level. The emphasis scale was therefore replaced by a more general scale, namely appropriateness of prosody (see Page 185 of this thesis). This scale was designed to parallel the other scales. Also regarding accuracy, in section 2.6, it was suggested that choice of correct pronunciation is not given in the speech generated by TTS synthesis systems. It was therefore decided that it was desirable to consider this aspect of the quality of the speech generated by TTS synthesis systems in addition to the aspects suggested by the SLA literature and the literature on best practice in LL&T. None of the *MOS-X* scales appeared to address choice of pronunciation. A scale was therefore developed to parallel the other scales to address this variable (see Page 185 of this thesis).

Regarding naturalness, which is defined here as human-like, as presented in section 6.3, it is also believed that different CALL setups place demands on the naturalness of the speech generated by TTS synthesis systems at different levels. None of the *MOS-X* scales appear to address naturalness at the phonetic level. The scale naturalness of phonemes was added to address this (see Page 185 of this thesis). The wording of this scale reflects the definition of naturalness used here. Regarding naturalness at the prosodic level, two scales appear to address this, namely scale 10, rhythm, and scale 11, intonation. As mentioned earlier in this section, although it would be interesting to get participants to evaluate these individual aspects of prosody, MOS places high cognitive demands on participants (Goldstein, 1995). It was therefore decided that these scales should be replaced by one general scale, namely naturalness of prosody (see Page 185 of this thesis). Two of the other scales which make up *MOS-X* also address naturalness, namely scale 6, voice naturalness, and scale 7, humanlike voice. More specifically, they both address the naturalness of the voice of the TTS synthesis system. It was believed that CALL applications may also place demands on this aspect of the speech generated by TTS synthesis systems. It was therefore decided to keep one of these scales, namely voice naturalness. This scale was, however, slightly adapted to reflect our definition of naturalness (see Page 185 of this thesis). It was also renamed naturalness of voice to parallel the names of the other scales.

Regarding expressiveness, used here to refer to the ability of the voice to express emotion and attitudes expressed in the text, the social impression scales address emotion and attitude. Trust, confidence, enthusiasm, and persuasiveness are, however, just a few of many emotions and attitudes. As presented above, *MOS-X* is already places high cognitive demands on participants (Goldstein, 19995). It is therefore argued that it is not appropriate to evaluate each emotion and attitude individually. Rather, a single general scale was proposed for the evaluation of expressiveness (see Page 185 of this thesis). It was designed to parallel the other scales.

Regarding register, none of the *MOS-X* scales appear to address this variable. It was therefore necessary to develop a scale for the evaluation of this variable. This scale labelled appropriateness of register was also designed to parallel the other scales (see Page 185 of this thesis).

Regarding the remaining scales which make up *MOS-X*, scale 3, speech sound articulation, corresponds closely to intelligibility as it is generally used in TTS synthesis circles (see section 2.3.2). Intelligibility is the most evaluated aspect of TTS synthesis systems. As presented in section 6.3, several tests have been developed for its evaluation. This suggested to us that most applications are believed to place demands on the intelligibility of the speech generated by TTS synthesis systems. This scale was therefore retained. It is, however, argued that the scale for the evaluation of intelligibility should parallel the one used for the evaluation of comprehensibility. Scale 3, was therefore renamed intelligibility and adapted to parallel the scale used for the evaluation of comprehensibility (see Page 185 of this thesis).

The scales which make up the social impression factor, namely trust, confidence, enthusiasm and persuasiveness, address aspects of the quality of the speech generated by TTS synthesis systems which may have an effect on learners' affective filter (see sections 5.3 and 5.3.1.3). We believe that learners should evaluate these aspects of the quality of the speech rather than teachers. These scales were therefore not included in *MOS-CALL*, the name we have given to the modified version of *MOS-X* used in this study. During the analysis phase, it was decided that pleasantness of voice like the social impression factors should also be evaluated by learners rather than by teachers. While a scale for the evaluation of pleasantness of voice was included in *MOS-CALL* (see Page 185 of this thesis), the data collected for this IV are therefore not considered in the analysis in section 6.3.2.

All scales were translated into French in order to keep the participants thinking in French (see Appendix 3).

1. *Listening Effort*: Please rate the degree of effort you had to make to understand the message.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **IMPOSSIBLE EVEN WITH MUCH EFFORT** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **NO EFFORT REQUIRED** |

2. *Comprehensibility Problems*: Were single words hard to understand?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **ALL WORDS HARD TO UNDERSTAND** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **ALL WORDS EASY TO UNDERSTAND** |

3. *Speech Sound Articulation*: Were the speech sounds clearly distinguishable?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **NOT AT ALL CLEAR** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **VERY CLEAR** |

4. *Precision*: Was the articulation of speech sounds precise?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **SLURRED OR IMPRECISE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **PRECISE** |

5. *Voice Pleasantness*: Was the voice you heard pleasant to listen to?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **VERY UNPLEASANT** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **VERY PLEASANT** |

6. *Voice Naturalness*: Did the voice sound natural?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **VERY UNNATURAL** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **VERY NATURAL** |

7. *Humanlike Voice*: To what extent did this voice sound like a human?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **NOTHING LIKE A HUMAN** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **JUST LIKE A HUMAN** |

8. *Voice Quality*: Did the voice sound harsh, raspy, or strained?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **SUGNIFICANTLY HARSH/RASPY** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **NORMAL QUALITY** |

9. *Emphasis*: Did emphasis of important words occur?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **INCORRECT EMPHASIS** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **EXCELLENT USE OF EMPHASIS** |

10. *Rhythm*: Did the rhythm of the speech sound natural?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **UNNATURAL OR MECHANICAL** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **NATURAL RHYTHM** |

11. *Intonation*: Did the intonation pattern of sentences sound smooth and natural?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **ABRUPT OR ABNORMAL** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **SMOOTH OR NORMAL** |

12. *Trust*: Did the voice appear to be trustworthy?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **NOT AT ALL TRUSTWORTHY** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **VERY TRUSTWORTHY** |

13. *Confidence*: Did the voice suggest a confident speaker?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **NOT AT ALL CONFIDENT** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **VERY CONFIDENT** |

14. *Enthusiasm*: Did the voice seem to be enthusiastic?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **NOT AT ALL ENTHUSIASTIC** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **VERY ENTHUSIASTIC** |

15. *Persuasiveness*: Was the voice persuasive?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **NOT AT ALL RESUASIVE** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **VERY PERSUASIVE** |

*MOS-X Scales*
Overall: Average items 1 – 15
Intelligibility: Average items 1 – 4
Naturalness: Average items 5 – 8
Prosody: Average items 9 – 11
Social Impression: Average items 12 – 15

**Figure 5 *MOS-X* (Polkosky and Lewis, 2003: 176-7)**

1. Acceptance (*Do you think that this voice could be used for such an information service by telephone?*) 1: yes, 2: no.
2. Overall impression (*How do you rate the quality of the sound of what you have just heard?*) 1: excellent, 2: good, 3: fair, 4: poor, 5: bad.
3. Listening effort (*How would you describe the effort you were required to make in order to understand the message?*) 1: complete relaxation possible, no effort required, 2: attention necessary, no appreciable effort required, 3: moderate effort required, 5: no meaning understood without any feasible effort.
4. Comprehensibility problems (*Did you find certain words hard to understand?*) 1: never, 2: rarely, 3: occasionally, 4: often, 5: no, not at all.
5. Articulation (*Were the sounds discriminable?*) 1: yes, very clear, 2: yes, clear enough, 3: fairly clear, 4: no, not very clear, 5: no, not at all.
6. Pronunciation (*Did you notice any anomalies in pronunciation?*) 1: no, 2: yes, but not annoying, 3: yes, annoying, 4: yes, very annoying
7. Speaking rate (*What do you think of the average speed of delivery?*) 1: much faster than preferred, 2: faster than preferred, 3: preferred, 4: slower than preferred, 5: much slower than preferred.
8. Voice pleasantness (*how would you describe the voice?*) 1: very pleasant, 2: pleasant, 3: fair, 4: unpleasant, 5: very unpleasant.

**Figure 6 *ITU-T Overall Quality Test* (van Bezooijen and van Heuven, 1997: 562-3)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Adequacy and acceptability of the speech*** | | | | | | | | | |
| Adequacy | Is the speech adequate for use as a reading machine (in comparison with other media)? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Not at all adequate | | | | | | | | Very adequate |
| Acceptability | Is the speech acceptable for use as a reading machine (when it is not possible to use other media)? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Very unacceptable | | | | | | | | Very acceptable |
| ***Quality of the speech*** | | | | | | | | | |
| Comprehensibility | Is the message easy to understand? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Very difficult | | | | | | | | Very easy |
| Intelligibility | Are the individual phonemes/sounds and words easy to recognise (and discriminate one from another)? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Very difficult | | | | | | | | Very easy |
| Choice of pronunciation | Is the pronunciation correct? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Incorrect | | | | | | | | Correct |
| Precision of phonemes | Was the articulation of the phonemes/sounds precise? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Very imprecise | | | | | | | | Very precise |
| Appropriateness of prosody | Was the prosody (music) of the utterance appropriate? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Very inappropriate | | | | | | | | Very appropriate |
| Naturalness of phonemes | Do the phonemes/sounds sound natural/human? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Not at all natural/human | | | | | | | | Very natural/human |
| Naturalness of prosody | Does the prosody (music) sound natural/human? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Not at all natural/human | | | | | | | | Very natural/human |
| Voice naturalness | Does the voice sound natural/human? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Not at all natural/human | | | | | | | | Very natural/human |
| Expressiveness | Was emotion expressed well? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Very badly expressed | | | | | | | | Very well expressed |
| Appropriateness of register | Was the register appropriate? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Very inappropriate | | | | | | | | Very appropriate |
| Voice pleasantness | Was the voice pleasant to listen to? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Very unpleasant | | | | | | | | Very pleasant |

**Figure 7** ***MOS-CALL***

### 6.3.1.4 Procedure

Originally, it was intended that all participants would participate on-line at a distance and thus be run individually. Due to the low response rate reported in section 6.3.1.2, during the period of data collection, it was decided that it might be better to offer to be present while the participants completed the investigation. Of the 17 participants 8 chose to do this, and 3 of which were run at the same time.

Those who participated on-line were sent a brief introduction to the investigation. In this e-mail they were firstly reminded that they had volunteered to participate in an investigation relating to the use of TTS synthesis in CALL. They were then told that three main roles of TTS synthesis in CALL had been identified, namely as an RM, a PM, and a CP, that our goal was to develop tests which could be used to determine the suitability of TTS synthesis systems for use in CALL in these three different roles, and that the goal of the current investigation was to identify the requirements that these different roles impose on the quality of the speech generated by TTS synthesis systems. The conditions under which the experiment ought to be completed, namely in a quiet office where they are unlikely to be disturbed, and the equipment that would be needed, namely a computer with sound card connected to the internet, and headphones or speakers, were then explained. The participants were then told that they should open *Internet Explorer* and go to the Website where the experiment was hosted. Finally, before beginning the investigation in earnest, the participants were told that they had been divided into two groups, and each participant was told which group they had been assigned to. This information was presented verbally by the investigator to those participants who completed the experiment in her presence.

On arrival at the Website where the experiment was hosted, which is presented in flat form in Appendix 5, the participants were presented a brief introduction to the use of speech synthesis in CALL. This included an introduction to speech synthesis, and TTS synthesis more specifically, a presentation of the advantages of the use of TTS synthesis in CALL, and a presentation of the 3 roles in which TTS synthesis is being used in CALL applications and with respect to which the participants were going to be asked to evaluate the speech generated by the TTS synthesis systems in the investigation. Next, the participants were reminded of the goal of the experiment. Then, the participants were asked to check that the volume of the

speech output was set at a comfortable level, by listening to an audio clip. If the volume required adjusting, instructions on how to do this were provided. Once the participants were happy that the volume was at a comfortable level, the procedure of the investigation was explained to them. Specifically, they were told that, for each of 6 TTS synthesis systems,[54] they would be presented, a brief passage to familiarise themselves with the speech generated by the TTS synthesis system, one example of its use as an RM in CALL, two examples of its use as a PM, one at the phonetic level and one at the prosodic level, and one example of its use as a CP. Regarding the examples of the use of the TTS synthesis systems in the 4 different roles, the participants were told that, their task was to rate the adequacy, acceptability, and quality of the speech generated by the TTS synthesis systems with respect to its use in each on the scales provided either on the online form or in their printed answer booklet. Next participants were asked to follow the hyperlink which corresponded to the experimental group to which they had been assigned. The first TTS synthesis system to be evaluated was then presented. For those participants assigned to group 1, this was S1. As indicated in the explanation of the experimental procedure, in order to familiarise themselves with the voice of the TTS synthesis system, the participants were asked to listen to a short passage of speech generated by the TTS synthesis system. To listen to this passage, they were told that they should click on each of its constituent sentences. Next, the 4 examples of the use of S1 in CALL were presented in the following order: RM, phonetic PM, prosodic PM, CP. For each of these examples, the participants were asked to listen to the example and rate the adequacy, acceptability and quality of the speech generated by the TTS synthesis system with respect to its use in the role indicated and presented in the screenshot on the scales provided either on the online form or in their printed answer booklet. In order to listen to the examples, the participants were told that they should click on each of their constituent sentences. The participants in group 1 were then asked to repeat this procedure for S2 and S3. Regarding S4, S5 and S6, as presented in section 6.3.1.3.1, it was not possible to download the speech generated by these TTS synthesis systems. The procedure of the investigation therefore differed slightly for these TTS synthesis systems. Specifically, the participants were asked to go to the site of the online demonstrations and synthesise the examples themselves.

[54] The participants were not made aware of the fact that two of the TTS synthesis systems were in fact different voices offered by the same TTS synthesis system. They may, however, have been able to establish this for themselves because this was one of the systems that did not permit us to download the speech that it generated and the participants were directed to the site where the demonstration was hosted.

Instructions on how to do this were provided (see Appendix 5). The experimental procedure was the same for all participants. All that differed was the order of presentation of the synthesisers and the CALL contexts (the orders in which the synthesisers and CALL contexts were presented to the different groups of participants are presented in section 6.3.1.2).

Finally, participants were asked to fill in a questionnaire which probed the variables which it was thought might affect their ratings of the TTS synthesis systems in this investigation, namely, age, sex and occupation, their linguistic background, their experience of teaching and CALL, and their experience of speech synthesis. In this questionnaire, the participants were also asked to indicate how ready they thought TTS synthesis was for use in CALL in general prior to participating in the investigation and how ready they now think that TTS synthesis is for use in each of the 4 roles, RM, phonetic PM, prosodic PM and CP.

### 6.3.2 Results

For the 4 French TTS synthesis systems considered here, missing data constituted less than 1% of the total data.

As mentioned in section 6.3.1.1, the DVs were the quality of the speech generated by the different TTS synthesis systems with respect to their use in each of the four roles, specifically the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, appropriateness of register, and pleasantness of voice of the speech generated by the TTS synthesis systems, the adequacy and acceptability of the speech generated by the TTS synthesis systems for use in the 4 roles, and the readiness of TTS synthesis in general for use in the 4 roles.

Of these variables, the accuracy and naturalness of the speech generated by the TTS synthesis systems with respect to its use in the different roles were composite scores:

- a measure of the accuracy of the speech generated by each TTS synthesis system was calculated for each role for each participant by taking the mean of their ratings of the following aspects of the quality of the speech generated by the TTS synthesis systems in each role: precision of phonemes and appropriateness of prosody; and,
- a measure of the naturalness of the speech generated by each TTS synthesis system was calculated for each role for each participant by taking the mean of their ratings of

the following aspects of the quality of the speech generated by the TTS synthesis systems in each role: naturalness of phonemes and naturalness of prosody.

Histograms of the participants' ratings of the speech generated by the TTS synthesis systems for use in the different roles that TTS synthesis may assume within CALL applications were plotted. These revealed that not all the data fitted the normal distribution. They therefore did not all meet all the assumptions of parametric tests of significance (Dancey and Reidy, 2002). It was therefore not possible to use parametric tests to analyse the data, as hoped (see section 6.3.1.2). Rather, it was necessary to use non-parametric tests of significance to analyse the data. Non-parametric tests are less powerful than parametric tests, i.e. if there is a difference a parametric test is more likely to find it than a non-parametric test (Dancey and Reidy, 2002). When interpreting the results, it should therefore be borne in mind that Type II errors, i.e. failure to detect reliable differences that exist (Harris, 2002), are more likely than when using non-parametric tests. Following conventions in psychology research, an alpha level of 0.05 (i.e. $p < 0.05$) is used in all tests of significance (Everitt and Hay, 1992).

Regarding the measures of central tendency used to describe the data, the aforementioned histograms revealed that, in general, the distributions of participants' ratings were not skewed and extreme ratings were rare. It was therefore possible to use the mean to describe the central tendency of the data collected in this study in most cases (Heiman, 2001; Dancey and Reidy, 2002). In order to permit the comparison of the results obtained for different variables, the mean was therefore used throughout the analysis presented here.

As presented in section 6.3, the main hypotheses of the investigation were:

- CALL setups in general place demands on the following aspects of the quality of TTS synthesis systems: comprehensibility, intelligibility, accuracy, naturalness, smoothness, expressiveness and register;
- the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems;
- the speech generated by different TTS synthesis systems differs in quality;
- different TTS synthesis systems will be suitable for use in different roles in CALL applications;

- the speech generated by different voices offered by the same TTS synthesis system differs in quality; and,
- different voices will be suitable for use in different roles in CALL applications.

In section 6.3.2.1.1, the results are analysed with respect to the first and second of these hypotheses; in section 6.3.2.2, they are analysed with respect to the third; and, in section 6.3.2.3, they are analysed with respect to the fourth. As mentioned in section 6.3.1.2, the data will be analysed with respect to the final two hypotheses at a later date. In addition to answering the aforementioned questions, the data collected also provide an insight into the readiness of TTS synthesis for use in CALL and aspects of the quality of the speech generated by TTS synthesis systems which require improvement for TTS synthesis systems to be ready for use in CALL. The data are analysed with respect to these questions in sections 6.3.2.4 and 6.3.2.5 respectively.

### 6.3.2.1 On what aspects of the quality of the speech generated by TTS synthesis systems do CALL applications place demands?

In this section, the participants' ratings of the quality of the speech generated by each of the TTS synthesis systems, S1 (see section 6.3.2.1.1), S2 (see section 6.3.2.1.2), S3 (see section 0), and S6 (see section 6.3.2.1.4), are analysed in turn first with respect to the hypothesis that CALL applications in general impose requirements on the comprehensibility, accuracy, naturalness, expressiveness and register of the speech generated by TTS synthesis and then with respect to the hypothesis that the different roles that TTS synthesis systems may assume within CALL applications impose different requirements on the quality of the speech that they generate. Also, for each TTS synthesis system, the ratings of the adequacy and acceptability of the speech generated are considered; as presented in section 6.3, it is believed that if indeed the different roles do place different demands on the quality of the speech generated by TTS synthesis systems, then ratings of adequacy and acceptability will differ across the roles. Similarly, it is believed that if indeed the different roles do place different demands on the quality of the speech generated by TTS synthesis systems, then the participants' ratings of the readiness of TTS synthesis in general having participated in the investigation will differ across the roles. This data is therefore also analysed in section 6.3.2.1.5.

#### *6.3.2.1.1 S1*

As presented in section 6.3, it is believed that aspects of the speech generated by TTS synthesis upon which CALL applications place demands will not receive top ratings, i.e. ratings of 6 or 7.

The mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of the speech generated by S1 with respect to its use in each of the four different roles were therefore calculated across participants. The results are presented in Table 12.

**Table 12 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Comprehensibility** | 4.53 | 4.24 | 4.82 | 4.47 |
| **Intelligibility** | 4.53 | 3.88 | 4.65 | 4.24 |
| **Choice of pronunciation** | 4.59 | 4.12 | 4.53 | 4.29 |
| **Accuracy** | 3.76 | 3.56 | 3.53 | 3.47 |
| **Naturalness** | 3.82 | 3.68 | 3.15 | 3.29 |
| **Naturalness of voice** | 4.00 | 3.53 | 3.59 | 3.53 |
| **Expressiveness** | 3.24 | 3.12 | 2.35 | 2.65 |
| **Appropriateness of register** | 5.00 | 4.76 | 4.53 | 4.53 |

The descriptive statistics, presented in Table 12, show that for all the roles none of the aspects of the speech generated by S1 considered received top ratings, i.e. ratings of 6 or 7. This appears to suggest that all the roles place demands on all of the aspects considered.

Regarding the hypothesis that the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated, as presented in section 6.3, it was believed that if indeed the different roles did place different demands on the quality of the speech generated then the ratings of some of the aspects of the quality of the speech generated by the TTS synthesis systems with respect to their use in the different roles would differ across the different roles. The actual mean ratings presented in Table 12 show that, as predicted, all aspects of the speech generated by S1 considered so far differed across the 4 roles, with the exception of the naturalness of voice of S1 for use as a phonetic PM and a CP, and the appropriateness of register of S1 for use as a prosodic PM and a CP which were rated equally. Analysis of the data using the Friedman test revealed that the differences in the

following aspects of the speech generated by S1 across the 4 roles were statistically significant: naturalness and expressiveness (see Table 13). The differences in the following aspects of the speech generated by S1 across the 4 roles were, on the other hand, not found to be statistically significant: comprehensibility, intelligibility, accuracy, choice of pronunciation, naturalness of voice, and appropriateness of register (see Table 13)

**Table 13 Significance of differences among the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1**

| | $X^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Comprehensibility** | 3.508 | 3 | 0.320 | No |
| **Intelligibility** | 6.902 | 3 | 0.075 | No |
| **Choice of pronunciation** | 2.328 | 3 | 0.507 | No |
| **Accuracy** | 0.152 | 3 | 0.985 | No |
| **Naturalness** | 12.000 | 3 | 0.007 | Yes |
| **Naturalness of voice** | 5.313 | 3 | 0.150 | No |
| **Expressiveness** | 10.500 | 3 | 0.015 | Yes |
| **Appropriateness of register** | 7.244 | 3 | 0.065 | No |

Regarding individual aspects of the speech generated by the TTS synthesis systems, it was hypothesised that use of TTS synthesis as a phonetic PM would place greatest demands on accuracy and naturalness at the phonetic level and that use of TTS synthesis as a prosodic PM would place greatest demands on accuracy and naturalness at the prosodic level. The mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes and naturalness of prosody, the variables that were combined to arrive at measures of accuracy and naturalness, were therefore considered individually. They are presented in Table 14.

**Table 14 Mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Precision of phonemes** | 4.11 | 3.65 | 4.24 | 4.06 |
| **Appropriateness of prosody** | 3.41 | 3.47 | 2.82 | 2.88 |
| **Naturalness of phonemes** | 4.06 | 3.65 | 3.41 | 3.53 |
| **Naturalness of prosody** | 3.59 | 3.71 | 2.88 | 3.06 |

Very generally, the descriptive statistics presented in Table 14 show that all four of these aspects of the speech generated by S1 differed across the four roles. The data were analysed using the Friedman test. The results of this analysis, which are presented in Table 15, indicated that these differences were only statistically significant for naturalness of phonemes and naturalness of prosody.

**Table 15 Significance of differences among the precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1 across the 4 roles**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Precision of phonemes** | 2.546 | 3 | 0.467 | No |
| **Appropriateness of prosody** | 6.610 | 3 | 0.085 | No |
| **Naturalness of phonemes** | 8.460 | 3 | 0.037 | Yes |
| **Naturalness of prosody** | 15.189 | 3 | 0.002 | Yes |

More specifically, regarding the hypothesis that the role of phonetic PM will place higher demands on accuracy and naturalness at the phonetic level than the other roles, in line with this hypothesis, precision of phonemes was rated lower in this role than in the other three roles (see Table 14). The data were analysed using the Wilcoxon test in order to determine whether these differences were statistically significant. The results of this analysis, which are presented in Table 16, indicated that the differences between the ratings of precision of phonemes for the roles of phonetic PM and prosodic PM were statistically significant, but that those for the roles of phonetic PM and RM and phonetic PM and CP were not.

**Table 16 Significance of differences between the precision of phonemes of S1 for use as a phonetic PM and for use in the other three roles**

| Difference between phonetic PM and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 0.963 | 5 | 0.168 | One-tailed | No |
| **Prosodic PM** | 1.682 | 4 | 0.046 | One-tailed | Yes |
| **CP** | 1.087 | 5 | 0.139 | One-tailed | No |

Also in line with this hypothesis, naturalness of phonemes was rated lower for this role than for the roles of RM and CP. The data were analysed using the Wilcoxon test in order to determine whether these differences were statistically significant. The results of this analysis, presented in Table 17, indicated that neither the differences between the ratings of naturalness of phonemes for the roles of phonetic PM and CP, nor those between the ratings of naturalness of phonemes for the roles of phonetic PM and RM were statistically significant.

**Table 17 Significance of differences between the naturalness of phonemes of S1 for use as a phonetic PM and for use in the roles of RM and CP**

| Difference between phonetic PM and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 1.427 | 5 | 0.077 | One-tailed | No |
| **CP** | 0.557 | 8 | 0.282 | One-tailed | No |

In addition, naturalness of phonemes was rated higher for this role than for the roles of prosodic PM (see Table 14).

Regarding the hypothesis that the role of prosodic PM places higher demands on accuracy and naturalness at the prosodic level than the other roles, in line with this hypothesis, appropriateness of prosody and naturalness of prosody were rated lower in this role than in the other three roles (see Table 14). The data were analysed using the Wilcoxon test in order to determine whether these differences were statistically significant. The results of this analysis for appropriateness of prosody are presented in Table 18. They indicated that the differences between the ratings of appropriateness of prosody for the roles of prosodic PM and RM, prosodic PM and phonetic PM were statistically significant, but that those for the roles of prosodic PM and CP were not. The results of this analysis for naturalness of prosody are presented in Table 19. They indicated that the differences between the ratings of naturalness of prosody for the roles of prosodic PM and RM and prosodic PM and phonetic PM were statistically significant, but that those for the roles of prosodic PM and CP were not.

**Table 18 Significance of differences between the appropriateness of prosody of S1 for use as a prosodic PM and for use in the other three roles**

| Difference between prosodic PM and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 1.653 | 9 | 0.049 | One-tailed | Yes |
| **Phonetic PM** | 2.145 | 7 | 0.016 | One-tailed | Yes |
| **CP** | 0.277 | 7 | 0.391 | One-tailed | No |

**Table 19 Significance of differences between the naturalness of prosody of S1 for use as a prosodic PM and for use in the other three roles**

| Difference between prosodic PM and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 2.351 | 5 | 0.010 | One-tailed | Yes |
| **Phonetic PM** | 2.914 | 5 | 0.003 | One-tailed | Yes |
| **CP** | 1.134 | 13 | 0.138 | One-tailed | No |

Returning to the more general hypothesis that the different roles that TTS synthesis may assume within CALL applications place different demands on the speech generated, as presented in section 6.3, it is also believed that if indeed the different roles do place different requirements on the quality of the speech generated by TTS synthesis systems, then ratings of the adequacy and acceptability of the quality of the speech generated by the TTS synthesis systems will differ across the roles. The mean ratings of the adequacy and acceptability of the

speech generated by S1 for use in each of the four different roles were therefore calculated across participants. These results are presented in Table 20.

**Table 20 Mean ratings of the adequacy and the acceptability of S1for each of the four roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Adequacy** | 4.53 | 4.12 | 4.06 | 4.18 |
| **Acceptability** | 4.88 | 4.29 | 4.24 | 4.41 |

The descriptive statistics presented in Table 20 show that, as predicted, both the adequacy and the acceptability of the speech generated by S1 differed across the 4 roles. Analysis of the data using the Friedman test revealed that these differences were, however, not statistically significant ($\chi^2_r = 2.352$, df = 3, $p = 0.503$; $\chi^2_r = 6.616$, df = 3, $p = 0.085$, respectively).

#### *6.3.2.1.2 S2*

As for S1, the mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S2 with respect to its use in each of the four different roles that TTS synthesis systems may assume within CALL applications were calculated across participants. The results are presented in Table 21.

**Table 21 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S2**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Comprehensibility** | 5.06 | 5.41 | 5.41 | 5.59 |
| **Intelligibility** | 4.89 | 5.24 | 5.35 | 5.41 |
| **Choice of pronunciation** | 5.00 | 5.18 | 5.06 | 5.29 |
| **Accuracy** | 3.91 | 4.91 | 3.97 | 4.38 |
| **Naturalness** | 3.63 | 4.53 | 3.69 | 3.84 |
| **Naturalness of voice** | 3.94 | 5.00 | 4.18 | 4.41 |
| **Expressiveness** | 3.06 | 4.53 | 2.41 | 2.82 |
| **Appropriateness of register** | 4.59 | 5.12 | 4.76 | 4.88 |

The descriptive statistics, presented in Table 21, show that for all the roles none of the aspects of the speech generated by S2 considered received top ratings, as was the case for S1. This, as said for S1, appears to suggest that all the roles that TTS synthesis systems may assume within CALL applications place demands on all of the aspects considered.

In order to investigate the hypothesis that the different roles that TTS synthesis systems may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, as for S1, the mean ratings of the different aspects of the quality of the speech generated by S2 with respect to its use in the different roles that TTS synthesis may assume within CALL applications presented in Table 21 were analysed for differences across the four roles. In line with this hypothesis, the ratings of all the aspects of speech considered differed across the four roles with the exception of comprehensibility for use as a phonetic PM as a prosodic PM which were rated equally. The data were analysed using the Friedman test in order to determine whether these differences were statistically significant. The results presented in Table 22 show that while the differences in accuracy, naturalness of voice, expressiveness, and appropriateness of register across the roles were found to be statistically significant, the differences in comprehensibility, intelligibility, naturalness, and choice of pronunciation across the roles were not found to be statistically significant.

**Table 22 Significance of differences among the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S2**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Comprehensibility** | 4.486 | 3 | 0.214 | No |
| **Intelligibility** | 2.243 | 3 | 0.523 | No |
| **Choice of pronunciation** | 3.026 | 3 | 0.388 | No |
| **Accuracy** | 13.360 | 3 | 0.004 | Yes |
| **Naturalness** | 7.617 | 3 | 0.055 | No |
| **Naturalness of voice** | 14.258 | 3 | 0.003 | Yes |
| **Expressiveness** | 20.937 | 3 | < 0.0005 | Yes |
| **Appropriateness of register** | 10.920 | 3 | 0.012 | Yes |

As for S1, in order to permit the investigation of the hypotheses that use of TTS synthesis as a phonetic PM places greater demands on accuracy and naturalness at the phonetic level than the other roles and use of TTS synthesis as a prosodic PM places greater demands on accuracy and naturalness at the prosodic level than the other roles, the mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody were considered individually. They are presented in Table 23.

**Table 23 Mean ratings of the precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S2**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Precision of phonemes** | 4.47 | 5.18 | 4.94 | 5.18 |
| **Appropriateness of prosody** | 3.29 | 4.75 | 2.94 | 3.47 |
| **Naturalness of phonemes** | 4.18 | 4.71 | 4.35 | 4.19 |
| **Naturalness of prosody** | 3.06 | 4.53 | 3.24 | 3.65 |

Very generally, the descriptive statistics presented in Table 23 show that these aspects of the quality of the speech generated by S2 differed across the 4 roles with the exception of the precision of phonemes for use as a phonetic PM and a CP which received the same mean rating. The data were analysed using the Friedman test. The results of this analysis, which are presented in Table 24, revealed that these differences were statistically significant for precision of phonemes, appropriateness of prosody, naturalness of prosody and naturalness of voice, but not for naturalness of phonemes.

**Table 24 Significance of differences among precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S2 across the 4 roles**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Precision of phonemes** | 7.875 | 3 | 0.49 | Yes |
| **Appropriateness of prosody** | 21.429 | 3 | < 0.0005 | Yes |
| **Naturalness of phonemes** | 3.161 | 3 | 0.368 | No |
| **Naturalness of prosody** | 18.287 | 3 | < 0.0005 | Yes |

Regarding the specific hypothesis that the role of phonetic PM places higher demands on accuracy and naturalness at the phonetic level than the other roles, contrary to this hypothesis, both precision of phonemes and naturalness of phonemes were higher for the role of phonetic PM than for the other three roles.

As regards the hypothesis that the role of prosodic PM places higher demands on accuracy and naturalness at the prosodic level than the other roles, in line with this hypothesis appropriateness of prosody was lower for the role of prosodic PM than for the other three roles. The data were analysed using the Wilcoxon test in order to determine whether the differences were statistically significant. The results presented in Table 25 indicate that the differences between the ratings of appropriateness of prosody for the roles of prosodic PM and phonetic PM and prosodic PM and CP were statistically significant, but that those for the roles of prosodic PM and RM were not.

**Table 25 Significance of differences between the appropriateness of prosody of S2 for use as a prosodic PM and for use in the other three roles**

| Difference between prosodic PM and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 1.610 | 6 | 0.054 | One-tailed | No |
| **Phonetic PM** | 3.225 | 3 | 0.0005 | One-tailed | Yes |
| **CP** | 2.008 | 8 | 0.023 | One-tailed | Yes |

Also in line with this hypothesis, naturalness of prosody was lower for the role of prosodic PM than for the roles of phonetic PM and CP (see Table 23). The data were analysed using the Wilcoxon test in order to determine whether the differences were statistically significant. The results of this analysis presented in Table 26 indicated that the differences were statistically significant.

**Table 26 Significance of differences between the naturalness of prosody of S2 for use as a prosodic PM and for use in the roles of phonetic PM and CP**

| Difference between prosodic PM and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **Phonetic PM** | 3.099 | 3 | 0.001 | One-tailed | Yes |
| **CP** | 1.811 | 8 | 0.035 | One-tailed | Yes |

Contrary to this hypothesis, however, naturalness of prosody was higher for the role of prosodic PM than for the role of RM.

Returning to the general hypothesis that the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, the mean ratings of the adequacy and acceptability of the speech generated by S2 for use in each of the four different roles were calculated across participants. As for S1, the results, presented in Table 27 show that, as predicted, both the adequacy and the acceptability of the speech generated by S2 differed across the 4 roles. Analysis of the data using the Friedman test revealed that these differences were significant for adequacy ($\chi^2_r$ = 8.010, df = 3, $p$ = 0.046), but not for acceptability ($\chi^2_r$ = 6.303, df = 3, $p$ = 0.098).

**Table 27 Mean ratings of the adequacy and acceptability of S2 for each of the four roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Adequacy** | 4.76 | 5.00 | 4.41 | 4.65 |
| **Acceptability** | 5.12 | 5.41 | 4.82 | 5.06 |

#### *6.3.2.1.3 S3*

The mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of the speech generated by S3 with respect to its use in the 4 roles across participants are presented in Table 28.

**Table 28 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S3**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Comprehensibility** | 4.12 | 3.71 | 4.77 | 4.59 |
| **Intelligibility** | 3.88 | 3.18 | 4.59 | 4.35 |
| **Choice of pronunciation** | 4.00 | 3.71 | 4.41 | 4.35 |
| **Accuracy** | 3.32 | 2.85 | 3.79 | 3.62 |
| **Naturalness** | 2.53 | 2.38 | 3.09 | 2.97 |
| **Naturalness of voice** | 2.24 | 1.94 | 2.82 | 2.71 |
| **Expressiveness** | 2.18 | 2.24 | 2.88 | 2.94 |
| **Appropriateness of register** | 4.12 | 3.76 | 4.24 | 4.18 |

As for S1 and S2, the descriptive statistics presented in Table 28 show that for all the roles none of the aspects of the speech generated by S3 considered attained top ratings. This, as said for S1 and S2, appears to suggest that all the roles place demands on all of the aspects of speech considered.

Regarding the hypothesis that the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, as for S1 and S2, the mean ratings of the different aspects of the quality of the speech generated by S3 with respect to its use in the different roles that TTS synthesis may assume within CALL applications presented in Table 28 were analysed for differences across the four roles. They show that as predicted the comprehensibility, intelligibility, accuracy, naturalness, choice of pronunciation, naturalness of voice, expressiveness, and appropriateness of register of the speech generated by S3 differed across the 4 roles. Analysis of the data using the Friedman test revealed that the differences in all the variables across the 4 roles were statistically significant with the exception of appropriateness of register (see Table 29).

**Table 29 Significance of differences among the comprehensibility, intelligibility, accuracy, naturalness, choice of pronunciation, naturalness of voice, expressiveness, and appropriateness of register of S3**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Comprehensibility** | 14.714 | 3 | 0.002 | Yes |
| **Intelligibility** | 23.575 | 3 | < 0.0005 | Yes |
| **Choice of pronunciation** | 9.393 | 3 | 0.025 | Yes |
| **Accuracy** | 15.866 | 3 | 0.001 | Yes |
| **Naturalness** | 9.063 | 3 | 0.028 | Yes |
| **Naturalness of voice** | 14.143 | 3 | 0.003 | Yes |
| **Expressiveness** | 8.286 | 3 | 0.040 | Yes |
| **Appropriateness of register** | 4.534 | 3 | 0.209 | No |

In order to investigate the hypotheses that of all the roles considered the roles of phonetic PM and prosodic PM place greatest demands on accuracy and naturalness at the phonetic and prosodic levels respectively, as for S1 and S2, the mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody were considered individually. These descriptive statistics are presented in Table 30.

**Table 30 Mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S3**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Precision of phonemes** | 3.76 | 2.82 | 4.12 | 3.88 |
| **Appropriateness of prosody** | 2.88 | 2.88 | 3.47 | 3.35 |
| **Naturalness of phonemes** | 2.76 | 2.29 | 3.18 | 2.94 |
| **Naturalness of prosody** | 2.29 | 2.47 | 3.00 | 3.00 |

Very generally, the descriptive statistics presented in Table 30 show that these aspects of the speech generated by S3 differed across the roles with the exception of the appropriateness of prosody for use as an RM and as a phonetic PM which were rated equally and the naturalness of prosody for use as a prosodic PM and a CP which were also rated equally. The data were analysed using the Friedman test in order to determine whether the differences were statistically significant. The results which are presented in Table 31 show that the differences in the following aspects of the speech generated by S3 were statistically significant: precision of phonemes, naturalness of phonemes, naturalness of prosody. The differences in appropriateness of prosody, on the other hand, were not found to be statistically significant.

**Table 31 Significance of differences among precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S3 across the 4 roles**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Precision of phonemes** | 19.387 | 3 | < 0.0005 | Yes |
| **Appropriateness of prosody** | 6.551 | 3 | 0.088 | No |
| **Naturalness of phonemes** | 8.298 | 3 | 0.040 | Yes |
| **Naturalness of prosody** | 8.232 | 3 | 0.041 | Yes |

Regarding the specific hypothesis that the role of phonetic PM places higher demands on accuracy and naturalness at the phonetic level than the other roles, in line with this hypothesis, both precision of phonemes and naturalness of phonemes were rated lower in this role than in the other three roles. The data were analysed using the Wilcoxon test in order to determine whether these differences were statistically significant. The results of this analysis for precision of phonemes are presented in Table 32. They indicate that the differences for precision of phonemes are statistically significant. The results of this analysis for naturalness of phonemes are presented in Table 33. They indicate that the differences between naturalness of phonemes for the roles of phonetic PM and prosodic PM and phonetic PM and CP are statistically significant, while those for phonetic PM and RM are not.

**Table 32 Significance of differences between the precision of phonemes of S3 for use as a phonetic PM and for use in the other three roles**

| Difference between phonetic PM and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 2.499 | 2 | 0.006 | One-tailed | Yes |
| **Prosodic PM** | 1.816 | 4 | 0.035 | One-tailed | Yes |
| **CP** | 2.970 | 4 | 0.002 | One-tailed | Yes |

**Table 33 Significance of differences between the naturalness of phonemes of S3 for use as a phonetic PM and for use in the other three roles**

| Difference between phonetic PM and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 1.565 | 5 | 0.059 | One-tailed | No |
| **Prosodic PM** | 2.437 | 5 | 0.008 | One-tailed | Yes |
| **CP** | 1.990 | 5 | 0.024 | One-tailed | Yes |

Regarding the specific hypothesis that the role of prosodic PM places higher demands on accuracy and naturalness at the prosodic level than the other roles, contrary to this hypothesis appropriateness of prosody and naturalness of prosody were higher for the role of prosodic PM than for the other three roles.

Returning to the general hypothesis that the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, the mean ratings of the adequacy and acceptability of the speech generated by S3 for use in each of the four different roles were calculated across participants. As for S1 and S2, the results, presented in Table 34 show that, as predicted, both the adequacy and the acceptability of the speech generated by S3 differed across the 4 roles. The data were again analysed using the Friedman test. Like for S1, the test revealed that the differences were not statistically significant for either adequacy or acceptability ($\chi^2_r = 3.467$, df = 3, $p = 0.325$; $\chi^2_r = 3.194$, df = 3, $p = 0.363$, respectively).

**Table 34 Mean ratings of the adequacy and acceptability of S3 for each of the four roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Adequacy** | 3.76 | 3.59 | 3.94 | 4.05 |
| **Acceptability** | 4.18 | 3.82 | 4.35 | 4.29 |

#### *6.3.2.1.4 S6*

The mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of the speech generated by S6 with respect to its use in the four roles across participants are presented in Table 35.

**Table 35 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S6**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Comprehensibility** | 5.65 | 5.88 | 5.94 | 6.47 |
| **Intelligibility** | 5.41 | 6.12 | 5.82 | 6.29 |
| **Choice of pronunciation** | 5.71 | 5.76 | 5.71 | 6.47 |
| **Accuracy** | 5.38 | 5.71 | 5.29 | 5.82 |
| **Naturalness** | 5.38 | 5.60 | 5.38 | 5.78 |
| **Naturalness of voice** | 5.53 | 5.69 | 5.53 | 6.06 |
| **Expressiveness** | 4.94 | 5.24 | 4.88 | 5.18 |
| **Appropriateness of register** | 5.47 | 5.76 | 5.41 | 5.65 |

The descriptive statistics presented in Table 35 show that for the roles of RM and prosodic PM, none of the aspects of the quality of the speech generated by S6 considered attained top ratings. This would appear to suggest that these roles place demands on all of the aspects of

the quality of the speech generated by TTS synthesis systems considered. Regarding the role of phonetic PM, the descriptive statistics presented in Table 35 show that the following aspects of the quality of the speech generated by S6 did not receive top ratings: comprehensibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register. This would appear to suggest that the role of phonetic PM places demands on these aspects of the quality of the speech generated by TTS synthesis systems. Finally, regarding the role of CP, the descriptive statistics presented in Table 35 show that the following aspects of the quality of the speech generated by S6 did not receive top ratings: accuracy, naturalness, expressiveness and appropriateness of register. This would appear to suggest that the role of CP places demands on these aspects of the quality of the speech generated by TTS synthesis systems.

Regarding the hypothesis that the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, as for S1 and S2, the mean ratings of the different aspects of the quality of the speech generated by S6 with respect to its use in the different roles that TTS synthesis may assume within CALL applications were analysed for differences across the four roles. Presented in Table 35, they show that, as expected, the aspects of the speech generated by S1 differed across the 4 roles, with the exception of choice of pronunciation for use as an RM and as a prosodic PM, which were rated equally. The data were analysed using the Friedman test in order to determine whether the differences in these aspects of speech across roles were statistically significant. The results are presented in Table 36. These results reveal that the differences across the roles were statistically significant for comprehensibility, intelligibility, and choice of pronunciation, but not for accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register.

**Table 36 Significance of differences among the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S6**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Comprehensibility** | 10.107 | 3 | 0.018 | Yes |
| **Intelligibility** | 15.406 | 3 | 0.002 | Yes |
| **Choice of pronunciation** | 13.527 | 3 | 0.004 | Yes |
| **Accuracy** | 3.678 | 3 | 0.298 | No |
| **Naturalness** | 6.750 | 3 | 0.080 | No |
| **Naturalness of voice** | 5.902 | 3 | 0.116 | No |
| **Expressiveness** | 2.918 | 3 | 0.404 | No |
| **Appropriateness of register** | 4.765 | 3 | 0.190 | No |

Regarding the hypotheses that, of all the four roles, the roles of phonetic and prosodic PM place greatest demands on accuracy and naturalness at the phonetic and prosodic levels respectively, the mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes and naturalness of prosody were considered individually (see Table 37).

**Table 37 Mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S6**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Precision of phonemes** | 5.71 | 5.82 | 5.41 | 6.24 |
| **Appropriateness of prosody** | 5.06 | 5.59 | 5.18 | 5.41 |
| **Naturalness of phonemes** | 5.59 | 5.82 | 5.44 | 6.00 |
| **Naturalness of prosody** | 5.18 | 5.47 | 5.24 | 5.65 |

Very generally, these aspects of the speech generated by S6 differed across the four roles. The data were analysed using the Friedman test in order to determine whether the differences were statistically significant. The results presented in Table 38 indicate that the differences were statistically significant for precision of phonemes only.

**Table 38 Significance of differences among, precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S6 across the 4 CALL setups**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Precision of phonemes** | 9.072 | 3 | 0.028 | Yes |
| **Appropriateness of prosody** | 3.639 | 3 | 0.303 | No |
| **Naturalness of phonemes** | 7.255 | 3 | 0.064 | No |
| **Naturalness of prosody** | 4.165 | 3 | 0.244 | No |

More specifically, regarding the hypothesis that the role of phonetic PM places higher demands on accuracy and naturalness at the phonetic level than the other three roles, in line with this hypothesis precision of phonemes and naturalness of phonemes were lower in this role than for the role of CP. Analysis of the data using the Wilcoxon test revealed that these differences were statistically significant for precision of phonemes (z = 1.933, N-ties = 10, $p$ = 0.027, one-tailed), but not for naturalness of phonemes (z = 0.866, N-ties = 9, $p$ = 0.418, one-tailed). Also contrary to the hypothesis, precision of phonemes and naturalness of phonemes were higher for the role of phonetic PM than for the roles of RM and prosodic PM.

Regarding the hypothesis that the role of prosodic PM places higher demands on accuracy and naturalness at the prosodic level than the other three roles, in line with this hypothesis appropriateness of prosody and naturalness of prosody were lower for the role of prosodic PM than for the roles of phonetic PM and CP. The data were analysed using the Wilcoxon test to determine whether these differences were statistically significant. The results of this analysis for appropriateness of prosody are presented in Table 39. They indicate that the differences for appropriateness of prosody were not statistically significant. The results of this analysis for naturalness of prosody are presented in Table 40. They indicate that the differences for naturalness of prosody were statistically significant.

**Table 39 Significance of differences between the appropriateness of prosody of S6 for use as a prosodic PM and for use as a phonetic PM and a CP**

| Difference between prosodic PM and ... | z value | N-ties | $p$ value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **Phonetic PM** | 1.218 | 7 | 0.112 | One-tailed | No |
| **CP** | 1.155 | 5 | 0.124 | One-tailed | No |

**Table 40 Significance of differences between the naturalness of prosody of S6 for use as a prosodic PM and and for use as a phonetic PM and a CP**

| Difference between prosodic PM and ... | z value | N-ties | $p$ value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **Phonetic PM** | 2.673 | 6 | 0.004 | One-tailed | Yes |
| **CP** | 1.706 | 6 | 0.044 | One-tailed | Yes |

Contrary to this hypothesis, precision of phonemes and naturalness of phonemes were, however, higher for the role of prosodic PM than in the role of RM.

Returning to the general hypothesis that the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, the mean ratings of the adequacy and acceptability of the speech generated by S6 for use in each of the four different roles were calculated across participants. Like for the other three TTS synthesis systems, the results, presented in Table 41 show that, as predicted, the adequacy of the speech generated by S6 differed across the 4 roles. Regarding the acceptability of S6 for use in the 4 roles, unlike for the other three TTS synthesis systems, differences were not found across all the four roles: the mean ratings of the acceptability of S6 for use as a phonetic PM and as a prosodic PM were equal. Analysis of the data using the Friedman test revealed that the differences were statistically significant for adequacy ($\chi^2_r$ = 8.063, df = 3, $p$ = 0.045), but not for acceptability ($\chi^2_r$ = 5.547, df = 3, $p$ = 0.163).

**Table 41 Mean ratings of the adequacy and acceptability of S6 for each of the four roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Adequacy** | 5.35 | 5.65 | 5.24 | 5.82 |
| **Acceptability** | 6.00 | 5.94 | 5.94 | 6.29 |

#### *6.3.2.1.5 TTS synthesis in general*

Regarding the hypothesis that the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, as presented in section 6.3, if indeed they do place different demands on the quality of the speech generated by TTS synthesis systems, it is believed that the participants' ratings of the readiness of TTS synthesis in general having participated in the investigation will also differ across the four roles. The mean ratings of the readiness of the TTS synthesis systems in general for use in each of the four roles were therefore calculated across the participants. The results, presented in Table 42 show that, as predicted, readiness differed across all four roles. Analysis of the data using the Friedman test revealed that these differences were statistically significant ($\chi^2_r$ = 13.244, df = 3, $p$ = 0.004).

**Table 42 Mean ratings of the readiness of TTS synthesis in general for in the four roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Readiness** | 4.53 | 4.71 | 3.71 | 3.88 |

### 6.3.2.2 Does the speech generated by different TTS synthesis systems differ in quality?

In this section, the participants' ratings of the quality of the speech generated by the TTS systems are analysed for each of the roles that TTS synthesis may assume within CALL applications, namely RM (see section 6.3.2.2.1), phonetic PM (see section 6.3.2.2.2), prosodic PM (see section 6.3.2.2.3) and CP (see section 6.3.2.2.4), in turn with respect to the hypothesis that the speech generated by different TTS synthesis systems differs in quality.

#### *6.3.2.2.1 RM*

As presented in section 6.3, it is believed that, if indeed the speech generated by the different TTS synthesis systems does differ in quality, then participants' ratings of the different aspects of the speech generated by the TTS synthesis systems considered in this investigation will differ across the TTS synthesis systems.

The mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of the speech generated by S1, S2, S3 and S6 with respect to its use as an RM were therefore calculated across participants. The results are presented in Table 43.

**Table 43 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as an RM**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Comprehensibility** | 4.53 | 5.06 | 4.12 | 5.65 |
| **Intelligibility** | 4.53 | 4.89 | 3.88 | 5.41 |
| **Choice of pronunciation** | 4.59 | 5.00 | 4.00 | 5.71 |
| **Accuracy** | 3.76 | 3.88 | 3.32 | 5.38 |
| **Naturalness** | 3.82 | 3.62 | 2.53 | 5.38 |
| **Naturalness of voice** | 4.00 | 3.94 | 2.24 | 5.53 |
| **Expressiveness** | 3.24 | 3.06 | 2.18 | 4.94 |
| **Appropriateness of register** | 5.00 | 4.59 | 4.12 | 5.47 |

The descriptive statistics presented in Table 43 show that, as predicted, intelligibility, accuracy, naturalness, choice of pronunciation, naturalness of voice, expressiveness, and appropriateness of register for use as an RM differed across the TTS synthesis systems. The data were analysed using the Friedman test in order to determine whether these differences were statistically significant. The results presented in Table 44 show that these differences were statistically significant.

**Table 44 Significance of differences among the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as an RM**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Comprehensibility** | 17.493 | 3 | 0.001 | Yes |
| **Intelligibility** | 15.521 | 3 | 0.001 | Yes |
| **Choice of pronunciation** | 18.568 | 3 | < 0.0005 | Yes |
| **Accuracy** | 26.809 | 3 | < 0.0005 | Yes |
| **Naturalness** | 34.107 | 3 | < 0.0005 | Yes |
| **Naturalness of voice** | 34.788 | 3 | < 0.0005 | Yes |
| **Expressiveness** | 31.591 | 3 | < 0.0005 | Yes |
| **Appropriateness of register** | 17.894 | 3 | < 0.0005 | Yes |

More specifically, regarding the different types of TTS synthesis systems considered in this investigation, as presented in section 6.3, it was expected that the systems based on USS, namely S1 and S6, would be more accurate and natural at the prosodic level and hence more comprehensible than the systems based on concatenative synthesis, namely S2 and S3, the individual scales that were combined in order to arrive at measures of accuracy and naturalness were therefore also analysed separately. The mean ratings of these aspects of the speech generated by S1, S2, S3, and S6 with respect to its use as an RM across participants are presented in Table 45.

**Table 45 Mean ratings of the, precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1, S2, S3, and S6 for use as an RM**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Precision of phonemes** | 4.11 | 4.47 | 3.76 | 5.71 |
| **Appropriateness of prosody** | 3.41 | 3.29 | 2.88 | 5.06 |
| **Naturalness of phonemes** | 4.06 | 4.18 | 2.76 | 5.59 |
| **Naturalness of prosody** | 3.59 | 3.06 | 2.29 | 5.18 |

The mean ratings of these aspects of the speech also differed across the TTS synthesis systems. Analysis of the data using the Friedman test revealed that these differences were statistically significant (see Table 46).

**Table 46 Significance of differences among the precision of phonemes, appropriateness of prosody, and naturalness of phonemes, naturalness of prosody of S1, S2, S3, and S6 for use as an RM**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Precision of phonemes** | 19.107 | 3 | < 0.0005 | Yes |
| **Appropriateness of prosody** | 31.159 | 3 | < 0.0005 | Yes |
| **Naturalness of phonemes** | 32.197 | 3 | < 0.0005 | Yes |
| **Naturalness of prosody** | 34.655 | 3 | < 0.0005 | Yes |

Regarding the specific hypothesis that the systems based on USS would be more accurate and natural at the prosodic level than the systems based on concatenative synthesis, in line with this hypothesis, the descriptive statistics presented in Table 45, show that the appropriateness of the prosody and the naturalness of the prosody were higher for both S1 and S6 than for S2 and S3 for this role. Analysis of the data using the Wilcoxon test revealed that the differences in appropriateness of prosody and naturalness of prosody between S1 and S3 (see Tables 47 and 49) and between S6 and S2 and S3 (see Tables 48 and 50) were statistically significant, but that those between S1 and S2 were not (see Tables 47 and 49).

**Table 47 Significance of differences between the appropriateness of the prosody of S1 and S2 and S3 for use as an RM**

| Difference between S1 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 0.000 | 8 | 0.500 | One-tailed | No |
| S3 | 2.081 | 10 | 0.019 | One-tailed | Yes |

**Table 48 Significance of differences between the appropriateness of the prosody of S6 and S2 and S3 for use as an RM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 3.345 | 3 | 0.0005 | One-tailed | Yes |
| S3 | 3.558 | 1 | < 0.0005 | One-tailed | Yes |

**Table 49 Significance of differences between the naturalness of the prosody of S1 and S2 and S3 for use as an RM**

| Difference between S1 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 1.140 | 8 | 0.127 | One-tailed | No |
| S3 | 2.758 | 3 | 0.003 | One-tailed | Yes |

**Table 50 Significance of differences between the naturalness of the prosody of S6 and S2 and S3 for use as an RM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 3.449 | 2 | 0.0005 | One-tailed | Yes |
| S3 | 3.641 | 0 | < 0.0005 | One-tailed | Yes |

Regarding the specific hypothesis that the speech generated by the systems based on USS would be more comprehensible than that generated by the systems based on concatenative

synthesis, in line with this hypothesis, the descriptive statistics presented in Table 45 show that the comprehensibility of the speech generated by S6 with respect to its use as an RM was higher than that of the speech generated by S2 and S3, and that that of the speech generated by S1 was higher than that of the speech generated by S3. Contrary to this hypothesis, the comprehensibility of the speech generated by S1 with respect to it use as an RM was, however, lower than that of the speech generated by S2. The data were analysed using the Wilcoxon test in order to determine whether, in the cases where the speech generated by the systems based on USS was more comprehensible than the speech generated by the systems based on concatenative synthesis, the differences were statistically significant. The results presented in Tables 51 and 52 indicate that the differences in comprehensibility between S6 and S2 and S3 were statistically significant, but that those between S1 and S3 were not.

**Table 51 Significance of differences between the comprehensibility of S1 and S3 for use as an RM**

| Difference between S1 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S3 | 0.964 | 4 | 0.168 | One-tailed | No |

**Table 52 Significance of differences between the comprehensibility of S6 and S2 and S3 for use as an RM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 2.066 | 7 | 0.015 | One-tailed | Yes |
| S3 | 3.2.09 | 0 | 0.0005 | One-tailed | Yes |

#### *6.3.2.2.2 Phonetic PM*

The mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of the speech generated by S1, S2, S3 and S6 with respect to its use as a phonetic PM across participants are presented in Table 53.

**Table 53 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a phonetic PM**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Comprehensibility** | 4.24 | 5.41 | 3.71 | 5.88 |
| **Intelligibility** | 3.88 | 5.24 | 3.18 | 6.12 |
| **Choice of pronunciation** | 4.12 | 5.18 | 3.71 | 5.76 |
| **Accuracy** | 3.59 | 4.91 | 2.84 | 5.66 |
| **Naturalness** | 3.68 | 4.62 | 2.38 | 5.65 |
| **Naturalness of voice** | 3.53 | 5.00 | 1.94 | 5.69 |
| **Expressiveness** | 3.12 | 4.53 | 2.24 | 5.24 |
| **Appropriateness of register** | 4.76 | 5.12 | 3.76 | 5.76 |

As for the role of RM, as predicted, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register for use as a phonetic PM differed across the TTS synthesis systems. Analysis of the data using the Friedman test revealed that these differences were statistically significant (see Table 54).

**Table 54 Significance of differences among the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a phonetic PM**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Comprehensibility** | 32.460 | 3 | < 0.0005 | Yes |
| **Intelligibility** | 34.846 | 3 | < 0.0005 | Yes |
| **Choice of pronunciation** | 28.280 | 3 | < 0.0005 | Yes |
| **Accuracy** | 37.480 | 3 | < 0.0005 | Yes |
| **Naturalness** | 43.660 | 3 | < 0.0005 | Yes |
| **Naturalness of voice** | 42.345 | 3 | < 0.0005 | Yes |
| **Expressiveness** | 36.395 | 3 | < 0.0005 | Yes |
| **Appropriateness of register** | 29.068 | 3 | < 0.0005 | Yes |

As for the role of RM, in order to permit the investigation of the hypothesis that the systems based on USS would be more accurate and natural at the prosodic level and hence more comprehensible than the systems based on concatenative synthesis, the mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody were considered individually. They are presented in Table 55.

**Table 55 Mean ratings of the precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1, S2, S3, and S6 for use as a phonetic PM**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Precision of phonemes** | 3.65 | 5.18 | 2.82 | 5.82 |
| **Appropriateness of prosody** | 3.47 | 4.75 | 2.88 | 5.59 |
| **Naturalness of phonemes** | 3.65 | 4.71 | 2.29 | 5.82 |
| **Naturalness of prosody** | 3.71 | 4.53 | 2.47 | 5.47 |

Like for the role of RM, the mean ratings of these aspects of the speech also differed across the TTS synthesis systems (see Table 55). Also as for the role of RM, analysis of the data using the Friedman test revealed that these differences were statistically significant (Table 56).

**Table 56 Significance of differences among the precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1, S2, S3, and S6 for use as a phonetic PM**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Precision of phonemes** | 39.656 | 3 | < 0.0005 | Yes |
| **Appropriateness of prosody** | 29.056 | 3 | < 0.0005 | Yes |
| **Naturalness of phonemes** | 43.981 | 3 | < 0.0005 | Yes |
| **Naturalness of prosody** | 40.188 | 3 | < 0.0005 | Yes |

Regarding the specific hypothesis that the systems based on USS would be more accurate and natural at the prosodic level than the systems based on concatenative synthesis, in line with this hypothesis, like for the role of RM, the descriptive statistics presented in Table 55, show that the appropriateness of the prosody and the naturalness of the prosody for S6 were higher than for S2 and S3 for this role. Analysis of the data using the Wilcoxon test revealed that these differences were statistically significant (see Tables 57 and 58).

**Table 57 Significance of differences between the appropriateness of the prosody of S6 and S2 and S3 for use as a phonetic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 2.066 | 6 | 0.020 | One-tailed | Yes |
| **S3** | 3.459 | 1 | 0.0005 | One-tailed | Yes |

**Table 58 Significance of differences between the naturalness of the prosody of S6 and S2 and S3 for use as a phonetic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 2.588 | 9 | 0.005 | One-tailed | Yes |
| **S3** | 3.639 | 0 | < 0.0005 | One-tailed | Yes |

Also in line with this hypothesis, like for the role of RM, the appropriateness of prosody and naturalness of prosody for S1 were higher than for S3. Analysis of the data using the Wilcoxon test revealed that the differences for naturalness of prosody were statistically significant (z = 2.986, N-ties = 4, $p$ = 0.002, one-tailed), but that those for appropriateness of prosody were not (z = 1.585, N-ties = 3, $p$ = 0.057, one-tailed). Unlike for the role of RM, contrary to this hypothesis, however, the appropriateness of prosody and naturalness of prosody for S1 were lower than for S2.

Regarding the hypothesis that the systems based on USS would be more comprehensible than the systems based on concatenative synthesis, like for the role of RM, in line with this hypothesis, the mean rating of the comprehensibility of the speech generated by S6 for use as a phonetic PM was higher than that of S2 and S3 and that of S1 was higher than that of S3. Contrary to this hypothesis, like for the role of RM, the mean rating of the comprehensibility of the speech generated by S1 for use as a phonetic PM was lower than that of S2. The data were analysed using the Wilcoxon test in order to determine whether, in the cases where the speech generated by the systems based on USS was more comprehensible than the speech generated by the systems based on concatenative synthesis, the differences were statistically significant. The results presented in Tables 59 and 60 indicate that the differences in comprehensibility between S1 and S3 and S6 and S3 were statistically significant, but that those between S6 and S2 were not.

**Table 59 Significance of differences between the comprehensibility of S1 and S3 for use as a phonetic PM**

| Difference between S1 and ... | z value | N-ties | $p$ value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S3** | 1.853 | 5 | 0.032 | One-tailed | Yes |

**Table 60 Significance of differences between the comprehensibility of S6 and S2 and S3 for use as a phonetic PM**

| Difference between S6 and ... | z value | N-ties | $p$ value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 1.347 | 6 | 0.089 | One-tailed | No |
| **S3** | 3.509 | 0 | < 0.0005 | One-tailed | Yes |

#### *6.3.2.2.3 Prosodic PM*

The mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of the speech

S1, S2, S3 and S6 with respect to its use as a prosodic PM across participants are presented in Table 61.

**Table 61 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a prosodic PM**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Comprehensibility** | 4.82 | 5.41 | 4.77 | 5.94 |
| **Intelligibility** | 4.65 | 5.35 | 4.59 | 5.82 |
| **Choice of pronunciation** | 4.53 | 5.06 | 4.41 | 5.71 |
| **Accuracy** | 3.53 | 3.94 | 3.79 | 5.29 |
| **Naturalness** | 3.09 | 3.75 | 3.03 | 5.38 |
| **Naturalness of voice** | 3.59 | 4.18 | 2.82 | 5.53 |
| **Expressiveness** | 2.35 | 2.41 | 2.88 | 4.88 |
| **Appropriateness of register** | 4.53 | 4.76 | 4.24 | 5.41 |

Like for the roles of RM and phonetic PM, the descriptive statistics presented in Table 61 show that, as predicted, the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register for use as a prosodic PM differed across the TTS synthesis systems. Also like for the roles of RM and phonetic PM, analysis of the data using the Friedman test revealed that these differences were statistically significant (see Table 62).

**Table 62 Significance of differences among the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a prosodic PM**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Comprehensibility** | 20.791 | 3 | < 0.0005 | Yes |
| **Intelligibility** | 24.773 | 3 | < 0.0005 | Yes |
| **Choice of pronunciation** | 17.887 | 3 | < 0.0005 | Yes |
| **Accuracy** | 26.456 | 3 | < 0.0005 | Yes |
| **Naturalness** | 35.664 | 3 | < 0.0005 | Yes |
| **Naturalness of voice** | 31.213 | 3 | < 0.0005 | Yes |
| **Expressiveness** | 28.717 | 3 | < 0.0005 | Yes |
| **Appropriateness of register** | 19.455 | 3 | < 0.0005 | Yes |

In order to permit the investigation of the hypothesis that the systems based on USS would be more accurate and natural at the prosodic level and hence more comprehensible than the systems based on concatenative synthesis, the mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody were considered individually. They are presented in Table 64.

**Table 63 Mean ratings of the precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1, S2, S3, and S6 for use as a prosodic PM**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Precision of phonemes** | 4.24 | 4.94 | 4.12 | 5.41 |
| **Appropriateness of prosody** | 2.82 | 2.94 | 3.47 | 5.18 |
| **Naturalness of phonemes** | 3.41 | 4.35 | 3.18 | 5.44 |
| **Naturalness of prosody** | 2.88 | 3.24 | 3.00 | 5.24 |

Very generally, like for the roles of RM and phonetic PM, the descriptive statistics presented in Table 63 show that these aspects of the quality of the speech generated also differed across the TTS synthesis systems. Analysis of data using the Friedman test revealed that these differences were statistically significant (see Table 64).

**Table 64 Significance of differences among the precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1, S2, S3, and S6 for use as a prosodic PM**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Precision of phonemes** | 18.956 | 3 | < 0.0005 | Yes |
| **Appropriateness of prosody** | 29.934 | 3 | < 0.0005 | Yes |
| **Naturalness of phonemes** | 33.644 | 3 | < 0.0005 | Yes |
| **Naturalness of prosody** | 32.628 | 3 | < 0.0005 | Yes |

Regarding the specific hypothesis that the systems based on USS would be more accurate and natural at the prosodic level and hence more comprehensible than the systems based on concatenative synthesis, in line with this hypothesis, as for the roles of RM and phonetic PM, the descriptive statistics presented in Table 63, show that the appropriateness of the prosody and the naturalness of the prosody for S6 were higher than for S2 and S3 for this role. Also like for the roles of RM and phonetic PM, analysis of the data using the Wilcoxon test revealed that these differences were statistically significant (see Tables 65 and 66).

**Table 65 Significance of differences between the appropriateness of the prosody of S6 and S2 and S3 for use as a prosodic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 3.328 | 3 | 0.0005 | One-tailed | Yes |
| **S3** | 3.231 | 2 | 0.0005 | One-tailed | Yes |

**Table 66 Significance of differences between the naturalness of the prosody of S6 and S2 and S3 for use as a prosodic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 3.559 | 1 | < 0.0005 | One-tailed | Yes |
| S3 | 3.462 | 2 | 0.0005 | One-tailed | Yes |

Also in line with this hypothesis, like for the roles of RM and phonetic PM, the appropriateness of prosody and naturalness of prosody for S1 were higher than for S3. Analysis of the data using the Wilcoxon test revealed that these differences were not, however, statistically significant (z = 1.563, N-ties = 5, $p$ = 0.059, one-tailed; z = 0.288, N-ties = 5, $p$ = 0.387, one-tailed, respectively). Also contrary to this hypothesis, like for the role of phonetic PM, the appropriateness of prosody and naturalness of prosody for S1 were lower than for S2.

Regarding the hypothesis that the systems based on USS would be more comprehensible than the systems based on concatenative synthesis, in line with this hypothesis, like for the roles of RM and phonetic PM, the mean rating of the comprehensibility of the speech generated by S6 for use as a prosodic PM was higher than that of S2 and S3 and that of S1 was higher than that of S3. Contrary to this hypothesis, however, also like for the roles of RM and phonetic PM, the mean rating of the comprehensibility of the speech generated by S1 for use as a prosodic PM was lower than that of S2. The data were analysed using the Wilcoxon test in order to determine whether, in the cases where the speech generated by the systems based on USS was more comprehensible than the speech generated by the systems based on concatenative synthesis, the differences were statistically significant. The results of this analysis presented in Tables 67 and 68 indicate that the differences in comprehensibility between S6 and S2 and S3 were statistically significant, but that those between S1 and S3 were not.

**Table 67 Significance of differences between the comprehensibility of S1 and S3 for use as a prosodic PM**

| Difference between S1 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S3 | 0.108 | 11 | 0.457 | One-tailed | No |

**Table 68 Significance of differences between the comprehensibility of S6 and S2 and S3 for use as a prosodic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 2.496 | 7 | 0.007 | One-tailed | Yes |
| S3 | 2.989 | 6 | 0.002 | One-tailed | Yes |

#### *6.3.2.2.4 CP*

The mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of the speech S1, S2, S3 and S6 with respect to its use as a CP across participants are presented in Table 69.

**Table 69 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a CP**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Comprehensibility** | 4.47 | 5.59 | 4.59 | 6.47 |
| **Intelligibility** | 4.24 | 5.41 | 4.35 | 6.29 |
| **Choice of pronunciation** | 4.29 | 5.29 | 4.35 | 6.47 |
| **Accuracy** | 3.47 | 4.32 | 3.62 | 5.82 |
| **Naturalness** | 3.19 | 3.84 | 2.88 | 5.78 |
| **Naturalness of voice** | 3.53 | 4.41 | 2.71 | 6.06 |
| **Expressiveness** | 2.65 | 2.82 | 2.94 | 5.18 |
| **Appropriateness of register** | 4.53 | 4.88 | 4.18 | 5.65 |

Like for the other three roles, the descriptive statistics presented in Table 69 show that, as predicted, the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register for use as a CP differed across the TTS synthesis systems. Analysis of the data using the Friedman test revealed the differences were significant for all of the variables considered (see Table 70).

**Table 70 Significance of differences among the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a CP**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Comprehensibility** | 33.921 | 3 | < 0.0005 | Yes |
| **Intelligibility** | 36.894 | 3 | < 0.0005 | Yes |
| **Choice of pronunciation** | 31.650 | 3 | < 0.0005 | Yes |
| **Accuracy** | 35.481 | 3 | < 0.0005 | Yes |
| **Naturalness** | 35.163 | 3 | < 0.0005 | Yes |
| **Naturalness of voice** | 37.411 | 3 | < 0.0005 | Yes |
| **Expressiveness** | 29.942 | 3 | < 0.0005 | Yes |
| **Appropriateness of register** | 22.916 | 3 | < 0.0005 | Yes |

In order to permit the investigation of the hypothesis that the systems based on USS would be more accurate and natural at the prosodic level and hence more comprehensible than the systems based on concatenative synthesis, the mean ratings of precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody were considered individually. They are presented in Table 71.

**Table 71 Mean ratings of the precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1, S2, S3, and S6 for use as a CP**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Precision of phonemes** | 4.06 | 5.18 | 3.88 | 6.24 |
| **Appropriateness of prosody** | 2.88 | 3.47 | 3.35 | 5.41 |
| **Naturalness of phonemes** | 3.53 | 4.19 | 2.94 | 6.00 |
| **Naturalness of prosody** | 3.06 | 3.65 | 3.00 | 5.65 |

Very generally, like for the other three roles, these aspects of also differed across the TTS synthesis systems (see Table 71). Analysis of the data using the Friedman test revealed that these differences were statistically significant (see Table 72).

**Table 72 Significance of differences among the precision of phonemes, appropriateness of prosody, naturalness of phonemes, and naturalness of prosody of S1, S2, S3, and S6 for use as a CP**

| | $\chi^2_r$ | df | *p* value | Significant |
|---|---|---|---|---|
| **Precision of phonemes** | 36.258 | 3 | < 0.0005 | Yes |
| **Appropriateness of prosody** | 28.714 | 3 | < 0.0005 | Yes |
| **Naturalness of phonemes** | 36.380 | 3 | < 0.0005 | Yes |
| **Naturalness of prosody** | 31.469 | 3 | < 0.0005 | Yes |

Regarding the specific hypothesis that the systems based on USS would be more accurate and natural at the prosodic level and hence more comprehensible than the systems based on concatenative synthesis, in line with this hypothesis, like for the other three roles, the descriptive statistics presented in Table 71, show that the appropriateness of the prosody and the naturalness of the prosody for S6 were higher than for S2 and S3 for this role. Also like for the role of phonetic PM, analysis of the data using the Wilcoxon test revealed that these differences were statistically significant (see Tables 73 and 74).

**Table 73 Significance of differences between the appropriateness of the prosody of S6 and S2 and S3 for use as a CP**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 3.349 | 1 | 0.0005 | One-tailed | Yes |
| **S3** | 3.593 | 1 | < 0.0005 | One-tailed | Yes |

**Table 74 Significance of differences between the naturalness of the prosody of S6 and S2 and S3 for use as a CP**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 3.260 | 2 | 0.0005 | One-tailed | Yes |
| **S3** | 3.441 | 2 | 0.0005 | One-tailed | Yes |

Also in line with this hypothesis, like for the other three roles, the appropriateness of prosody and naturalness of prosody for S1 were higher than for S3. Analysis of the data using the Wilcoxon test revealed that these differences were, however, not statistically significant (z = 1.327, N-ties = 4, $p$ = 0.093, one-tailed; z = 0.225, N-ties = 4, $p$ = 0.411, one-tailed, respectively). Also contrary to this hypothesis, like for the roles of phonetic and prosodic PM, the appropriateness of prosody and naturalness of prosody for S1 were lower than for S2.

Regarding the hypothesis that the systems based on USS would be more comprehensible than the systems based on concatenative synthesis, in line with this hypothesis, like for the other three roles, the mean rating of the comprehensibility of the speech generated by S6 for use as a CP was higher than that of S2 and S3. Analysis of the data using the Wilcoxon test revealed that these differences were statistically significant (see Table 75).

**Table 75 Significance of differences between the comprehensibility of S6 and S2 and S3 for use as a CP**

| Difference between S6 and ... | z value | N-ties | $p$ value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 3.095 | 3 | 0.001 | One-tailed | Yes |
| **S3** | 3.443 | 2 | 0.0005 | One-tailed | Yes |

Contrary to this hypothesis, however, the comprehensibility of the speech generated by S1 for use as a CP was lower than that of S2 and S3.

### 6.3.2.3 Are different TTS synthesis systems suitable for use in different roles in CALL applications?

In this section as said, the data collected are explored with respect to the hypothesis that different TTS synthesis systems will be suitable for use in different roles. There are two ways of interpreting this hypothesis. In the first interpretation, it predicts that, of all the TTS synthesis systems considered, different TTS synthesis systems will turn out to be the most suitable for use in the different roles, for example that, of all the TTS synthesis systems considered, SA is most suitable for use in role X, SB is most suitable for use in role Z, and SC is most suitable for use in role Y. In the second interpretation, it predicts that, different TTS synthesis systems will turn out to be more suitable for use in different roles, for example SA will be more suitable for use in role Z than in role X and SB will be more suitable for use in role Y than in role Z. The data were analysed with respect to both of these hypotheses.

In order to explore these hypotheses, the mean ratings of the adequacy and acceptability of S1, S2, S3, and S6 for use in each of the 4 roles were calculated across participants. The mean ratings for adequacy are presented in Table 76. Those for acceptability are presented in Table 77.

**Table 76 Mean ratings of the adequacy of S1, S2, S3, and S6 for use in the 4 roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **S1** | 4.53 | 4.12 | 4.06 | 4.18 |
| **S2** | 4.76 | 5.00 | 4.41 | 4.65 |
| **S3** | 3.76 | 3.59 | 3.94 | 4.05 |
| **S6** | 5.35 | 5.65 | 5.24 | 5.82 |

**Table 77 Mean ratings of the acceptability of S1, S2, S3, and S6 for use in the 4 roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **S1** | 4.88 | 4.29 | 4.24 | 4.41 |
| **S2** | 5.12 | 5.41 | 4.82 | 5.06 |
| **S3** | 4.18 | 3.82 | 4.35 | 4.29 |
| **S6** | 6.00 | 5.94 | 5.94 | 6.29 |

Contrary to the first hypothesis, the descriptive statistics presented in Table 76, show that for all roles S6 is the most adequate of the TTS synthesis systems, S2 is the second most adequate, S1 the third most adequate and S3 the least adequate. This trend can be seen more clearly in the following table in which the TTS synthesis systems have been ranked according the mean rating for adequacy that they received for each role in turn.

**Table 78 Ranking of TTS synthesis systems with respect to adequacy for use in the different roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **S1** | 3 | 3 | 3 | 3 |
| **S2** | 2 | 2 | 2 | 2 |
| **S3** | 4 | 4 | 4 | 4 |
| **S6** | 1 | 1 | 1 | 1 |

Also contrary to this first hypothesis, the descriptive statistics presented in Table 77, show that for all roles S6 is the most acceptable of the TTS synthesis system and S2 is the second most acceptable. In line with this hypothesis, on the other hand, while S1 is the third most acceptable of the TTS synthesis systems for use in the roles of RM, phonetic PM and CP, S3 is

the third most acceptable for use in the roles prosodic PM. This pattern can be seen more clearly in the following table in which the TTS synthesis systems have been ranked according to the mean rating for acceptability that the received for each role in turn.

**Table 79 Ranking of TTS synthesis systems with respect to acceptability for use in the different roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **S1** | 3 | 3 | 4 | 3 |
| **S2** | 2 | 2 | 2 | 2 |
| **S3** | 4 | 4 | 3 | 4 |
| **S6** | 1 | 1 | 1 | 1 |

Further regarding this first possibility, as presented in section 6.3, it was hypothesised that the systems based on USS, namely S1 and S6, synthesis would more suitable for use in all roles than the systems based on concatenative synthesis, namely S2 and S3. The descriptive statistics presented in Tables 76 and 77 show that as predicted S6 was rated more adequate and acceptable for use in all roles than S2 and S3. Analysis of the data using the Wilcoxon test revealed that these differences were statistically significant (see Table 80, Table 81, Table 82, Table 83, Table 84, Table 85, Table 86, and Table 87).

**Table 80 Significance of differences between the adequacy of S6 and S2 and S3 for use as an RM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 1.955 | 5 | 0.026 | One-tailed | Yes |
| **S3** | 3.537 | 1 | < 0.0005 | One-tailed | Yes |

**Table 81 Significance of differences between the adequacy of S6 and S2 and S3 for use as a phonetic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 2.178 | 6 | 0.015 | One-tailed | Yes |
| **S3** | 3.359 | 3 | 0.0005 | One-tailed | Yes |

**Table 82 Significance of differences between the adequacy of S6 and S2 and S3 for use as a prosodic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **S2** | 2.435 | 5 | 0.008 | One-tailed | Yes |
| **S3** | 2.898 | 5 | 0.002 | One-tailed | Yes |

**Table 83 Significance of differences between the adequacy of S6 and S2 and S3 for use as a CP**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 3.557 | 3 | < 0.0005 | One-tailed | Yes |
| S3 | 3.351 | 3 | 0.0005 | One-tailed | Yes |

**Table 84 Significance of differences between the acceptability of S6 and S2 and S3 for use as an RM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 2.517 | 4 | 0.006 | One-tailed | Yes |
| S3 | 3.050 | 2 | 0.001 | One-tailed | Yes |

**Table 85 Significance of differences between the acceptability of S6 and S2 and S3 for use as a phonetic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 1.941 | 7 | 0.026 | One-tailed | Yes |
| S3 | 3.316 | 3 | 0.0005 | One-tailed | Yes |

**Table 86 Significance of differences between the acceptability of S6 and S2 and S3 for use as a prosodic PM**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 2.709 | 6 | 0.004 | One-tailed | Yes |
| S3 | 3.108 | 5 | 0.001 | One-tailed | Yes |

**Table 87 Significance of differences between the acceptability of S6 and S2 and S3 for use as a CP**

| Difference between S6 and ... | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| S2 | 3.439 | 3 | 0.0005 | One-tailed | Yes |
| S3 | 3.436 | 2 | 0.0005 | One-tailed | Yes |

Also, as predicted S1 was rated more adequate for use in all roles than S3 and more acceptable for use in the roles of RM, phonetic PM and CP than S3. Analysis of the data using the Wilcoxon test revealed that the differences between the ratings of the adequacy of S1 and S3 for use as an RM and a phonetic PM were statistically significant, but that the differences between the ratings of the adequacy of S1 and S3 for use as a prosodic PM and a CP were not

(see Table 88). It also revealed that those between the ratings of the acceptability of S1 and S3 for use as an RM were statistically significant, but those between the ratings of the acceptability of S1 and S3 for use as a phonetic PM and a CP were not (see Table 89).

**Table 88 Significance of differences between the adequacy of S1 and S3 for use in the four different roles**

| | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 2.266 | 9 | 0.012 | One-tailed | Yes |
| **Phonetic PM** | 1.793 | 6 | 0.037 | One-tailed | Yes |
| **Prosodic PM** | 0.454 | 6 | 0.325 | One-tailed | No |
| **CP** | 0.431 | 10 | 0.333 | One-tailed | No |

**Table 89 Significance of differences between the acceptability of S1 and S3 for use in the roles of RM, phonetic PM and CP**

| | z value | N-ties | *p* value | Hypothesis | Significant |
|---|---|---|---|---|---|
| **RM** | 2.145 | 7 | 0.016 | One-tailed | Yes |
| **Phonetic PM** | 1.628 | 7 | 0.052 | One-tailed | No |
| **CP** | 0.702 | 10 | 0.241 | One-tailed | No |

Also contrary to this hypothesis, S1 was rated less acceptable for use in the role of prosodic PM than S3. Moreover, S1 was also rated less adequate and acceptable for use in all roles than S2.

It was not considered appropriate to analyse the data with respect to the second interpretation because significant differences were not found among the ratings of adequacy and acceptability across the different roles in most cases (see section 6.3.2.1).

### 6.3.2.4 Is TTS synthesis ready for use in CALL applications?

As said, the ratings of the adequacy and acceptability of the TTS synthesis systems for use in the different roles also give an insight into the readiness of the systems for use in the different roles. In this section, we look again at the data collected for adequacy and acceptability, which are replicated in Tables 90 and 91, with respect to this question.

**Table 90 Mean ratings of the adequacy of S1, S2, S3, and S6 for use in the 4 roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **S1** | 4.53 | 4.12 | 4.06 | 4.18 |
| **S2** | 4.76 | 5.00 | 4.41 | 4.65 |
| **S3** | 3.76 | 3.59 | 3.94 | 4.05 |
| **S6** | 5.35 | 5.65 | 5.24 | 5.82 |

**Table 91 Mean ratings of the acceptability of S1, S2, S3, and S6 for use in the 4 roles**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **S1** | 4.88 | 4.29 | 4.24 | 4.41 |
| **S2** | 5.12 | 5.41 | 4.82 | 5.06 |
| **S3** | 4.18 | 3.82 | 4.35 | 4.29 |
| **S6** | 6.00 | 5.94 | 5.94 | 6.29 |

The descriptive statistics presented in Table 90 show that none of the TTS synthesis systems achieved top ratings for adequacy for use in any of the roles. The ratings of the adequacy of S6 for use in all four roles, in particular the rating of the adequacy of S6 for use as a CP, are, however, not far off. Similarly S1, S2, and S3 do not achieve top ratings for acceptability for use (see Table 91). S6 on the other hand does achieve top ratings for two of the roles, namely for RM and CP, suggesting that the quality of speech that it generated is acceptable for use in those roles. The ratings of the acceptability of the speech generated by S6 for use in the roles of phonetic and prosodic PM are not much lower than those of the speech generated by S6 for use in the roles of RM and CP.

#### 6.3.2.5 What aspects of the quality of the speech generated by TTS synthesis systems require improvement for TTS synthesis to be ready for use in CALL?

As said, the ratings of the quality of the speech generated by the different TTS synthesis systems with respect to their use in the different roles that TTS synthesis systems may assume within CALL applications also give an insight into what aspects of the quality of the speech generated by TTS synthesis systems require improvement in order to render TTS synthesis ready for use in CALL. In this section, we therefore look again at the mean ratings of the quality of the speech generated by the TTS synthesis systems with respect to their use in the different roles that TTS synthesis may assume within CALL applications. The mean ratings of the different aspects of the quality of the speech generated by S1 with respect to its use in each

of the four different roles that TTS synthesis systems may assume within CALL applications are presented in Table 92.

**Table 92 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Comprehensibility** | 4.53 | 4.24 | 4.82 | 4.47 |
| **Intelligibility** | 4.53 | 3.88 | 4.65 | 4.24 |
| **Choice of pronunciation** | 4.59 | 4.12 | 4.53 | 4.29 |
| **Accuracy** | 3.76 | 3.56 | 3.53 | 3.47 |
| **Naturalness** | 3.82 | 3.68 | 3.15 | 3.29 |
| **Naturalness of voice** | 4.00 | 3.53 | 3.59 | 3.53 |
| **Expressiveness** | 3.24 | 3.12 | 2.35 | 2.65 |
| **Appropriateness of register** | 5.00 | 4.76 | 4.53 | 4.53 |

As presented in section 6.3.2.1.1, the descriptive statistics presented in Table 92 show that none of the aspects of the quality of the speech generated by S1 received top ratings for any of the roles. This appears to suggest that none of the aspects of the quality of the speech generated by S1 considered meet the requirements of any of the roles that TTS synthesis systems may assume within CALL applications and hence that all aspects of these require improvement. More specifically, of all the aspects of the quality of the speech generated by S1 considered, for the role of RM, expressiveness received the lowest mean rating. This would appear to suggest that S1 is furthest from meeting the requirements placed on expressiveness by the role of RM. This would also appear to be the case for the roles of phonetic PM, prosodic PM and CP: for these roles, of all the aspects of the quality of the speech generated by S1 considered, expressiveness also received the lowest mean rating. The mean ratings of the accuracy and naturalness of the speech generated by S1 with respect to its use as an RM were also very low, under 4, the mid-point of our scales, suggesting that S1 is also far from meeting the requirements placed on them in the role of RM. This was also the case for the role of phonetic PM, but in addition, the mean ratings of the intelligibility and naturalness of voice of the speech generated by S1 with respect to its use in this role were also very low suggesting that S1 is far from meeting the requirements placed on accuracy, naturalness, intelligibility, and naturalness of voice in addition to expressiveness in the role of phonetic PM. Regarding the role of prosodic PM, in addition to expressiveness, the mean ratings of the accuracy, naturalness, and naturalness of voice of the speech generated by S1 with respect to its use in

this role were very low suggesting that S1 is far from meeting the requirements placed on accuracy, naturalness and naturalness of voice in addition to expressiveness by the role of prosodic PM. This was also the case for the role of CP.

The mean ratings of the different aspects of the quality of the speech generated by S2 with respect to its use in each of the four different roles that TTS synthesis systems may assume within CALL applications are presented in Table 93.

**Table 93 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S2**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Comprehensibility** | 5.06 | 5.41 | 5.41 | 5.59 |
| **Intelligibility** | 4.89 | 5.24 | 5.35 | 5.41 |
| **Choice of pronunciation** | 5.00 | 5.18 | 5.06 | 5.29 |
| **Accuracy** | 3.91 | 4.91 | 3.97 | 4.38 |
| **Naturalness** | 3.63 | 4.53 | 3.69 | 3.84 |
| **Naturalness of voice** | 3.94 | 5.00 | 4.18 | 4.41 |
| **Expressiveness** | 3.06 | 4.53 | 2.41 | 2.82 |
| **Appropriateness of register** | 4.59 | 5.12 | 4.76 | 4.88 |

As for S1, the descriptive statistics presented in Table 93 show that none of the aspects of the quality of the speech generated by S2 received top ratings for any of the roles. This appears to suggest that none of the aspects of the quality of the speech generated by S2 considered meet the requirements of any of the roles that TTS synthesis systems may assume within CALL applications either and hence that all aspects of the quality of the speech generated by S2 also require improvement. More specifically, as for S1, of all the aspects of the quality of the speech generated by S2 considered, for all four roles, expressiveness received the lowest mean rating. This would appear to suggest that the requirements placed on expressiveness are the furthest from being met for all four roles. Also as for S1, the mean ratings of the accuracy and naturalness of the speech generated by S2 with respect to its use as an RM were also very low suggesting that S2 is far from meeting the requirements placed on accuracy and naturalness in addition to those placed on expressiveness in the role of RM. So too were the mean ratings of the naturalness of voice of the speech generated by S2 with respect to its use as an RM suggesting that S2 is also far from meeting the requirements that this role places on naturalness of voice. Regarding the role of phonetic PM, none of the aspects of the quality of the speech generated by S2 including expressiveness received very low ratings suggesting that

S2 is not as far from meeting the requirements that this role places on the quality of the speech generated by TTS synthesis systems. As for the role of RM for this TTS synthesis system and for all of the roles for S1, in addition to expressiveness, accuracy and naturalness received very low ratings for the role of prosodic PM suggesting that the speech generated by S2 is far from meeting the demands placed on accuracy and naturalness in addition to those placed on expressiveness. Regarding the role of CP, the mean ratings of the naturalness of the speech generated by S2 in addition to those of the expressiveness of the speech generated by S2 received very low ratings suggesting that S2 is far from meeting the demands placed on naturalness in addition to those placed on expressiveness.

The mean ratings of the different aspects of the quality of the speech generated by S3 with respect to its use in each of the four different roles that TTS synthesis systems may assume within CALL applications are presented in Table 94.

**Table 94 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S3**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Comprehensibility** | 4.12 | 3.71 | 4.77 | 4.59 |
| **Intelligibility** | 3.88 | 3.18 | 4.59 | 4.35 |
| **Choice of pronunciation** | 4.00 | 3.71 | 4.41 | 4.35 |
| **Accuracy** | 3.32 | 2.85 | 3.79 | 3.62 |
| **Naturalness** | 2.53 | 2.38 | 3.09 | 2.97 |
| **Naturalness of voice** | 2.24 | 1.94 | 2.82 | 2.71 |
| **Expressiveness** | 2.18 | 2.24 | 2.88 | 2.94 |
| **Appropriateness of register** | 4.12 | 3.76 | 4.24 | 4.18 |

As for S1 and S2, the descriptive statistics presented in Table 94 show that none of the aspects of the quality of the speech generated by S3 received top ratings for any of the roles. This appears to suggest that none of the aspects of the quality of the speech generated by S3 considered meet the requirements of any of the roles that TTS synthesis systems may assume within CALL applications either and hence that all aspects of the quality of the speech generated by S3 also require improvement. More specifically, as for S1 and S2, of all the aspects of the quality of the speech generated by S3 considered, for the role of RM, expressiveness received the lowest mean rating. This would appear to suggest that, as for S1 and S2, the requirements placed on expressiveness are the furthest from being met for this role for this TTS synthesis system. Naturalness of voice, of all of the aspects of the quality of the

speech generated by S3, received the lowest mean rating for the other three roles. This would appear to suggest that the requirements placed on naturalness of voice are furthest from being met for the roles of phonetic PM, prosodic PM, and CP for this TTS synthesis system. Returning to the role of RM, the mean ratings of the intelligibility, accuracy, naturalness and naturalness of voice, in addition to those of the expressiveness, of the speech generated by S3 with respect to use in this role received very low ratings. This would appear to suggest that S3 is far from meeting the demands placed on intelligibility, accuracy, naturalness and naturalness of voice in addition to those placed on expressiveness by this role. Regarding the role of phonetic PM, all aspects of the speech generated by S3 received very low ratings with respect to use in this role which would appear to suggest that S3 is far from meeting the demands placed on all aspects of the quality of the speech generated by the systems considered. Finally, regarding the role of CP, the mean ratings of accuracy, naturalness, and expressiveness, in addition to those of naturalness of voice, of the speech generated by S3 received very low ratings. This would appear to suggest that S3 is far from meeting the demands placed on accuracy, naturalness and expressiveness in addition to those placed on naturalness of voice by this role.

The mean ratings of the different aspects of the quality of the speech generated by S6 with respect to its use in each of the four different roles that TTS synthesis systems may assume within CALL applications are presented in Table 95.

**Table 95 Mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S6**

| | RM | Phonetic PM | Prosodic PM | CP |
|---|---|---|---|---|
| **Comprehensibility** | 5.65 | 5.88 | 5.94 | 6.47 |
| **Intelligibility** | 5.41 | 6.12 | 5.82 | 6.29 |
| **Choice of pronunciation** | 5.71 | 5.76 | 5.71 | 6.47 |
| **Accuracy** | 5.38 | 5.71 | 5.29 | 5.82 |
| **Naturalness** | 5.38 | 5.60 | 5.38 | 5.78 |
| **Naturalness of voice** | 5.53 | 5.69 | 5.53 | 6.06 |
| **Expressiveness** | 4.94 | 5.24 | 4.88 | 5.18 |
| **Appropriateness of register** | 5.47 | 5.76 | 5.41 | 5.65 |

Unlike for the other TTS synthesis systems, none of the aspects of the quality of the speech generated by the TTS synthesis systems considered including expressiveness received very low ratings. In fact most aspects of the quality of the speech generated by the TTS synthesis

systems considered received close on top ratings. This would appear to suggest that S6 is not far from meeting the requirements placed on all of the aspects of the quality of the speech generated by TTS synthesis considered including those placed on expressiveness for any of the roles.

### 6.3.3 Interpretation

In this section, we assess whether the results of this investigation as a whole over the four TTS synthesis systems and the four different roles that TTS synthesis systems considered support our hypotheses and also look at other possible explanations for the results that we have obtained.

Specifically, in section 6.3.3.1, the results are assessed with respect to the hypotheses that:

- CALL setups in general place demands on the following aspects of the quality of TTS synthesis systems: *comprehensibility*, *intelligibility*, *accuracy*, *naturalness*, *smoothness*, *expressiveness* and *register*; and,
- the different roles that TTS synthesis may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems.

In section 6.3.3.2, the results are assessed with respect to the hypothesis that the speech generated by different TTS synthesis systems differs in quality. In section 6.3.3.3, the results are assessed with respect to the hypothesis that different TTS synthesis systems will be suitable for use in different roles in CALL applications.

As mentioned in section 6.3.2, the results of this investigation also provide insights into the readiness of TTS synthesis for use in CALL and aspects of the quality of the speech generated by the TTS synthesis systems which require improvement in order to render TTS synthesis ready for use in CALL. In section 6.3.3.4, the results are assessed with respect to the question of whether TTS synthesis is ready for use in CALL applications. And, in section 6.3.3.5, the results are assessed with respect to the question of which aspects of the quality of the speech generated by TTS synthesis systems require improvement in order to render TTS synthesis ready for use in CALL.

### 6.3.3.1 On what aspects of the quality of the speech generated by TTS synthesis systems do CALL applications place demands?

Regarding the question of what demands CALL applications place on the quality of the speech generated by TTS synthesis systems, with few exceptions (intelligibility for the role of phonetic PM and comprehensibility, intelligibility, choice of pronunciation, and naturalness of voice for the role of CP for S6) none of the aspects of the quality of the speech considered received top ratings, i.e. 6 or 7, for any of the TTS synthesis systems. On the whole, the results therefore appear to suggest that all the roles that TTS synthesis systems may assume within CALL applications place demands on all aspects of the speech generated by TTS synthesis systems considered, not just comprehensibility, accuracy, naturalness, expressiveness and register as the literature review and the preliminary exploratory investigation suggest. A limitation of the method employed in this investigation is that it is subjective; it is based on what teachers *think* the requirements of the different roles that TTS synthesis may assume within CALL applications are. *Actual* requirements may be different. Further investigation is therefore required to validate this finding and the findings of this investigation with respect to the demands that CALL applications place on the quality of the speech generated by TTS synthesis systems in general.

Regarding the hypothesis that the different roles that TTS synthesis may assume within CALL applications impose different demands on the quality of speech generated, consistent with this hypothesis, statistically significant differences were found among the ratings of some of the aspects of the quality of the speech generated by each of the TTS synthesis systems evaluated in this investigation across the different roles. Also consistent with this hypothesis, statistically significant differences were found among the ratings of the adequacy of S2 and S6 across the different roles. Contrary to this hypothesis however, statistically significant differences were not found among either the ratings of the adequacy of S1 and S3 across the different roles or the ratings of the acceptability of S1, S2, S3, and S6 across the different roles. A possible explanation for this is that the aspects of the quality of the speech generated by the four TTS synthesis systems for which statistically significant differences were found across the different roles only play a minor role in determining the adequacy and acceptability of the speech generated by TTS synthesis systems for use in the different roles. The investigation presented here does not permit us to establish how large a role each of the different aspects of the quality of the speech generated by the TTS synthesis systems play in determining the adequacy and acceptability of the speech for use in CALL applications. Another set of ratings that it was

believed might provide insights into whether the different roles that TTS synthesis systems may assume within CALL applications place different demands on the quality of the speech generated were the ratings of the readiness of TTS synthesis in general for use in the different roles. In line with the hypothesis, statistically significant differences were found among the ratings of the readiness of TTS synthesis in general across the four different roles. While this result appears to support our hypothesis, it may seem inconsistent with the findings for adequacy and acceptability. A possible explanation consistent with our findings for adequacy and acceptability is that there are differences in adequacy and acceptability across the roles for each of the TTS synthesis systems. These differences are, however, small and only their cumulative effect on participants' impressions of the readiness of TTS synthesis for use in CALL is detectable given the limited power of the statistical test that it was necessary to employ. Another possible explanation for the inconsistencies in the results is that the participants were not able to reliably discriminate between the different roles and the requirements that they place on the quality of the speech generated by TTS synthesis systems. Yet another possible explanation for the findings is that the roles overlap – although not suggested in the literature on the use of TTS synthesis in CALL, learners might use the speech provided in talking dictionaries, talking texts and by conversational partners as pronunciation models to imitate for example. Whichever explanation is correct, the implications for the evaluation of the quality of the speech generated by TTS synthesis systems for use in CALL applications and for the use of TTS synthesis systems in CALL applications are the same (see section 6.4.1).

Regarding the more specific hypothesis that use of TTS synthesis as a phonetic PM would place greatest demands on accuracy and naturalness at the phonetic level, precision of phonemes was statistically lower for the role of phonetic PM than for: the role of prosodic PM for S1, the roles of RM, prosodic PM and CP for S3, and the role of CP for S6. Precision of phonemes was also lower for the role of phonetic PM than for the roles of RM and CP for S1. The differences were, however, not statistically significant. On the other hand, precision of phonemes was higher for the role of phonetic PM than for: the roles of RM, prosodic PM and CP for S2, and for the roles of RM and prosodic PM for S6. Regarding naturalness of phonemes, it was statistically lower for the role of phonetic PM than for the roles of prosodic PM and CP for S3. It was also lower for the role of phonetic PM that for: the roles of RM and CP for S1, and the role of RM for S3. The differences in these cases were, however, not

statistically significant. On the other hand, naturalness of phonemes was higher for the role of phonetic PM that for: the role of prosodic PM for S1, the roles of RM, prosodic PM and CP for S2, and the roles of RM and prosodic PM for S6. In summary, the results are mixed for both accuracy of phonemes and naturalness of phonemes. There are several possible explanations for such mixed results: there are differences but the differences are small and hence not detectable due to the power of the statistical test used; there are differences but the participants cannot reliably discriminate between the different roles and the demands that they place on the quality of the speech generated by TTS synthesis systems; and, the different roles that TTS synthesis systems may assume within CALL applications overlap. Whichever explanation is correct, the implications for the evaluation of the quality of the speech generated by TTS synthesis systems for use in CALL applications and for the use of TTS synthesis systems in CALL applications are the same (see section 6.4.1).

Finally, regarding the hypothesis that use of TTS synthesis as a prosodic PM would place greatest demands on accuracy and naturalness at the prosodic level, appropriateness of prosody was statistically lower for the role of prosodic PM than: for the roles of RM and phonetic PM for S1, and for the roles of phonetic PM and CP for S2. It was also lower for the role of prosodic PM than for: the role of CP for S1, the role of RM for S2, and the roles of phonetic PM and CP for S6. The differences in these cases were, however, not statistically significant. On the other hand, appropriateness of prosody was higher for the role of prosodic PM than for: the roles of RM, phonetic PM and CP for S3, and the role of RM for S6. Regarding naturalness of prosody, it was statistically lower for the role of prosodic PM than: for the roles of RM and phonetic PM for S1, for the roles of phonetic PM and CP for S2, and for the roles of phonetic PM and CP for S6. It was also lower for the role of prosodic PM than: for the role of CP for S1, and for the role of RM for S3. The differences in these cases were, however, not statistically significant. On the other hand, naturalness of prosody was higher for the role of prosodic PM than: for the role of RM for S2, and for the role of RM for S6. In summary, the results are mixed for both appropriateness of prosody and naturalness of prosody. The possible explanations for these results are the same as those for the results of the analysis of the data with respect to the hypothesis that use of TTS synthesis as a phonetic PM would place greater demands on accuracy and naturalness at the phonetic level than use of TTS synthesis as an RM, prosodic PM, or CP.

### 6.3.3.2 Does the speech generated by different TTS synthesis systems differ in quality?

Regarding the general hypothesis that the speech generated by different TTS synthesis systems differs in quality, consistent with this hypothesis, statistically significant differences were found among the ratings of all aspects of speech considered across all four TTS synthesis systems for each role. This, we believe, is the most likely explanation for the findings.

Regarding the more specific hypothesis that the speech generated by the systems based on USS, namely S1 and S6, would be more accurate and natural at the prosodic level and hence more comprehensible than that generated by the systems based on concatenative synthesis, namely S2 and S3, in line with this hypothesis appropriateness of prosody, naturalness of prosody and comprehensibility for S6 were statistically higher than for S2 and S3 for the roles of RM, prosodic PM and CP. Also in line with this hypothesis appropriateness of prosody and naturalness of prosody for S6 were statistically higher than for S2 and S3 for the role of phonetic PM – comprehensibility, while higher for S6 than for S2 and S3 for this role was not statistically higher. Regarding the other system based on USS, namely S1, also in line with the hypothesis under examination, appropriateness of prosody and naturalness of prosody were statistically higher for S1 than for S3 for the role of RM – comprehensibility while higher for S1 than for S3 was not significantly higher, appropriateness of prosody and comprehensibility was statistically higher for S1 than for S3 for the role of phonetic PM. Regarding the roles of prosodic PM and CP, appropriateness of prosody, naturalness of prosody and comprehensibility while higher for S1 than for S3 were not significantly higher. Similarly, appropriateness of prosody and naturalness of prosody while higher for S1 than for the other system based on concatenative synthesis, namely S2, were not statistically significant for the role of RM. Contrary to this hypothesis, however, comprehensibility was lower for S1 than for S2 for the role of RM and both appropriateness of prosody, naturalness of prosody and comprehensibility were lower for S1 than for S2 for the roles of phonetic PM, prosodic PM, and CP. In summary, overall the data for S6 appear to support the hypothesis, but the data for S1 do not. There are two possible explanations for this finding. The first possible explanation is that the techniques employed in TTP conversion, in particular in prosody generation, are poorer for S1 than they are for S2 and S3. The second possible explanation is that the database of speech segments upon which S1 is based is of poorer quality than those upon which S2 and S3 are based. We therefore maintain that, all other aspects of the systems being equal, ratings of the appropriateness of prosody, naturalness of prosody and comprehensibility of systems

based on USS will be higher than those of systems based on concatenative synthesis. For a number of TTS synthesis systems, comparison of two different versions of the systems one employing concatenative synthesis and one employing USS based on the same database of speech segments is required to validate this hypothesis.

#### 6.3.3.3 Are different TTS synthesis systems suitable for use in different roles in CALL applications?

As presented in section 6.3.2.3, there are two ways of interpreting the hypothesis that different TTS synthesis systems will be suitable for use in different roles. In the first interpretation, it predicts that, of all the TTS synthesis systems considered, different TTS synthesis systems will turn out to be the most suitable for use in the different roles, for example that, of all the TTS synthesis systems considered, SA is most suitable for use in role X, SB is most suitable for use in role Z, and SC is most suitable for use in role Y. In the second interpretation, it predicts that, different TTS synthesis systems will turn out to be more suitable for use in different roles, for example SA will be more suitable for use in role Z than in role X and SB will be more suitable for use in role Y than in role Z. The data were analysed with respect to both of these hypotheses.

Contrary to the first interpretation of the hypothesis, for all roles S6 was found to be the most adequate of the TTS synthesis systems, S2 is the second most adequate, S1 the third most adequate and S3 the least adequate. In addition, for all roles S6 was found to be the most acceptable and S2 was found to be the second most acceptable, and for three out of four of the roles S1 was found to be the third most acceptable and S3 was found to be the least acceptable. On the whole, the results therefore appear to suggest that the same TTS synthesis system is most suitable for use in the different roles. This is consistent with the finding that there are only small differences in the requirements across the roles and the fact that for most aspects of speech considered for most of the roles S6 received the highest ratings, S1 the second highest rating, S2 the third highest rating and S3 the lowest rating (see Tables 96, 97, 98 and 99 in which the mean ratings of the different aspects of speech considered are ranked across the TTS synthesis systems from lowest (1) to highest (4)).

**Table 96 Ranking of mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as an RM across the TTS synthesis systems**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| Comprehensibility | 2 | 3 | 1 | 4 |
| Intelligibility | 2 | 3 | 1 | 4 |
| Choice of pronunciation | 2 | 3 | 1 | 4 |
| Accuracy | 2 | 3 | 1 | 4 |
| Naturalness | 3 | 2 | 1 | 4 |
| Naturalness of voice | 3 | 2 | 1 | 4 |
| Expressiveness | 3 | 2 | 1 | 4 |
| Appropriateness of register | 2 | 3 | 1 | 4 |

**Table 97 Ranking of mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a phonetic PM across the TTS synthesis systems**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| Comprehensibility | 2 | 3 | 1 | 4 |
| Intelligibility | 2 | 3 | 1 | 4 |
| Choice of pronunciation | 2 | 3 | 1 | 4 |
| Accuracy | 2 | 3 | 1 | 4 |
| Naturalness | 2 | 3 | 1 | 4 |
| Naturalness of voice | 2 | 3 | 1 | 4 |
| Expressiveness | 2 | 3 | 1 | 4 |
| Appropriateness of register | 2 | 3 | 1 | 4 |

**Table 98 Ranking of mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a prosodic PM across the TTS synthesis systems**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| Comprehensibility | 2 | 3 | 1 | 4 |
| Intelligibility | 2 | 3 | 1 | 4 |
| Choice of pronunciation | 2 | 3 | 1 | 4 |
| Accuracy | 1 | 3 | 2 | 4 |
| Naturalness | 2 | 3 | 1 | 4 |
| Naturalness of voice | 2 | 3 | 1 | 4 |
| Expressiveness | 1 | 2 | 3 | 4 |
| Appropriateness of register | 2 | 3 | 1 | 4 |

**Table 99 Ranking of mean ratings of the comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness, and appropriateness of register of S1, S2, S3, and S6 for use as a CP across the TTS synthesis systems**

| | S1 | S2 | S3 | S6 |
|---|---|---|---|---|
| **Comprehensibility** | 2 | 3 | 1 | 4 |
| **Intelligibility** | 1 | 3 | 2 | 4 |
| **Choice of pronunciation** | 1 | 2 | 2 | 4 |
| **Accuracy** | 1 | 1 | 2 | 4 |
| **Naturalness** | 2 | 3 | 1 | 4 |
| **Naturalness of voice** | 3 | 1 | 2 | 4 |
| **Expressiveness** | 1 | 2 | 3 | 4 |
| **Appropriateness of register** | 3 | 1 | 2 | 4 |

Further regarding this first possibility, as presented in section 6.3, it was hypothesised that the systems based on USS, namely S1 and S6, synthesis would be more suitable for use in all the four roles than the systems based on concatenative synthesis, namely S2 and S3. In line with this hypothesis, the ratings of the adequacy and the acceptability of the speech generated by S6 were found to be statistically higher than those or the speech generated by S2 and S3 for all roles. Also in line with this hypothesis, the ratings of the adequacy and acceptability of the speech generated by S1 were found to be statistically higher than those of the speech generated by S3 for use as an RM. In addition, the ratings of the adequacy of the speech generated by S1 for use as a phonetic PM were also found to be statistically higher than those of the speech generated by S3. The ratings of the adequacy of the speech generated by S1 for the roles of prosodic PM and CP were higher than those of the speech generated by S3, but the differences were not statistically significant. Similarly, the ratings of the acceptability of the speech generated by S1 for use as a phonetic PM and a CP were higher than those of the speech generated by S3, but the differences were not statistically significant. Also contrary to this hypothesis, the ratings of the acceptability of the speech generated by S1 for use as a prosodic PM were lower than those of the speech generated by S3. Moreover, the ratings of the ratings of the adequacy and acceptability of the speech generated by S1 for use in all four roles were lower than those of the speech generated by S2. These findings are consistent with the fact that overall across the four roles S1 did not achieve statistically higher ratings in appropriateness of prosody, naturalness of prosody and comprehensibility than S2 and S3.

The implications of these findings with respect to the use of TTS synthesis in CAL are presented in section 6.4.2.

Regarding the second interpretation of the main hypothesis considered in this section, it was not considered appropriate to analyse the data with respect to this interpretation because significant differences were not found among the ratings of adequacy and acceptability across the different roles in most cases (see section 6.3.2.1).

#### 6.3.3.4 Is TTS synthesis ready for use in CALL applications?

Regarding the question of whether TTS synthesis is ready for use in CALL applications, none of the TTS synthesis systems achieved top ratings for adequacy for use in any of the roles. This would appear to suggest that none of the systems investigated are adequate for use in CALL applications. As presented in section 6.3.1.3.2, the TTS synthesis systems that were evaluated were selected so as to cover the range of different types of TTS synthesis systems currently available on the market. The results therefore would appear to suggest that French TTS synthesis systems in general are not adequate for use in CALL applications – we do not believe that it is possible to generalise to other languages because TTS synthesis for different languages poses different problems (see section 2.4) and more is known about some languages than others. The ratings of the adequacy of S6 for use in all four roles, in particular the rating of the adequacy of S6 for use as a CP, are, however, not far off. This would appear to suggest that S6 and hence some but not the majority of French TTS synthesis systems on the market are nearly adequate for use in CALL applications. Similarly S1, S2, and S3 do not achieve top ratings for acceptability for use. This would appear to suggest that these systems are not acceptable for use in CALL applications either. S6 on the other hand does achieve top ratings for two of the roles, namely for RM and CP, and the ratings of the acceptability of the speech generated by S6 for use in the roles of phonetic and prosodic PM are not much lower than those of the speech generated by S6 for use in the roles of RM and CP. This would appear to suggest that the speech generated by S6 is nearly acceptable for use in all four roles. Regarding French TTS synthesis systems in general these results would appear to suggest that some, but not the majority, are acceptable for use in CALL in all roles. The implications of these results with respect to the use of state-of-the-art TTS synthesis systems in CALL applications are presented in section 6.4.2.

#### 6.3.3.5 What aspects of the quality of the speech generated by TTS synthesis systems require improvement for TTS synthesis to be ready for use in CALL?

Regarding the question of what aspects of the quality of the speech generated by TTS synthesis systems require improvement for TTS synthesis systems to be ready for use in the different roles that TTS synthesis systems may assume within CALL applications, the mean ratings of the expressiveness, naturalness of voice, naturalness, and accuracy of the speech generated by both S1 and S3 with respect to use in all four roles that TTS synthesis systems may assume within CALL applications were very low, 4 or lower, as were the mean ratings of the expressiveness, naturalness, and accuracy of the speech generated by S2 with respect to use in three out of four of the roles that TTS synthesis systems may assume within CALL applications, namely the roles of RM, prosodic PM and CP. On the whole, the results would therefore appear to suggest that the following aspects of the quality of the speech generated by most French TTS synthesis systems need to be improved in order to be ready for use in CALL applications: expressiveness, naturalness of voice, naturalness and accuracy. This is consistent with what is known about the quality of the speech generated by TTS synthesis systems based on concatenative synthesis and USS in general: as presented in section 2.5.2.4 such systems do not provide control over voice quality (Edgington, 1997) and hence expressiveness (Edgington, 1997; Bailly *et al.*, 2003b); inadequacies of the approaches mean that the speech generated does not sound entirely natural, namely distortions at concatenation points (Huang *et al.*, 2001), the inability to model changes at the segmental level which accompany changes at the suprasegmental level (Campbell and Black, 1997), and the fact that speech segments are typically extracted from corpora of prosodically neutral speech (*ibid.*); and regarding accuracy, as presented in section 2.5.1.4, methods for determining the prosodic specification of utterances are inadequate (Dutoit, 1997; Rodman, 1999; Henton, 2002).

### 6.3.4 Limitations

In section 6.3.3, a number of limitations of the investigation have already been highlighted. Specifically, regarding the internal validity of the investigation (see section 4.6.1.1.1), it was noted that:

- the investigation is subjective based on what teachers and CALL researchers *think* the requirements of CALL applications are and *actual* requirements may be different;
- there may be differences in the requirements impose on the quality of the speech generated by TTS synthesis systems across the different roles that they may assume

within CALL applications, but the power of the statistical techniques used to analyse the results does not permit their detection; and,

- there may be differences in the requirements impose on the quality of the speech generated by TTS synthesis systems across the different roles that they may assume within CALL applications, but the participants may not be able to discriminate between the different roles and their requirements.

Regarding the external validity of the investigation, a further limitation of the investigation is that the sample was a small convenience sample. Consequently it does not truly reflect the population of French teachers and CALL researchers. The most significant bias is that none of the participants were secondary or primary school teachers even though attempts were made to recruit such participants. As regards the other features of the sample, without statistics on the constitution of the population it is impossible to make comment. Further investigation with a more representative sample is therefore recommended.

Yet another limitation of the investigation with respect to its external validity is that it only considers French TTS synthesis systems. As presented in section 2.4, TTS synthesis for different languages presents different problems. Moreover, more is known about some languages, in particular English, than others, in particular minority languages. It is therefore questionable whether the results of this investigation, in particular those concerning the readiness of TTS synthesis for use in CALL applications, would generalise to TTS synthesis for other languages. It is therefore recommended that similar investigations are conducted for other languages.

Regarding the practicality of the investigation, the investigation was too long. As presented in section 6.3.1.2, a number of participants dropped out or only partially completed the investigation because they did not have time to complete it. For a number of reasons it has been suggested that further investigation is required. In order to be able to recruit enough participants for these investigations, it is therefore necessary to look for ways in which the method of investigation can be made more efficient.

## 6.4 Recommendations

In section 6.4.1, assuming that the results obtained in the investigations presented here are corroborated in further investigations we make recommendations regarding how the quality of

the speech generated by TTS synthesis systems ought to be evaluated for use in CALL applications. Then, parting from the same assumption, recommendations on how state-of-the-art TTS synthesis systems ought to be used in CALL are made in section 6.4.2.

### 6.4.1 Evaluation of TTS synthesis for CALL purposes

In section 4.9, given the number of levels of evaluation that ought to be conducted (see section 4.5.3), in order to overcome the cost of evaluation, it was proposed that what CALL researchers need are benchmark tests. Regarding the development of a benchmark test for the evaluation of the adequacy of the quality of the speech generated by TTS synthesis for use in CALL applications to which it was intended that the results of the investigations presented in this thesis would contribute, the results of the investigations allow us to make a few recommendations, namely that such benchmark tests should not discriminate between the different roles that TTS synthesis may assume within CALL applications, i.e. that TTS synthesis systems should be evaluated once for use in CALL applications in general, and that such benchmark tests should address the full range of aspects of the quality of the speech generated by TTS synthesis systems evaluated in the investigations presented here. Regarding the first point, namely that benchmark tests should not discriminate between the different roles, it is believed that: if there are differences among the roles, but the differences are only small, there is little gain in making the difference; if teachers and CALL researchers cannot discriminate between the different roles and their requirements, then it is not possible to make the difference; and, if the roles overlap, then the difference should not be made. We are, however, still a long way off developing benchmark tests for the evaluation of the adequacy of the quality of the speech generated by TTS synthesis systems for use in CALL applications. Through these investigations it was not possible to establish the contribution of these different aspects to the adequacy and acceptability of TTS synthesis systems for use in CALL applications. Moreover, once the aspects of the quality of the speech generated by TTS synthesis systems upon which CALL applications place demands are established and a test is developed it will be necessary to establish acceptance levels for performance in that test. Until such tests are available and results in those test conducted by an independent body are published for all TTS synthesis systems on the market, it will be necessary for CALL researchers to devise and conduct their own adequacy evaluations. The same recommendations apply: they should not attempt to discriminate between the different roles that TTS synthesis systems may assume within CALL applications, and if they wish to provide a collaborator who is developing a TTS synthesis system with information on how to

render their system more suitable for use in CALL they should address all aspects of the quality of the speech generated by TTS synthesis systems considered in the investigations presented here. That there is no need to discriminate between the different roles that TTS synthesis may assume within CALL applications is particularly advantageous in the absence of published results in benchmark tests because it will significantly reduce the length of time that an evaluation takes to conduct. A significant limitation of such evaluations, like the investigations presented in this thesis (see section 6.3.4), is that they are subjective, the results indicate whether teachers and CALL researchers *think* that TTS synthesis systems have met the demands that CALL applications place on the quality of the speech generated. *Actual* requirements may be different to what teachers and CALL researchers *think* they are. It is therefore recommended that CALL researchers build say one example of each type of exercise that they wish to provide and submit those exercises to the remaining levels of evaluation of the proposed infrastructure for the evaluation of CALL applications integrating TTS synthesis before building the full-scale application. It is intended that the findings of these and further investigations of the requirements involving teachers and CALL researchers as participants are validated by actually building CALL applications integrating TTS synthesis systems which meet the requirements that they think that CALL applications place on TTS synthesis systems and evaluating those applications with learners. Once a benchmark test is available, it will therefore be less important to adopt this approach to developing CALL applications integrating TTS synthesis.

### 6.4.2 Use of TTS synthesis in CALL

The finding that there are either only small differences between the different roles that TTS synthesis systems may assume within CALL applications with respect to the requirements that they place on the quality of the speech generated, or steachers and CALL researchers cannot discriminate between the different roles and the requirements that they place on the quality of the speech generated, or the roles overlap also has implications for the use of TTS synthesis in CALL. Specifically, it means that one TTS synthesis system can be used for all applications, if it is sufficiently flexible, i.e. provides options over all of the features presented in section 5.4: if there are differences among the roles, but the differences are only small, there is little gain in using different TTS synthesis systems; if teachers and CALL researchers cannot discriminate between the different roles and their requirements, then it is not possible to select different TTS synthesis system; and, if the roles overlap, then a TTS synthesis system which is suitable for use in all roles ought to be used. This is clearly advantageous in cases where

CALL researchers wish to develop several different applications integrating TTS synthesis in which it assumes different roles as it will reduce the cost of development.

As presented in section 6.3.2, regarding the use of TTS synthesis in CALL, the results of the investigations presented in this thesis also provide an insight into the readiness of TTS synthesis for use in CALL applications. Specifically, as presented in section 6.3.3.4, they suggest that the quality of the speech generated by some French TTS synthesis systems has reached acceptability levels for use in CALL applications. This suggests that, if the flexibility were also there (see section 5.4), some French TTS synthesis systems would already be ready for use in applications which are not possible to provide through the use of other speech output technologies such as digitised speech, namely those applications in which the generative nature of TTS synthesis, its manipulability, and its capacity to generate new types of speech models (see sections 3.2 and 3.3). Regarding the use of French TTS synthesis systems in applications which could be provided through the use of other speech output technologies such as digitised speech, as presented in section 6.3.3.4, those French TTS synthesis systems that have reached acceptability levels are also nearing adequacy levels for use in CALL applications. Moreover, advances are being made with respect to the provision of the options which it is believed that TTS synthesis systems for use in CALL applications ought to provide (see van Santen *et al.* (2002), for example, for work on the manipulation of voice quality and the provision of options over voice; and see Olinsky and Cummins (2002), for example, for work on the provision of options over accent). It should therefore not be long before it is possible to use TTS synthesis to provide all of the applications presented in section 3.3, including those which it is already possible to provide through the use of other speech output technologies.

Regarding the use of TTS synthesis in CALL for other languages, as presented in section 2.4, TTS synthesis for different language presents different problems. Moreover, more is known about some languages, in particular English, than others, in particular minority languages. Hence as presented in section 6.3.4, it is therefore questionable whether the results of this investigation, in particular those concerning the readiness of TTS synthesis for use in CALL applications, will generalise to other languages. Whether or not they are ready, however, if TTS synthesis is being used to provide services which learners might encounter when the telephone or visit the countries in which the TL is spoken, then we believe that ready or not

for use for the provision of applications presented in section 3.3 learners should be exposed to TTS synthesis in CALL and language learning in general.

# 7 Conclusion

As presented in the introduction, the research presented in this thesis was motivated by the fact that despite the fact that some of the benefits that TTS synthesis could bring to CALL were identified more than 20 years ago (Sherwood, 1981), TTS synthesis is still hardly used in CALL. An in-depth review of the state-of-the-art of TTS synthesis and its use in CALL made the fact that it has still not made an impact on CALL all the more surprising. Regarding the quality of the speech generated by TTS synthesis systems, this review showed that TTS synthesis systems have long been able to generate highly intelligible speech and since the introduction of concatenative synthesis and in particular its successor USS considerable gains in naturalness have been made. As regards the use of TTS synthesis in CALL, this review showed that not only does TTS synthesis have the potential to improve on the possibilities that other speech output technologies such as digitised speech brought to CALL, but TTS synthesis itself also has the potential to bring new possibilities to CALL, i.e. has the potential to add value to CALL, specifically to generate a number of new types of speech models that might be particularly useful in LL&T and the ability to generate speech models on demand.

Following the failure of the much heralded but unproven language laboratory, as presented in the introduction, teachers became sceptical about the introduction of new technologies, in particular those which were not proven, i.e. had not been adequately evaluated for the purposes of LL&T. We therefore believed that the most likely reason why TTS synthesis had still not made an impact on CALL, was that it has not been adequately evaluated for the purposes of CALL. The first aim of the research presented in this thesis was to determine whether this was indeed the case. In order to achieve this goal, it was first necessary to establish what types of evaluation ought to be conducted. Through the synthesis of the infrastructures proposed for the evaluation of SALTs and CALL applications, an infrastructure for the evaluation of CALL applications integrating TTS synthesis was put forward and the evaluations of CALL applications integrating TTS synthesis which had been conducted to date were assessed with respect to the proposed infrastructure. It was found that TTS synthesis has indeed not been adequately evaluated for the purposes of CALL: only one evaluation of TTS synthesis has addressed the adequacy of TTS synthesis for use in CALL and this evaluation only looked at one TTS synthesis system, namely the *SSI 263 Spanish TTS chip*, for use in one application, namely for the presentation of grammar exercises in a language laboratory (Stratil *et al.*, 1987b); only one product-oriented evaluation of learners' performance in a teacher-

planned task using a CALL application integrating TTS synthesis has been conducted and, like the adequacy evaluation mentioned above, this evaluation only looked at the use of one TTS synthesis system, namely the *KTH Swedish TTS synthesis system*, in one task, namely a task aimed at teaching the lexical stress of English words which have Swedish cognates (Hincks, 2002); only one process-oriented evaluation of learners' performance using a CALL application integrating TTS synthesis has been conducted and this evaluation, like the other two evaluations mentioned above, also only looked at one CALL application, namely the TWP *Composition* (Cohen, 1993); and, only one evaluation of the impact of CALL applications integrating TTS synthesis was conducted and this evaluation only looked at one TTS synthesis system, namely the *SSI 263 Spanish TTS chip*, in two applications, namely for the presentation of grammar exercises and pronunciation rules in a language laboratory (Stratil *et al.*, 1987a).

Even though CALL applications integrating TTS synthesis are already available on the market, we believe that it is necessary to go back and conduct evaluations of the adequacy of TTS synthesis for use in CALL applications, because if it is omitted, considerable time and resources may be wasted integrating the technology into applications for which it is not suitable. Regarding the reasons why evaluation has been omitted, we believe that it is because it is costly in terms of time and resources. One way in which this problem has been overcome in the field of software evaluation more generally is through the use of benchmark tests, "efficient, easily administered ... tests, that can be used to express the performance of a ... system in numerical terms" (van Bezooijen and van Heuven, 1997: 497). When the performance of state-of-the-art systems in such tests are made publicly available they make the process of evaluation even more efficient; all the CALL researcher would need to do if such results were made publicly available is to obtain the publication and look up the results and compare them with their requirements. The development of a benchmark test for the evaluation of the adequacy of TTS synthesis for use in CALL applications was therefore proposed.

Regarding the development of such a benchmark test, the evaluations of CALL applications integrating TTS synthesis conducted to date were also analysed with respect to best practice in evaluation. This analysis revealed that all of the evaluations had omitted an essential stage in the evaluation process, namely requirements analysis. Moreover, such an analysis was not to

be found elsewhere in the literature. Before the aforementioned benchmark test can be developed, the requirements that CALL applications place on TTS synthesis systems must be identified. This was the second aim of the research presented in this thesis. This goal was approached from two perspectives. First the literature was reviewed for insights into what the requirements might be. Then empirical investigations were conducted in order to validate the findings of the literature review. Regarding the bodies of literature that might provide insights into the demands that CALL applications place on TTS synthesis, CALL in general draws on a diverse range of fields. A significant limitation of the research presented in this thesis is that only one of these fields is considered, namely SLA. It is, however, believed that the findings of SLA research ought to be the primary consideration when evaluating CALL applications (Pederson, 1987; Chapelle, 1997, 1998, 2001). The findings of this review which looked at both the goals of language learning and models of SLA suggested that CALL applications place demands on the quantity, quality and flexibility of the speech generated by TTS synthesis systems. More specifically, regarding quantity it suggested that TTS synthesis systems for use in CALL need to be able to generate large quantities of TL input. Regarding quality, it suggested that the speech generated by TTS synthesis systems for use in CALL needs to be comprehensible, i.e. understandable, accurate, natural and smooth. And, regarding flexibility, the literature suggested that TTS synthesis systems for use in CALL need to provide options over voice (gender, age, etc.), accent, register, SR, and duration. No one would question the fact that learners require exposure to large quantities of TL input, i.e. that TTS synthesis systems for use in CALL need to be able to generate large quantities of speech. The requirements that the SLA literature suggest that CALL applications place on the quality of the speech generated by TTS synthesis systems on the other hand require validation. The results of two investigations which asked teachers and CALL researchers of French to rate a range of aspects of the quality of the speech generated by a selection of French TTS synthesis systems representative of the current state-of-the-art with respect to their use in CALL in the roles of RM, phonetic PM, prosodic PM and CP – it was believed that the different roles may impose different requirements on the quality of the speech generated by TTS synthesis systems – revealed that they did indeed place demands on comprehensibility, accuracy and naturalness, but also on intelligibility, choice of pronunciation, naturalness of voice, expressiveness and register. In addition to the aforementioned aspects of the quality of the speech generated by the TTS synthesis systems, the participants were also asked to rate the adequacy and acceptability of the speech generated by each of the TTS synthesis systems with

respect to their use in each of the roles that they might assume within CALL applications and the readiness of TTS synthesis in general for use in each of those roles. Regarding the question of whether the different roles that TTS synthesis systems may assume within CALL applications place different demands on the quality of the speech generated by TTS synthesis systems, the findings were inconsistent: while statistically significant differences were found among some aspects of the quality of the speech generated by each of the TTS synthesis systems across the roles and the readiness of TTS synthesis in general across the roles, statistically significant differences were only found among the adequacy of the speech generated by two of the TTS synthesis systems evaluated across the roles and the differences in the acceptability of the speech generated by all the TTS synthesis systems evaluated across the roles were not found to be statistically significant. A number of possible explanations for these findings were put forward, namely that: there are differences across the roles with respect to the requirements that they impose on the quality of the speech generated by TTS synthesis systems and the participants can make those differences, but the differences are only small; there are differences across the roles with respect to the requirements that they impose on the quality of the speech generated by TTS synthesis systems, but the participants cannot make those differences; and, the roles overlap. Whichever the case the implications for the evaluation of the adequacy of the quality of the speech generated by TTS synthesis systems for use in CALL are the same, as are the implications for the use of TTS synthesis systems in CALL (see below). Smoothness, or rather the effects of a lack of smoothness, namely distraction, ought to be evaluated with learners. Smoothness was therefore not addressed in the investigations which constitute this piece of research and is a subject for further research. Before it is possible to develop a benchmark test, it will also be necessary to determine the degree to which each of the aspects of the quality of the speech generated by a TTS synthesis system contribute to its adequacy and acceptability for use in CALL applications.

Regarding the demands that CALL applications place on the flexibility of the speech generated by TTS synthesis systems, further research is needed to determine what degree of control is required over voice, accent, register, SR and duration. Such research and the investigation of whether TTS synthesis systems for use in CALL also need to provide options over pause frequency and duration, clarity of articulation, phonology and intonation, and the ability to generate the whisper, humming, and reiterant speech was beyond the scope of the research presented in this thesis. Once the quality of the speech generated by TTS synthesis

systems has met the demands that CALL applications place on it, we believe that these requirements will be best investigated through CALL applications integrating TTS synthesis systems: TTS synthesis systems provide a good way of controlling speech and CALL applications provide a good way of controlling and monitoring SLA.

We are therefore still a long way off developing a benchmark test for the evaluation of the adequacy of TTS synthesis systems for use in CALL. The research presented here does, however, permit us to make a few recommendations regarding the development of such a benchmark test. First, it is believed that such a test should not attempt to discriminate between the different roles that TTS synthesis systems may assume within CALL applications: if there are differences among the roles, but the differences are only small, there is little gain in making the difference; if teachers and CALL researchers cannot discriminate between the different roles and their requirements, then it is not possible to make the difference; and, if the roles overlap, then the difference should not be made. Secondly, such a test should address all of the aspects of the quality of the speech generated by TTS synthesis systems considered in the investigations presented in this thesis, namely comprehensibility, intelligibility, choice of pronunciation, accuracy, naturalness, naturalness of voice, expressiveness and register. CALL developers who must devise their own evaluations in the mean time should also follow these recommendations. A significant limitation of such evaluations, like the investigations presented in this thesis, is that they are subjective, the results indicate whether teachers and CALL researchers *think* that TTS synthesis systems have met the demands that CALL applications place on the quality of the speech generated. *Actual* requirements may be different to what teachers and CALL researchers *think* they are. It is therefore recommended that CALL researchers build say one example of each type of exercise that they wish to provide and submit those exercises to the remaining levels of evaluation of the proposed infrastructure for the evaluation CALL applications integrating TTS synthesis before building the full-scale application. It is intended that the findings of these and further investigations of the requirements involving teachers and CALL researchers as participants are validated by actually building CALL applications integrating TTS synthesis systems which meet the requirements that they think that CALL applications place on TTS synthesis systems and evaluating those applications with learners. Once a benchmark test is available, it will therefore be less important to adopt this approach to developing CALL applications integrating TTS synthesis.

Regarding the use of TTS synthesis in CALL, the implications of the findings of the investigations conducted as part of this piece of research presented so far are that one TTS synthesis system can be used for all applications, if it is sufficiently flexible, i.e. provides options over all of the features presented earlier in this chapter. In addition, regarding the use of TTS synthesis in CALL, the research presented in this thesis also provides insights into the readiness of TTS synthesis for use in CALL applications. Specifically, it suggests that if the flexibility were also there – a number of the options which it is believed that TTS synthesis systems for use in CALL applications ought to provide are not yet available – some French TTS synthesis systems may already ready for use in applications which are not possible to provide through the use of other speech output technologies such as digitised speech, namely those applications in which the unique features of TTS synthesis are exploited. Regarding the use of French TTS synthesis systems in applications which could be provided through the use of other speech output technologies such as digitised speech, the quality of the speech generated by the aforementioned TTS synthesis systems is also nearing acceptance levels for use in those applications too. Moreover, advances are being made with respect to the provision of the options which it is believed that TTS synthesis systems for use in CALL applications ought to provide. It should therefore not be long before French TTS synthesis is ready for use in CALL in all its possible applications. Unfortunately, it is not possible to say whether this is the case for other languages: TTS synthesis for different languages poses different problems and more is known about some languages than others. Returning to CALL applications integrating French TTS synthesis, if learners are to profit from the benefits of these applications further evaluation is required. Without it teachers will remain sceptical about the use of TTS synthesis in CALL. And, unless teachers express a demand for such software publishers will not stock it (Ariew, 1990).

# References

Notes:

All URLs, unless otherwise stated, were still available on 28th September 2005.

Any incomplete references are marked with an asterisk.


Abercrombie, D. (1967). *Elements of General Phonetics.* Edinburgh: Edinburgh University Press.

Ahmad, K., Corbett, G., Rogers, M., and Sussex, R. (1985). *Computers, Language Learning and Language Teaching.* Cambridge: Cambridge University Press.

Ainsworth, W. A. (1973). A System for Converting English Text into Speech. *IEEE Transactions on Audio and Electroacoustics.* AU21 (3): 288-290

Akahane-Yamada, R., Tohkura, Y., Bradlow, A., and Pisoni, D. (1996). Does Training in Speech Perception Modify Speech Production? In *Procs. ICSLP 96* (pp. 606-609). Philadelphia, PA, October, 1996

Allen, J. (1973). Reading Machines for the Blind: The Technical Problems and the Methods Adopted for their Solution. *IEEE Transactions on Audio and Electroacoustics.* AU21 (3): 259-264

Allen, J. (1992). Overview of Text-to-Speech Systems. In Furui, S. and Sondhi, M. (eds.) (1992). *Advances in Speech Signal Processing* (pp. 741-791). New York: M. Dekker.

Allen, J., Hunnicutt, M. S., and Klatt, D. with Armstrong, R.C., and Pisoni, D.B. (1987). *From Text to Speech: The MITalk System.* Cambridge: Cambridge University Press.

Andersen, A. (2000). Spanish for Business Professionals. *CALICO Software Review.* Retrieved from http://calico.org/CALICO_Review/review/sbp00.htm

Anderson, A. and Lynch, T. (1998). *Listening.* Oxford: Oxford University Press.

ANSI (1989). *Method for Measuring the Intelligibility of Speech Over Communication* Systems (ANSI S3.2-1989 – A Revision of ANSI S3.2-1960). New York: Amercan Standards Association.

Ariew, R. (1990). Integrating Video and CALL in the Curriculum: The Role of the ACTFL Guidelines. In Flint Smith, Wm. (1988). *Modern Media in Foreign Language Education: Theory and Implementation* (pp. 1-66). Lincolnwood, Illinois: National Textbook Company.

Auralog (2002). *Talk to Me: The Conversation Method (French).* (Version 3.5) from Auralog http://www.auralog.fr

Babel Technologies (2003). *BrightSpeech.* Retrieved from http://www.babeltech.com/Products.php?s=76&m=75&f=70

Bachman, L. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bailly, G. (2001). The COST 258 signal generation test array. In Keller, E. (ed.) (2001). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 39-51). Chichester: John Wiley.

Bailly, G. (2002a). The COST 258 Signal Generation Test Array. In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 39-51). Chichester: John Wiley.

Bailly, G. (2002b). Towards More Versatile Signal Generation Systems. In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 18-21). Chichester: John Wiley.

Bailly, G., Bérar, M., Elisei, F. and Odisio, M. (2003a). Audiovisual Speech Synthesis. *International Journal of Speech Technology.* 6: 331-346

Bailly, G., Campbell, N. and Mobius, B. (2003b) ISCA Special Session: Hot Topics in Speech Synthesis. In *Procs. Eurospeech 2003* (pp. 37-40). Geneva.

Battye, A. and Hintze, M.-A. (1992). *The French Language Today.* London: Routledge.

Bax, S. (2003). CALL – Past, Present and Future. System. 31: 13-28

BCS (British Computer Society) (1995). A Glossary of Computing Terms. Burnt Mill, Harlow, Essex: Longman.

Beatty, K. (2003). *Teaching and Researching Computer-Assisted Language Learning.* London: Longman

Beebe, L. (1985). Input: Choosing the right Stuff. In Gass, S. and Madden, C. (eds.) (1985). *Input in Second Language Acquisition* (pp. 404-414). Rowley, MA: Newbury House.

Begag, A. (1990). *Les voleurs d'écriture.* Paris: Editions du Seuil.

Benoît, C. and Le Goff, B. (1998). Audio-Visual Speech Synthesis from French Text: Eight Years of Models Designs and Evaluation at the ICP. *Speech Communication.* 26: 117-129.

Benoît, C., Martin, J.-C., Pelachaud, C., Schomaker, L., and Suhm, B. (2000). Audio-Visual and Multi-Modal Speech-Based Systems. In Gibbon, D., Mertins, I., and Moore, R. K. (eds.) (2000). *Handbook of Multimodal Spoken Dialogue Systems* (102-203). London: Kluwer Academic Publishers.

Beskow, J. (1996). Talking Heads - Communication, Articulation and Animation. *TMH-QPSR.* 2/1996: 53-56

Beskow, J., Granström, B., House, D. and Lundeberg, M. (2000). Experiments with Verbal and Visual Conversational Signals for an Automatic Language Tutor. In *Procs. InSTIL 2000* (pp. 137-142). Dundee: University of Abertay Dundee.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (1999). The AT&T Next-Gen TTS system. In *Procs. Joint Meeting of the ASA, EAA, and DAGA.* Berlin, Germany.*

Bhaskararo, P. (1994). Subphonemic Inventories for Concatenative Speech Synthesis. In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 69-85). Chichester: John Wiley.

Bickerton, D., Stenton, T., and Temmerman, M. (2001). Criteria for the Evaluation of Authoring Tools in Language Education. In Chambers, A., and Davies, G. (2001). *ICT and Language Learning: A European Perspective* (pp. 53-66). Lisse: Swets and Zeitlinger.

Bickley, C. and Bruckert, E. (2002). Improvements in the Voice Quality of DECTALK. In In *Procs. IEEE 2002 Workshop on Speech Synthesis* (pp. 55-58). Santa Monica, California.

Bickley, C., Syrdal, A., and Schroeter, J. (1999) Speech Synthesis. In Pickett (ed.) (1999). *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology* (pp. 325-335). London: Allyn and Bacon.

Bingham Wesche, M. (1994). Input and Interaction in Second Language Acquisition. In Gallaway, C. and Richards, B. (eds.) (1994). *Input and Interaction in Language Acquisition* (pp. 219-249). Cambridge: Cambridge University Press.

Black, A. (1996). *CHATR, Version 0.8, a Generic Speech Synthesiser, System Documentation.* Kyoto, Japan: ATR-Interpreting Telecommunication Laboratories.

Black, A., and Taylor, P. (1994). CHATR: A Generic Speech Synthesis System. In *Procs. COLING 94, the 15th International Conference on Computational Linguistics* (pp. 983-986). Kyoto, Japan.

Bonneau, A., Camus, M., Laprie, Y., Colotte, V. (2004). A Computer-Assisted Learning of English Prosody for French Students. In *Procs. InSTIL/ICALL 2004* (pp. 119-121). Venice, Italy.

Bonneau, A., Laprie, Y., and Colotte, V. (2000). Towards Phonetic Tools for Speech Training. In *Procs. InSTIL 2000* (pp. 77-80). Dundee: University of Abertay Dundee.

Borden, G., Harris, K., and Lawrence, R. (1994). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech.* London: Williams and Wilkins.

Boula de Mareuil, P., Célérier, P., Cesses, T., Fabre, S., Jobin, C., Le Meur, P.-Y., Obadia, D., Soulage, B. and Toen, J. (2001). Elan text-to-speech: un système multilingue de synthèse de la parole à partir du texte. *Tratiement Automatique des Languages (TAL).* 42 (1): 1-30.

Bradlow, A., and Pisoni, D. (1999). Recognition of Spoken Words by Native and Non-Native Listeners: Talker-, Listener-, and Item-Related Factors. *Journal of the Acoustical Society of America.* 106: 2074-2085

Bradlow, A., Pisoni, D., Akahane-Yamada, R., and Tohkura, Y. (1997). Training Japanese Listeners to Identify English /r/ and /l/. IV. Some Effects of Perceptual Learning on Speech Production. *Journal of the Acoustical Society of America.* 96 (4): 2076-2087

Bradlow, A., Torretta, G., and Pisoni, D. (1996). Intelligibility of Normal Speech I: Global and Fine-Grained Acoustic-Phonetic Talker Characteristics. *Speech Communication.* 20: 255-272

Brown, A. (1992). Twenty Questions. In Brown, A. (ed.) (1992). Approaches to Pronunciation Teaching. London : MacMillan.

Brown, A. (1992). Twenty Questions. In Brown, A. (ed.) (1992). *Approaches to Pronunciation Teaching* (pp. 1-17). London: MacMillan.

Brown, D. H. (1994). *Principles of Language Learning and Teaching.* London: Prentice Hall.

Brown, G. (1977). *Listening to Spoken English.* London: Longman.

Cabré, M. (1998). *Terminology: Theory, Models and Applications.* Amsterdam: Benjamins.

Cai, J.-Y., Nerurkar, A., & Wu, M.-Y. (1998). Making Benchmarks Uncheatable. *Procs. IEEE International Computer Performance and Dependability Symposium* (IPDS '98) (pp. 216-226). Durham, NC.

Callamand, M. (1981). *Methodologie de l'enseignement de la prononciation.* Paris: Clé International.

Campbell, N. and Black, A. (1997). Prosody and the Selection of Source Units for Concatenative Synthesis. In van Santen, J., Sproat, R., Olive, J., and Hirschberg, J. (eds.) (1997). *Progress in Speech Synthesis* (pp. 279-292). London: Springer Verlag.

Canale, M. (1983). From Communicative Competence to Communicative Language Pedagogy. In Richards, J. and Schmidt, R. (eds.) (1983). *Language and Communication.* London: Longman.

Canale, M. and Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics.* 1: 1-47

Celce-Murcia, M., Brinton, D., and Goodwin, J. (1996). *Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages.* Cambridge: Cambridge University Press.

Chambers (1998). *The Chambers Dictionary.* Edinburgh: Chambers Harrap Publishers Ltd.

Chapelle, C. (1997). CALL in the Year 2000: Still in Search of Research Paradigms? *Language Learning and Technology.* 1 (1): 19-43. Retrieved from: http://llt.msu.edu/vol1num1/chapelle/default.html

Chapelle, C. (1998). Multimedia CALL: Lessons to be Learned from Research on Instructed SLA. *Language Learning and Technology.* 2 (1): 22-34. Retrieved from: http://llt.msu.edu/vol2num1/article1/index.html

Chapelle, C. (2001). *Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing and Research.* Cambridge: Cambridge University Press.

Chapelle, C. (2003). *English Language Learning and Technology: Lectures on Applied Linguistics in the Age of Information and Communication Technology.* Cambridge: Cambridge University Press.

Chapelle, C. and Jamieson, J. (1990). Internal and External Validity Issues in Research on CALL Effectiveness. In Dunkel, P. (ed.) (1990). *Computer Assisted Language Learning and Testing: Research Issues and Practice* (pp. 37-59). New York: Newbury House.

Charliac, L. and Motoron, A.-C. (1998). *Phonétique progressive du français avec 600 exercices.* Paris: Clé International.

Chaudron, C. (1988). *Second Language Classrooms: Research on Teaching and Learning.* Cambridge: Cambridge University Press.

Chomsky, N. (1975). *Reflections on Language.* New York: Pantheon.

Codling, S. (1998). *Benchmarking.* Aldershot, England: Gower.

Cohen, R. (1993). The Use of a Voice Synthesizer in the Discovery of the Written Language by Young Children. *Computers in Educations.* 21(1/2): 25-30.

Coker, C. H., Umeda, N. and Browman, C. P. (1973). Automatic Synthesis for Ordinary English Text. *IEEE Transactions on Audio and Electroacoustics.* AU21 (3): 293-298

Coleman, J. (1991). Interactive Multimedia. In Brierley, B. and Kemble, I. (eds.) (1991). *Computers as a Tool in Language Teaching* (pp. 87-111). London: Ellis Horwood.

Colpaert and Decoo (1999). The Role of Didactic Functions in CALL Design. In Cameron, K. (ed.) (1999). *CALL and the Learning Community* (pp. 65-74). Exeter: Elm Bank.

Conkie, A. (1999). Robust Unit Selection System for Speech Synthesis. In *Procs. Joint meeting of ASA, EAA, and DAGA.* Berlin. Germany.*

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Cutler, A., Dahan, D., and van Donselaar, W. (1997) Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech.* 40 (2): 141-201

D'Alessandro, C. (2001). 33 Ans de Synthèse de la Parole à Partir du Texte : Une Promenade Sonore (1968-2001). *Traitement Automatique des Langues (TAL).* 42 (1): 297-321.

D'Alessandro, C. and Liénard, J.-S. (1996). Synthetic Speech Generation. In Battista Varile, G., Zampoli, A., Cole, R., Mariani, J., Uszkoreit, H., and Zaenen, A. (eds.) (1996). *Survey of the State of the Art in Human Language Technology* (pp. 170-174). Cambridge: Cambridge University Press.

Dabène, M. (1974). L'écrit en question. *Le Français dans le Monde.* 14 (109): 6-9

Dalton, C. and Seidlhofer, B. (1994). *Pronunciation.* Oxford: Oxford University Press.

Dancey, C. and Reidy, J. (2002). *Statistics Without Maths for Psychology using SPSS for Windows*$^{TM}$. Prentice Hall: London.

Dansereau, D. (1995). Phonetics in the Beginning and Intermediate Oral Proficiency-Oriented French Classroom. *The French Review.* 68 (4): 638-651

Davies, G. and Higgins, J. (1985). *Using Computers in Language Learning: A Teacher's Guide.* London: CILT.

De la Fuente, M. J. (2002). Negotiation and Oral Acquisition of L2 Vocabulary. Studies in *Second Language Acquisition.* 24: 81-112

de Pijper, J. (1997). High-Quality Message-to-Speech Generation in a Practical Application. In van Santen, J., Sproat, R., Olive, J., and Hirschberg, J. (eds.) (1997). *Progress in Speech Synthesis* (pp. 575-588). London: Springer Verlag.

de Quincey, P. (1986). Stimulating activity: The role of computers in the language classroom. *CALICO Journal.* 4(1): 55-66.

de Saint-Exupéry, A. (1999). *Le Petit Prince.* Paris: Gallimard.

Delogu, C., Conte, S. and Sementina, C. (1998). Cognitive Factors in the Evaluation of Synthetic Speech. *Speech Communication.* 24: 153-168

Denes, P. and Pinson, E. (1993). *The Speech Chain: The Physics and Biology of Spoken Language.* New York: W.H. Freeman.

DISC (1999). *A survey of existing methods and tools for developing and evaluating of speech synthesis and commercial speech synthesis systems.* Retrieved from http://www.disc2.dk/tools/SGsurvey.html

Divay, M. and Vitale, A. (1997). Grapheme-Phoneme Transcription. *Association for Computational Linguistics.* 23 (4): 495-523.

Docherty, D. and Shockey, L. (1988). Speech Synthesis. In Jack, M. And Laver, J. (1988). *Aspects of Speech Technology* (pp. 144-183). Edinburgh : Edinburgh University Press.

Doughty, C. (1987). Relating Second-language Acquisition Theory to CALL Research and Application. In Flint Smith, Wm. (ed.). (1987). *Modern Media in Foreign Language Education: Theory and Implementation* (pp. 133-167). Lincolnwood, Illinois: National Textbook Company.

Dunkel, P. (1990). The Effectiveness Research on Computer-Assisted Instruction and Computer-Assisted Language Learning. In Dunkel, P. (ed.) (1990). *Computer-Assisted Language Learning and Teaching: Research Issues and Practice* (pp. 5-35). New York: Newbury House.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis.* London : Kluwer Academic Publishers.

EAGLES (Expert Advisory Group on Language Engineering Standards). (1999). *EAGLES Evaluation of Natural Language Processing Systems. Final Report.* EAGLES Document EAG-II-EWG-PR.1. Copenhagen: Center for Sprogteknologi.

Edgington, M. (1997). Investigating the Limitations of Concatenative Synthesis. In *Procs. Eurospeech '97* (pp. 1-4). Rhodes, Greece.*

Edgington, M., Lowry, A., Jackson, P., Breen, A. and Minnis., S. (1996a). Overview of Current Text-to-Speech Techniques: Part I – Text and Linguistic Analysis. *BT Technology Journal.* 14 (1): 68-83.

Edginton, M., Lowry, A., Jackson, P., Breen, A., and Minnis, s. (1996b). Overview of Current Text-to-Speech Techniques: Part II – Prosody and Speech Generation. *BT Technology Journal.* 14 (1): 84-99.

Edwards, A. (1991). *Speech Synthesis: Technology for Disabled People.* Baltimore, Maryland: Paul Chapman Publishing.

Egan, B. and LaRocca, S. (2000). Speech Recognition in Language Learning: A Must. In *Procs. InSTIL 2000* (pp. 4-9). Dundee, England: University of Abertay Dundee.

Ellis, R. (1985). *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

Ellis, R. (1994). *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Ellis, R. (1997). *SLA Research and Language Teaching.* Oxford: Oxford University Press.

Ellis, R. and He, X. (1999). The Roles of Modified Input and Output in the Incidental Acquisition of Word Meanings. *Studies in Second Language Acquisition.* 21: 285-301

Ellis, R., Tanaka, Y., and Yamazaki, A. (1994). Classroom Interaction, Comprehension, and the Acquisition L2 Word Meanings. *Language Learning.* 44: 449-491

ELSE (Evaluation in Language and Speech Engineering). (1999). *A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment* (Report no. D1.1). Retrieved from: http://m17.limsi.fr/TLP/ELSE/ (Site no longer available).

Esling, J. H. (1992). Speech technology systems in applied linguistics instruction. In M. Pennington and V. Stevens (eds.). (1992). *Computers in Applied Linguistics: An International Perspective* (pp. 244-272). Clevedon, UK: Multilingual Matters.

Fourcin, A. (1992). Assessment of Synthetic Speech. In G. Bailly, C. Benoit and Sawallis, T. (eds.) (1992). *Talking Machines – Theories, Models and Designs* (pp. 431-434). Amsterdam: Elsevier Science.

Fox, M. (1997). Beyond the Technocentric – Developing and Evaluating Content-Driven, Internet-Based Language Acquisition Courses. *Computer Assisted Language Learning.* 10 (5): 443-453

Francis, A., and Nusbaum, H. (1999). Evaluating the Quality of Synthetic Speech. In Gardner-Bonneau, D. (Ed.) (1999). *Human Factors and Voice Interactive Systems* (pp. 63-97). Boston: Kluwer Academic Publishers.

Francis, A., and Nusbaum, H. (1999). Evaluating the Quality of Synthetic Speech. In Gardner-Bonneau, D. (Ed.) (1999). *Human Factors and Voice Interactive Systems* (pp. 63-97). Boston: Kluwer Academic Publishers.

Gabioud, B. (1994). Articulatory Models in Speech Synthesis. In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 215-230). Chichester: John Wiley.

Garant-Viau, C. (1994). *La portée des sons.* Québec: Université Laval.

Garrett, N. (1995). ICALL and Second Language Acquisition. In Holland, M., Kaplan, J. and Sama, M. (eds.). *Intelligent Language Tutors: Theory Shaping Technology* (pp. 344-358). Hove, UK: Lawrence Erlbaum.

Gass, S. (1997). *Input, Interaction, and the Second Language Learner.* Mahwah, New Jersey: Lawrence Erlbaum.

Gass, S. and Varonis, E. (1994). Input, Interaction and Second Language production. *Studies in Second Language Acquisition Research.* 16: 283-302

Gattegno, C. (1972). *Teaching Foreign Languages in Schools: The Silent Way.* New York: Educational Solutions.

Gaudinat, G. and Wehrli, E. (1997). Analyse Syntaxique et Synthèse de la Parole: le Projet FIPSvox. *Traitement Automatique des Langues (TAL).* 38 (1): 121-134

Germain-Rutherford, A. and Martin, P. (2000). Présentation d'un Logiciel de Visualisation pour l'Apprentissage de l'Oral en Langue Seconde. *Apprentissage des Languages et Systèmes d'Information et de Communication (ALSIC).* 3 (1): 61-76.

Germain-Rutherford, A. and Martin, P. (2001). Perspectives Nouvelles dans l'Enseignement à Distance de l'Oral: Les Technologies de Visualisation et de Synthèse. In Cornaire, C. and Raymond, P. (eds.) (2001). *Regards sur la Didactique des Langues* (pp. 105-132). Les Editions Logiques.

Goldstein, M. (1995). Classification of Methods used for Assessment of Text-to-Speech Systems According to the Demands Placed on the Listener. *Speech Communication.* 16: 225-244

Goodfellow, R. (1999). Evaluating Performance, Approach and Outcome. In Cameron, K. (ed. (1999). *Computer-Assisted Language Learning (CALL): Media, Design and Applications* (pp. 109-140). Lisse: Swets & Zeitlinger.

Grace, R. (1996). *The Benchmark Book.* London: Prentice Hall.

Gray, T. (1984). Talking Computers in the Classroom. In Bristow, G. (ed.) (1984). *Electronic Speech Synthesis* (pp. 234-259). London : McGraw-Hill.

Griffiths, R. (1990). Speech Rate and NNS Comprehension: A Preliminary Study in Time-Benefit Analysis. *Language Learning.* 40 (3): 311-336

Groschel, B. (1979). Mündliche und Schriftliche Kommunikation – Autonomie und Wechselbeziehungen in Sprachlernprozessen. *Folia Linguistica.* 13 (3-4): 291-302.

Grosjean, F. (1972). *Le role joué par trios variables temporelles dans la comprehension orale de l'anglais étudié comme seconde langue et perception de la vitesse de lecteurs et de auditeurs.* Unpublished Doctoral Thesis. Université de Paris VII, Paris, France.

Grotjahn, R. (1991). The Research Programme Subjective theories: A New Approach in Second Language Research. *SSLA.* 13: 187-214

Hakanson, G. (1986). Quantitative Studies of Teacher Talk. In Kasper, G. (ed.) (1986). *Learning, Teaching and Communication in the Foreign Language Classroom.* Aarhus: Aarhus University Press.

Halliday, M. (1973). *Explorations in the Functions of Language.* London: Edward Arnold.

Hamburger, H. (1990). Evaluation of L2 Systems Learners and Theory. *CALL.* 1: 19-27

Hamel, M.-J. (2003a). *Re-using Natural Language Processing Tools in Computer Assisted Language Learning: The Experience of SAFRAN.* Unpublished Doctoral Thesis. UMIST, Manchester.

Hamel, M.-J. (2003b). FreeText: A "Smart" Multimedia Web-based Computer Assisted Language Learning Environment for Learners of French. In *Procs. m-ICTE2003*, (Vol. III, pp. 1661-1665). Badajoz, Spain.

Handley, Z and Hamel, M.-J. (2004). Investigating the Requirements of Speech Synthesis for CALL with a View to Developing a Benchmark. In *Procs. InSTIL/ICALL 2004* (pp. 71-74). Venice, Italy.

Handley, Z and Hamel, M.-J. (2005). Establishing a Methodology for Benchmarking Speech Synthesis for Computer-Assisted Language Learning (CALL). *Language Learning and Technology Journal.* 9 (3): 99-119. Retrieved from: http://llt.msu.edu/vol9num3/handley/default.html

Hardisty, D. and Windeatt, S. (1989). *CALL.* Oxford: Oxford University Press.

Harris, P. (2002). *Designing and Reporting Experiments in Psychology.* Open University Press: Buckingham.

Hart, R., Marty, F, and Fukada, A. (1988). Transcribing French Text into Speech. In Jung, U. (ed.) (1988). *Computers in Applied Linguistics and Language Teaching: A CALL Handbook* (pp. 137-146). New York: Peter Lang.

Haycroft, B. (1992). Sentence Stress – For More Meaningful Speech. In Brown, A. (ed.) (1992). *Approaches to Pronunciation Teaching* (pp. 57-72). London: Macmillan.

Heiman, G. (2001). *Understanding Research Methods and Statistics: An Integrated Introduction for Psychology.* Boston, Mass.: Houghton Mifflin.

Henton, C. (2002). Challenges and Rewards in Using Parametric or Concatenative Speech Synthesis. *International Journal of Speech Technology.* 5: 117-131.

Henzl, V. (1979). Foreigner Talk in the Classroom. *International Review of Applied Linguistics.* 17: 159-29

Henzl, V. (19873). Linguistic Register of Foreign Language Instruction. *Language Learning.* 23: 207-227

Hertz, S. R. (2002). Integration of Rule-Based Formant Synthesis and Wavefrom Concatenation: A Hybrid Approach to Text-to-Speech Synthesis. In *Procs. IEEE 2002 Workshop on Speech Synthesis* (pp. 87-90). Santa Monica, California.

Hertz, S., Younes, R., and Hoskins, S. (2000). Space, Speed, Quality, and Flexibility. In *Procs. AVOIS 2000* (pp. 217-227). San José, CA.

Heshusius, L. (1994). Freeing Ourselves from Objectivity: Managing Subjectivity or Turning Toward a participatory Mode of Consciousness? *Educational Researcher.* 23: 15-22

Hetzel, B. (1993). *Making Software Measurement Work: Building an Effective Measurement Program.* Boston: QED Publishing Group.

Higgins, J. (1983). Can Computers Teach? *CALICO Journal.* 1/2: 4-6

Higgins, J. (1988). *Language, Learners and Computers.* London: Longman.

Higgins, J. (1995). *Computers and English Language Learning.* Oxford: Intellect.

Hiller, S., Rooney, E., Vaughn, R., Eckert, M., Laver, J. and Jack, M. (1994) An automated system for computer-aided pronunciation learning. *Computer-assisted language learning.* 7 (1): 51-63

Hincks, R. (2002). Speech Synthesis for Teaching Lexical Stress. *TMH-QPSR.* 44: 153-165

Hirschman and Thompson (1996). Overview of Evaluation in Speech and Natural Language Processing. In Battista Varile, G., Zampoli, A., Cole, R., Mariani, J., Uszkoreit, H., and Zaenen, A. (eds.). (1996) *Survey of the State of the Art in Human Language Technology* (pp. 175-181). Cambridge University Press: Cambridge.

Hiyakumoto, L., Prevost, S., and Cassell, J. (1997). Semantic and Discourse Information for Text-to-Speech Intonation. In Alter, K., Pirker, H., Finker, W. (eds.) (1997). *Concept to Speech Generation Systems* (pp. 47-56). Madrid, Spain: Association for Computational Linguistics.

Holland, M. (1995). Introduction: The Case of Intelligent CALL. In Holland, M., Kaplan, J. and Sama, M. (eds.). *Intelligent Language Tutors: Theory Shaping Technology* (pp. vii-xvii). Hove, UK: Lawrence Erlbaum.

Horne, M. and Filipsson, M. (1997). Computational Extraction of Lexico-Grammatical Information for Generation of Swedish Intonation. In van Santen, J., Sproat, R., Olive, J., and Hirschberg, J. (eds.) (1997). *Progress in Speech Synthesis* (pp. 443-457). London: Springer Verlag.

House, A., Williams, C., Hecker, M., and Kryter, K. (1965). Articulation Testing Methods: Consonantal Differentiation with a Closed Response Set. *Journal of the Acoustical Society of America.* 37: 158-166

Howard-Jones, P. (1992). *SOAP, Speech Output Assessment Package. Version 4.0* (ESPRIT SAM-UCL-042). London: University College London.

Howatt, A. (1969). *Programmed Learning and the Language Teacher.* London: Longman.

Huang, X, Acero, X., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Upper Saddle River, New Jersey: Prentice Hall

Hubbard, P. (1987). Language Teaching Approaches, the Evaluation of CALL Software, and Design Implications. In Flint Smith, Wm. (ed.). (1987). *Modern Media in Foreign Language Education: Theory and Implementation* (pp. 227-254). Lincolnwood, Illinois: National Textbook Company.

ISO (1999) *Information Technology – Software Product Evaluation – Part 1: General Overview* (BS ISO/IEC 14598-1:1999). London: British Standards Institute (BSI).

JEIDA (Japanese Electronic Industry Development Association )(1995). *JEIDA Guideline for Speech Synthesizer Evaluation (Mar '95).* Retrieved from http://www.slt.atr.co.jp/cocosda/output/jeida2.txt

Izumi, S. and Bigelow, M. (2000). Does Output Promote Noticing and Second Language Acquisition? *TESOL Quarterly.* 34: 239-278

Izumi, S., Bigelow, M., Fujiwara, M., and Fearnow, S. (1999). Testing the Output Hypothesis: Effects of Output on Noticing and Second Language Acquisition. *Studies in Second Language Acquisition.* 21: 421-452

Jannedy, S., Poletto, R., & Weldon, T. L., (es.). (1994*). Language Files: Materials for an Introduction to Language and Linguistics* (6th ed.). Ohio: Ohio State University.

Jekosch U. (1992). The Cluster-Identification Test. In *Procs. ICSLP '92* (pp. 205-208). Banff, Alberta, Canada.

Jenkins, J. (2000). *The Phonology of English a an International Language.* Oxford: Oxford University Press.

Jenner, B. (1992). The English Voice. In Brown, A. (ed.) (1992). *Approaches to Pronunciation Teaching* (pp. 38-46). London: Macmillan.

Johnson, K. (2001). An Introduction to Foreign Language Learning and Teaching. Harlow: Longman.

Jones, C. (1986). It's not so much the Program, more what you do with it: The Importance of Methodology in CALL. *System.* 14 (2): 171-178

Jones, C. and Fortescue, S. (1987). *Using Computers in the Language Classroom.* Harlow: Longman.

Kumaravadivelu, B. (1994). Intake Factors and Intake Processes in Adult Language Learning. *Applied Language Learning.* 5: 33-71

Kelch, K. (1985). Modified Input as an Aid to Comprehension. *Studies in Second Language Acquisition.* 7: 81-90

Keller, E. (2002) Towards Greater Naturalness: Future Directions of research in Speech Synthesis. In Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 3-17). Chichester: John Wiley and Sons

Keller, E. and Zellner-Keller, B. (2000). Speech Synthesis in Language Learning: Challenges and Opportunities. In *Procs. InSTIL 2000* (pp. 109-116). Dundee, England: University of Abertay Dundee.

Kenworthy, J. (1987). *Teaching English Pronunciation.* London: Longman.

Klatt, D. (1987). Review of Text-to-Speech Conversion for English. *Journal of the Acoustical Society of America.* 82 (3): 73-793

Knowles, G. (1986). The Role of the Computer in the Teaching of Phonetics. In Leech, G. and Candlin, C. (eds.) (1986). *Computers in English Language Teaching and Research* (pp. 133-148). London: Longman.

Kommissarchick, J. and Kommissarchick, E. (2000). BetterAccent Tutor – Analysis and Visualization of speech Prosody. In *Procs. InSTIL 2000* (pp. 86-89). Dundee, England: University of Abertay Dundee.

Krashen, S. (1982). *Principles and Practice in Second Language Acquisition.* Oxford: Pergamon.

L'Haire, S. (2000). *L'Enseignement Assisté par Ordinateur et le Traitement Automatique du Langage Naturel.* Unpublished Masters Thesis. Faculté des Lettres, Université de Genève, Geneva, Switzerland.

L'Huilllier, M. (1990). Evaluation of CALL Programs for Grammar. *CALL.* 1: 79-86

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics.* London: Oxford University Press.

Lado, R. (1975). *Linguistics Across Cultures: Applied Linguistics for Language Teachers.* Ann Arbor: Michigan University Press.

Laroy, C. (1995). *Pronunciation.* Oxford: Oxford University Press.

Larsen-Freeman, D. (1976). An Explanation for the Morpheme Accuracy order of Learners of English as a Second Language. *Language Learning.* 26: 125-135

Larsen-Freeman, D. and Long, M. (1991). *An Introduction to Second Language Acquisition Research.* London: Longman.

Le Goff, B. and Benoît, C. (1996). A Text-to-Audiovisual-Speech Synthesizer for French. In *Procs. ICSLP ' 96* (pp. 2163-2166). Philadelphia.

LeBel, J.-G. (1990). *Traité de correction phonétique ponctuelle.* Université Laval: Les Editions de la Faculté des Lettres.

Lemmetty, S. (1999). *Review of Speech Synthesis Technology.* Unpublished Masters Thesis. Helsinki University of Technology, Helsinki, Finland. Retrieved from http://www.acoustics.hut.fi/~slemmett/dippa/contents.html

Léon, P. (1992). *Phonétisme et prononciations du français : avec des travaux pratiques d'applications et leurs corrigés.* Paris : Nathan.

Léon, P. and Léon, M. (1969). *Introduction à la phonétique corrective.* Paris: Hachette and Larousse.

Levelt, W. (1989). *Speaking: From Intention to Articulation.* London: MIT Press

Levy, M. (1997). *Computer-Assisted Language Learning: Context and Conceptualization.* Oxford: Clarendon.

Levy, M. (1999a). Design Processes in CALL: Integrating Theory, Research and Evaluation. In Cameron, K. (ed.) (1999). *CALL Media, Design & Applications* (pp. 83-107). Lisse: Swets & Zeitlinger.

Levy, M. (1999b). Responding to the Context of CALL: Directions for Research. *Prospect.* 14 (3): 24-31

Liberman, M. and Church, K. (1992). Text analysis and word pronunciation in text-to-speech synthesis. In Furui, S. and Sondhi, M. (eds.) (1992). *Advances in Speech Signal Processing* (pp. 791-831). New York: M. Dekker

Lightbown, P. and Spada, N. (1993). *How Languages are Learned.* Oxford: Oxford University Press.

Lindgaard, Gitte (1994). *Usability Testing and System Evaluation: A Guide for Designing Useful Computer Systems.* Chapman and Hall: London.

Linggard, R. (1985). *Electronic Speech Synthesis.* Cambridge: Cambridge University Press

Lively, S., Logan, J., and Pisoni, D. (1993). Training Japanese Listeners to Identify English /r/ and /l/. II: The Role of Phonetic Environment and Talker Variability in Learning new Perceptual Categories. *Journal of the Acoustical Society of America.* 94 (3): 1242-1255

Lively, S., Pisoni, D., Akahane-Yamada, R., Tohkura, Y., and Yamada, T. (1994). Training Japanese Listeners to Identify English /r/ and /l/. III. Long-Term Retention of New Phonetic Categories. *Journal of the Acoustical Society of America.* 96 (4): 2076-2087

Logan, J., Lively, S., and Pisoni, D. (1991). Trianing Japanese Listeners to Identify English /r/ and /l/: A First Report. *Journal of the Acoustical Society of America.* 89 (2): 874-886

Long, M. (1985). Input and Second Language Acquisition Theory. In Gass, S. and Madden, C. (eds.) (1985). *Input in Second Language Acquisition* (pp. 377-393). Rowley, MA: Newbury House.

Lopez de Ipina, K., Ezeiza, N., Bordel, G., and Gaña, M. (2002). Morphological Segmentation for Speech Processing in Basque. In *Procs. IEEE 2002 Workshop on Speech Synthesis* (pp. 187-190). Santa Monica, California.

Loschky, L. (1994). Comprehensible Input and Second Language Acquisition: What is the Relationship? *Studies in Second Language Acquisition.* 16: 303-324

MacCarthy, P. (1975). The *Pronunciation of French.* London: Oxford University Press.

Mackey, A. (1995). *Stepping up the Pace: Input, Interaction and Interlanguage Development, an Empirical Study of Questions in ESL.* Unpublished Doctoral Thesis. University of Sydney, Sydney, Australia.

Mackey, A. (1999). Input, Interaction and Second Language Development: An Empirical Study of Question Formation in ESL. *Studies in Second Language Acquisition.* 21: 557-88

Macon, M., Jesen-Link, L., Oliverio, J., Clements, M., and George, E. (1997). A Singing Voice Synthesis System Based on Sinusoidal Modeling. In *Procs. ICASSP '97* (Vol. 1, pp. 439-442). Munich, Germany.

MacWhinney, B. (1995). Evaluating Foreign Language Tutoring Systems. In Holland, M., Kaplan, J. and Sama, M. (eds.). *Intelligent Language Tutors: Theory Shaping Technology* (pp. 317-326). Hove, UK: Lawrence Erlbaum.

Massaro, D., and Cole, R. (2000). From "Speech is Special" to Talking Heads in Language Learning. In *Procs. InSTIL 2000* (pp. 153-161). Dundee: University of Abertay Dundee.

Massaro, D.W., Cohen, M.M., & Beskow, J. (1999). From theory to practice: rewards and challenges. *In Procs. International Conference of Phonetic Sciences* (pp. 1289-1292). San Francisco, CA: Regents of the University of California.

McGrath, J. (1995) Methodology Matters: Doing Research in the Behavioral and Social Sciences. In Baecker, R., Grundin, J., Buxton, W., and Greenberg, S. (eds.) (1995*). Readings in Human-Computer Interaction: Toward the Year 2000* (pp. 152-169). San Francisco, CA: Morgan Kaufman.

McLaughlin, B. (1987). *Theories of Second Language Acquisition.* London: Arnold.

McNaughton, D., Fallon, K., Tod, J., Weiner, F., and Neisworth, J. (1994). Effect of Repeated Listening Experiences on the Intelligibility of Synthesized Speech. *AAC Augmentative and Alternative Communication.* 10: 161-168

Mercier, G., Guyomard, M., Siroux, J., Bramoullé, A., Gourmelon, H., Guillou, A., & Lavannant, P. (2000). Courseware for Breton Spelling Pronunciation and Intonation Learning. In *Procs. InSTIL 2000* (pp. 145-148). Dundee, England: University of Abertay Dundee.

Mertens, P. Goldman, J.-P., Wehrli, E., and Gaudinat, A. (2001). La Synthèse de l'Intonation à Partir de Structures Syntaxiques Riches. *Traitement Automatique des Langues.* 42 (1): 142-195.

Mitchell, R., and Myles, F. (2004). *Second Language Learning Theories.* London: Arnold.

Moisa, T. and Ontanu, D. (1999). Learning Romanian Using Speech Synthesis. In Cumming, G., Toshio, O., and Louis, G. (eds.) (1999). *Advanced Research in Computers and Communications in Education: New Human Abilities for the Networked Society: Procs. ICCE '99, 7th International Conference on Computers in Education* (pp. 808-815). Oxford: IOS Press.

Möller, S., Jekosch, U. ETC ETC (2001). Auditory Assessment of Synthesized Speech in Application Scenarios: Two Case Studies. *Speech Communication.* 34: 229-246

Monaghan, A. (2002a). Mark-up for Speech Synthesis. In Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 307-319). Chichester: John Wiley and Sons.

Motteram, G. (1999). Changing the Research Paradigm: Qualitative Research Methodology and the CALL Classroom. In Debski, R. and Levy. M. (eds.) (1999). *WorldCALL* (pp. 201-213). Amsterdam: Swets & Zeitlinger.

Moulines, E. (1992). Synthesis Models: A Discussion. In Bailly, G. and Benoit, C. (eds.) (1992). *Talking Machines: Theories, Models, and Designs* (pp. 7-12). London: North-Holland.

Multitel (2005). eLite Documentation. Retrieved from http://www.multitel.be/TTS/layout.php?page=eLite_doc

Munro, M. and Derwing, T. (1999). Foreign Accent, Comprehensibility, and Intelligibiility in the Speech of Second Language Learners. In Leather, J. (ed.) (1999). *Phonological Issues in Language Learning* (pp. 285-310). Oxford: Blackwell.

Murray, G. (1999). Exploring Learners' CALL Experiences: A Reflection on Method. *Computer Assisted Language Learning.* 12 (3): 179-195

Murray, L. and Barnes, A. (1998). Boyond the "Wow" Factor – Evaluating Multimedia Language Learning Software from a Pedagogical Point of View. *System.* 26: 249-259

Myers, M. (2000). Voice Recognition Software used to Learn Pronunciation. In *Procs. InSTIL 2000* (pp. 97-100). Dundee, England: University of Abertay Dundee.

Nagano, K. and Ozawa, K. (1990) English Speech Training using Voice Conversion. In *Procs. ICSLP* (pp. 1169-1172). Kobe, Japan.

Nagata, N. (1998). Input vs. Output Practice in Educational Software for Second Language Acquisition. *Language Learning and Technology.* 1 (2): 23-40. Retrieved from: http://llt.msu.edu/vol1num2/article1/default.html

Nass C., and Min Lee, K. (2001). Does Computer-Synthesized Speech Manifest Personality? Experimetal Tests of Recognition, Similarity, and Consistency-Attraction. *Journal of Experimental Psychology: Applied.* 7 (3): 171-181

Nelson, T. and Oliver, W. (1999). Murder on the Internet. *CALICO Journal.* 17 (1): 101-114

Nobuyoshi, J. and Ellis, R. (1993). Focused Communication Tasks and Second Language Acuqisition. *ELT Journal.* 47: 203-210

Nye, P., and Gaitenby, J. (1974). The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences. *Haskins Laboratories Status Report on Speech research.* 37/38: 169-190

O'Connor, J. (1973). *Phonetics.* Harmondsworth, Middlesex, England: Penguin.

O'Shaughnessy, D. (1987). *Speech Communication: Human and Machine.* Wokingham, England: Addison-Wesley.

O'Shaughnessy, D. (1992). Spectral Transitions in Rule-Based and Diphone Synthesis. In Bailly, G. and Benoît, C. (eds.) (1992). *Talking Machines: Theories, Models, and Designs* (pp. 77-91). London: North-Holland.

Offord, M. (1990). *Varieties of Contemporary French.* London: Macmillan.

Olinsky, C. and Cummins, F. (2002). Iterative English Accent Adaptations in a Speech Synthesis System. In *Procs. IEEE 2002 Workshop on Speech Synthesis* (pp. 79-82). Santa Monica, California.

Olive, J., van Santen, J., Möbius, B., and Shih, C. (1998). Synthesis. In Sproat, R. (ed.) (1998). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach* (pp. 191-228). London: Kluwer Academic Publishers.

Oxford (2003). *Oxford Hachette French Dictionary.* Oxford: Oxford University Press

Oxford, R. (1993). FORUM: Intelligent Computers for Learning Languages: The View for Language Acquisition and Instructional Methodology. I. *CALL.* 6 (2): 173-188

Oxford, R. (1995). When Emotion Meets (Meta)Cognition in Language Learning histories. *International Journal of Educational Research.* 23: 581-94

Oxford, R., Holland, M., and Alosh, M. (1993). Intelligent Computers for Learning Languages: The View for Language Acquisition and Instructional Methodology. *Computer Assisted Language Learning.* 6 (2): 173-188.

Oxford-Hachette (2003). Oxford-*Hachette French Dictionary on CD ROM* (Version 2.0). Oxford: Oxford University Press

Peckels, J. and Rossi, M. (1973). Le test de diganostic par paires minimales: Adaptation au français du diagnostic rhyme test de W. D. Voiers. *Revue d'acoustique.* 27: 245-262

Pederson, K. (1987). Research on CALL. In Flint Smith, Wm. (ed.) (1987). *Modern media in Foreign Language Education: Theory and Implementation* (pp. 99-131). Lincolnwood, Illinois: National Textbook Company.

Pennington, M. (1996). *Phonology in English Language Teaching.* London: Longman.

Pennington, M. (1999). Computer-Aided Pronunciation Pedagogy: Promise, Limitations, and Directions. *Computer Assisted Language Learning.* 12 (5): 427-440

Pennington, M. and Esling, J. (1996). Computer-Assisted Development of Spoken Language Skills. In Pennington, M. (ed.) (1996). *The Power of CALL* (pp. 153-189). Houston, USA: Athelstan Publications.

Pfister, B. and Traber, C. (1994).Text-to-Speech Synthesis : An Introduction and Case Study. . In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 87-107). Chichester: John Wiley.

Phillips, M. (1987). Potential Paradigms and Possible Problems for CALL. *System.* 15 (3): 275-287

Pica, T., Young, R., and Doughty, C. (1987). The Impact of Interaction on Comprehension. *TESOL Quarterly.* 21: 737-758

Pisoni, D. (1978/9). Some Measures of Intelligibility and Comprehension. In Allen, J., Hunnicutt, M. S., and Klatt, D. with Armstrong, R.C., and Pisoni, D.B. (1987). *From Text to Speech: The MITalk System* (pp. 151-171). Cambridge: Cambridge University Press.

Pitt, I., and Edwards, A. (2003). *Design of Speech-Based Devices.* London: Springer.

Polkosky, M. and Lewis, J. (2003). Expanding the MOS: Development and Psychometric Evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology.* 6: 161-182

Portele, T., Höfer, F., and Hess, W. (1997). A Mixed Inventory Structure for German Concatenative Synthesis. In van Santen, J., Sproat, R., Olive, J., and Hirschberg, J. (eds.) (1997). *Progress in Speech Synthesis* (pp. 263-277). London: Springer Verlag.

Probst, K., Ke, Y., and Eskenazi, M. (2002). Enhancing Foreign Language Tutors – In Search of the Golden Speaker. *Speech Communication.* 37: 161-173

Prudon, R.., d'Alessandro, C., and Boula de Mareuïl, P. (2002). Prosody Synthesis by Unit Selection and Transplantation on Diphones. In *Procs. IEEE 2002 Workshop on Speech Synthesis* (pp. 119-122). Santa Monica, California.

Ralston, A., Reilly, E. D., & Hemmdinger, D. (Eds.). (2000). *Encyclopedia of Computer Science.* London: Nature Publishing Group.

Rank, E. (2002). Concatenative speech synthesis using SRELP. . In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 76-85). Chichester: John Wiley.

Reeser, T. (2002). Review of Tell Me More – French. *CALICO Journal.* 19 (2): 419-428.

Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat Computers, Televisions, and New Media like Real People and Places.* Cambridge: Cambridge University Press.

Reynolds, M., Bond, Z., and Fucci, D. (1996). Synthetic Speech Intelligibility: Comparison of Native and Non-native Speakers of English. *AAC Augmentative and Alternative Communication.* 12: 32-36

Reynolds, M., Isaacs-Duvall, C., Sheward, B., and Rotter, M. (2000). Examination of the Effects of Listening Practice on Synthesired Speech Comprehensibility. *AAC Augmentative and Alternative Communication.* 16: 250-259

Rochet, B. (1990). Training Non-Native Speech Contrasts on the Macintosh. In Craven, M.-L. Signor, R. and Paramskas, D. (eds.) (1990). *CALL: Papers and Reports* (pp. 119-126). La Jolla, CA: Athelstan.

Rodman, R. (1999). *Computer Speech Technology.* London: Artech House.

Rodriguez Banga, E., Garcia Mateo, C. and Frenandez Salgado, X. (2002). Concatenative Text-to-Speech Synthesis Based on Sinusoidal Modelling. In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 52-63). Chichester: John Wiley.

Rost, M. (2001). *Teaching and Researching Listening.* London: Longman.

Sachs, J. (1977). The Adaptive Significance of Linguistic Input to Prelinguistic Infants. In Snow and Ferguson (eds.) (1977). *Talking to Children: Language Input and Acquisition.* Cambridge: Cambridge University Press.

Sagisaka, Y. (1990). Speech Synthesis from Text. *IEEE Communications Magazine*. January 1990: 35-41

Sato, C. (1985). Task Variation in Interlanguage Phonology. In Gass, S. and Madden, C. (eds.) (1985). *Input in Second Language Acquisition* (pp.181-96). Newbury House, Rowley, Mass.

Sato, C. (1991). Sociolinguistic Variation and Language Attitudes in Hawaii. In Cheshire, J. (ed.) (1991). *English Around the World: Sociolinguistic Perspectives* (pp. 647-663). Cambridge: Cambridge University Press

Schmidt, R. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*. 11 (2): 129-158

Schmidt, R. and Frota, S. (1986). Developing Basic Conversational Ability in a Second Language: A Case Study of an Adult Learner of Portuguese. In Day, R. (ed.). *Talking to Learn: Conversation in Second Language Acquisition* (pp. 237-326). Rowley, Mass.: Newbury House.

Schmidt-Nielsen, A. (1995). Intelligibility and Acceptability Testing for Speech Technology. In Syrdal, A., Bennett, R., and Greenspan, S. (eds.) *Applied Speech Technology* (pp. 195-231). Boca Raton: CRC.

Scholfield, P. and Ypsiladis, G. (1992). Evaluating Computer Assisted Language Learning from the Learners' Point of View. In Graddol, D. and Swann, J. (eds.) (1992). *Evaluating Language.* Clevedon: Multilingual Matters Ltd.

Schomaker, L., Nijtmans, J., Camurri, Lavagetto, A.F., Morasso, P., Benoît, C., Guiard-Marigny, T., Le Goff, B., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K., and Blauert, J. (1995). Audio-Visual Speech Synthesis. In Schomaker, L., Nijtmans, J., Camurri, Lavagetto, A.F., Morasso, P., Benoît, C., Guiard-Marigny, T., Le Goff, B., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K., and Blauert, J. (eds.) (1995). *A Taxonomy of Multimodal Interaction in the Human Information Processing System* (ESPRIT Project 8575 WP1). Retrieved from http://hwr.nici.kun.nl/~miami/taxonomy/taxonomy.html

Schroeter, J. (2001). The Fundamental of Text-to-Speech Synthesis. *Voice XML Review*. 1 (3). Retrieved from http://www.voicexmlreview.org/Mar2001/features/tts.html

Schroeter, J., Conkie, A., Syrdal, A., Beutnagel., M., Juka, M., Strom, V., Kim, M-J., Kang, H.-G., and Kapllow, D. (2002). A Perspective on the Next Challenges for TTS Research. In *Procs. IEEE 2002 Workshop on Speech Synthesis* (pp. 211-214). Santa Monica, California.

Searle, J. (1976). A Classification of Illocutionary Acts. *Language in Society.* 5: 1-23

Seneff, S., Wang, C., and Zhang, J. (2004). Spoken Conversational Interaction for Language Learning. In *Procs. InSTIL/ICALL 2004 – NLP and Speech Technologies in Language Learning Systems* (pp. 151-154). Venice, Italy.

Shadle, C. H. and Damper, R. I. (2002) Prospects for Articulatory Synthesis: A Position Paper. In *Procs. 4th ISCA Workshop on Speech Synthesis* (pp. 121-126). Pitlochry, Scotland.

Sharwood Smith, M. (1991). Speaking to Many Minds: On the Relevance of Different Types of Language Information for the L2 Learner. *Second Language Research.* 7: 118-132

Sherwood, B. (1981). Speech Synthesis Applied to Language Teaching. *Studies in Language Learning.* 3: 175-181

Sim, S. E., Easterbrook, S., & Holt, R. C. (1998). Using Benchmarking to Advance Research: A Challenge to Software Engineering. In *Proceedings of the 25th International Conference on Software Engineering* (pp. 74-83). Portland, OR.

Sioufi, N. (2000). Lessons from Auralog, Auralang Lessons. In *Procs. InSTIL 2000* (pp. 117-119). Dundee, England: University of Abertay Dundee.

Skehan, P. (1998). *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

Skrelin, P. and Volskaya, N. (1998). Application of New Technologies in the Development of Education Programs. In Jager, S., Nerbonne, J., and van Essen, A. (eds.) (1998). *Language Teaching and Language Technology* (pp. 21-24). Lisse, The Netherlands: Swets and Zeitlinger.

Slobin, D. (1973). Cognitive Prerequisites for the Development of Grammar. In Ferguson, C. and Slobin, D. (eds.) (1973). *Studies of Child Language Development.* New York: Holt, Reinhart, and Winston.

Smith, L. and Nelson, C. (1985). International Intelligibility of English: Directions and Resources. *World Englishes.* 4: 333-342

Sobkowiak, W. (1998). Speech in EFL CALL. In Cameron, K (ed.) (1998). *Multimedia CALL: Theory and Practice.* Exeter: Elm Bank.

Sobkowiak, W. (2005). Pronunciation in EFL CALL. *Teaching English with Technology A Journal for Teachers of English.* 5 (1) Retrieved from: http://www.iatefl.org.pl/call/j_article20.htm#sob

Sondhi, M. (2002). Articulatory Modelling: A Possible Role in Concatenative Text-to-Speech Synthesis. In *Procs. 2002 IEEE Workshop on Speech Synthesis* (pp. 73-78). Santa Monica, California.

Spada, N., and Lightbown, P. (1993). Instruction and the Development of Questions in L2 Classrooms. *Studies in Second Language Acquisition.* 15: 205-224

Sparck Jones, K., & Galliers, J. R. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review.* London: Springer.

Spencer, Richard H. (1985). *Computer Usability Testing and Evaluation.* Prentice Hall: London.

Spiegel, M., Altom, M., Machhi, M., and Wallace, K. (1990). Comprehensive Assessment of the Intelligibility of Synthesized and Natural Speech. *Speech Communication.* 9: 279-291

Sproat, R. (1996). Text Interpretation for TtS Synthesis. In Battista Varile, G., Zampoli, A., Cole, R., Mariani, J., Uszkoreit, H., and Zaenen, A. (eds.) (1996). *Survey of the State of the Art in Human Language Technology* (pp. 175-180). Cambridge: Cambridge University Press.

Sproat, R., Black, A., Chen, S., Kumar, S., Ostendorf, M. and Richards, C. (2001). Normalization of Non-Standard Words. *Computer Speech and Language.* 1: 287-333.

Sproat, R., Möbius, B., Maeda, K. and Tzoukerman, E. (1998). Multilingual Text Analysis. In Sproat, R. (ed.) (1998). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach* (pp. 31-87). London: Kluwer Academic Publishers.

Stevens, K. N. (1992). Speech Synthesis Methods: Homage to Dennis Klatt. In Bailly, G. and Benoît, C. (eds.) (1992). *Talking Machines: Theories, Models, and Designs* (pp. 3-6). London: North-Holland.

Stevens, K. N. (2002). Toward Formant Synthesis with Articulatory Controls. In *Procs. 2002 IEEE Workshop on Speech Synthesis* (pp. 67-70). Santa Monica, California.

Stevens, V. (1989). A Direction for CALL: From Behaviouristic to Humanistic Courseware. In Pennington, M. (ed.) (1989). *Teaching Languages With Computers: The Sate of the Art* (pp. 31-43). La Jolla, CA: Athelstan.

Strange, W. (1999a). Perception of Vowels: Dynamic Constancy. In Pickett, J. (ed.) (1999). *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology* (pp. 153-165). London: Allyn and Bacon.

Strange, W. (1999b). Perception of Consonants: From Variance to Invariance. In Pickett, J. (ed). (1999). *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology* (pp. 166-182). London: Allyn and Bacon.

Stratil, M., Burkhardt, D., Jarratt, P., & Yandle, J. (1987a) Computer-Aided Language Learning with Speech Synthesis: User Reactions. *Programmed Learning and Educational Technology*. 24(4): 309-316.

Stratil, M., Weston, G., & Burkhardt, D. (1987b). Exploration of Foreign Language Speech Synthesis. *Literary and Linguistic Computing*. 2(2): 116-119.

Styger, T. and Keller, E. (1994). Formant Synthesis. In Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds.) (2002). *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (pp. 109-128). Chichester: John Wiley.

Swain, M. (1985). Communicative Competence: Some Roles of Comprehensible Input and Comprehensive Output in Its Development. In Gass, S. and Madden, C. (eds.) (1985). *Input in Second Language Acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Swartz, M., Kostyla, S., Hanfling, S., and Holland, M. (1990). Preliminary Assessment of a Foreign Language Learning Environment. *CALL*. 1: 51-64

Sweet, H. (1877). *A Handbook of Phonetics*. Oxford: Clarendon Press.

Takeda, K., Abe, K., and Sagisaka, Y. (1992). On the Basic Scheme and Algorithms in Non-Uniform Unit Speech Analysis. In Bailly, G. and Benoît, C. (eds.) (1992). *Talking Machines: Theories, Models, and Designs* (pp. 93-105). London: North-Holland

Tatham, M. (1993). Voice Output for Man-Machine Interaction. In Baber, C. and Noyes, J. (eds.) (1993). *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers* (pp. 25-35). London: Taylor and Francis.

Taylor, M. (ed.) (1980). The *Computer in the School: Tutor, Tool, Tutee*. New York: Teacher's College Press, New York.

Taylor, M. and Perez, L. (1989). Something to do on Monday. La Jolla, CA: Athelstan.

Tench, P. (1992). Phonetic Symbols in the Dictionary and in the Classroom. In Brown, A. (ed.) (1992). *Approaches to Teaching Pronunciation* (pp. 90-102). London: Macmillan.

Tench, P. (1996). *The Intonation Systems of English*. London: Cassell.

Theune, M. (2000). *From Data to Speech: Language Generation in Context.* Eindhoven: University of Eindhoven.

Thomas, C., Levinson, M., and Lessard, G. (2004). Experiments in Prosody for the Generation of Oral French. In *Procs. InSTIL/ICALL 2004 – NLP and Speech Technologies in Language Learning Systems* (pp. 123-126). Venice, Italy.

Thorndike, R. and Hage, E. (1977) *Measurement and Evaluation in Psychology and Education* (4th ed.). New York: Wiley.

TMA Associates (2003). *Nuance: US English.* Retrieved from http://www.tmaa.com/tts/Nuance_USEng.htm

Todaka, Y. (1990). *An Error Analysis of Japanese Students' Intonation its Pedagogical Applciations.* Unpublished Masters Thesis. University of California, Los Angeles, USA.

Trask, R. (1995). *Language: The Basics.* London: Routledge.

Underhill, A. (1985). Working with the Monolingual Learners' Dictionary. In Ilson, R. (ed.) (1985). *Dictionaries, Lexicography and Language Learning.* London: Pergamon/British Council.*

van Bezooijen, R., & van Heuven, V. J. (1997). Assesment of Synthesis Systems. In D. Gibbon, R. Moore, & R. Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems* (pp. 481-563). New York: Mouton de Gruyter.

van Bezooijen, R., and van Heuven, V. (1997) Assessment of Synthesis Systems. In Gibbon, D. Moore, R. and Winski, R. (eds.) (1997). *Handbook of Standards and Resources for Spoke Language Systems* (pp. 481-563). New York: Walter de Gruyter Publishers.

van Santen, J. (1993). Perceptual Experiments for Diagnostic Testing of Text-to-Speech Systems. *Computer Speech and Language.* 7: 49-100

van Santen, J., Wouters, J., Kain, A. (2002). Modifications of Speech: A Tribute to Mike Macon. In *Procs. IEEE Workshop on Speech Synthesis* (pp. 1-6). Santa Monica, California.

VanPatten, B. and Cadierno, T. (1993). Explicit Instruction and Input Processing. Studies in *Second Language Acquisition.* 15: 225-243

VanPatten, B. and Santz, C. (1995). From Input to Output: Processin Instruction and Communicative Tasks. In Eckman, F., Highland, D., Lee, P., Mileham, J., and Weber, R. (eds.) (1995). *Second Language Acquisition Theory and Pedagogy* (pp. 169-186). Hillsdale, NJ: Lawrence Erlbaum.

Venkatagiri, H. (1994). Effect of Sentence Length and Exposure on the Intelligibility of Synthesized Speech. *AAC Augmentative and Alternative Communication.* 10: 96-104

Venkatagiri, H. (2003). Segmental Intelligibility of Four Currently used Text-to-Speech Synthesis Methods. *Journal of the Acoustical Society of America.* 113 (4): 2095-2104

Voiers, W. (1983). Evaluating Processed Speech Using the Diagnostic Rhyme Test. *Speech Technology.* Jan/Feb: 0-9

Wallace, C. (1992). *Reading.* Oxford: Oxford University Press.

Warschauer, M. (1996). Computer-Assisted Language Learning: An Introduction. In Fotos, S. (ed.) (1996). *Multimedia Language Teaching* (pp. 3-20). Tokyo: Logos International.

Warschauer, M. and Healey, D. (1998). Computers and Language Learning: An Overview. *Language Teaching.* 31: 57-71

Weizenbaum, J. (1976). *Computer Power and Human Reason.* San Francisco: W H Freeman

Wessels, C. and Lawrence, K. (1992). Using Drama Voice Techniques in the Teaching of Pronunciation. In Brown, A. (ed.) (1992). *Approaches to Pronunciation Teaching* (pp. 29-37). London: Macmillan.

White, J. (2003). How to Evaluate Machine Translation. In Somers, H. (ed.) (2003). *Computers and Translation: A Translator's Guide* (pp. 211-242). Amsterdam: John Benjamins.

Witten, I. (1982). *Principles of Computer Speech.* London: Academic Press.

Wode, H. (1981). *Learning a Second Language 1: An Integrated View of Language Acquisition.* Tübingen: Gunter Narr.

Wyatt, D. (1988). Applying Pedagogical Principles to CALL Courseware Development. In Flint Smith, Wm. (ed.) (1988). *Modern Media in Foreign Language Education: Theory and Implementation* (pp. 86-98). Lincolnwood, Illinois: National Textbook Company.

Yarowsky, D. (1997). Homograph Disambiguation in Text-to-Speech Synthesis. In van Santen, J., Sproat, R., Olive, J., and Hirschberg, J. (eds.) (1997). *Progress in Speech Synthesis* (pp. 159-175). London: Springer Verlag.

Yoram, M. and Hirose, K. (1996). Language Training System Utilizing Speech Modification. In *Procs. ICSLP* (pp. 1449-1452). Phliadelphia.

Zue, V., Cole, R., and Ward, W. (1996). Speech Recognition. In Battista Varile, G., Zampoli, A., Cole, R., Mariani, J., Uszkoreit, H., and Zaenen, A. (eds.) (1996). *Survey of the*

*State of the Art in Human Language Technology* (pp. 3-9). Cambridge: Cambridge University Press.

# Appendix 1 The CEF

In this appendix the levels of general communicative competence distinguished in the CEF are presented.

Table 100 presents the levels of general communicative competence distinguished in the CEF.

**Table 100 Levels of general communicative competence in the CEF framework (Council of Europe, 2001: 24).**

| | | |
|---|---|---|
| **Proficient User** | **C2** | *Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning in more complex situations.* |
| | **C1** | *Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.* |
| **Independent User** | **B2** | *Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.* |
| | **B1** | *Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.* |
| **Basic User** | **A2** | *Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring simple and direct exchange of information on familiar routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.* |
| | **A1** | *Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.* |

# Appendix 2 Corpora

In this appendix, the corpora used in the main investigation is presented. The familiarisation passage is presented in section A2.1, the RM corpus in section A2.2, the phonetic PM corpus in section A2.3, the prosodic PM corpus in section A2.4, and the CP corpus in section A2.5.

## A2.1 Familiarisation Passage

On disait dans le livre : « Les serpents boas avalent leur proie tout entière,

[ɔ□ dizɛ da□ lәlivr lesɛrpa□ bɔa avallœrprwatuta□ ɛr

sans la mâcher. Ensuite ils ne peuvent plus bouger et ils dorment pendant les

sa□ lamaʃe a□ sɥit ilnәpøvplybuʒe eildɔrm pa□ da□ le

six mois de leur digestion. »

simwa dәlœrdiʒɛstjɔ□

J'ai alors beaucoup réfléchi sur les aventures de la jungle et à mon tour j'ai

ʒealɔr bokurefleʃi syrlezava□ tyr dәlaʒ{œ□ /ɔ□ }gl eamɔ□ tur ʒe

réussi, avec un crayon de couleur, à tracer mon premièr dessin.

reysi avɛkœ□ krɛjɔ□ dәkulœr atrase mɔ□ prәmjɛdesɛ□ ]

Source: *Le Petit Prince* (de Saint-Exupéry, 1999: 13)

## A2.2 RM corpus

Je suis très vieux. Plus de deux fois centenaire, je demeure néanmoins très solide, parce que bâti de fer recouvert de bronze. Mon existence n'est pas sans aventures. Mon premier maître fut un capitaine de bateau qui faisait la navette entre l'Europe et l' Asie.

Source: *Le Vieux Lit, FreeText* (Hamel, 2003b).

## A2.3 Phonetic PM corpus

[R] dehors, intérieur, rat, tiroir, règle ...

[ ɛ ] elle, mais, ver, verres, verre ...

[ ɛ ] voudrais, éclairs, lait, vrai, paquet ...

[f] parfois, fais, suffira, affaire, boeuf ...

[e] et, marié, assez, manger, habillés ...

[ʃ] supermarché, chez, boucher, brochettes, achèterons ...

[w] droit, reçois, envoi, boîte, recevoir ...

[d] vide, grande, regarde, indiscrète ...

[a] va, baskets, parle, appartement, jardin ...

[ɛ□ ] vingt, invités, combien, maintenant, besoin ...

Source: *Phonetics Exercises, Talk To Me: The Conversation Method (French)* (Auralog, 2002)

## A2.4 Prosodic PM corpus

Je ne m'en souviens pas !

Jérémie, tu viens ?

L'ensemble résidentiel ?

Non, je suis très lente.

Effectivement, je préfère le bus.

Vous appelez ça être en vacances !

J'ai une mauvaise toux.

Elle est dans l'allée suivante.

Est-elle enceinte ?

Deux livres. Je voudrais un sac s'il vous plaît.

Source: *Sentence Pronunciation, Talk To Me: The Conversation Method (French)* (Auralog, 2002).

## A2.5 CP corpus

Vous êtes assise à votre bureau, l'air très sérieuse ... Que faites-vous ?

- J'écris une lettre !

> Vous écrivez souvent ?

- J'écris seulement pour Noël et le jour de l'an.

> Vous avez assez d'enveloppes ?

- J'ai encore un gros paquet.

> J'espère que vous avez assez de timbres !

Source: *Faire son courier*, *Dialogue: Comprehension*, *Talk To Me: The Conversation Method (French)* (Auralog, 2002).

# Appendix 3 MOS-CALL

In this appendix, the French translation of *MOS-CALL* is presented.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ***L'adéquation et l'acceptabilité de la parole*** | | | | | | | | | |
| Adéquation | Est-ce que la parole de synthèse est adéquate dans son utilisation comme lecteur de texte (par rapport à d'autres médias)? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Pas du tout adéquate | | | | | | | | Très adéquate |
| Acceptabilité | Est-ce que la parole de synthèse est acceptable dans son utilisation comme lecteur de texte (lorsqu'il n'est pas possible d'utiliser d'autres médias)? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Très inacceptable | | | | | | | | Très acceptable |
| ***La qualité de la parole*** | | | | | | | | | |
| Compréhensibilité | Le message lu, est-il facile à comprendre? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Très difficile | | | | | | | | Très facile |
| Intelligibilité | Est-ce que les phonèmes/sons et mots individuels sont faciles à reconnaître (et à discriminer les uns des autres) ? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Très difficile | | | | | | | | Très facile |
| Choix de Prononciation | Est-ce que la prononciation est juste? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Incorrecte | | | | | | | | Correcte |
| Précision des phonèmes | Est-ce que l'articulation des phonèmes/sons est précise? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Très imprécise | | | | | | | | Très précise |
| Prosodie | Est-ce que la prosodie (musicalité) de la phrase est appropriée? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Très inappropriée | | | | | | | | Très appropriée |
| Caractère naturel des phonèmes/sons | Est-ce que les phonèmes/sons sonnent naturels/humains? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Pas du tout naturels/humains | | | | | | | | Très naturels/humains |
| Caractère naturel de la prosodie | Est-ce que la prosodie (musicalité) sonne naturelle/humaine? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Pas du tout naturelle/humaine | | | | | | | | Très naturelle/humaine |
| Caractère naturel de la voix | Est-ce que la voix sonne naturelle/humaine? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Pas du tout naturelle/humaine | | | | | | | | Très naturelle/humaine |
| Expressivité | Est-ce que les émotions sont bien exprimées? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Très mal exprimées | | | | | | | | Très bien exprimées |
| Convenance du registre | Est-ce que le registre est approprié? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Très inapproprié | | | | | | | | Très approprié |
| Caractère agréable de la voix | Est-ce que la voix est agréable à écouter? | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | Très désagréable | | | | | | | | Très désagréable |

# Appendix 4 Questionnaire

In this appendix, the questionnaire that was used to probe the variables that it was thought might affect the participants' ratings of TTS synthesis for use in CALL applications is presented.

***Vos impressions et votre expérience de la synthèse de parole***

1. Avant de participer à cette expérience, connaissiez-vous la synthèse de parole ?

   ☐ Oui

   ☐ Non (Passez à question 3)

2. Si oui, pensiez-vous que la synthèse de parole en général était prête pour utilisation dans les logiciels d'ALAO ?

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Non, pas du tout prête | | | | | | | | Oui, tout à fait prête |

3. Maintenant que vous avez fait cette expérience, comprenez-vous mieux ce qu'est la synthèse de parole ?

   ☐ Oui

   ☐ Non

4. Maintenant que vous avez complété t cette expérience, pensez-vous que la synthèse de parole en général est prête pour utilsation dans les logiciels d'ALAO dans les contextes suivants :

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|---|
| Lecteur de texte | Non, pas du tout prête | | | | | | | | Oui, tout à fait prête |
| Modèle de prononciation au niveau segmental (du son/mot) | Non, pas du tout prête | | | | | | | | Oui, tout à fait prête |
| Modèle de prononciation au niveau suprasegmental (de la phrase) | Non, pas du tout prête | | | | | | | | Oui, tout à fait prête |
| Partenaire de conversation | Non, pas du tout prête | | | | | | | | Oui, tout à fait prête |

5. Avez-vous participé à notre première expérience ?

   ☐ Oui
   ☐ Non

6. Avez-vous déjà participé à d'autres évaluations sur la parole de synthèse ?

   ☐ Oui
   ☐ Non

7. Aviez-vous déjà été en contact avec (écouté) la parole de synthèse ?

   ☐ Oui (Passez à question 8)
   ☐ Non (Passez à question 10)

8. Si oui, dans quel(s) contexte(s) ou quelle(s) application(s)?

- ☐ Annuaires automatisés
- ☐ Journaux vocaux
- ☐ Livres parlants
- ☐ Jeux parlants
- ☐ Serveurs vocaux d'informations (p.ex. horaires de train/de cinéma, etc.)
- ☐ Domotique (appareils domiciliaires automatisés
- ☐ Autre (veuillez spécifier)

_____

_____

9. Et, en moyenne, a quel rythme utilisez-vous ces applications?

- ☐ Tous les jours
- ☐ Quelques fois par semaine
- ☐ Une fois par semaine
- ☐ Quelques fois par mois
- ☐ Une fois par mois
- ☐ Moins d'une fois par mois
- ☐ Jamais

***Vos impressions et votre expérience de la synthèse de parole en ALAO***

10. Connaissez-vous des logiciels d'ALAO qui emploient de la synthèse de parole à partir du texte?

- ☐ Oui (Passez à question 11)
- ☐ Non (Passez à question 12)

11. Si oui, quels sont ces logiciels?

_____

_____

_____

12. Qu'est-ce que vous faites dans la vie?

- ☐ Professeur de français langue étrangère/seconde (Passez à question 13)
- ☐ Ingénieur de recherche en ALAO (Passez à question 19)
- ☐ Autre (veuillez spécifier) (Passez à question 19)

13. Aux apprenants de quelle(s) langue(s) maternelle(s) enseignez-vous le français ?

- ☐ Allemand
- ☐ Anglais
- ☐ Espagnol
- ☐ Grec
- ☐ Italien
- ☐ Portugais
- ☐ Autre (veuillez spécifier)

_____

_____

14. A quel(s) niveau(x) enseignez-vous le français?

- ☐ Débutant
- ☐ Faux debutant
- ☐ Moyen
- ☐ Avancé
- ☐ Très avancé
- ☐ Tous les niveaux
- ☐ Autre (veuillez spécifier)

_____
_____

15. Utilisez-vous des logiciels d'ALAO dans vos cours de français?

- ☐ Oui (Passez à question 16)
- ☐ Non (Passez à question 17)

16. Utilisez-vous des logiciels d'ALAO qui exploitent la synthèse de parole à partir du texte dans vos cours de français?

- ☐ Oui
- ☐ Non

17. Avez-vous déjà recommandé à vos élèves l'utilisation de logiciels d'ALAO pour leur pratique du français (hors de la classe)?

- ☐ Oui (Passez à question 18)
- ☐ Non (Passez à question 19)

18. Avez-vous déjà recommandé à vos élèves l'utilisation de logiciels d'ALAO qui exploitent la synthèse de parole pour leur pratique (hors de la classe)?

- ☐ Oui
- ☐ Non

19. Pouvez-vous penser à d'autres emplois potentiels de la synthèse de parole en ALAO/en apprentissage du français dont on n'a pas fait mention dans cette expérience? Veuillez spécifier.

_____
_____
_____
_____
_____
_____

***Vos détails personnels***

20. Age :

- ☐ Moins de 25
- ☐ 25-34
- ☐ 35-44
- ☐ 45-55
- ☐ Plus que 55

21. Sexe :

- ☐ M
- ☐ F

22. Quelle est votre langue première?

☐ Français
☐ Autre (Veuillez spécifier)

_____

23. D'où venez-vous?

☐ Belgique
☐ Canada
☐ France
☐ Etats-Uni
☐ Royaume Uni
☐ Autre (veuillez spécifier)

_____

***Vos commentaries***

24. Avez-vous des commentaires?

_____
_____
_____
_____
_____
_____
_____
_____
_____

# Appendix 5 On-line presentation of the investigation

In this appendix, the website where the investigation was hosted is presented in flat-form.

# Evaluation de la synthèse de parole pour emploi dans les logiciels d'ALAO

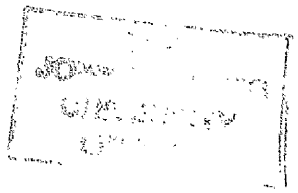
# Expérience #2

Merci beaucoup de vous être porté(e) volontaire pour participer à cette expérience qui a pour but d'identifier et de classer en ordre d'importance les qualités que les applications d'Apprentissage des Langues Assisté par Ordinateur (ALAO) exigent de la synthèse de parole.

Dans les pages qui suivent :

- Nous vous présentons une brève introduction à l'utilisation de la synthèse de parole en ALAO.
- Nous vous demanderons d'évaluer la parole produite par 4 systèmes de synthèse de parole pour utilisation dans les logiciels d'ALAO.
- Nous vous demanderons de remplir un questionnaire sur vos impressions de la synthèse de parole pour utilisation en ALAO.

Page suivante -->



# Qu'est-ce que la synthèse de parole?

De manière générale, les systèmes de synthèse de parole sont des logiciels qui permettent la génération automatique de messages oraux nouveaux/originaux (qui n'ont pas été pré-enrégistrés).

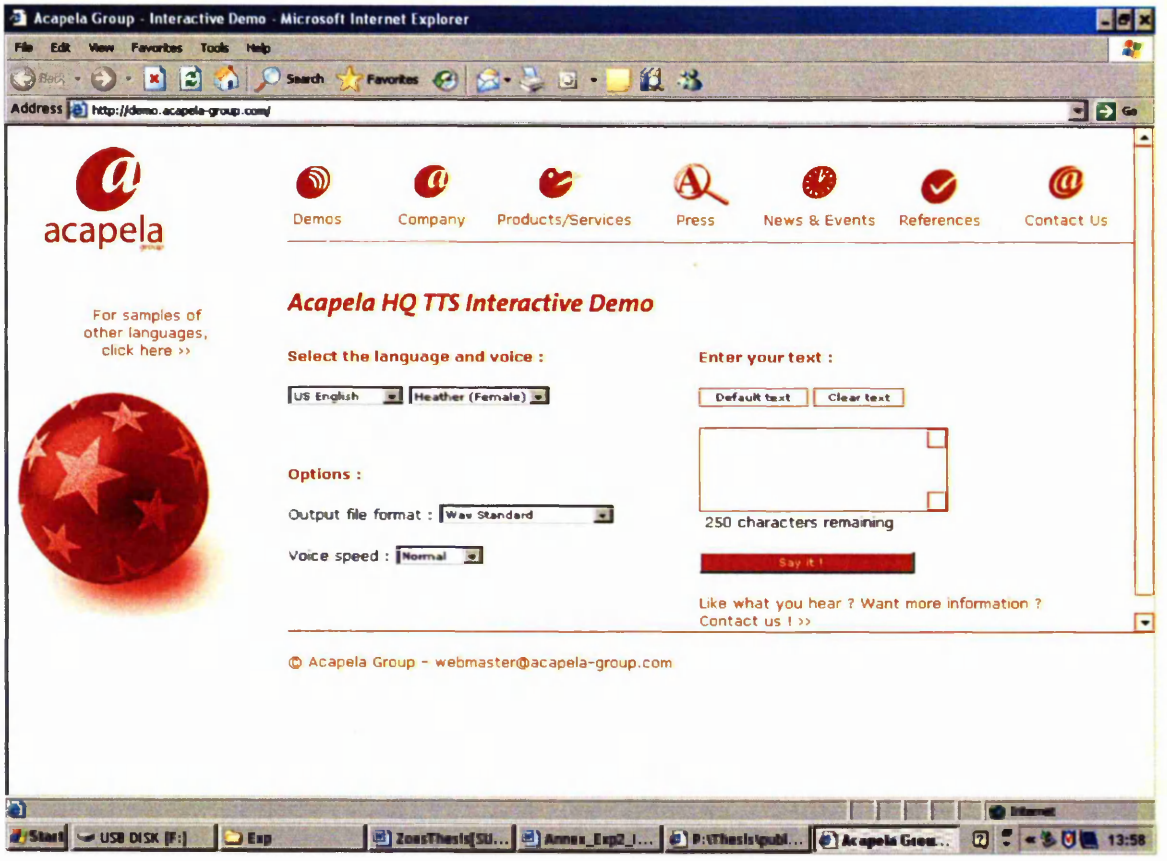
Nous nous intéressons en particulier aux systèmes de synthèse de parole à partir du texte. Ce sont des logiciels qui permettent la production des sons de la parole à partir d'une représentation orthographique du message. On les appelle des machines à parler.

Pour mieux comprendre ce que nous entendons par la synthèse de parole à partir du texte, vous pouvez explorer un des sites suivants et testez un des systèmes.

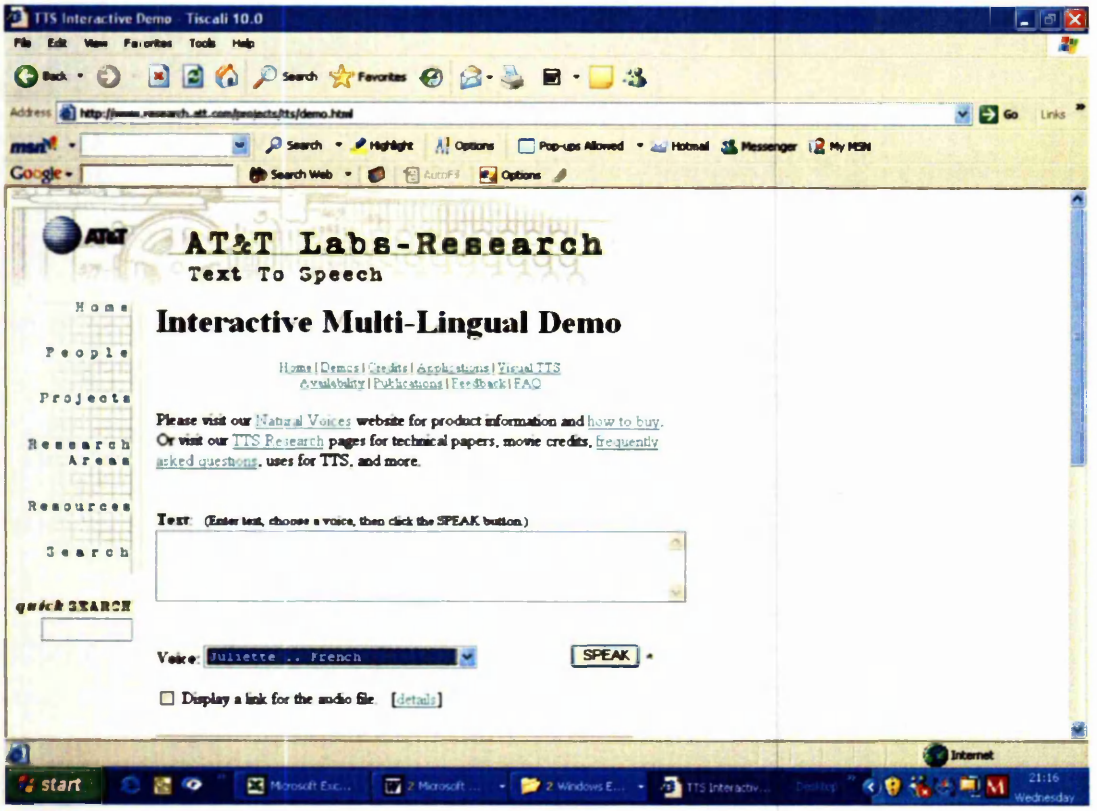
- Acapela HQ TTS
  (http://demo.acapela-group.com/)
- AT&T TTS
  (http://www.research.att.com/projects/tts/demo.html)
- Nuance Vocalizer
  (http://www.nuance.com/prodserv/demo_vocalizer.html)

<-- Page précédente Page suivante -->

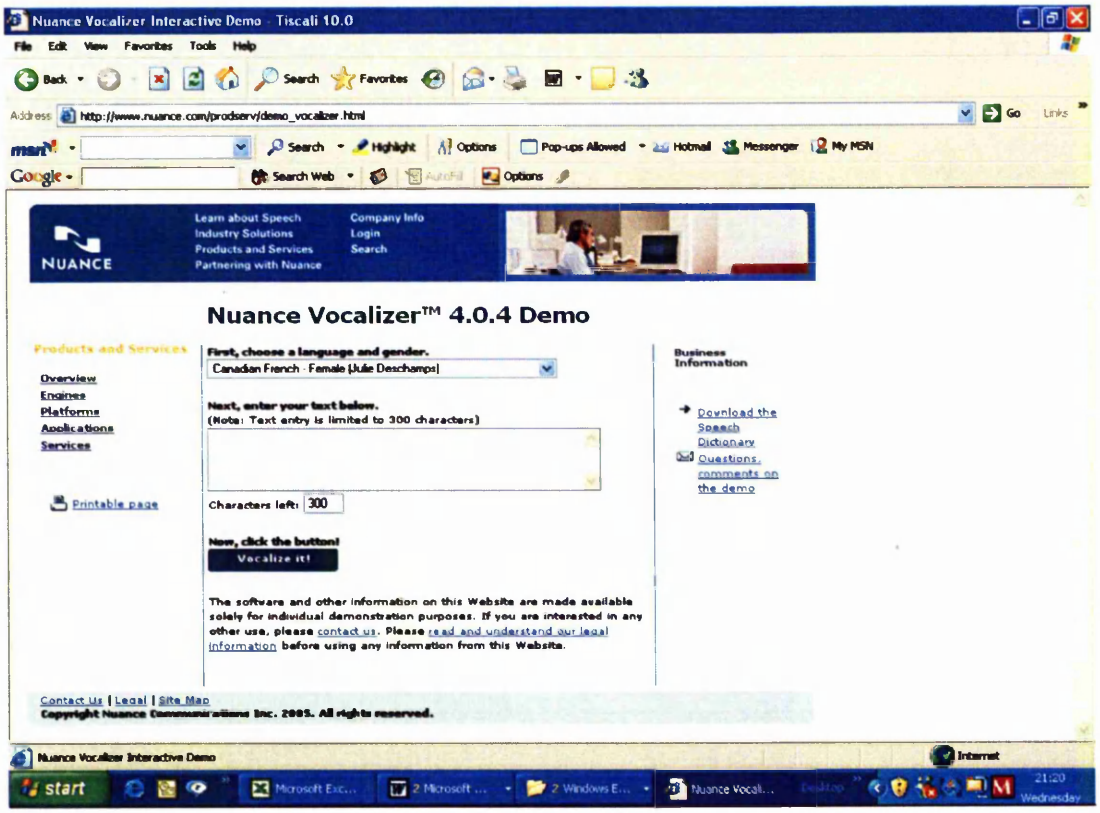
# Acapela HQ TTS



# AT&T TTS


TTS Interactive Demo - Tiscali 10.0
AT&T Labs-Research
Text To Speech
Interactive Multi-Lingual Demo
Please visit our Natural Voices website for product information and how to buy.
Or visit our TTS Research pages for technical papers, movie credits, frequently asked questions, uses for TTS, and more.
TEXT: (Enter text, choose a voice, then click the SPEAK button.)
Voice: Juliette .. French
SPEAK
Display a link for the audio file. [details]

# Nuance Vocalizer



# Pourquoi utiliser la synthèse de parole à partir du texte dans les logiciels d'ALAO ?

L'emploi des systèmes de synthèse de parole à partir du texte (plutôt que d'enregistrements de locuteurs natifs, par exemple) dans les systèmes d'ALAO est intéressant pour de nombreuses raisons.

De manière générale, la synthèse de parole augmente la flexibilité de tels systèmes permettant aux apprenants de pratiquer la langue parlée à leur propre rythme.

De manière plus spécifique, en ce qui attrait au professeur, l'emploi de la synthèse de parole à partir du texte dans les systèmes-auteurs (logiciels qui permettent la saisie d'exercices) représente une économie de temps de préparation de cours. Il est en effet plus facile et plus rapide de saisir du texte et de le modifier que d'enregistrer et de modifier des fichiers sons.

<-- Page précédente

Page suivante -->

# Applications de la synthèse de parole à partir du texte dans les logiciels d'ALAO

Il existe toute une gamme d'outils, activités et exercices exploitant la synthèse de parole notamment : des dictionnaires parlants, des textes parlants, des logiciels de traitement de texte parlants, des dictées, des activités de discrimination auditive, de compréhension orale, de prononciation, de simulation de dialogue, etc.

Nous avons identifié que dans ces outils, activités, et exercices, les systèmes de synthèse de parole à partir de texte sont amenés à jouer un de ces trois rôles principaux :

1. Lecteur de texte
2. Modèle de prononciation
3. Partenaire de conversation.

## 1. Lecteur de texte

La synthèse de parole à partir du texte est employée en tant que lecteur de texte dans les applications suivantes:

- Dictionnaires parlants
- Textes parlants
- Logiciels de traitement de texte parlants
- Dictées
- Activités de compréhension orale.

L'ultilisation de la synthèse de parole à partir du texte dans ces applications a l'avantage de permettre à l'apprenant d'écouter n'importe quel texte sur demande (la synthèse de la parole permet la génération automatique de la parole).

En annexe vous trouverez des copies d'écran de telles applications :

- Dictionnaire parlant
  (impression-écran 1 page 5)
- Texte parlant
  (impressions-écrans 2 et 3 page 5)
- Dictée
  (impression-écran 4 page 7)

## 2. Modèle de prononciation

La synthèse de parole à partir du texte est employée en tant que modèle de prononciation dans les activités suivantes :

- Exercices de discrimination auditive

- Exercices de prononciation au niveau segmental (c.-à-d. au niveau du phonème/son) et au niveau de la phrase.

L'emploi de la synthèse de parole à partir du texte dans ces applications possède plusieurs avantages :

- Les modèles de prononciation produits par les systèmes de synthèse de parole à partir du texte permettent à l'apprenant de se concentrer sur un aspect précis de la prononciation enseigné. Ces modèles sont plus formalisés et plus neutres.
- Les systèmes de synthèse de parole à partir du texte permettent la génération de modèles de prononciation qui rendent plus saillants certains aspects de la langue orale.
- En ce qui concerne la nature de ces modèles, en plus des modifications qu'un professeur de français ferait dans ses cours pour faire remarquer certains aspects de la langue parlée tel que la modification du débit de la parole et l'exagération de l'intonation à ses apprenants, la synthèse de parole permet la production des modèles de prononciation que le professeur aurait de la difficulté à produire telles que des phrases avec intonation mais sans rythme.
- La synthèse de la parole rend possible la génération automatique et sur demande d'une rétroaction (feedback).

En annexe vous trouverez des copies d'écran de telles applications :

- Exercice de prononciation au niveau segmental (du phonème/son)
  (impression-écran 5 page 8)
- Exercice de prononciation au niveau suprasegmental (de la phrase)
  (impression-écran 6 page 9)

# 3. Partenaire de conversation

La synthèse de parole à partir du texte est employée en tant que partenaire de conversation dans les simulations de dialogue.

L'emploi de la synthèse de parole à partir du texte dans ce type d'application a l'avantage de permettre la simulation de dialogues plus ouverts et ainsi plus authentiques (la synthèse de parole à partir du texte permet la génération sur demande d'une rétroaction (feedback) et de tours de parole).

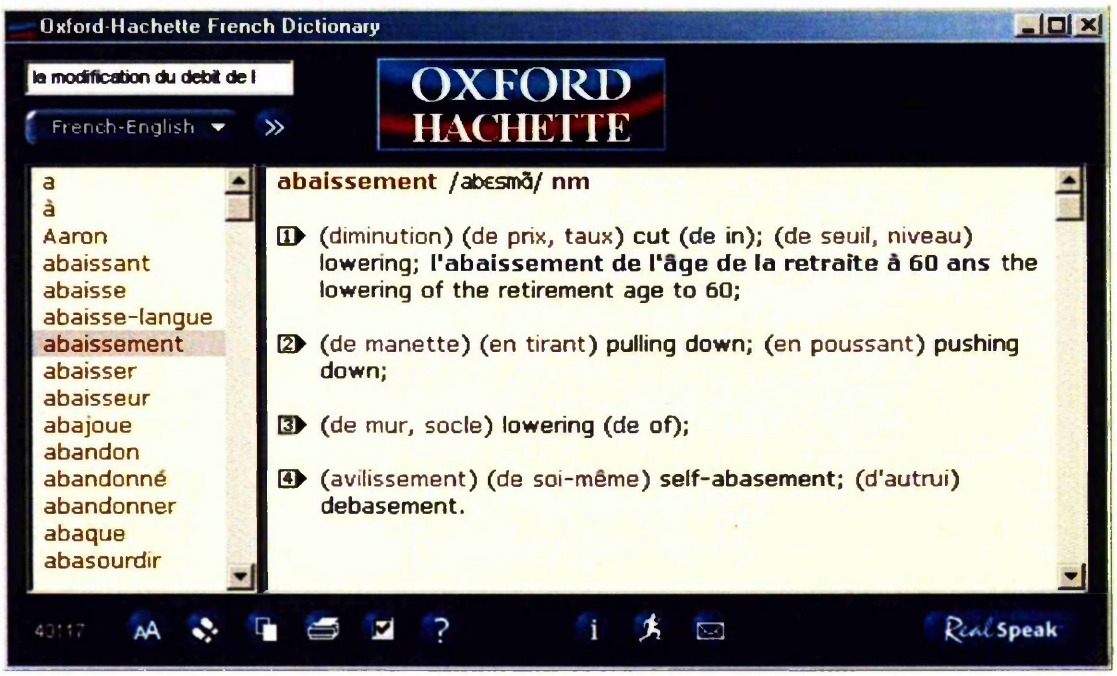
En annexe vous trouverez des copies d'écran de telles applications :

- Simulation de conversation
  (impression-écran 7 page 10)

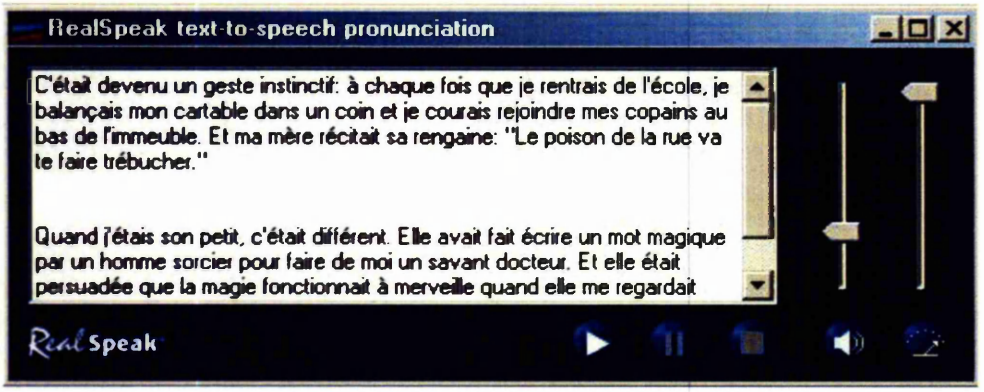
<-- Page précédente

Page suivante -->

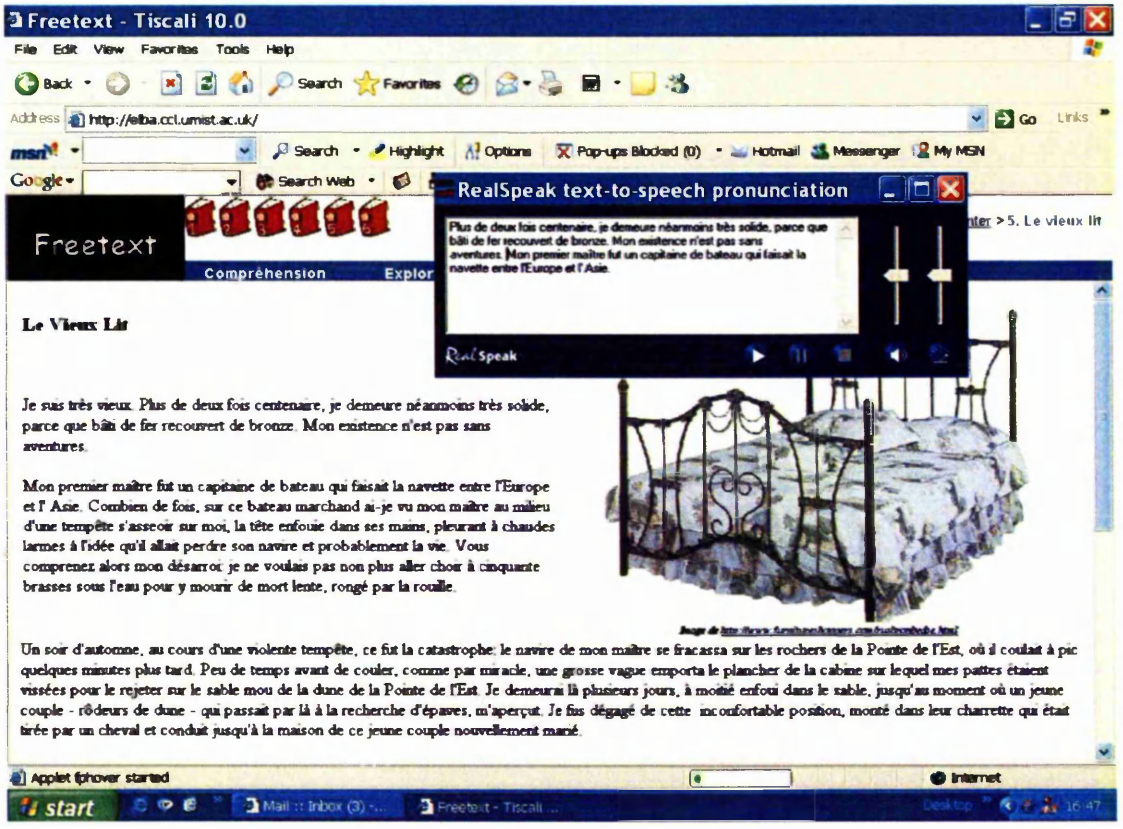
## Impression-écran 1 : Dictionnaire parlante



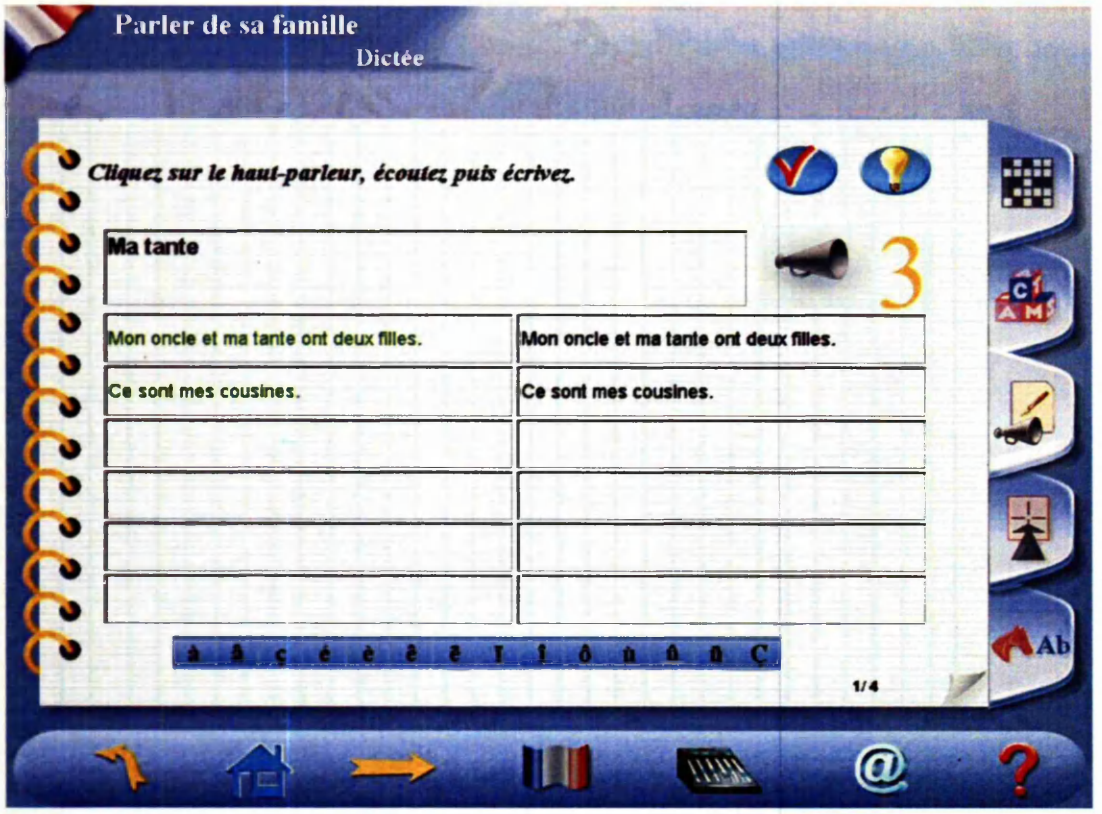
## Impression-écran 2 : Texte parlant 1



# Impression-écran 3 : Texte parlant 2



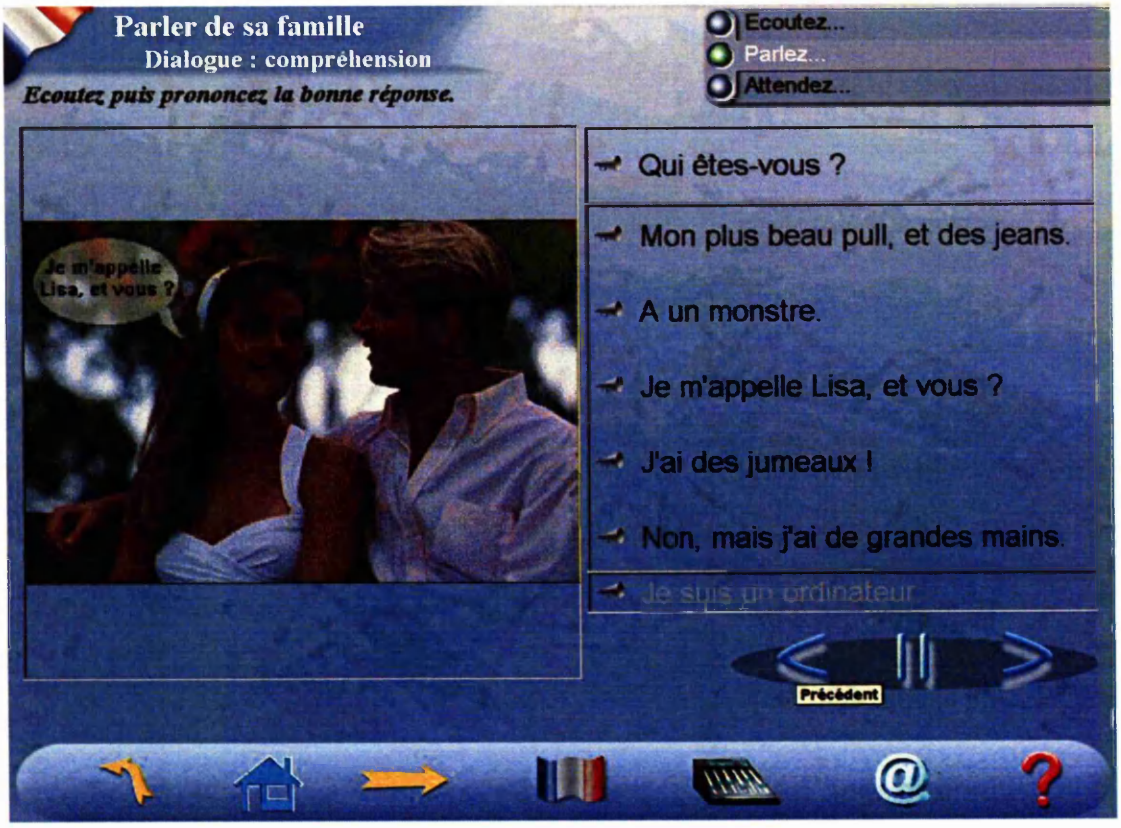
## Impression-écran 4 : Dictée



**Impression-écran 5 : Exercice de prononciation au niveau segmental (du phonème/son)**

**Impression-écran 6 : Exercices de prononciation au niveau segmental (c.-à-d. au niveau du phonème/son) et au niveau de la phrase.**

## Impression-écran 7 : Simulation de conversation



# Un besoin d'évaluer

Aujourd'hui, on trouve sur le marché de plus en plus d'exemples de logiciels d'ALAO qui intègrent des systèmes de synthèse de parole à partir du texte. Cependant, peu d'évaluations de ces systèmes dans ces environnements d'apprentissage ont été effectuées. Nous ne savons donc pas s'ils sont véritablement prêts pour emploi dans les contextes que nous avons présentés. Pour répondre à cette question, de telles évaluations sont nécessaires.

Le but de cette recherche est par conséquent le développement d'une méthode d'évaluation des systèmes de synthèse de parole pour déterminer si ceux-ci sont prêts pour emploi dans les logiciels d'ALAO.

En ce moment, nous sommes en train de développer une batterie de tests pour déterminer si les systèmes de synthèse du français à partir du texte sont prêts pour emploi dans les logiciels d'ALAO qui visent à l'apprentissage du français. Pour ce faire, il nous faut conduire quelques expériences avec des professeurs comme vous. Le but de notre expérience est ici d'identifier et de classer en ordre d'importance les qualités exigées de la synthèse de parole dans les 3 contexte d'ALAO présentés.

<-- Page précédente

Page suivante -->

# Avant de commencer, assurez-vous que le volume est au bon niveau...

- Avant de procéder à l'expérience, veuillez vous assurer que le volume est au bon niveau.
- La phrase peut-êre entendu en cliquant dessus.

Ceci est un test. Veuillez assurer que le volume est à un bon niveau.

- Le volume est-il au bon niveau ? Oui Non

<-- Page précédente

# Pour régler le volume...

- Cliquez sur le bouton 'Start'/'Démarrer'
- Sélectionez 'Programs'/'Programmes', 'Accessories'/'Accessoires', 'Entertainment'/'Divertissement', 'Volume Control'/'Contrôle du volume'
- Réglez le volume et réécouter la phrases pour vérifier le niveau sonore

  Ceci est un test. Veuillez assurer que le volume est à un bon niveau.

- Une fois que vous avez réglé le volume à un bon niveau cliquez sur le lien 'Page suivante'.

<-- Page précédente

Page suivante -->

# Instructions

Comme nous l'avons expliqué, les systèmes de synthèse de parole jouent trois rôles principaux dans les logiciels d'ALAO : (1) lecteur de texte, (2) modèle de prononciation, et (3) partenaire de conversation.

Votre tâche est :

- D'**évaluer l'adéquation et l'acceptabilité de la parole** produite par 4 systèmes de synthèse de parole indépendamment dans chacun des trois rôles présentés
- D'**évaluer la qualité de la parole** produite par 4 systèmes de synthèse de parole indépendamment par rapport à leur utilisation dans chacun des 3 rôles présentés.

<-- Page précédente

Page suivante -->

# Groupes

1. Si vous ne pouvez pas soumettre vos réponses par courrier éléctronique, télédécharger une feuille de réponse en format MS Word d'après le lien correspondant au groupe auquel nous vous avons assigné (notre adresse est fourni à la find de l'expérience):
    - Groupe 1
    - Groupe 2

2. Suivez le lien qui correspond au groupe auquel nous vous avons assigné.
    - Groupe 1
    - Groupe 2

Si vous n'avez pas été assigné à un groupe, veuillez nous contacter par courriel (zoe.handley@postgrad.manchester.ac.uk) et nous vous assignons à un des groupes.

<-- Page précédente

# Synthétiseur 1 : Texte de familiarisation

- Familiarisez-vous avec la voix du synthétiseur en écoutant le texte suivant.
- Chaque phrase peut-être entendue en cliquant dessus.

On disait dans le livre : "Les serpents boas avalent leur proie tout entière, sans la mâcher.
Ensuite ils ne peuvent plus bouger et ils dorment pendant les six mois de leur digestion."

J'ai alors beaucoup réfléchi sur les aventures de la jungle et à mon tour j'ai réussi, avec un crayon de couleur, à tracer mon premier dessin.

Source : De Saint-Exupéry (1999) *Le Petit Prince.* France: Gallimard. Page 13.

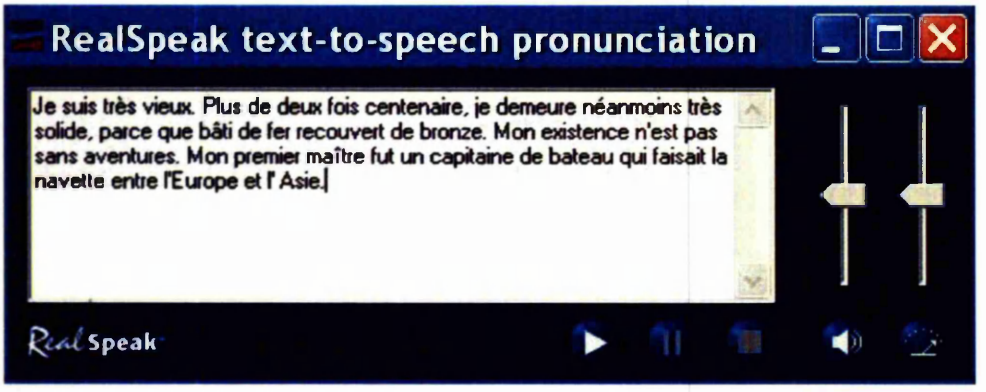
<-- Page précédente

Page suivante -->

# Synthétiseur 1 : Lecteur de texte

- Veuillez **évaluer les niveaux d'adéquation et d'acceptabilité de la parole** produite par le synthétiseur **pour utilisation comme lecteur de texte** d'après les échelles fournies.
- Ensuite, veuillez **évaluer la qualité de la parole** produite par le synthétiseur **par rapport à son utilisation comme lecteur de texte** d'après les échelles fournies.
- Cliquez sur les phrases une par une.

**Exemple**




Je suis très vieux.
Plus de deux fois centenaire, je demeure néanmoins très solide, parce que bâti de fer recouvert de bronze.
Mon existence n'est pas sans aventures.
Mon premier maître fut un capitaine de bateau qui faisait la navette entre l'Europe et l'Asie.

<-- Page précédente

Page suivante -->

# Synthétiseur 1 : Modèle de prononciation au niveau segmental (du son/mot)

- Veuillez **évaluer les niveaux d'adéquation et d'acceptabilité de la parole** produite par le synthétiseur **pour utilisation comme modèle de prononciation au niveau segmental (du phonème/mot)** d'après les échelles fournies.
- Ensuite, veuillez **évaluer la qualité de la parole** produite par le synthétiseur **par rapport à son utilisation comme modèle de prononciation au niveau segmental (du phonème/mot)** d'après les échelles fournies.
- Cliquez sur les phrases une par une.

## Exemple


dehors, intérieur, rat, tiroir, règle...
elle, mais, ver, verres, verre...
voudrais, éclairs, lait, vrai, paquet...
parfois, fais, suffira, affaire, boeuf...
et, marié, assez, manger, habillés...
supermarché, chez, boucher, brochettes, achèterons...
droit, reçois, envoi, boîte, recevoir...
vide, grande, regarde, indiscrète...
va, baskets, parle, appartement, jardin...
vingt, invités, combien, maintenant, besoin...

<-- Page précédente

Page suivante -->

# Synthétiseur 1 : Modèle de prononciation niveau suprasegmental (de la phrase)

- Veuillez **évaluer les niveaux d'adéquation et d'acceptabilité de la parole** produite par le synthétiseur **pour utilisation comme modèle de prononciation au niveau suprasegmental (de la phrase)** d'après les échelles fournies.
- Ensuite, veuillez évaluer **la qualité de la parole** produite par le synthétiseur **par rapport à son utilisation comme modèle de prononciation au niveau suprasegmental (de la phrase)** d'après les échelles fournies.
- Cliquez sur les phrases une par une.

**Exemple**


Je ne m'en souviens pas !
Jérémie, tu viens ?
L'ensemble résidentiel ?
Non, je suis très lente.
Effectivement, je préfère le bus.
Vous appelez ça être en vacances !
J'ai une mauvaise toux.
Elle est dans l'allée suivante.
Est-elle enceinte ?
Deux livres. Je voudrais un sac s'il vous plaît.

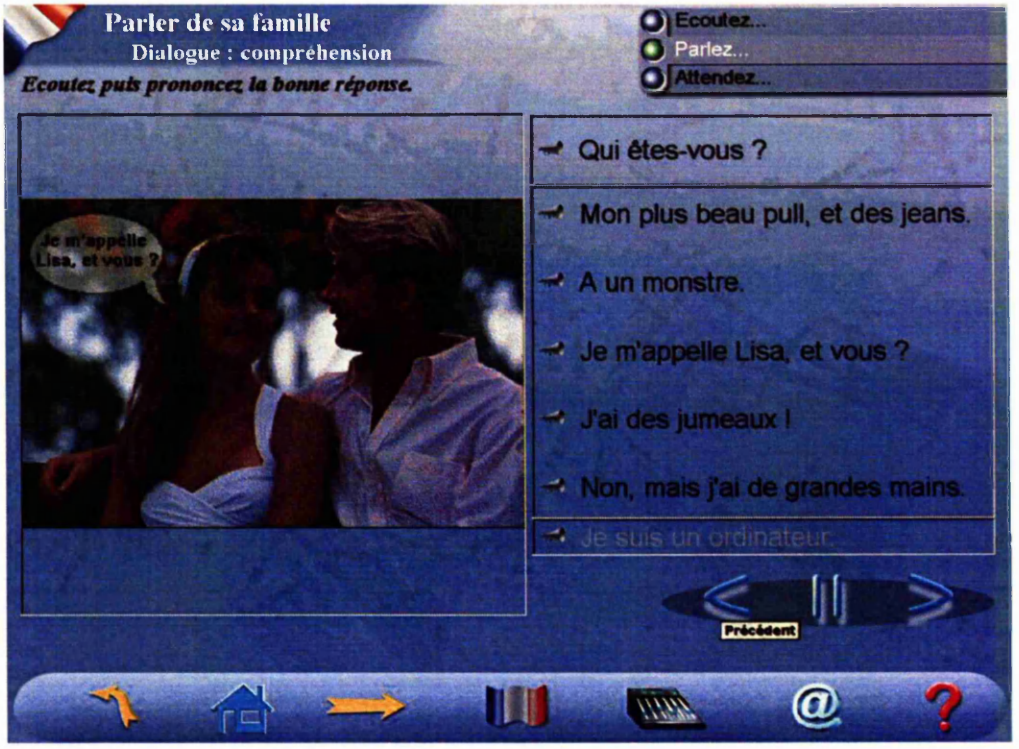
<-- Page précédente

Page suivante -->

# Synthétiseur 1 : Partenaire de conversation

- Veuillez **évaluer les niveaux d'adéquation et d'acceptabilité de la parole** produite par le **synthétiseur pour utilisation comme partenaire de conversation** d'après les échelles fournies.
- Ensuite, veuillez **évaluer la qualité de la parole** produite par le synthétiseur **par rapport à son utilisation comme partenaire de conversation** d'après les échelles fournies.
- Cliquez sur les phrases une par une.

## Exemple




Vous êtes assise à votre bureau, l'air très sérieuse...
Que faites-vous ?
J'écris une lettre !
Vous écrivez souvent ?
J'écris seulement pour Noël et le Jour de l'An.
Vous avez assez d'enveloppes ?
J'ai encore un gros paquet.
J'espère que vous avez assez de timbres !

<-- Page précédente

Page suivante -->

# Synthétiseur 4 : Texte de familiarisation

- Familiarisez-vous avec la voix du synthétiseur en écoutant le texte suivant.
- Cliquez sur le lien indiqué pour ouvrir la démonstration en-ligne du synthétiseur indiqué.
- Choisissez la voix suivante dans les options : **French – Julie**
- (Ne changez pas d'autres options)
- Copiez les phrases bloc par bloc dans la démonstration et entendre la parole produite.
- Ne fermez pas la démonstration quand vous avez terminé. Il vous faudra l'utiliser pour écouter les phrases sur les quatre pages suivantes.

Démonstration enligne du synthétiseur

**Bloc 1 :**
On disait dans le livre : "Les serpents boas avalent leur proie tout entière, sans la mâcher.
Ensuite ils ne peuvent plus bouger et ils dorment pendant les six mois de leur digestion."

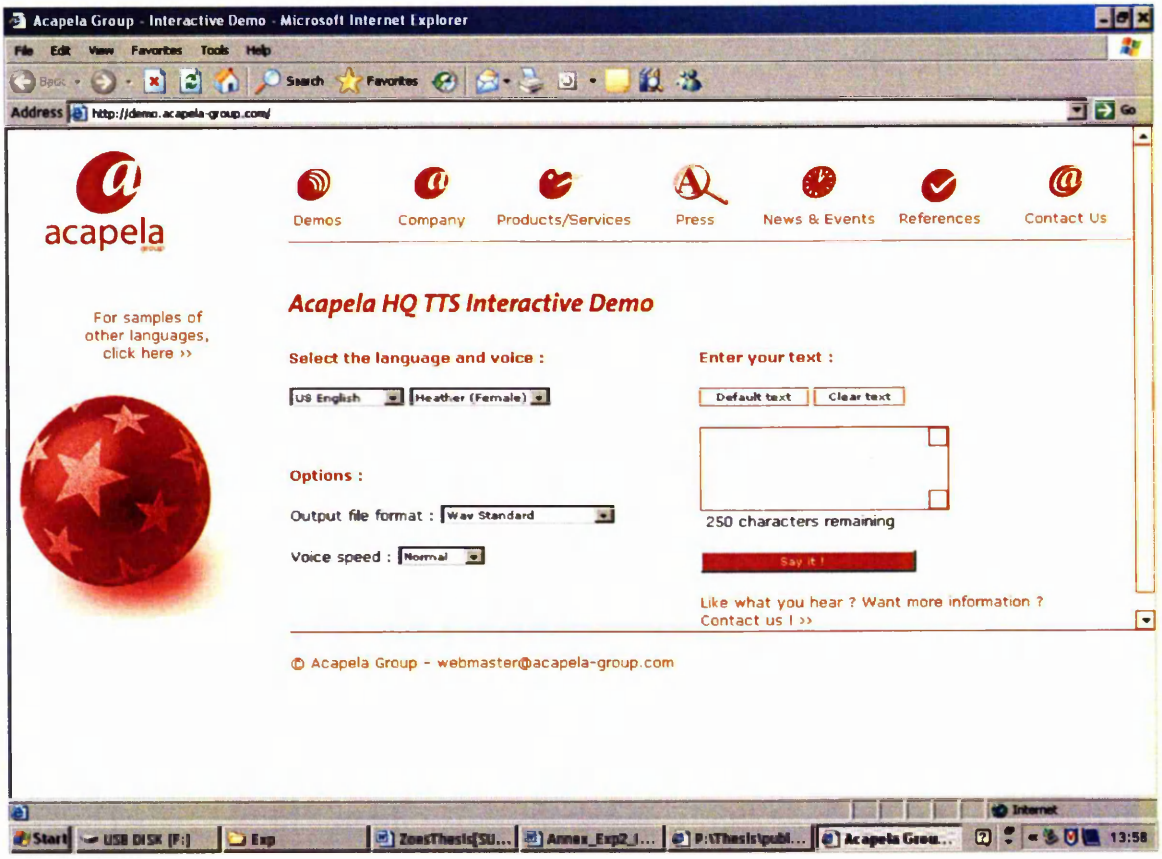
**Bloc 2 :**
J'ai alors beaucoup réfléchi sur les aventures de la jungle et à mon tour j'ai réussi, avec un crayon de couleur, à tracer mon premier dessin.

Source : De Saint-Exupéry (1999) *Le Petit Prince.* France: Gallimard. Page 13.

<-- Page précédente

Page suivante -->

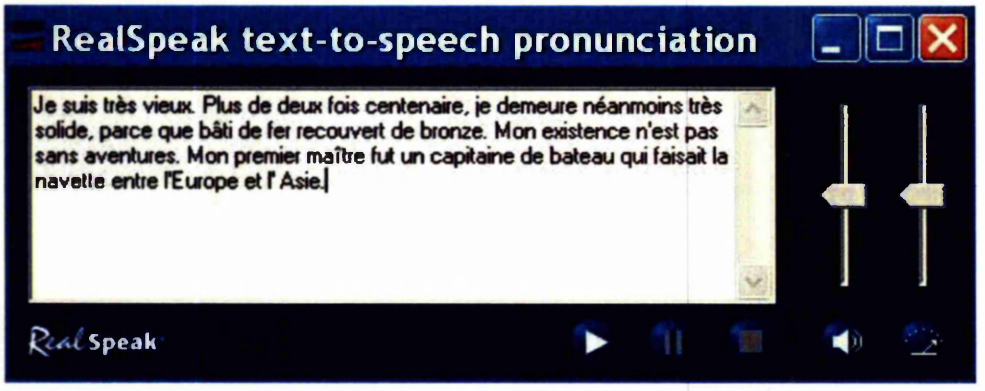
# On-line TTS synthesis demonstration



# Synthétiseur 4 : Lecteur de texte

- Veuillez **évaluer les niveaux d'adadéquat et d'acceptabilité de la parole** produite par le synthétiseur **pour utilisation comme lecteur de texte** d'après les échelles fournies.
- Ensuite, veuillez **évaluer la qualité de la parole** produite par le synthétiseur **par rapport à son utilisation comme lecteur de texte** d'après les échelles fournies.
- Cliquez sur le lien indiqué pour ouvrir la démonstration en-ligne du synthétiseur indiqué.
- Choisissez la voix suivante dans les options : **French - Julie**
  (Ne changez pas d'autres options)
- Copiez les phrases bloc par bloc dans la démonstration et entendre la parole produite.
- N'oubliez pas de fermer la démonstration quand vous avez terminé.

**Exemple**




Démonstration enligne du synthétiseur.

**Bloc 1 :**
Je suis très vieux.
Plus de deux fois centenaire, je demeure néanmoins très solide, parce que bâti de fer recouvert de bronze.

**Bloc 2 :**
Mon existence n'est pas sans aventures.
Mon premier maître fut un capitaine de bateau qui faisait la navette entre l'Europe et l'Asie.

<-- Page précédente

Page suivante -->

# Synthétiseur 4 : Modèle de prononciation au niveau segmental (du son/mot)

- Veuillez **évaluer les niveaux d'adéquation et d'acceptabilité de la parole** produite par le synthétiseur **pour utilisation comme modèle de prononciation au niveau segmental (du phonème/mot)** d'après les échelles fournies.
- Ensuite, veuillez **évaluer la qualité de la parole** produite par le synthétiseur **par rapport à son utilisation comme modèle de prononciation au niveau segmental (du phonème/mot)** d'après les échelles fournies.
- Cliquez sur le lien indiqué pour ouvrir la démonstration en-ligne du synthétiseur indiqué.
- Choisissez la voix suivante dans les options : **French - Julie**
- (Ne changez pas d'autres options)
- Copiez les phrases bloc par bloc dans la démonstration et entendre la parole produite.

## Exemple


Démonstration enligne du synthétiseur.

**Bloc 1 :**
dehors, intérieur, rat, tiroir, règle...
elle, mais, ver, verres, verre...
voudrais, éclairs, lait, vrai, paquet...
parfois, fais, suffira, affaire, boeuf...
et, marié, assez, manger, habillés...

**Bloc 2 :**
supermarché, chez, boucher, brochettes, achèterons...
droit, reçois, envoi, boîte, recevoir...
vide, grande, regarde, indiscrète...
va, baskets, parle, appartement, jardin...
vingt, invités, combien, maintenant, besoin...

<-- Page précédente

Page suivante -->

# Synthétiseur 4 : Modèle de prononciation au niveau suprasegmental (de la phrase)

- Veuillez évaluer **les niveaux d'adéquation et d'acceptabilité de la parole** produite par le synthétiseur **pour utilisation comme modèle de prononciation au niveau suprasegmental (de la phrase)** d'après les échelles fournies.
- Ensuite, veuillez **évaluer la qualité de la parole** produite par le synthétiseur **par rapport à son utilisation comme modèle de prononciation au niveau suprasegmental (de la phrase)** d'après les échelles fournies.
- Cliquez sur le lien indiqué pour ouvrir la démonstration en-ligne du synthétiseur indiqué.
- Choisissez la voix suivante dans les options : **French - Julie**
- (Ne changez pas d'autres options)
- Copiez les phrases bloc par bloc dans la démonstration et entendre la parole produite.

## Exemple


Démonstration enligne du synthétiseur.

**Bloc 1:**
Je ne m'en souviens pas !
Jérémie, tu viens ?
L'ensemble résidentiel ?
Non, je suis très lente.
Effectivement, je préfère le bus.

**Bloc 2 :**
Vous appelez ça être en vacances !
J'ai une mauvaise toux.
Elle est dans l'allée suivante.
Est-elle enceinte ?
Deux livres. Je voudrais un sac s'il vous plaît.

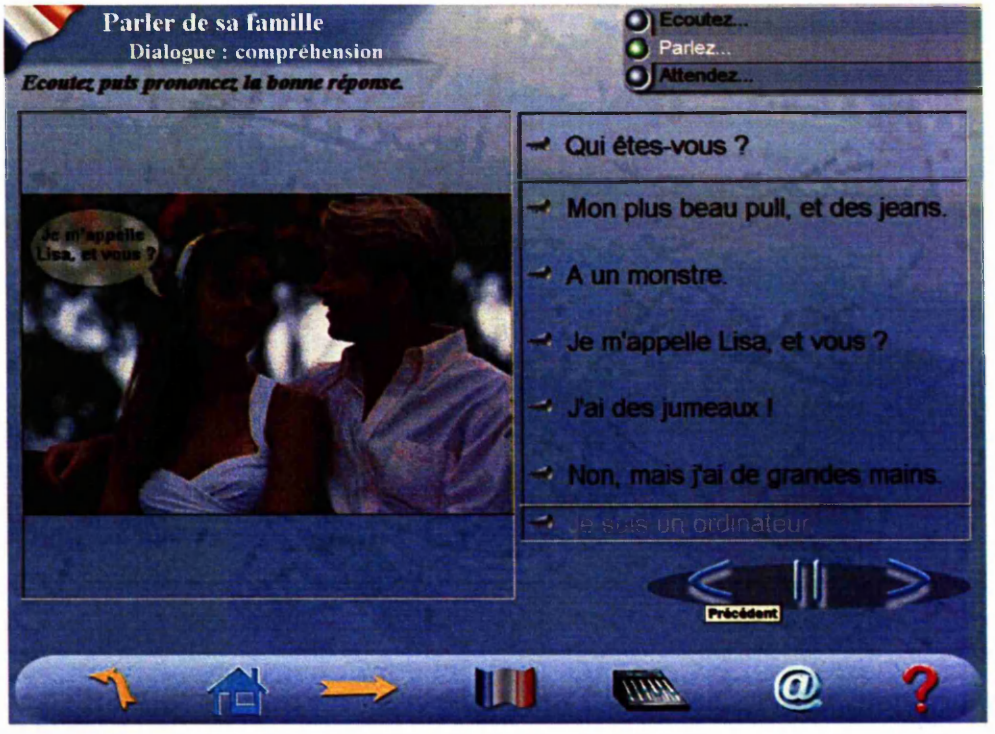
<-- Page précédente

Page suivante -->

# Synthétiseur 4 : Partenaire de conversation

- Veuillez **évaluer les niveaux d'adéquation et d'acceptabilité de la parole** produite par le synthétiseur **pour utilisation comme partenaire de conversation** d'après les échelles fournies.
- Ensuite, veuillez **évaluer la qualité de la parole** produite par le synthétiseur **par rapport à son utilisation comme partenaire de conversation** d'après les échelles fournies.
- Cliquez sur le lien indiqué pour ouvrir la démonstration en-ligne du synthétiseur indiqué.
- Choisissez la voix suivante dans les options : **French - Julie**
- (Ne changez pas d'autres options)
- Copiez les phrases bloc par bloc dans la démonstration et entendre la parole produite.

## Exemple




Démonstration enligne du synthétiseur.

**Bloc 1 :**
Vous êtes assise à votre bureau, l'air très sérieuse...
Que faites-vous ?
J'écris une lettre !

Vous écrivez souvent ?
J'écris seulement pour Noël et le Jour de l'An.

**Bloc 2 :**
Vous avez assez d'enveloppes ?
J'ai encore un gros paquet.
J'espère que vous avez assez de timbres !

<-- Page précédente Page suivante -->