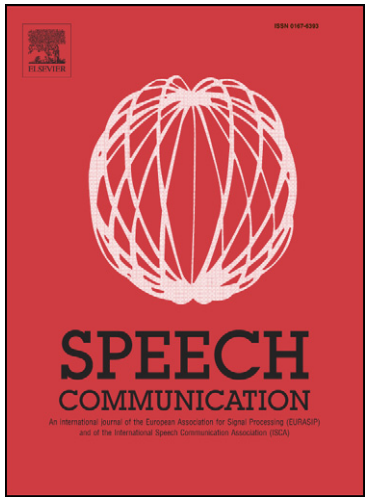

Is Text-to-Speech Synthesis Ready for Use in Computer-Assisted Language Learning?

Zöe Handley

PII: S0167-6393(08)00184-2
DOI: 10.1016/j.specom.2008.12.004
Reference: SPECOM 1770

To appear in: *Speech Communication*

Received Date: 30 June 2008
Revised Date: 10 December 2008
Accepted Date: 17 December 2008

Please cite this article as: Handley, Z., Is Text-to-Speech Synthesis Ready for Use in Computer-Assisted Language Learning?, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.12.004

Is Text-to-Speech Synthesis Ready for Use in Computer-Assisted Language Learning?[1]

Zöe Handley[a,2]

[a] *School of Computer Science, The University of Manchester, Lamb Building, Booth St East, Manchester, UK. M13 9EP*[3]

ABSTRACT

Text-to-Speech (TTS) synthesis, the generation of speech from text input, offers another means of providing spoken language input to learners in Computer-Assisted Language Learning (CALL) environments. Indeed, many potential benefits (ease of creation and editing of speech models, generation of speech models and feedback on demand, etc.) and uses (talking dictionaries, talking texts, dictation, pronunciation training, dialogue partner, etc.) of TTS synthesis in CALL have been put forward. Yet, the use of TTS synthesis in CALL is not widely accepted and only a few applications have found their way onto the market. One potential reason for this is that TTS synthesis has not been adequately evaluated for this purpose. Previous evaluations of TTS synthesis for use in CALL, have only addressed the comprehensibility of TTS synthesis. Yet, CALL places demands on the comprehensibility, naturalness, accuracy, register and expressiveness of the output of TTS synthesis. In this paper, the aforementioned aspects of the quality of the output of four state-of-the-art French TTS synthesis systems are evaluated with respect to their use in the three different roles that TTS synthesis systems may assume within CALL applications, namely: 1) reading machine, 2) pronunciation model and 3) conversational partner (Handley and Hamel, 2005). The results of this evaluation suggest that the best TTS synthesis systems are ready for use in applications in which they 'add value' to CALL, i.e. exploit the unique capacity of TTS synthesis to generate speech models on demand. An example of such an application is a dialogue

---

[1] Portions of the work presented in this paper will be presented at CALL 2008 (August 30th – September 1st 2008, Antwerp, Belgium).

[2] Present address: Learning Sciences Research Institute, The University of Nottingham, Exchange Building, Jubilee Campus, Wollaton Road, Nottingham, NG7 1BB. Tel.: +44 115 846 6561. Fax: +44 115 846 7931.

E-mail addresses: zoe.handley@nottingham.ac.uk (Z. Handley).

[3] The research was actually carried out in The School of Informatics, which has since merged with The School of Computer Science.

partner. In order to fully meet the requirements of CALL, further attention needs to be paid to accuracy and naturalness, in particular at the prosodic level, and expressiveness.

Keywords: CALL, speech synthesis, TTS synthesis, evaluation

1 Introduction

In very simple terms, speech synthesis is the process of making the computer talk. Unlike other methods of providing the computer with a voice, such as the digital recording of human speakers, Text-to-Speech (TTS) synthesis systems, which generate speech from text input, have the unique ability to generate speech models, which can be exploited for the provision of talking text facilities (Hamel, 2003a), the automated generation of exercises with spoken language support (de Pijper, 1997), and the generation of feedback (Sherwood, 1981) and conversational turns (Egan and LaRocca, 2000) on demand to unanticipated learner interactions. Yet, the use of TTS synthesis in Computer-Assisted Language Learning (CALL) is not very widely accepted (Egan and LaRocca, 2000; Sobkowiack, 1998) and the number of commercial applications which integrate TTS is quite limited (Handley and Hamel, 2005).

One possible reason for this is that the suitability and benefits of the use of TTS synthesis in CALL have not been proven. One way in which this can be achieved is through evaluation. However, as we shall see, TTS synthesis has been only partially evaluated for use in CALL applications. Specifically, only one of the requirements placed on the quality of the output of TTS synthesis systems has been addressed, namely comprehensibility (see section 4.3). Yet, CALL applications also place demands on the naturalness and accuracy of the output of TTS synthesis systems (Handley and Hamel, 2005) (see section 4.2). Moreover, the majority of the evaluations that have been conducted are out-of-date given the advances in TTS synthesis of the last few years (see section 2).

In this paper, I report an evaluation of the readiness of a selection of state-of-the-art TTS synthesis systems with respect to their use in the three different roles that they may assume within CALL applications: (1) reading machine, (2) pronunciation

model, and (3) dialogue partner (Handley and Hamel, 2005). According to Handley and Hamel, CALL applications place demands on both the quality and the flexibility of the speech generated by TTS synthesis systems. The current evaluation focuses on criteria relating to the quality of the speech generated by the TTS systems.

2 Speech Synthesis

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. More formally, speech synthesis systems are defined as systems which:

> allow 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).

According to van Bezooijen and van Heuven, this definition includes *limited domain synthesis*, that is systems in which individually stored words are substituted into information slots in carrier sentences (Black and Lenzo, 2000). Examples of applications for which limited domain speech synthesis systems have been developed include talking clocks, fixed weather reports and dialogue systems for booking flights, hotels, and so on (*ibid.*).

While limited domain speech synthesis systems should generate better quality speech than systems which permit the synthesis of unrestricted texts because the utterances to be synthesised are necessarily similar to the recordings in the systems' speech database (Black and Lenzo, 2000), limited domain synthesis will not be considered further here because the main attraction of the use of speech synthesis in CALL is the ability to generate *any* utterance on demand (see section 5). Moreover, the development of limited domain synthesis specifically for the purposes of CALL is not feasible given the low levels of funding that CALL research and development receives. A notable exception is CALL for military training (Johnson *et al.*, 2002).

A further distinction is made between Text-to-Speech (TTS) synthesis systems and Concept-to-Speech (CTS) synthesis systems. These two types of speech synthesis

system are distinguished by the type of input that they support. TTS synthesis systems take text as input and mimic the human process of reading. CTS synthesis systems take a conceptual or semantic representation as input. 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).

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 therefore 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 (van Bezooijen and van Heuven, 1997). For example, a CTS synthesis system should know the syntactic and semantic structures and the communicative purpose of the utterances to be generated (van Bezooijen and van Heuven, 1997; Rodman, 1999). 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 better quality speech than TTS synthesis, TTS synthesis is more frequently (re-)used for CALL purposes. TTS synthesis is hence the focus of this paper. TTS synthesis will therefore be considered in more detail in the next section.

2.1 TTS Synthesis

TTS synthesis systems are typically composed of two modules: 1) a Text-to-Phoneme (TTP) module (or NLP module), and 2) a Phoneme-to-Speech (PTS) module (or DSP

module). The goal of the TTP module is to generate an unambiguous narrow phonetic transcription of the input text augmented with specifications for the generation of prosody. The goal of the PTS module is to transform these control parameters into waveforms, i.e. speech. As yet, no standard architecture for the TTP module has been established – it may be implemented using simple heuristics, fully-fledged linguistically motivated grammars/parsers, or stochastic (statistically-driven) grammars/parsers. On the other hand, two standard techniques are used to implement the PTS module, namely formant and concatenative synthesis. Up until the turn of the century, formant synthesis was the dominant technology. Formant synthesisers produce speech by electronically modelling the acoustic characteristics of speech sounds, in particular their characteristic formant patterns. Concatenative synthesis, the current dominant technology, is based on the recombination of segments of pre-recorded human speech. There are several different approaches to concatenative synthesis. The choice of approach has a significant effect on the quality of the speech generated. Approaches to concatenative synthesis therefore merit further consideration.

Concatenation-based synthesisers consist in a database of segments which have been extracted from a corpus of recordings of human speakers. The first such synthesisers were based on the concatenation of diphones. A diphone consists in the second half of an allophone, a contextually conditioned variant of a phoneme, and the first half of the following allophone (Dutoit, 1997; Huang et al., 2001). The optimal segment for concatenative synthesis:

- Leads to low distortion at concatenation points,[4]
- Is generalisable, i.e. permits the synthesis of unrestricted text, and,
- Captures inter-allophonic effects such as coarticulation (Dutoit, 1997; Huang et al., 2001).

While diphones are generalisable, capture inter-allophonic effects, and lead to low distortion at concatenation points, they are not optimal; they only capture coarticulatory effects which span adjacent phonemes and give rise to a large number

[4] 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).

of concatenation points (Dutoit, 1997; see footnote 4). A number of other types of segment of different sizes have therefore been suggested for use in systems based on concatenative synthesis (see Dutoit (1997) or Huang et al (2001) for a review of the different types of segment which have been suggested). None, however, is optimal. Consequently, state-of-the-art TTS synthesis systems often use a combination of the different types of segment (Dutoit, 1997). This is referred to as non-uniform concatenative synthesis. Regarding the number of instances of each segment used, the first TTS synthesis systems based on concatenative synthesis only used one instance of each diphone. Consequently, the diphones did not always fit the prosody of the utterance to be generated. In order to overcome this, the database of state-of-the-art TTS synthesis systems often includes multiple instances of each segment, each extracted from a different prosodic context (Black and Taylor, 1994; Campbell and Black, 1997; Conkie, 1999), from which, on synthesis, the system selects the most appropriate segment given the prosody of the utterance to be generated. This is referred to as Unit Selection Synthesis (USS). Unit selection is often combined with the use of different types of segment. Systems which combine the two techniques are referred to as non-uniform USS systems (Schroeter, 2001; Henton, 2002).

3 TTS Synthesis in CALL

CALL applications integrating speech technology have emerged from the general need in language learning and teaching for "self-paced interactive learning environments" which provide "controlled interactive speaking practice outside the classroom" (Ehsani and Knodt, 1998, p. 45). Though little heard of in CALL until recently, it was identified that TTS synthesis could play a role in responding to this need over twenty five years ago (Sherwood, 1981). Specifically, Sherwood made the observation that typing/editing text is easier than (re-)recording voice and that navigating through a textual database is easier than retrieving recorded samples from an audiotape. He also observed that TTS synthesis has the capacity to generate speech models on demand, and that this capacity could be exploited in CALL to provide learners with personalised feedback. A decade or so later, the same advantages were again put forward, this time by the technology specialists themselves (Dutoit, 1997; Keller and Zellner-Keller, 2000). They see TTS synthesis as an “indefatigable substitute native speaker” (Keller and Zellner-Keller 2000: 111) which because it is not human is perceived as non-judgemental.

It has been suggested that the advantages of TTS synthesis presented above could be exploited in a number of different CALL applications. Regarding the evaluation of TTS synthesis for use in CALL, different setups or operational contexts often impose different requirements and therefore require different methods of evaluation (Sparck-Jones and Galliers, 1996). In the following sections, the proposed applications are therefore classified according to the operational context, or role, in which TTS synthesis is used. Three roles have been identified: 1) reading machine, 2) pronunciation model, and 3) conversational partner (Handley and Hamel, 2005).

3.1 Reading Machine

Applications in which TTS synthesis assumes the role of a reading machine include: talking dictionaries, talking texts and dictations. A talking dictionary is an electronic dictionary which integrates either digital recordings of human speakers or speech synthesis for the oral presentation of dictionary entries. An example of a commercially available talking dictionary integrating TTS synthesis is the *Oxford-Hachette French Dictionary on CD-ROM*. A talking text is 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. The *Oxford Hachette French Dictionary on CD-ROM* also integrates such a facility as does *FreeText*, a CALL program for advanced learners of French (Hamel, 2003b), and the *Appeal* ("A pleasant personal environment for adaptive learning") system (de Pijper, 1997). Dictation is a traditional writing activity in which the teacher reads aloud a text which the learner is asked to transcribe. *DICTOR* (Santiago-Oriola, 1999) and *Ordictée* (Mercier et al., 2000) are examples of CALL applications dedicated to dictation.

3.2 Pronunciation Model

It has also been suggested that TTS synthesis could be used as a pronunciation model in exercises focusing on both the segmental (practice of individual and combined phonemes) and supra-segmental (practice of intonation and prosody) levels. At the segmental level, it is typically used to present individual and combined sounds to the learner, sounds which are retrieved from a database in which they are stored in textual format. The experimental pronunciation tutor *SAFexo*, a module of the CALL system

SAFRAN (Système d'Apprentissage du FRANçais; Hamel, 1998, 2003a), focuses on this kind of practice. An example of a CALL application that uses TTS synthesis in the teaching of prosody is Mercier et al.'s (2000) prosodic tutor for Breton.

3.3 Conversational Partner

Since responses in dialogues are unpredictable and may be infinite in number, it is difficult to both predict and store all possible responses in the form of digitally recorded human speech to learner utterances. TTS synthesis with its unique ability to generate spoken utterances from text on demand provides part of the solution to this problem. Examples of spoken dialogue systems which integrate TTS synthesis that are currently being developed for use in language learning include the *Let's Go Spoken Dialogue System* (SDS) (Raux and Eskenazi, 2004) and *SCILL* (Spoken Conversational Interaction for Language Learning) system (Seneff et al., 2004).

4 Evaluation of TTS synthesis for CALL purposes

The evaluation of TTS synthesis for CALL purposes is important because general purpose TTS synthesis systems are being re-used in roles in which they are not used in general applications, namely as a pronunciation model and as a conversational partner. Moreover, while TTS synthesis is used as a reading machine in applications outside CALL such as reading machines for the blind, screen readers for people with visual impairments and learning disabilities such as dyslexia and aphasia, and talking word processor, the CALL setup differs from these operational contexts. The most important difference is that the main users of CALL are learners, that is non-native speakers of the language spoken by the TTS synthesis system. In the following sections, I consider the levels of evaluation that ought to be conducted and the requirements that ought to be addressed by those evaluations before going on to review the evaluations of TTS synthesis for CALL purposes that have been conducted to date.

4.1 Evaluation Infrastructure

According to Handley and Hamel (2005), the following levels of evaluation ought to be conducted:

- Basic research evaluation of TTS synthesis for use in CALL: An evaluation of the viability and potential benefits of the use of TTS synthesis in the intended CALL application.
- Technology evaluation of TTS synthesis for use in CALL: An evaluation of the extent to which the selected TTS synthesis system(s) meets the requirements of the intended CALL application.
- Judgemental evaluation of the CALL application: An evaluation of the potential of the CALL application to provide conditions which promote Second Language Acquisition (SLA).
- Judgemental evaluation of the teacher planned activity: An evaluation of the potential of the teacher-planned activity to provide conditions which promote SLA.
- Usage evaluation of the teacher planned activity: An evaluation of learners' performance in the teacher-planned activity.
- Program evaluation: An evaluation of the success of the funding program.

This is a combination of the levels of evaluation recommended by Chapelle (2001a, 2001b) for the evaluation of CALL activities and by ELSE (1999) for the evaluation of Speech And Language Technologies (SALT).

4.2 Requirements

Based on a review of 'ideal' conditions for SLA (Chapelle, 1998), research on teacher talk, the register of speech that teachers use to address learners, and best practice in language learning and teaching, Handley and Hamel (2005) suggested that CALL applications place demands first and foremost on the quality of the output of TTS synthesis systems. Specifically, they suggested that the quality of the output should be such that it is as comprehensible, natural, and accurate as possible. Comprehensibility refers to the ease with which a listener can understand a speaker's intended message (Francis and Nusbaum, 1999). Natural speech is speech which sounds as if it was produced by a native speaker. Accurate speech is speech which is free from error. In addition to quality, Handley and Hamel (2005) suggested that CALL applications place demands on the flexibility of TTS synthesis. Specifically, they suggested that TTS synthesis systems for use in CALL should provide control over voice, style

(familiar or formal), speech rate and pitch. Handley and Hamel's recommendations were based on the fact that CALL applications should provide what Chapelle (2001a, 2001b) refers to as language learning potential. Language learning potential concerns whether features of the target language can actually be learned from a CALL activity as it is designed, and whether the activity provides plenty of opportunities to focus on linguistic form. In a case study involving a research French TTS synthesis system, Handley and Hamel (2005) established that CALL applications do indeed place demands on the comprehensibility, naturalness and accuracy of the speech generated by TTS synthesis, but that they also place demands on register and expressiveness. The demands placed on the flexibility of TTS synthesis systems remain to be validated.

4.3 State of the Art

Our review of the literature reveals that very few "formal" evaluations of TTS synthesis for the specific purposes of CALL have been conducted (Stratil et al., 1987a; Stratil et al., 1987b; Cohen, 1993; Santiago-Oriola, 1999; Hincks, 2002). Moreover, general purpose tools for the evaluation of speech synthesis systems such as the *ITU-T Overall Quality Test* (Schmidt-Nielson, 1995; van Bezooijen and van Heuven, 1997) which is exploited in the *Blizzard Challenge* (Bennett, 2005; Black and Tokuda, 2005), a speech synthesis comparative evaluation campaign, do not address some of the criteria which are believed to be important for language learning applications, such as naturalness, expressiveness and register (see section 4.2).

Regarding evaluations of TTS synthesis for the specific purposes of CALL, identification of the potential benefits TTS could bring to CALL could be considered to fulfil the function of basic research evaluation. However, regarding the next stage of evaluation recommended by Hamel and Handley (2005), namely technology evaluation, only one report of an evaluation of the adequacy of TTS for use in CALL was found in the literature. In this evaluation, the quality, specifically the comprehensibility, of the output of a Spanish TTS synthesizer was evaluated to determine whether it was suitable for use as a reading machine for the presentation of grammar exercises in a language laboratory setting (Stratil et al.,1987a). In other words, only one of the requirements of CALL has been addressed in one role for one TTS synthesis system – and that TTS synthesis system was out of date with respect to

recent developments (see section 2.1). Further evaluation of the adequacy of TTS synthesis for use in CALL is important. If we do not conduct a more complete evaluation of the adequacy of TTS synthesis for use in CALL, we run the risk of wasting time and money integrating TTS synthesis into CALL applications for which it is not suited. Further evaluations should address all three roles which TTS synthesis may assume within CALL because different roles may impose different requirements on TTS synthesis, and a range of TTS synthesis systems should be evaluated because, as presented above, the quality of the output generated by different TTS synthesis systems varies significantly.

5 An Evaluation of Four State-of-the-Art French TTS Synthesis Systems

In order to go some way towards redressing the inadequacy of technology evaluations conducted to date, in this paper, I present an evaluation of a selection of state-of-the-art French TTS synthesis systems for use in the three roles which TTS synthesis may assume within CALL applications: 1) reading machine, 2) pronunciation model, and 3) conversational partner.

Through this evaluation I hope to answer the following questions:

- Is TTS synthesis ready for use in CALL?
- What aspects of TTS synthesis require improvement in order to fully meet the requirements of CALL?
- Do the different roles that TTS synthesis may assume within CALL applications have different requirements with respect to quality of speech generated?

Regarding the use of technology in CALL, I take the stance that technologies:

> 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 are replacing (Stevens, 1989: 33).

For this reason, two levels of readiness are distinguished in this study:

- *Acceptability*, readiness for use in applications in which CALL 'adds value', that is applications which are not possible without TTS synthesis and which exploit its unique capacities, specifically the ability to generate speech models on demand. Examples of such applications include, talking texts and dialogue partners.
- *Adequacy*, readiness for use in applications which it is already possible to provide through the use of other media such as digitised speech and do not exploit the unique capacities of TTS synthesis. Examples of such applications include talking dictionaries,[5] dictation,[6] and pronunciation training.

For this same reason, in this study, I do not compare the speech generated by TTS synthesis systems with that of a human control subject.

5.1 Method

One voice offered by each of four different state-of-the-art French TTS synthesis systems was evaluated with respect to its use in CALL as a: 1) reading machine, 2) pronunciation model at the phonetic level (henceforth phonetic pronunciation model), 3) pronunciation model at the prosodic level (henceforth prosodic pronunciation model) and 4) conversational partner. In this study I decided to treat phonetic and prosodic pronunciation models separately. I believed that they may place different demands on the quality of the output of TTS synthesis systems because different aspects of the quality of the speech are the focus of instruction and the models typically differ in length and complexity (see Handley and Hamel (2005)), prosodic pronunciation models being longer and more complex than phonetic pronunciation models.

5.1.1 Design

The investigation had a related design. The independent variables were the different TTS synthesis systems (see section 5.1.3.2 for details) and the following roles that

[5] Contrary to this, the *Oxford-Hachette French Dictionary* on *CD-ROM* does exploit TTS synthesis in a way that adds value; it exploits the capacity of TTS synthesis to generate speech models on demand to read aloud definition and examples of usage in addition to headwords.

[6] The dictation system *Ordictée* (Mercier et al., 2000) also exploits TTS synthesis in a way that adds value; it exploits the manipulability of TTS synthesis to adapt the speed of delivery of the dictations to the rate at which the learners types their responses.

TTS synthesis systems may assume within CALL applications: 1) reading machine, 2) phonetic pronunciation model, 3) prosodic pronunciation model, and 4) conversational partner. The dependent variables 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*, *expressiveness*, and *appropriateness of register* of the speech generated by the TTS synthesis systems, and the *acceptability* and *adequacy* of the speech generated by the TTS synthesis systems for use in the four roles. *Accuracy* and *naturalness* were composite measures. Ratings of *accuracy* were based on those of *precision of phonemes* and *appropriateness of prosody*. Ratings of *naturalness* were based on those of *naturalness of phonemes* and *naturalness of prosody*. All measures were obtained using metrics which consisted in a question and a 7-point MOS scale. A detailed description of the questionnaire is provided below.

5.1.2 Participants

Participants 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 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 the 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).

Participants were therefore a convenience sample of 17 French teachers and CALL researchers. 6 were male and 11 were female. Their ages ranged from under 25 to over 55 years old. 13 spoke French as their mother tongue. 10 were French teachers and 2 were CALL researchers. 4 out of the 10 teachers had experience of CALL. 7 of the participants understood what TTS is prior to taking part in the experiment. These participants were asked to rate the readiness of TTS synthesis for use in CALL on a scale of 1 (not at all ready) to 7 (entirely ready). The mean rating of readiness across

these participants was 4 with a standard deviation of 1.41. In other words, as a group, the participants were not biased for or against the use of TTS synthesis in CALL prior to taking part in the study. 14 of the participants believed they had used speech synthesis prior to participating in this experiment. Their frequency of use of applications integrating TTS synthesis ranged from less than once a month to once a week.

In previous evaluations of TTS synthesis systems, it has been observed that the ability to recognise synthetic speech improves after only short periods of exposure (Pisoni, 1978-9; Francis and Nusbaum, 1999). The TTS synthesis systems and roles were therefore presented to each participant in a different randomized order.

5.1.3 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.

5.1.3.1 Corpus

As said, the ability to recognise synthetic speech improves after only short periods of exposure (Pisoni, 1978-9; Francis and Nusbaum, 1999), for each TTS synthesis system evaluated, before beginning the experiment in earnest, the participants were therefore presented a short passage to familiarise themselves with the voice of the TTS synthesis system. This passage was taken from *Le Petit Prince* (de Saint-Exupery, 1999).

The test corpus consisted of 10 sentences that might be presented to learners in the roles of reading machine, prosodic pronunciation model, and conversational partner, and 10 lists of 5 words that might be presented to learners in the role of phonetic pronunciation model. The reading machine corpus was taken from the text *Le Vieux Lit* proposed to learners in *FreeText* (Hamel, 2003b). The remaining corpora were all taken from *Talk to Me: The Conversation Method (French)* (Auralog, 2002), specifically from the following modules: ‘Phonetic Exercises’, ‘Sentence Pronunciation’ and ‘Dialogues: Comprehensibility’.

In order to provide the contextualisation which it is believed is necessary to permit participants to make reliable judgements of the acceptability, adequacy and quality of the speech generated by TTS synthesis systems for use in the different roles they may assume in CALL applications (Handley and Hamel, 2005), the corpora were accompanied by a screenshot of a CALL application integrating TTS synthesis in that role.

5.1.3.2 TTS synthesis systems

The TTS synthesis systems were selected in order to cover a range of different synthesis techniques, and hence qualities of output, and to include a range of varieties of French and a balance of male and female voices (see Table 1). The systems that were used were: *AT&T Next-Gen*, *Bright Speech* from Babel Technologies, *eLite* from Multitel, and *Nuance Vocalizer*. *AT&T Next-Gen*[7] is based on USS (Beutnagel *et al.*, 1999). Two voices are provided for Parisian French, one male (Alain) and one female (Julie). The female voice was used in the experiment reported here. *Nuance Vocalizer*[8] is based on concatenative synthesis (TMA Associates, 2003). One female voice is available for French (Julie Deschamps). *eLite*[9] (Enhanced Linguistically-based Text-to-speech synthesis) (Multitel, 2005) from Multitel is based on concatenative synthesis. Four voices are provided for French, two male (Vincent and Thierry), and two female (Anne-Carole and Céline). Vincent was used in this experiment. *BrightSpeech* from Babel Technologies, now owned by Acapela Group and marketed as *HQ TTS*,[10] 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.

[7] http://www.research.att.com/projects/tts/demo.html
[8] http://www.nuance.com/prodserv/demo_vocalizer.html
[9] http://www.multitel.be/TTS/layout.php?page=eLite_demo
[10] The on-line interactive demonstration of the TTS synthesis system is now available at: http://demo.acapela-group.com/

6.1.3.3 Questionnaire

The questionnaire employed in this investigation, presented in Figure 1, 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). All scales were translated into French in order to keep the participants thinking in French (see Appendix A for French translation).

6.2 Procedure

On arrival at the Website where the experiment was hosted, participants were presented a brief introduction to the use of TTS synthesis in CALL which included an overview of TTS synthesis, a presentation of the proposed advantages of the use of TTS synthesis in CALL, and a presentation of the proposed uses of TTS synthesis in CALL. Next, the participants were reminded of the goal of the experiment, namely to determine what requirements CALL imposes on TTS synthesis and whether TTS synthesis is ready for use in CALL. Then, the participants were presented a short audio clip and asked to adjust the volume of the speech output to a comfortable level. Next, the procedure of the investigation was explained. Specifically, the participants were told that, for each of 4 TTS synthesis systems, they would be presented a brief passage to familiarise themselves with the speech generated by the TTS synthesis system followed by an example of its use in CALL in each of four roles: 1) reading machine, 2) phonetic pronunciation model, 3) prosodic pronunciation model, and 4) conversational partner. The participants were told that, their task was to rate the quality of the speech generated by the TTS synthesis systems with respect to its use in the role indicated on the scales provided. The experimental procedure was the same for all participants. All that differed was the order of presentation of the synthesisers.

5.3 Analysis

In the following sections I consider the results of this evaluation with respect to each of the research questions in turn, namely:

- Is TTS synthesis ready for use in CALL?
- What aspects of TTS synthesis require improvement in order to fully meet the requirements of CALL?

- Do the different roles that TTS synthesis may assume within CALL applications have different requirements with respect to quality of speech generated?

5.3.1 Is TTS synthesis ready for use in CALL?

The mean ratings of the adequacy and acceptability of the speech generated by the TTS synthesis systems for use in each of the four roles was calculated across participants. The results are presented in Table 2 and Table 3 respectively.

The descriptive statistics presented in Table 2 show that none of the TTS synthesis systems achieved top ratings, i.e. ratings of 6 or 7, for adequacy for use in any of the roles. The ratings of the adequacy of system 4 for use in all four roles, in particular the rating of the adequacy of system 4 for use as a conversational partner, are, however, not far off. Similarly, systems 1, 2, and 3 do not achieve top ratings for acceptability for use (see Table 3). System 4 on the other hand does achieve top ratings for two of the roles, namely reading machine and conversational partner, suggesting that the quality of speech that it generated is acceptable for use in those roles, i.e. ready for use in applications to which TTS synthesis adds value. The ratings of the acceptability of the speech generated by system 4 for use in the roles of phonetic and prosodic pronunciation model are not much lower than those of the speech generated by system 4 for use in the roles of reading machine and conversational partner.

This suggests that, of the four systems evaluated, only one system, system 4, is ready for use in CALL, and, that that system is only ready for use as a reading machine and a conversational partner in applications in which it adds value, that is exploit its capacity to generate speech models on demand. The quality of the speech generated is not sufficient for it to be used in applications which could be provided through the use of digitised speech.

5.3.2 What improvements are needed for TTS synthesis to more fully meet the requirements of CALL?

The mean ratings of the different aspects of the quality of the speech generated by TTS synthesis systems with respect to their use in each of the four roles was calculated across participants. The results for system 1 are presented in Table 4. The

descriptive statistics presented in Table 4 show that none of the aspects of the quality of the speech generated by system 1 received top ratings, i.e. ratings of 6 or 7, for any of the roles. More specifically, of all the aspects of the quality of the speech considered expressiveness received the lowest mean rating for all four roles. The mean ratings of the accuracy and naturalness of the speech with respect to its use as a reading machine were also very low, under 4, negative on the original scale used in the questionnaire. This was also the case for the role of phonetic pronunciation model, but in addition, the mean rating of the intelligibility of the speech with respect to its use in this role were also very low. Regarding the role of prosodic pronunciation model, in addition to expressiveness, the mean ratings of the accuracy and naturalness of the speech with respect to its use in this role were very low. This was also the case for the role of conversational partner.

As said, accuracy and naturalness were composite measures. The results for the individual measures that were combined to arrive at measures of accuracy and naturalness are presented in Table 5. They show that ratings of accuracy and naturalness at the prosodic level were lower than ratings of accuracy and naturalness at the phonetic level; while ratings of accuracy and naturalness at the phonetic level were sometimes more than 4, i.e. positive on our original scale, ratings of accuracy and naturalness at the prosodic level were less than 4.

The results for system 2 are presented in Table 6 and Table 7. As for system 1, the descriptive statistics presented in Table 6 show that none of the aspects of the quality of the speech generated by system 2 received top ratings for any of the roles. More specifically, as for system 1, of all the aspects of the quality of the speech considered, for all four roles, expressiveness received the lowest mean rating. Also, as for system 1, the mean ratings of the accuracy and naturalness of the speech with respect to its use as reading machine were also very low. Regarding the role of phonetic pronunciation model, none of the aspects of the quality of the speech including expressiveness received very low ratings. As for the role of reading machine for this TTS synthesis system and for all of the roles for system 1, in addition to expressiveness, accuracy and naturalness received very low ratings for the role of prosodic pronunciation model. Regarding the role of conversational partner, the mean

ratings of the naturalness of the speech, in addition to those of the expressiveness of the speech received very low ratings.

Regarding the measures that were combined to arrive at measures of accuracy and naturalness, as for system 1, ratings of accuracy and naturalness at the prosodic level were lower than ratings of accuracy and naturalness at the phonetic level, with ratings of accuracy and naturalness at the prosodic level in general below 4 and ratings of accuracy and naturalness at the phonetic level above 4 in some cases.

The results for system 3 are presented in Table 8 and Table 9. As for systems 1 and 2, the descriptive statistics presented in Table 8 show that none of the aspects of the quality of the speech generated by system 3 received top ratings for any of the roles. More specifically, as for systems 1 and 2, for all four roles, expressiveness received the lowest mean rating. Returning to the role of reading machine, intelligibility, accuracy, and naturalness also received very low ratings. Regarding the role of phonetic pronunciation model, all aspects of the speech received very low ratings with respect to use in this role. Finally, regarding the role of conversational partner, in addition to expressiveness, accuracy and naturalness received very low ratings.

Regarding the measures that were combined to arrive at measures of accuracy and naturalness, with one exception, precision of phonemes for use as a pronunciation model at the prosodic level, ratings were below 4.

The results for system 4 are presented in Table 10 and Table 11. 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 system 4 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. The picture is the same when we look in more detail at the measures which make up the ratings of accuracy and naturalness (see Table 11).

In summary, expressiveness received the lowest rating for all four roles, for three out of four of the TTS synthesis systems evaluated. Accuracy and naturalness also received low ratings for all for roles for three out of four of the TTS synthesis systems evaluated. 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, accuracy and naturalness.

These results are consistent with what is known about the quality of the speech generated by TTS synthesis systems based on concatenative and USS synthesis. Concatenative synthesis systems do not provide control over voice quality (Edgington, 1997) and hence expressiveness (Edgington, 1997; Bailly *et al.*, 2003). Regarding naturalness, there are several factors which give rise to unnatural sounding speech, namely:

- Distortions at concatenation points (Huang *et al.*, 2001)
- The inability to model changes at the phonetic level which accompany changes at the prosodic level (Campbell and Black, 1997)
- The fact that speech segments are typically extracted from corpora of prosodically neutral speech (*ibid.*).

Regarding accuracy at the prosodic level, methods for determining the prosodic specification of utterances are inadequate (Dutoit, 1997; Rodman, 1999; Henton, 2002).

5.3.3 Do the different roles of TTS synthesis in CALL have different requirements?

In order to answer this question, I looked again at the ratings of the adequacy (see Table 2) and the acceptability (see Table 3) of the speech generated by the TTS synthesis systems for use in the different roles that it may assume within CALL applications. The descriptive statistics presented in Table 2 show that adequacy differed across the roles for all four TTS synthesis systems. The same was true for acceptability, with the exception of ratings of the acceptability of S4 for use as a pronunciation model at the phonetic and prosodic levels (see Table 3). Analysis of the data using the Friedman test revealed that, while the differences in adequacy were

significant for systems 2 and 4 ($\chi^2_r = 8.010$, df = 3, $p = 0.046$; $\chi^2_r = 8.063$, df = 3, $p = 0.045$, respectively), they were not statistically significant for systems 1 and 3 ($\chi^2_r = 2.352$, df = 3, $p = 0.503$; $\chi^2_r = 3.467$, df = 3, $p = 0.325$; $\chi^2_r = 3.194$, respectively). The differences in acceptability were not statistically significant for any of the TTS synthesis systems (system 1 $\chi^2_r = 6.616$, df = 3, $p = 0.085$, system 2 $\chi^2_r = 6.303$, df = 3, $p = 0.098$, system 3 $\chi^2_r = 3.194$, df = 3, $p = 0.363$, and system 4 $\chi^2_r = 5.547$, df = 3, $p = 0.163$). A similar result was obtained when the Friedman test was run on the various aspects of the quality of the speech evaluated (see Handley, 2006). It is therefore not possible to draw any clear conclusions as to whether the different roles impose different requirements on the quality of the speech generated by TTS synthesis systems. A possible explanation for these results is that the different roles do impose different requirements on the quality of the speech generated by TTS synthesis systems, but the differences in requirements are only small. Another possible explanation for 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 – learners might use the speech provided in talking dictionaries, talking texts and by conversational partners as a model to imitate, for example. Whichever explanation is correct, the implications for the use of TTS synthesis in CALL applications are that one TTS synthesis system can be used in all applications, if it is sufficiently flexible (see Handley and Hamel (2005)); if there are differences in requirements across the roles, but the differences are only small, there is little gain in using a different TTS synthesis system for each of the different roles; if teachers and CALL researchers cannot discriminate between the different roles and their requirements, then it is not possible to select different TTS synthesis systems for use in different applications; and, if the roles overlap, then a TTS synthesis systems which is suitable for all roles should be used.

6 Summary and future work

In response to the need for further evaluation of TTS synthesis for use in CALL, in this paper, the quality of the speech generated by four state-of-the-art French TTS synthesis systems was evaluated with respect to its use in the three different roles which TTS synthesis systems may assume within CALL applications, namely (1) reading machine, (2) pronunciation model, and (3) conversational partner (Handley

and Hamel, 2005). Regarding the readiness of TTS synthesis for use in CALL, the results of this evaluation suggested that the speech generate by some French TTS synthesis systems is ready for use in applications to which it adds value. The majority of French TTS synthesis systems evaluated, however, did not meet these requirements. In particular, they did not meet the requirements imposed on expressiveness. The good news is that expressive speech synthesis is the focus of much of current research into TTS synthesis (see for example Campbell et al. (2006)). It is, however, not possible to say whether this is the case for other languages; different language pose different challenges to TTS synthesis. Moreover, further research into the requirements imposed on the flexibility (Handley and Hamel, 2005) of TTS synthesis is necessary before we can draw any general conclusions about the readiness of TTS synthesis for use in CALL.

The results of this study also have implications for future evaluations of TTS synthesis. Regarding the distinction that was made between the different roles that TTS synthesis systems may play in CALL applications, whatever the cause, the fact that consistent differences were not found in the readiness of TTS synthesis for use in the different roles suggests that evaluations should not discriminate between the different roles that TTS synthesis may assume within CALL applications, i.e. TTS synthesis systems should be evaluated for use in CALL applications in general; 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. Regarding evaluation criteria, this study highlights the importance to CALL of two characteristics which are not currently addressed by general purpose evaluation tools, such as the *ITU-T Overall Quality Test* (Schmidt-Nielson, 1995; van Bezooijen and van Heuven, 1997), namely naturalness and expressiveness. It is recommended that general purpose tools for the evaluation of TTS synthesis and comparative evaluation campaigns address these criteria. If such general purpose tools for the evaluation of TTS synthesis and comparative evaluation campaigns do not address criteria which are important for CALL, improvements to TTS synthesis for CALL purposes are unlikely to be made because, as mentioned in section 2, in general, it is not feasible to build TTS synthesis systems for the specific purposes of CALL.

In conclusion, while further research is necessary to determine whether the results of our investigation generalise to other languages and to address the requirements imposed on flexibility, our results suggest that it will not be long before learners will be able to benefit from the support of an untiring non-judgemental substitute native speaker 24/7 in CALL applications.

ACKNOWLEDGMENTS

I would like to thank all the French teachers and CALL researchers at The University of Manchester, UK and Dalhousie University, Nova-Scotia, Canada who kindly participated in our study. I would also like to thank Dr Marie-Josée Hamel (Dalhousie University) for valuable discussions and suggestions and acknowledge the comments of two anonymous reviewers.

REFERENCES

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

Bailly, G., Campbell, N. and Mobius, B. (2003) ISCA Special Session: Hot Topics in Speech Synthesis. In *Procs. Eurospeech 2003* (pp. 37-40). Geneva.

Bennett, C. (2005). Large Scale Evaluation of Corpus-Based Synthesizers: Results and Lessons from the Blizzard Challenge 2005. In *Procs. INTERSPEECH 2005* (pp. 105-108). Lisbon, Portugal.

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.

Black, A., and Lenzo, K. (2000). Limited Domain Synthesis. In *Procs. ICSLP 2000.* Beijing, China.

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.

Black, A and Tokuda, K (2005). The Blizzard Challenge – 2005: Evaluating Corpus-Based Speech Synthesis on Common Datasets. In *Procs. INTERSPEECH 2005* (pp. 77-80). Lisbon, Portugal.

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.

Campbell, N., Hamza, W., Höge, H., Tao, J., and Bailly, G. (2006). Special Section on Expressive Speech Synthesis. *IEEE Transactions on Audio, Speech and Language Processing.* 14 (4).

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. (2001a). Innovative language learning: Achieving the vision. *ReCALL,* 23(10), 3-14

Chapelle, C. (2001b). *Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing and Research*. Cambridge: Cambridge University Press.

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.

Conkie, A. (1999). Robust Unit Selection System for Speech Synthesis. In *Procs. Joint meeting of ASA, EAA, and DAGA*. Berlin. Germany.

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 Saint-Exupéry, A. (1999). *Le Petit Prince*. Paris: Gallimard.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. London : Kluwer Academic Publishers.

Edgington, M. (1997). Investigating the Limitations of Concatenative Synthesis. In *Procs. Eurospeech '97* (pp. 1-4). Rhodes, Greece.

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.

Ehsani, B. K. and Knodt, E. (1998). Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm. *Language Learning & Technology*, 2(1), 45-60. Retrieved January 31, 2005, from http://llt.msu.edu/vol2num1/article3/

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).

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.

Hamel, M.-J. (1998). Les outils de TALN dans SAFRAN [NLP tools in SAFRAN]. *ReCALL*, 10(1), 79-85

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. (2006). *Evaluating Text-To-Speech (TTS) Synthesis for use in Computer-Assisted Language Learning (CALL).* Unpublished doctoral thesis. The University of Manchester.

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

Henton, C. (2002). Challenges and Rewards in Using Parametric or Concatenative Speech Synthesis. *International Journal of Speech Technology*. 5: 117-131.

Hincks, R. (2002). Speech Synthesis for Teaching Lexical Stress. *TMH-QPSR*. 44: 153-165

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

Johnson, W., Narayanan, S., Whitney, R., Das, R., Bulut, M., and LaBore, C. (2002). Limited Domain Synthesis of Expressive Military Speech for Animated Characters. *Proceedings of the 7th International Conference on Spoken Language Processing*. Denver, Colorado, USA.

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.

Mercier, G., Guyomard, M., Siroux, J., Bramoullé, A., Gourmelon, H., Guillou, A., and Lavannant, P. (2000). Courseware for Breton Spelling Pronunciation and Intonation Learning. In *Procs. InSTIL 2000* (pp. 145-148). Dundee, England: University of Abertay Dundee.

Multitel (2005). *eLite Documentation*. Retrieved from http://www.multitel.be/TTS/layout.php?page=eLite_doc

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.

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

Raux, A., and Eskenazi, M. (2004). Using task-oriented spoken dialogues for language learning: Potential, practical application and challenges. In R. Delmonte, P. Delcloque, and S. Tonelli (Eds.), *Proceedings of the InSTIL/ICALL 2004 Symposium* (pp. 147-150). Venice, Italy.

Rodman, R. (1999). *Computer Speech Technology*. London: Artech House.

Santiago-Oriola, C. (1999). Vocal synthesis in a computerized dictation exercise. In Proceedings of *EUROSPEECH'99* (Vol. 1, pp. 191-194), Budapest.

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.

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.

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.

Sherwood, B. (1981). Speech Synthesis Applied to Language Teaching. *Studies in Language Learning.* 3: 175-181

Sobkowiak, W. (1998). Speech in EFL CALL. In Cameron, K. (ed.) (1998). *Multimedia CALL: Theory and Practice.* Exeter: Elm Bank.

Sparck Jones, K., and Galliers, J. R. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review.* London: Springer.

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.

Stratil, M., Burkhardt, D., Jarratt, P., and 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., and Burkhardt, D. (1987b). Exploration of Foreign Language Speech Synthesis*. Literary and Linguistic Computing.* 2(2): 116-119.

TMA Associates (2003). *Nuance: US English.* Retrieved from http://www.tmaa.com/tts/Nuance_USEng.htm

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.

APPENDIX A French Translation of MOS-CALL

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)?								
		-3	-2	-1	0	1	2	3	
	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)?								
		-3	-2	-1	0	1	2	3	
	Très inacceptable								Très acceptable
La qualité de la parole									
Compréhensibilité	Le message lu, est-il facile à comprendre?								
		-3	-2	-1	0	1	2	3	
	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) ?								
		-3	-2	-1	0	1	2	3	
	Très difficile								Très facile
Choix de Prononciation	Est-ce que la prononciation est juste?								
		-3	-2	-1	0	1	2	3	
	Incorrecte								Correcte
Précision des phonèmes	Est-ce que l'articulation des phonèmes/sons est précise?								
		-3	-2	-1	0	1	2	3	
	Très imprécise								Très précise
Prosodie	Est-ce que la prosodie (musicalité) de la phrase est appropriée?								
		-3	-2	-1	0	1	2	3	
	Très inappropriée								Très appropriée
Caractère naturel des phonèmes/sons	Est-ce que les phonèmes/sons sonnent naturels/humains?								
		-3	-2	-1	0	1	2	3	
	Pas du tout naturels/humains								Très naturels/humains
Caractère naturel de la prosodie	Est-ce que la prosodie (musicalité) sonne naturelle/humaine?								
		-3	-2	-1	0	1	2	3	
	Pas du tout naturelle/humaine								Très naturelle/humaine
Expressivité	Est-ce que les émotions sont bien exprimées?								
		-3	-2	-1	0	1	2	3	
	Très mal exprimées								Très bien exprimées
Convenance du registre	Est-ce que le registre est approprié?								
		-3	-2	-1	0	1	2	3	
	Très inapproprié								Très approprié

Figure 1 ***MOS-CALL***

VITAE

Zöe Handley is currently a research fellow in the Learning Sciences Research Institute (LSRI), The University of Nottingham, UK. She completed the research reported here whilst studying for a PhD in the Centre for Computational Linguistics, UMIST, UK and later in the School of Informatics, The University of Manchester, UK. She is interested in the use of Speech And Language Technologies (SALT), including speech synthesis, speech recognition and visual displays, in CALL. Her particular interest lies in the evaluation of these technologies and their use in perception and pronunciation training.

FIGURES

Adequacy and acceptability of the speech									
Adequacy	Is the speech adequate for use as a reading machine (in comparison with other media)?								
		-3	-2	-1	0	1	2	3	
	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)?								
		-3	-2	-1	0	1	2	3	
	Very unacceptable								Very acceptable
Quality of the speech									
Comprehensibility	Is the message easy to understand?								
		-3	-2	-1	0	1	2	3	
	Very difficult								Very easy
Intelligibility	Are the individual phonemes/sounds and words easy to recognise (and discriminate one from another)?								
		-3	-2	-1	0	1	2	3	
	Very difficult								Very easy
Choice of pronunciation	Is the pronunciation correct?								
		-3	-2	-1	0	1	2	3	
	Incorrect								Correct
Precision of phonemes	Was the articulation of the phonemes/sounds precise?								
		-3	-2	-1	0	1	2	3	
	Very imprecise								Very precise
Appropriateness of prosody	Was the prosody (music) of the utterance appropriate?								
		-3	-2	-1	0	1	2	3	
	Very inappropriate								Very appropriate
Naturalness of phonemes	Do the phonemes/sounds sound natural/human?								
		-3	-2	-1	0	1	2	3	
	Not at all natural/human								Very natural/human
Naturalness of prosody	Does the prosody (music) sound natural/human?								
		-3	-2	-1	0	1	2	3	
	Not at all natural/human								Very natural/human
Expressiveness	Was emotion expressed well?								
		-3	-2	-1	0	1	2	3	
	Very badly expressed								Very well expressed
Appropriateness of register	Was the register appropriate?								
		-3	-2	-1	0	1	2	3	
	Very inappropriate								Very appropriate

Figure 1 ***MOS-CALL***

TABLES

Table 1 Summary of the features of the TTS synthesis systems used in the experiment

	Synthesiser	Sex	Voice	Variety	Method of PTS conversion
System 1	*AT&T Next-Gen*	M	Alain	Parisian French	Non-uniform USS
System 2	*Nuance Vocalizer*	F	Julie Deschamps	Canadian French	Concatenative synthesis
System 3	*eLite*	M	Vincent	French	Diphone-based concatenative synthesis
System 4	*BrightSpeech*	F	Julie	French	Non-uniform USS

Table 2 Mean ratings of the adequacy of the speech generated by the TTS synthesis systems for use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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
S4	5.35	5.65	5.24	5.82

Table 3 Mean ratings of the acceptability of the speech generated by the TTS synthesis systems for use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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
S4	6.00	5.94	5.94	6.29

Table 4 Mean ratings of the quality of the speech generated by system 1 with respect to its use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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
Expressiveness	3.24	3.12	2.35	2.65
Appropriateness of register	5.00	4.76	4.53	4.53

Table 5 Mean ratings of the accuracy and naturalness of the speech generated by system 1 with respect to its use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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

Table 6 Mean ratings of the quality of the speech generated by system 2 with respect to its use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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
Expressiveness	3.06	4.53	2.41	2.82
Appropriateness of register	4.59	5.12	4.76	4.88

Table 7 Mean ratings of the accuracy and naturalness of the speech generated by system 2 with respect to its use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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

Table 8 Mean ratings of the quality of the speech generated by system 3 with respect to its use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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
Expressiveness	2.18	2.24	2.88	2.94
Appropriateness of register	4.12	3.76	4.24	4.18

Table 9 Mean ratings of the accuracy and naturalness of the speech generated by system 3 with respect to its use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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

Table 10 Mean ratings of the quality of the speech generated by system 4 with respect to its use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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
Expressiveness	4.94	5.24	4.88	5.18
Appropriateness of register	5.47	5.76	5.41	5.65

Table 11 Mean ratings of the accuracy and naturalness of the speech generated by system 4 with respect to its use in CALL

	Reading machine	Phonetic pronunciation model	Prosodic pronunciation model	Conversational partner
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