ESTABLISHING A METHODOLOGY FOR BENCHMARKING SPEECH SYNTHESIS FOR COMPUTER-ASSISTED LANGUAGE LEARNING (CALL)

Zöe Handley
The University of Manchester, UK

Marie-Josée Hamel
Dalhousie University, Canada

ABSTRACT

Despite the new possibilities that speech synthesis brings about, few Computer-Assisted Language Learning (CALL) applications integrating speech synthesis have found their way onto the market. One potential reason is that the suitability and benefits of the use of speech synthesis in CALL have not been proven. One way to do this is through evaluation. Yet, very few formal evaluations of speech synthesis for CALL purposes have been conducted. One possible reason for the neglect of evaluation in this context is the fact that it is expensive in terms of time and resources. An important concern given that there are several levels of evaluation from which such applications would benefit. Benchmarking, the comparison of the score obtained by a system with that obtained by one which is known, to guarantee user satisfaction in a standard task or set of tasks, is introduced as a potential solution to this problem. In this article, we report on our progress towards the development of one of these benchmarks, namely a benchmark for determining the adequacy of speech synthesis systems for use in CALL. We do so by presenting the results of a case study which aimed to identify the criteria which determine the adequacy of the output of speech synthesis systems for use in its various roles in CALL with a view to the selection of benchmark tests which will address these criteria. These roles (reading machine, pronunciation model, and conversational partner) are also discussed here. An agenda for further research and evaluation is proposed in the conclusion.

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 generation of feedback (Sherwood, 1981) and conversational turns (Egan & LaRocca, 2000) to unanticipated learner interactions, and the automated generation of exercises with spoken language support (de Pijper, 1997). Yet, the use of TTS in computer-assisted language learning (CALL) is not very widely accepted (Egan & LaRocca, 2000; Sobkowiack, 1998) and the number of commercial applications which integrate TTS is quite limited.[1]

One possible reason for this is the fact that the suitability and benefits of the use of TTS in CALL have not been proven. One way in which this can be achieved is through evaluation. Ideally, CALL applications integrating TTS would benefit from six stages of evaluation. The objects of these six stages of evaluation are:

1) the viability and potential benefits of the use of TTS in CALL,
2) the adequacy of TTS for use in CALL,

3) the potential of the CALL program to provide (ideal) conditions for Second Language Acquisition (SLA),
4) the potential of the teacher-planned CALL-based activity to provide (ideal) conditions for SLA,
5) learners' performance in the CALL activity, and
6) the success of the funding program.

However, as we shall see, TTS has been only partially evaluated for use in CALL applications. Moreover, the majority of the evaluations that have been conducted are out-of-date given the advances in TTS of the last few years.

One possible reason for the neglect of evaluation in this context, as in many others, is the fact that it is costly in terms of both time and resources, which could otherwise be devoted to further development (Hirschman & Thompson, 1996). In order to achieve a balance between development and evaluation, benchmarking -- testing a system in a standard task (or set of tasks) in order to determine whether it is suitable for use in a given application in terms of both performance and/or usability -- is commonly used in software evaluation in general (Ralston, Reilly, & Hemmdinger, 2000), and in the evaluation of speech and language technologies (SALTs; Sparck Jones & Galliers, 1996) more specifically.

In this paper, we argue for the use of benchmarking in the evaluation of TTS for CALL purposes as a solution to this limitation as well as a means to achieve consistency and comparability of evaluation (Sparck Jones & Galliers, 1996). Having introduced the notion of benchmarking, presented the benefits and limitations of benchmarking, and given concrete examples of what such benchmarks might look like in this context, we report on our progress towards the development of a benchmark for determining the adequacy of speech synthesis for use in CALL. More specifically, we report the results of a literature review and an experiment aimed at the identification of the criteria which determine the adequacy of speech synthesis for use in CALL with a view to the selection of benchmark tests which will address these criteria. We focus on criteria relating to the quality of the speech and are particularly interested in determining whether the different functions/roles impose different requirements on the quality of the output.

In relation to the last point, it is noted, in the literature on the evaluation of SALTs, that different 'setups', operational contexts (the overall purpose of the application, and the function of the technology within the application, other tools which are available to end users, and the end users) often impose different requirements and therefore different criteria and methods of evaluation (Sparck Jones & Galliers, 1996). The first stage in the evaluation process should therefore be to understand speech synthesis, the goals of CALL applications that integrate it, and its functions or roles within those applications.

SPEECH SYNTHESIS

There are two major classes of speech synthesis. Distinguished by the type of input supported, these are concept-to-speech (CTS) and text-to-speech (TTS). CTS synthesis, also referred to as message-to-speech synthesis, takes concepts, for example, information from a database on train travel, as its input and aims not only to speak but also to generate the phrases to utter. TTS synthesis takes raw text as input and aims to mimic the human process of reading. Potential and actual uses of these different forms of speech synthesis in CALL are presented in the next section.

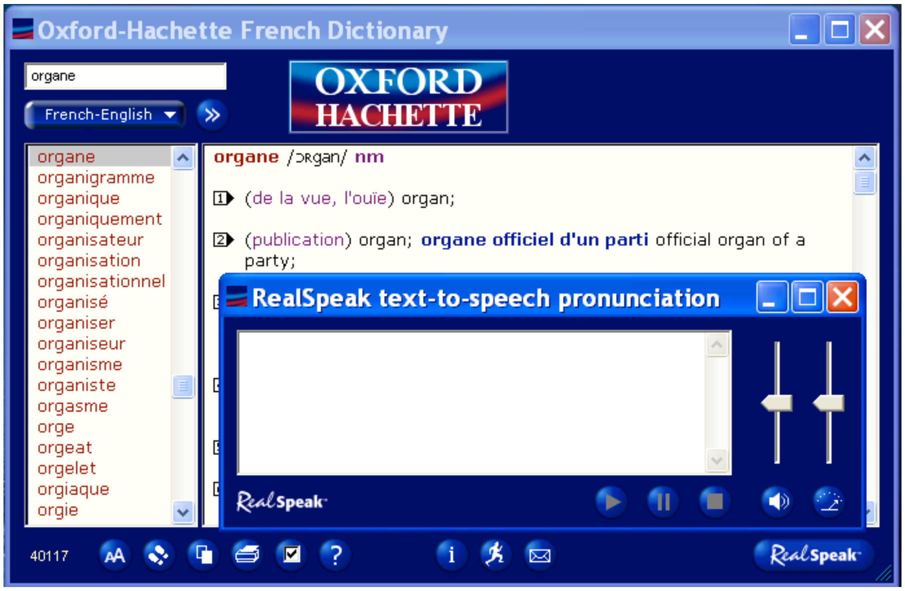
SPEECH SYNTHESIS IN CALL

CALL applications integrating speech technology have emerged from the general need in language learning and teaching (LLT) for "self-paced interactive learning environments" which provide "controlled interactive speaking practice outside the classroom" (Ehsani & Knodt, 1998, p. 45). Both forms of speech synthesis presented in the previous section could contribute to the provision of such an environment. Of the two forms of speech synthesis, it has been suggested that CTS might be more appropriate than TTS for CALL because it permits the generation of less monotonous speech with more human-like prosody (Thomas, Levinson, & Lessard, 2004). Yet, very few CALL systems (re-)use CTS synthesis. One example of a CTS system developed for use in CALL is *VINCI* (Thomas, Levinson, & Lessard, 2004). It is suggested that this system could be used to present dictations, pronunciation and auditory discrimination exercises focusing on intonation, and as a tool for teaching phonetic transcription (Thomas et al., 2004). Rather, most of the literature on the use of speech synthesis in CALL focuses on the (re-)use of TTS systems.[2] The use of TTS in CALL therefore merits further consideration. Applications of TTS in CALL are presented next, followed by a review of the benefits TTS brings to CALL and an analysis of the functions or roles that TTS may play within CALL applications.

Applications (Re-)Using TTS Synthesis

Talking Dictionaries

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. Due to the storage requirements of digital speech, these usually provide only one pronunciation model per entry, typically the headword. The unique ability of speech synthesis to permit the generation of models on demand from text, with considerably lower storage requirements than digital recordings, makes it possible to provide learners with models of all the inflectional forms of a word, its derivations, synonyms, and so forth, as well as examples of usage and definitions. Automatic access to spoken output provides the learner with an instant pronunciation model that can be imitated. This is recognized by Myers (2000) as a good form of lexical reinforcement. Talking dictionaries are also used to develop a more conscious awareness of the relationship between both the graphic and the phonic form of lexical items, a relationship which may not be straightforward in a given language and/or for some learners. An example of a commercially available talking dictionary integrating speech synthesis is the *Oxford-Hachette French Dictionary on CD-ROM* (2003). This dictionary, which integrates the *RealSpeak* TTS synthesizer, contains 360,000 French words and phrases. The learner can access pronunciations of words and phrases simply by clicking on them. Opening the talking text facility (see Figure 1) gives the learner access to volume and speech rate controls.



organe /ɔʀgan/ nm

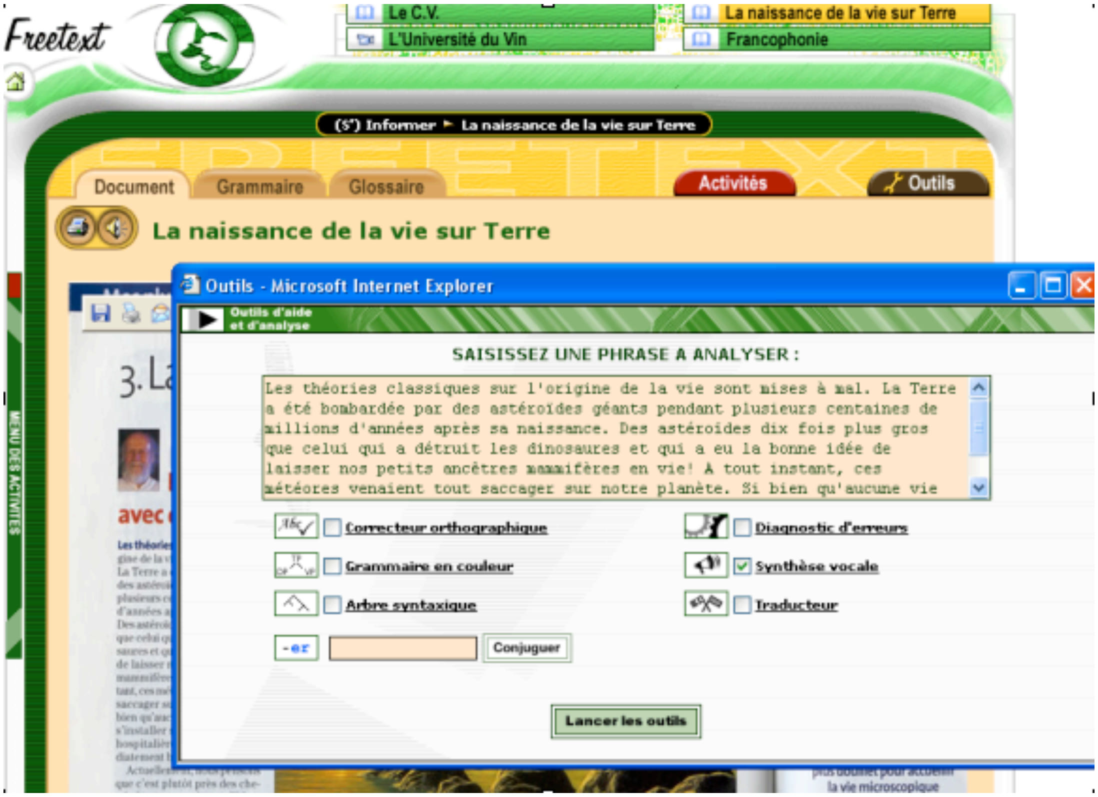
1 (de la vue, l'ouïe) organ;

2 (publication) organ; **organe officiel d'un parti** official organ of a party;

Figure 1. Interface of the talking text facility provided in the *Oxford-Hachette French Dictionary on CD-ROM*

Talking Texts

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. It can be used by the learner to support his or her reading comprehension activities and/or to check the pronunciation of individual words, expressions and/or full sentences (Hamel, 2003a). The *Oxford-Hachette French Dictionary on CD-ROM* also integrates such a facility as does *FreeText* (see Figure 2) a CALL program for advanced learners of French (Hamel, 2003b) which reuses the TTS system *FIPSvox* (Gaudinat & Werhli, 1997).


Freetext
Le C.V.
L'Université du Vin
La naissance de la vie sur Terre
Francophonie
(S') Informer ▸ La naissance de la vie sur Terre
Document
Grammaire
Glossaire
Activités
Outils
La naissance de la vie sur Terre
Outils - Microsoft Internet Explorer
Outils d'aide et d'analyse
MENU DES ACTIVITES
3. La
avec
la vie microscopique

SAISISSEZ UNE PHRASE A ANALYSER :

```
Les théories classiques sur l'origine de la vie sont mises à mal. La Terre
a été bombardée par des astéroïdes géants pendant plusieurs centaines de
millions d'années après sa naissance. Des astéroides dix fois plus gros
que celui qui a détruit les dinosaures et qui a eu la bonne idée de
laisser nos petits ancêtres mammifères en vie! A tout instant, ces
météores venaient tout saccager sur notre planète. Si bien qu'aucune vie
```

□ Correcteur orthographique
□ Diagnostic d'erreurs
□ Grammaire en couleur
☑ Synthèse vocale
□ Arbre syntaxique
□ Traducteur
-er [ ] Conjuguer
Lancer les outils

Figure 2. TTS in CALL program *FreeText*

Another use of talking texts in CALL programs is for the presentation of prompts (Gray, 1984). An example of a CALL system that integrates speech synthesis specifically for the purpose of reading aloud prompts and providing pronunciation models of sentences in grammar exercises in Dutch is the *Appeal* ("A pleasant personal environment for adaptive learning") system (de Pijper, 1997), which generates grammar exercises "on the fly" in response to individual learner requirements "according to predefined models" (p. 581). This is made possible by the unique feature of TTS to generate speech models on demand.

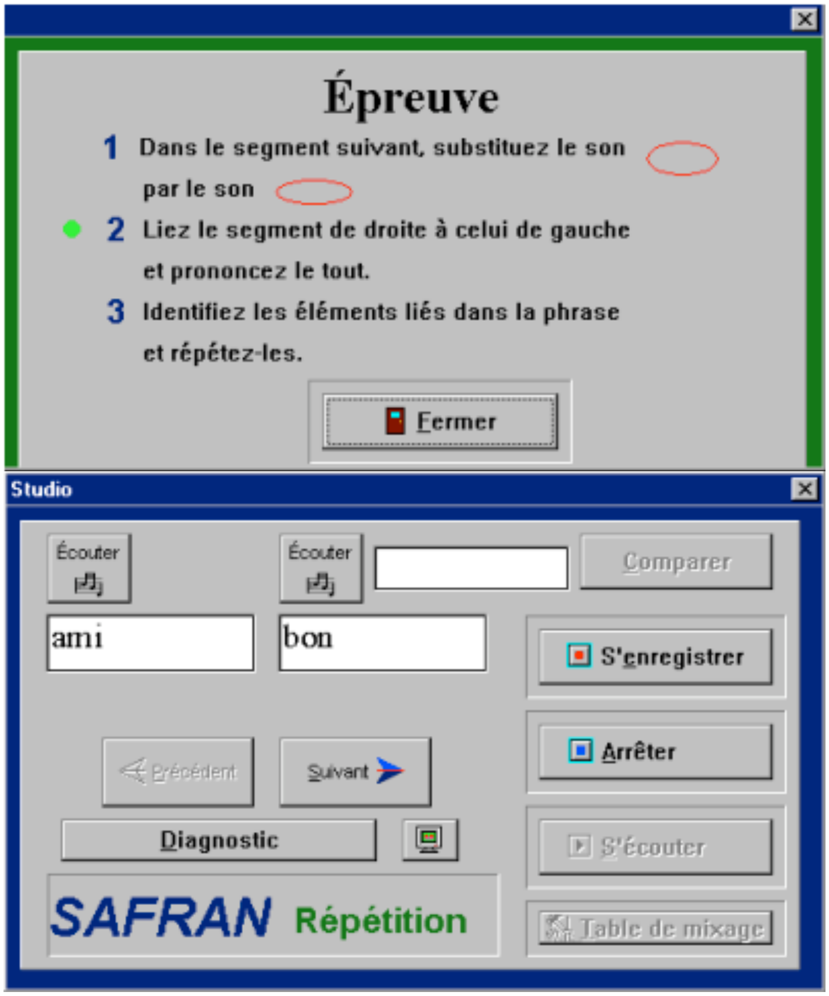
Dictation

Dictation is a traditional writing activity in which the teacher reads aloud a text which the learner is asked to transcribe. This activity, which focuses on the learner's perception, comprehension, and spelling (Ur, 1984), becomes quick and simple for the teacher to create when TTS is used. Exercises can be created simply by typing in the text or copying and pasting it from another electronic source. *DICTOR* (Santiago-Oriola, 1999) and *Ordictée* (Mercier, Guyomard, Siroux, Bramoullé, Gourmelon, Guillou, & Lavannant, 2000) are examples of CALL applications dedicated to dictation. Both of these systems integrate automatic error detection systems. In addition, *Ordictée* is unique in that it adapts the rate of presentation of the dictation, that is, the speech rate, to the rate at which the learner types in his/her transcription of the text.

Pronunciation Training

Speech synthesis can be exploited for pronunciation training at both the segmental (practice of individual and combined phonemes) and supra-segmental (practice of intonation and prosody) levels.

Practice of Individual and Combined Phonemes. 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; see Figure 3; Hamel, 1998, 2003a), which also reuses the TTS system FIPSvox (Gaudinat & Wehrli, 1997), focuses on this kind of practice. Three main types of activities are proposed by SAFexo: auditory discrimination, repetition, and phoneme/segment manipulation. In all three cases, the speech synthesizer is used as a model to imitate and a model with which a learner can compare his or her own pronunciation.



Épreuve

1 Dans le segment suivant, substituez le son par le son

● 2 Liez le segment de droite à celui de gauche et prononcez le tout.

3 Identifiez les éléments liés dans la phrase et répétez-les.

ami

bon

Figure 3. TTS in the experimental pronunciation tutor *SAFexo*

Practice of Intonation and Prosody. An example of a CALL application that uses TTS in the teaching of prosody is Mercier et al.'s (2000) prosodic tutor for Breton. The use of TTS enables teachers to create pronunciation exercises simply by typing in the orthographic transcription of the pronunciation models to be presented.

According to Skrelin and Volskaya (1998), "the generalized [prosodic] contours free from additional emotional colouring are used as models" (p. 24) typical of TTS make it particularly suitable for CALL because they allow the learner to focus on prosody without the distraction of emotional coloring. The utility of synthetic speech for teaching pronunciation was in fact demonstrated much earlier by Feldman (1977) described in Knoerr (2000). In an experiment in which he compared learners' ability to discriminate among intonation patterns produced by a speech synthesis system with their ability to discriminate among the same intonation patterns produced by a native speaker, he found that learners found it easier to discriminate between simplified examples produced with speech synthesis compared with examples produced by native speakers (Knoerr, 2000).

Dialogue 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. The dialogues proposed in systems use digital recordings of human speakers whether presented as open -- "learners have to come up with a response entirely on their own" (Ehsani & Knodt, 1998, p. 55) -- or closed -- learners select responses from a pre-defined list, dialogues are closed in the sense that the range of possible responses that the system can recognize and respond to are predetermined and hence limited (Ehsani & Knodt). Speech 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 speech synthesis that are currently being developed for use in language learning include the *Let's Go Spoken Dialogue System (SDS)* (Raux & Eskenazi, 2004) and *SCILL (Spoken Conversational Interaction for Language Learning)* system (Seneff, Wang, & Zhang, 2004).

Advantages of Using TTS in CALL

Through the presentation of the potential uses of TTS in CALL, several of the benefits of the use of TTS in this context have been presented:

- the low storage requirements of TTS,
- the ability of TTS to generate speech models on demand,
- the ease of creation/modification of exercises, and
- the suitability of the generalized prosodic contours characteristic of TTS for teaching pronunciation.

Most of the applications described are in fact already provided in CALL programs through the integration of digital recordings of human speakers. The use of TTS brings improvement to the provision of these activities. Improvement is of course interesting. But, what is really interesting is the value that a technology such as TTS can add to CALL, that is, the new possibilities that a technology can bring about. The new possibility that TTS brings to CALL is the ability to generate speech models on demand. In addition to improving dialogue simulation, by permitting the provision of more open dialogues (Egan & LaRocca, 2000), this unique feature of TTS makes it possible to produce oral feedback on demand to unanticipated learner responses in computer-assisted pronunciation training (CAPT; Sherwood, 1981) and other exercises, and to generate speech output for any text on demand, thus the provision of talking text facilities (Hamel, 2003a, 2003b), and the automation of the creation of grammar exercises and so forth with spoken language support (de Pijper, 1997).

Roles of TTS in CALL

Different setups or operational contexts often impose different requirements and therefore require different methods of evaluation (Sparck Jones & Galliers, 1996). TTS is used in three different roles within CALL applications. In talking dictionaries, talking texts and dictation systems it is used as a *reading machine.* In CAPT systems it is used to provide *pronunciation models* both at the segmental and suprasegmental level. And, in dialogue systems, it is used to provide the voice of a *conversational* *partner*. Of these roles, the most common role that TTS assumes outside the CALL context is that of a reading machine. Examples of applications in which it assumes this role include reading machines for the blind, screen readers for people with visual impairments and learning disabilities such as dyslexia and aphasia, and talking word processors. Although the use of TTS as a reading machine outside CALL is already widely accepted, the CALL setup differs from these operational contexts, the most important difference being that the main users of CALL are learners, that is non-native speakers of the language being spoken by the TTS synthesizer. It is therefore not right to assume that TTS is suitable for use in CALL as a reading machine without evaluation of the TTS in that specific operational context. Nor should we assume that, if TTS is suitable for use as a reading machine in CALL, it is also suitable as a pronunciation model and as the voice of a conversational partner in that context. Evaluation of TTS in these specific operational contexts is therefore also necessary. In the following section, evaluations of TTS for CALL purposes are reviewed.

EVALUATING SPEECH SYNTHESIS FOR CALL PURPOSES

An Infrastructure for the Evaluation of Speech Synthesis for CALL Purposes

According to Chapelle (2001a, 2001b), CALL software would benefit from three stages of evaluation. In the first stage, she recommends the judgmental evaluation of the CALL application for its potential to provide (ideal) conditions that promote SLA. These conditions are presented later in the paper when we look at the requirements of TTS for CALL. Then, in the second stage, she recommends that a similar evaluation of the activities that teachers plan around the CALL software be carried out. And, finally, in her third stage of evaluation, she recommends that learners' performance in those activities be empirically evaluated. There are two aspects to this level of evaluation: assessment of learning outcomes and assessment of the interactional processes in which learners engage.

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

- *Basic research evaluation* determines 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* determines whether a system meets its objectives, and is typically achieved through measurement of the performance of the system in a control task.
- *Usage evaluation* determines whether a system fulfils its function in a given operational context, determines whether end-users find the system useful and easy to use, and whether the system meets its objectives.
- *Impact evaluation* assesses the effects of the system beyond its primary function, such as the socio-economic effects of its use.
- *Program evaluation* determines whether a funding program was worthwhile, that is, whether the investment resulted in progress.

Of these levels of evaluation (near) equivalents to both technology evaluation and usage evaluation are found in the other infrastructures.[3] Basic research evaluation, impact evaluation, and program evaluation are 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 & 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, that is, where a system fails to meet user requirements (Hirschman & Thompson). 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 taxonimization of possible inputs, that is the production of a profile of a system's performance, and technology evaluation (Hirschman & Thompson).

When a new technology, such as TTS, is being considered for integration into CALL software, we suggest that Chapelle's (2001a, 2001b) framework for the evaluation of CALL software should be extended to include two further stages of evaluation, namely *basic research evaluation* and *adequacy evaluation,* used here to refer to the combination of adequacy evaluation and formative evaluation. Generally, when a funding program has been involved in the development of the software, program evaluation should also be conducted. Regarding impact evaluation, as we shall present later in the paper, 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.

PURPOSES

Evaluation of TTS for CALL: State of the Art

Our review of the literature reveals that very few "formal" evaluations have been conducted. Identification of the potential benefits TTS could bring to CALL could be considered to fulfill the function of basic research evaluation. One report of an evaluation of the adequacy of TTS for use in CALL was found in the literature. In this evaluation, the quality 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, Weston, & Burkhardt, 1987). In addition, several evaluations of learners' performance in CALL activities have been conducted. Two of these evaluations focused on learning outcomes. The first compared learners' performance in French dictation exercises presented by the dictation system *DICTOR* integrating the TTS synthesizer *TELEVOX* with their performance in the same exercises presented by the same system but integrating digitized speech (Santiago-Oriola, 1999). And, the second measured the effectiveness of the use of the KTH (Kungliga Tekniska högskolan, Royal Institute of Technology) Swedish TTS synthesizer in conjunction with the speech editor *WaveSurfer* in the teaching of the lexical stress of English to speakers of Swedish (Hincks, 2002). One evaluation focused on the learning processes which child learners of French as a foreign language engaged in when working with the CALL program *Composition,* a program which allows learners to produce pictures by selecting words from a pre-programmed list and provides facilities for them to subsequently write stories about those pictures and receive "vocal" feedback on those productions (Cohen, 1993). And two final studies focused on reactions to CALL programs that integrate speech synthesis. The first evaluated user, teacher and learner, reactions to the use of the aforementioned Spanish TTS system for the presentation of grammar exercises in a language laboratory setting (Stratil, Burkhardt, Jarratt, & Yandle, 1987). And, the second assessed learners' reactions to aforementioned dictation training program, *DICTOR*, and compared their reactions to the output of the TTS system *TELEVOX* with those to the output of the digital audio system *ECHOVOX* (Santiago-Oriola).

In summary, not all levels of evaluation have been addressed. And, regarding adequacy and usage evaluation, evaluations have been conducted in only a few operational contexts, and only a few TTS systems have been evaluated. Moreover, the evaluations are out of date given recent progress in speech synthesis.

Yet, every speech synthesizer intended for use in CALL would benefit from all six levels of evaluation for each role in which it is to be used: The quality of the output of different speech synthesizers differs greatly (Huang, Acero, & Hon, 2001); and, as said, different operational contexts may impose different requirements.

Benchmarking TTS for CALL

In order to achieve a balance between development and evaluation, benchmarking, is commonly used in software evaluation in general (Grace, 1996; Lindgaard, 1994; Ralston et al., 2000), and the evaluation of SALTs (Sparck Jones & Galliers, 1996; van Bezooijen & 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, p. 7), benchmarking has a long history and today is used in a wide range of fields including surveying, management, economics, education, and computing. 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 compared the scores obtained by their systems with those obtained by their competitors, and potential end-users compared 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, and so forth, as well as to the performance (Cai, Nerurkar, & Wu,1998) and usability (ease of use) of computer software (Lindgaard, 1994). Benchmarking the usability of computer software commonly 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 organizations gets 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, Easterbrook, & Holt, 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 organization 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. (p. 497)

As pointed out by van Bezooijen and van Heuven (1997), benchmarking "is more efficient than pairwise or multiple testing of competing products" (p. 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 & Galliers, 1996). Benchmarking could therefore provide a good solution to the neglect of the evaluation of TTS for CALL purposes.

TOWARD BENCHMARKS FOR THE EVALUATION OF TTS FOR CALL

In this section, we present our progress towards the development of a benchmark for the evaluation of TTS for CALL. Specifically, we present our work towards the development of a benchmark for the evaluation of French TTS for use in CALL programs for learners of French as a foreign language.

The Evaluation Process

According to the EAGLES guidelines for the evaluation of SALTs (EAGLES, 1999), the first stage in the evaluation process consists in identifying the object and purpose, that is, basic research evaluation, adequacy evaluation, and so forth, of the evaluation. The next stage consists in the analysis of the requirements of the application and the identification of attributes of the system that can be reported in order to get at those requirements. And, the following stage consists in the selection of metrics (i.e., methods and scales for the measurement of the reportable attributes). Once metrics have been selected, it is necessary to define what constitutes a desirable, acceptable, and unacceptable score (i.e., to establish rating levels). Finally, before the evaluation is conducted, an evaluation plan must be drawn up (i.e., a description of the evaluation methods and schedule of the evaluation).

Here the object of evaluation is French TTS, and the purpose is to determine whether it is adequate for use in the different CALL contexts, that is as a reading machine, a pronunciation model, and the voice of a conversational partner.

Requirements Analysis

According to Chapelle (1998), the goal of CALL applications should be to provide "ideal" conditions for SLA. These ideal conditions are therefore the requirements of CALL.

While Chapelle's (2001a, 2001b) literature review focused on ideal conditions for the acquisition of grammar, a review of experiments in the field of spoken language acquisition (Colotte, Laprie, & Bonneau, 2001; Protopapas & Calhoun, 2000), research on teacher talk (Ellis, 1994) and best practice in LLT (Celce-Murcia, Brinton, & Goodwin, 1996; LeBel, 1990; Pennington, 1996) suggests that the same conditions are desirable for the acquisition of pronunciation and spoken language in general.

Of Chapelle's (2001a, 2001b) criteria, language learning potential was identified as the first requirement of CALL. 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. Regarding the use of TTS in CALL, in order to match this first criterion, the *quality of the output* should be such that it is as *comprehensible, natural,* and *accurate* as possible.

Comprehensibility, the ease with which a listener can understand a speaker's intended message (Francis & Nusbaum, 1999), is a central requirement of the quality of the pronunciation of speech synthesizers for use in CALL, because for the majority of learners the goal of language learning is to produce speech that is comfortably comprehensible (Kenworthy, 1987).

Naturalness, the ability to sound native-like, and *accuracy,* error-free speech, are additional requirements. Naturalness and accuracy are the goals of specific groups of learners such as those conducting business on equal terms with natives, and those who wish to become teachers of the foreign language (Kenworthy, 1987).

Regarding focus on linguistic form, one way in which focus on form can be achieved is through interactional modification (Chapelle, 2001a, 2001b). A TTS system for use in CALL should therefore also provide the means to achieve a certain level of interactional modification. Interactional can be achieved through *flexibility,* that is the possibility to manipulate features of the speech output including, and not restricted to, the voice, the style, the speech rate, and the pitch. The flexibility and quality of the output of TTS systems are tested in different ways. The case study presented here focuses on the latter.

Case Study

The object of the case study presented here is to compare the levels of *appropriateness* and *acceptability* found when the TTS synthesizer FIPSvox (Gaudinat & Wehrli, 1997) is used as (a) a reading machine, (b) a pronunciation model, and (c) the voice of a conversational partner for the teaching/learning of French as a foreign language with a view to determining whether these different roles impose different requirements on the quality of the output. It also looks at the *comprehensibility* and the *accuracy* of the output of the speech synthesizer. Specifically, it looks at the relationship between these features of the output of the speech synthesizer and the appropriateness and acceptability levels of the output for use in CALL with the objective of determining whether they are good indicators of the appropriateness and acceptability of the output of speech synthesizers for use in CALL applications.

Although not the goal of this case study, ratings of *appropriateness* and *acceptability* also give us an indication of the readiness of the particular speech synthesizer evaluated in this case study for use in the three different roles in CALL applications.

Method. Twelve French native speakers all involved in the teaching of French as a foreign/second language at university level and/or in CALL research were recruited to participate in the case study. Participants were run individually.[4]

Sixty utterances, 20 representative of each of the three roles that speech synthesis might assume in a CALL application, were synthesized using the research speech synthesizer *FIPSvox* (Gaudinat & Wehrli 1997). The output of the speech synthesizer was collected for presentation to participants as pre-recorded utterances by means of an MS Power Point presentation. The corpus of utterances was selected from existing CALL software and one LLT exercise manual. For the reading machine corpus, 20 sentences from the text *Les voleurs d'écriture* (Begag, 1990) exploited in *FreeText* (Hamel, 2003b) were used. For the pronunciation corpus, 20 utterances, each focusing on a different aspect of the pronunciation of French, were selected from *La Portée des Sons* (Garant-Viau, 1994). For the conversation corpus, 20 consecutive turns were selected from one of the simulated dialogues in the ASR-based (Automatic Speech Recognition) CALL program *Talk to Me French: The Conversation Method* by Auralog.

The utterances were presented one per slide and blocked by role, the order of presentation of which was randomized in order to overcome order effects including fatigue and familiarization with the voice of the speech synthesizer. Reflecting CALL principles, the case study was self-paced; participants could move on to the next utterance precisely when they were ready, and could also return to a previous utterance if desired.

In a first pass, participants were asked to rate the comprehensibility and the acceptability of the output for each utterance with respect to its use in the role indicated. Comprehensibility was to be rated on a scale of +3 (*very easy to understand*) to -3 (*very difficult to understand*), and acceptability rated on a scale +3 (*entirely acceptable*) to -3 (*not at all acceptable*). At the end of each block, participants were also asked whether the output of the speech synthesizer was appropriate for use in the role indicated, and to rate its appropriateness on a scale of +3 (*yes, very appropriate*) to -3 (*no, not at all appropriate*). When the participants had completed this first pass for all three roles, they were then asked to indicate for which if any of the roles the speech synthesis was most suitable, and for which if any the speech synthesis was least suitable.

On a second pass, participants were asked to highlight any local errors and underline any global errors in the output that they believed affected its suitability for use in the function indicated. At the end of each block, they were asked to indicate the types and frequency level of errors highlighted and underlined previously and to rate the seriousness of those classes of error with respect to the use of speech synthesis in that role in CALL applications. Participants were provided with a list of pre-defined error classes to which they were invited to add any additional classes of error which they had come across. The frequency of the different classes of error was rated on the scale *very frequent, quite frequent,* and *hardly frequent,* and their seriousness was rated on the scale +3 (*very serious*) to –3 (*not at all serious*). Finally, participants were asked to fill in a questionnaire on their familiarity with speech synthesis in general and in CALL.

Results. Regarding acceptability and comprehensibility, overall ratings of acceptability and comprehensibility of utterances in each function were calculated for each subject by taking the mean of their ratings of each utterance across each corpus.[5] The results are presented in Table 1. Figures 4 and 5 present the rating scales used.

Table 1. Mean ratings of overall appropriateness, acceptability, comprehensibility (n = 12)

	Reading	Pronunciation	Conversation
Appropriateness	0.17	–1.50	0.50
Acceptability	0.82	0.06	1.49
Comprehensibility	1.21	0.91	2.07

Yes, very appropriate	+3	+2	+1	0	-1	-2	-3	No, not at all appropriate

Figure 4. Appropriateness rating scale

Very acceptable	+3	+2	+1	0	-1	-2	-3	Not very acceptable
Very easy to understand	+3	+2	+1	0	-1	-2	-3	Very difficult to understand

Figure 5. Acceptability and comprehension rating scales

Regarding accuracy, we based our measure on the participants' ratings of the frequency and seriousness of the different classes of errors observed by the participants in the three corpora. It was calculated as follows: First, ratings of the frequency of each type of error were converted to a numerical scale (*not present* = 0, *hardly frequent* = 1, *quite frequent* = 2, *very frequent* = 3); then the ratings of the seriousness of each type of error were converted from a scale of +3 (*very serious*) to –3 (*not at all serious*) to a scale of 1 (*not at all serious*) to 7 (*very serious*); and finally, accuracy for each individual role was calculated by taking the mean of the sum of the product of the ratings of the frequency and seriousness of each type of error on these scales across participants. The results are presented in Table 2. Figure 6 presents the rating scales used.

Table 2. Mean ratings of overall accuracy (n = 12)

	Reading	Pronunciation	Conversation
Appropriateness	95.00	89.83	76.42

		Frequency of error			Seriousness of error								
Class	Tick if present	Very frequent	Quite frequent	Hardly frequent	Very serious	+3	+2	+1	0	-1	-2	-3	Not at all serious
Inappropriate rhythm													

Figure 6. Accuracy rating scales

The total number of times each role was rated to be most appropriate and least appropriate was also calculated across the 12 participants. These results are presented in Table 3. As said in the method, participants could select none, one or more than one role.

Table 3. Most and Least Appropriate Uses of Speech Synthesis in CALL (*n* = 12)

	Reading	Pronunciation	Conversation	None
Most Appropriate	4	1	5	2
Least Appropriate	3	8	3	0

In order to determine whether a relationship existed between comprehensibility and acceptability, and if so, to determine the nature of that relationship, each participant's overall rating of comprehensibility, the speech was plotted against their overall rating of the acceptability of the speech for each role (Figures 7, 8, and 9).

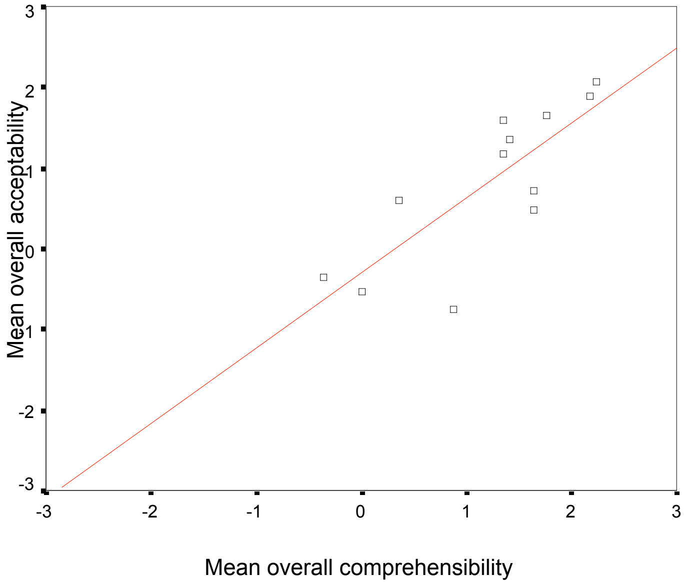


Figure 7. Scatter graph of ratings of comprehensibility against acceptability for the use of TTS as a reading machine

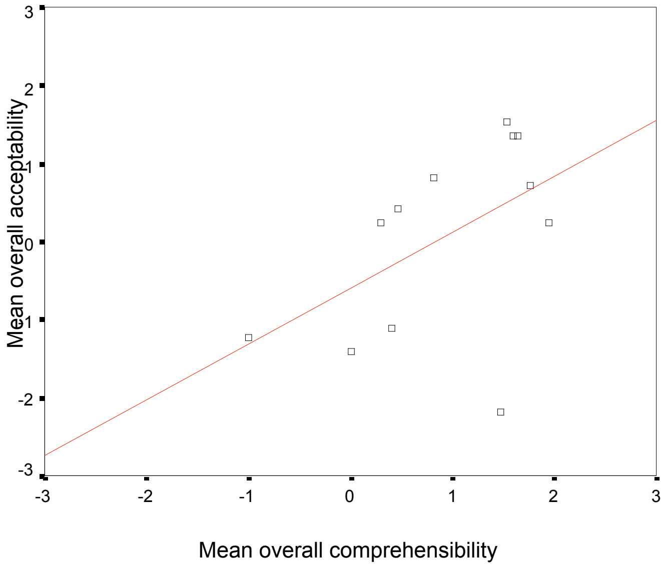


Figure 8. Scatter graph of ratings of comprehensibility against acceptability for the use of TTS as a pronunciation model

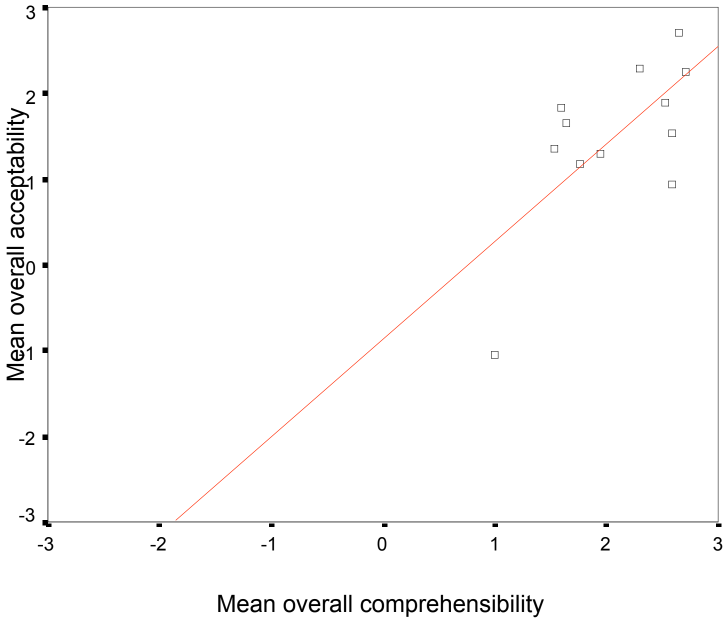


Figure 9. Scatter graph of ratings of comprehensibility against acceptability for the use of TTS as the voice of a conversational partner

Since the measures of accuracy and appropriateness were comparable, an analysis of this relationship was carried out. That is, each participant's rating of the accuracy the speech was plotted against their rating of the appropriateness of the speech for each role (Figures 10, 11, and 12).

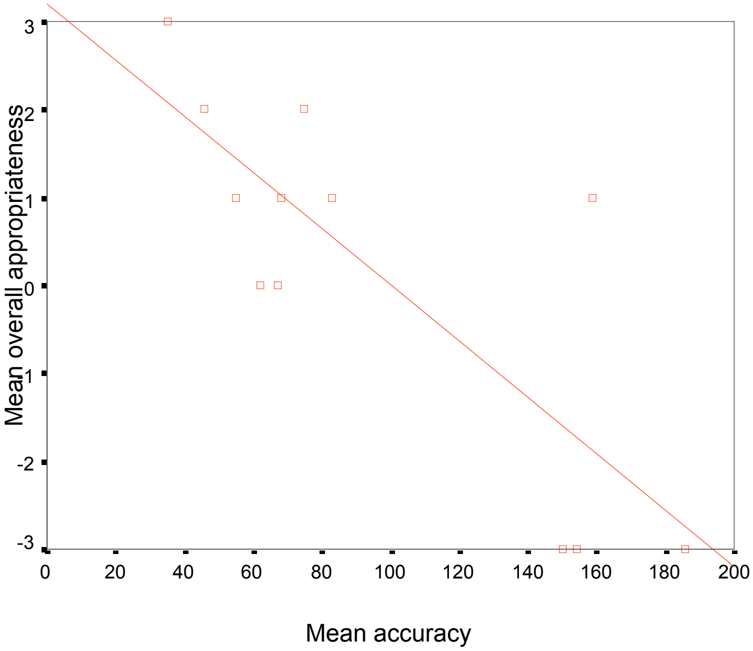


Figure 10. Scatter graph of ratings of accuracy against appropriateness for the use of TTS as a reading machine

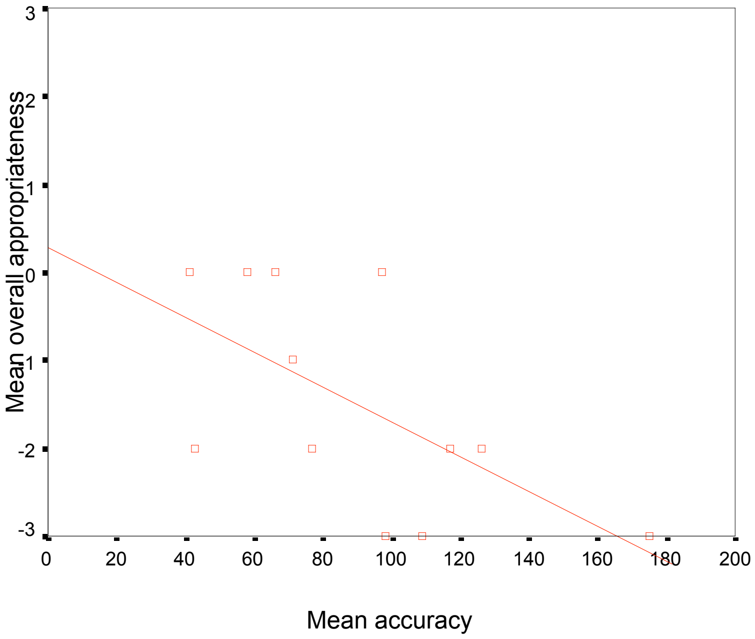


Figure 11. Scatter graph of ratings of accuracy against appropriateness for the use of TTS as a pronunciation model

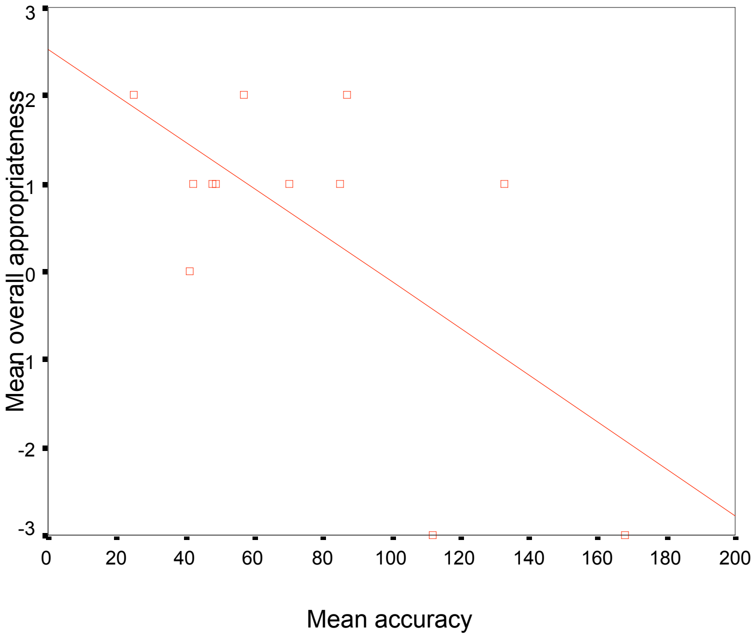


Figure 12. Scatter graph of ratings of accuracy against appropriateness for the use of TTS as the voice of a conversational partner

Discussion. Ratings of appropriateness and acceptability differ for the three different roles. In particular, the speech was found to be most appropriate and acceptable for use in the role of the voice of a conversational partner, and least comprehensible in the role of a pronunciation model. More specifically, the ratings of the appropriateness and the acceptability of the output of the speech synthesizer for use as a pronunciation model differ more from those of the appropriateness and the acceptability of the output for use as both a reading machine and the voice of a conversational partner, than those of the appropriateness and the acceptability of the output for use as both a reading machine and the voice of a conversational partner do from one another. Similarly, while the use of speech synthesis as a pronunciation model is rated the least appropriate use by participants, the use of speech synthesis as a reading machine and as the voice of a conversational partner are rated similarly, the former was rated by 4 participants to be the most appropriate use of speech synthesis in CALL, and the latter by 5 participants, and both were rated by 3 participants to be the least appropriate use of speech synthesis in CALL. This all suggests that, as hypothesized, the different functions of speech synthesis in CALL have different requirements and the technology is more suitable and ready for use in some functions than in others.

Regarding comprehensibility, ratings also differed across the three roles. Just as the speech was found to be most appropriate and acceptable for use in the role of the voice of a conversational partner, and least comprehensible in the role of a pronunciation model, so too the utterances in the conversation corpus were found to be the most comprehensible and those in the pronunciation corpus to be the least comprehensible. We would suggest that this is due to the fact that the utterances in the conversation corpus are in general shorter and less syntactically complex than those in the reading corpus and more predictable than those in the pronunciation corpus.

Regarding the contribution of comprehensibility to appropriateness and acceptability, as said due to missing data, it was only possible to analyze the relationship between comprehensibility and acceptability. As predicted, in Figure 7 a positive correlation can be seen between ratings of comprehensibility and acceptability for the use of the speech synthesizer as a reading machine. Contrary to our hypothesis, such a correlation is, however, not evident for the use of the speech synthesizer either as a pronunciation model (see Figure 8) or as the voice of a conversational partner (see Figure 9). We are therefore led to question whether comprehensibility has a role to play in determining the acceptability of speech synthesis for use as a pronunciation model and as the voice of a conversational partner. If it does not, we would have further evidence to support our other hypothesis, the hypothesis that the different functions of speech synthesis in CALL will have different requirements and consequently that the technology will be more suitable and ready for use in some functions than in others.

As said, a measure of the accuracy of the output was also obtained from the participants' ratings of the frequency and seriousness of the errors that they observed in the three corpora. Before we continue with our analysis of the results, we should remind ourselves that the higher this measure, the lower the accuracy of the output. Looking at the results we find that accuracy differs across the three roles. Specifically, the output was found to be most accurate for the conversation corpus, and least accurate for the reading corpus, with the reading and pronunciation corpora scoring similarly on average. One explanation for these results could be that the utterances to be synthesized were longest and most complex (with respect to syntax) in the reading corpus, and shortest and least complex in the conversation corpus.

Regarding the contribution of accuracy to appropriateness and acceptability, as said, due to missing data, it was only possible to investigate the relationship between accuracy and appropriateness. Contrary to our hypothesis, no correlations are evident between the ratings of accuracy and appropriateness for any of the roles (see Figures 10, 11, and 12). We are therefore led to question whether accuracy plays a role in determining the appropriateness of speech synthesis for use in CALL.

Conclusion. In order to be in a position to draw firmer conclusions about the tendencies observed in our case study, further investigation of the requirements is necessary involving several different speech synthesizers, a larger sample size and greater contextualization. Regarding contextualization, the purpose of adequacy evaluation is to avoid wasting time and resources integrating a technology into an application for which it is not suitable. We should therefore be careful not to confuse contextualization with integration. Contextualization will therefore be a challenge for further requirements analysis. Greater contextualization could be achieved through an explanation of TTS, its potential applications in CALL including mock screen shots of those applications, and the potential benefits of its use in those applications.

SUMMARY AND FURTHER WORK

In this article we presented our preliminary work towards the development of a benchmark for the evaluation of the adequacy of French TTS for use in CALL applications for teaching French as a foreign language. Specifically, the results of a case study designed to investigate the requirements of TTS for use in CALL was presented. The results of this study, which compared the acceptability and appropriateness of a research TTS system for use in CALL in the roles of reading machine, pronunciation model, and conversational partner, provided some preliminary evidence to suggest that these roles imposed different requirements on the quality of speech produced by the TTS system: The acceptability and appropriateness of the TTS system differed for the three roles.

Regarding the nature of the requirements that these different roles impose on the quality of the output of the TTS system, comprehensibility was found to correlate with acceptability for when the speech synthesizer was used as a reading machine, but not when it was used as either a pronunciation model or as a conversational partner. And, no correlations were found between the ratings of accuracy and appropriateness for any of the roles. We are therefore led to question whether the features identified in the literature do in fact determine the acceptability and appropriateness of TTS for use in CALL.

It was not, however, possible to draw any meaningful conclusions from this study due to the small sample size and the fact that only one speech synthesizer was evaluated. As said, further investigation of the requirements is necessary involving more speech synthesizers, a larger sample size and greater contextualization. Once the requirements have been identified, the next stage in the evaluation process will consist in identifying metrics that could be used to get at these requirements, that is the identification of benchmark tests.

We are currently in the process of conducting such an analysis of the requirements and hope to be able to provide suggestions for the selection of benchmark tests in another paper in the near future. Once selected, these tests will need to be conducted with large groups of participants in order to establish benchmark scores. As said, CALL also imposes requirements on the flexibility of TTS systems. Investigation of this requirement should also be carried out in order to permit the selection of benchmark tests and the establishment of benchmark scores for its evaluation. A benchmark for the evaluation of the adequacy TTS synthesis systems for use in CALL is therefore still a long way off. And, the field would benefit from the development of benchmarks for the evaluation of the potential of the CALL program to provide ideal conditions for SLA; the potential of the teacher-planned CALL activity to provide ideal conditions for SLA; and, the learner's performance in the CALL activity.

APPENDIX A
CALL Programs Integrating Speech Synthesis

Talking Dictionaries

Etaco Partner from Etaco (various languages), http://translatingtheworld.com/ectaco/ectacoau/html/Budget_c44.html

Spanish for Business Professionals by L. Kirk Hagen of the University of Houston-Downtown, http://www.dt.uh.edu/research/sbp/HOME.html

Net Dictionary from the Virtual Learning Centre (VLC), http://www.edict.com.hk/lexiconindex/

Oxford-Hachette French Dictionary on CD-ROM, Version 2.0, http://www.oup.com/

APPENDIX B
Online TTS Demonstrations

Research Systems

Festival from CMU (multilingual), http://www-2.cs.cmu.edu/~awb/festival_demos/

Fipsvox from Latl, Geneva (English and French), http://www.latl.unige.ch/french/projets/Synthetizer/synthetizer.html

KALI from the University of Caen (French), http://elsap1.unicaen.fr/KaliDemo.html

Commercial Systems

AT&T (multilingual), http://www.naturalvoices.att.com/demos/

Elan *Sayso* (English, French, German, Italian, and Spanish), http://sayso.elan.fr/interactive_vf.asp

Microsoft (English, French and German), http://www.microsoft.com/reader/downloads/tts.asp

Realspeak from Scansoft (multilingual), http://www.scansoft.com/realspeak/demo/

Rhetorical Systems (English [various accents], German, Greek, and Spanish), http://www.rhetoricalsystems.com/

NOTES

1. A list of CALL programmes that integrate speech synthesis is provided in Appendix A.

2. A list of on-line interactive TTS synthesis demonstrations is provided in Appendix B.

3. What ELSE refer to as *technology evaluation* is also known as *intrinsic evaluation* (Sparck Jones & Galliers, 1996), *performance evaluation* (Hirschman & Thompson, 1996), and *summative evaluation* (Hirschman & Thompson, 1996). And, what ELSE refer to as *usage evaluation* is also known as *extrinsic evaluation* (Sparck Jones & Galliers, 1996).

4. The Web platform WebCT was used for the distribution of the experiment protocol and the various experiment materials (PowerPoint presentation, .wav sound files, and MSWord response sheet). At the end of the experiment, participants submitted their response sheets by uploading them to the WebCT platform. Some participants did not have access to the Web and others were not familiar with WebCT. Hard copies of the experiment materials were posted to these participants who also returned their response sheets by post.

5. Three utterances in each sub-corpus were eliminated due to missing data (i.e., overall ratings of acceptability and comprehensibility for each role were calculated over 17 utterances).

ACKNOWLEDGEMENTS

We would like to thank Professor H. Somers for his valuable input and all the CALL researchers and French teachers who kindly participated in our case study.

ABOUT THE AUTHORS

Zöe Handley is a PhD student previously in the Centre for Computational Linguistics, UMIST, UK, now in the School of Informatics, The University of Manchester, UK. She is interested in the use speech technologies in CALL. Her research is focused on the evaluation of speech synthesis for use in CALL, specifically establishing a methodology for benchmarking speech synthesizers for use in CALL.

E-mail: zoe.handley@postgrad.manchester.ac.uk

Marie-Josée Hamel has been involved in CALL teaching, research and development since 1994. At the time of this case study, she was at the Centre for Computational Linguistics, UMIST, UK. She is now associate professor of Applied Linguistics at the University of Dalhousie, Nova Scotia, Canada. Her interests are in the reuse of Natural Language Processing (NLP) technologies in CALL and in the contribution of second language acquisition theories to CALL.

E-mail: marie.hamel@dal.ca

REFERENCES

Anderson, A. (2000, April). Spanish for business professionals. *The CALICO Review.* Retrieved January 31, 2005, from http://calico.org/CALICO_Review/review/sbp.htm

Begag, A. (1990). Les voleurs d'écriture [The stolen writings]. Paris: Editions du Seuil.

Cai, J.-Y., Nerurkar, A., & Wu, M.-Y. (1998). Making benchmarks uncheatable. *Proceedings of the IEEE International Computer Performance and Dependability Symposium* (IPDS '98; pp. 216-226). Durham, NC: IEEE Computer Society.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages.* Cambridge, England: Cambridge University Press.

Chapelle, C. A. (1998). Multimedia CALL: Lessons to be learned from instructed SLA. *Language Learning & Technology, 2*(1), 22-34. Retrieved January 31, 2005, from http://llt.msu.edu/vol2num1/article1/

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, England: Cambridge University Press.

Codling, S. (1998). *Benchmarking.* Aldershot, England: Gower.

Cohen, R. (1993). The use of voice synthesizer in the discovery of the written language by young children. *Computers in Educations, 21*(1/2), 25-30.

Colotte, V., Laprie, Y., & Bonneau, A. (2001). Perceptual experiments on enhanced and slowed down speech sentences for second language acquisition. *Proceedings of the European Conference on Speech Communication and Technology* (pp. 469-473). Bonn: ISCA.

de Pijper, J. R. (1997). High-quality message-to-speech generation in a practical application. In J. P. H. van Santen, R. W. Sproat, J. P. Olive, & J. Hirschberg (Eds.), *Progress in speech synthesis* (pp. 575-586). New York: Springer Verlag.

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.

Egan, B. K., & LaRocca, S. A. (2000). Speech recognition in language learning: A must. In proceedings of *InSTIL 2000* (pp. 4-9). Dundee, England: University of Abertay Dundee.

Ehsani, F., & 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/

Ellis, R. (1994). *The study of second language acquisition.* Oxford, England: Oxford University Press.

ELSE (Evaluation in Language and Speech Engineering). (1999, June). *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 January 31, 2005, from http://m17.limsi.fr/TLP/ELSE/

Feldman, D. M. (1977). Measuring auditory discrimination of suprasegmental features of Spanish. *IRAL, 11,* 195-209.

Francis, A. L., & Nusbaum, H. C. (1999). Evaluating the quality of synthetic speech. In D. Gardner-Bonneau (Ed.), *Human factors and voice interactive systems* (pp. 63-97). Boston, MA: Kluwer Academic Publishers.

Garant-Viau, C. (1994). *La portée des sons* [The significance/importance of sounds]. Québec: Université Laval.

Gaudinat, G., & Wehrli, E. (1997). Analyse syntaxique et synthèse de la parole: le projet FIPSvox [Syntactic analysis and speech synthesis: The FIPSvox project]. *TAL, 38*(1), 121-134

Grace, R. (1996). *The benchmark book.* London: Prentice Hall.

Gray, T. (1984). Talking computers in the classroom. In G. Bristow (Ed.), Electronic speech synthesis (pp. 234-259). London: McGraw-Hill.

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

Hamel, M.-J. (2003a). *Re-using natural language processing tools in CALL: The experience of SAFRAN.* Unpublished doctoral thesis. University of Manchester Institute of Science and Technology, UK.

Hamel, M.-J. (2003b). FreeText: A "Smart" multimedia Web-based computer assisted language learning environment for learners of French. In *Proceedings of m-ICTE2003*, volume III (pp. 1661-1665), Badajoz, Spain.

Hetzel, B. (1993). *Making software measurement work: Building an effective measurement program.* Boston: QED Publishing Group.

Hincks, R. (2002). *Speech synthesis for teaching lexical stress.* TMH-QPSR, 44, 153-165

Hirschman, L., & Thompson, H. S. (1996). Overview of evaluation in speech and natural language processing. In R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue (Eds.), *Survey of the state of the art in human language technology* (pp. 175-181). Cambridge, England: Cambridge University Press. http://cslu.cse.ogi.edu/HLTsurvey/

Huang, X, Acero, X., & Hon, H.-W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development.* Upper Saddle River, NJ: Prentice Hall.

Kenworthy, J. (1987). *Teaching English pronunciation.* London: Longman.

Knoerr, H. (2000). Pratique intonative et utilisation d'un logiciel de visualisation dans un cours de prononciation en Français langue seconde: Une etude descriptive [Intonation training using visualisation software in a pronunciation class for learners of French as a second language: A case study]. *Revue Canadienne de Linguistique Appliquée, 3*(1-2), 123-140

LeBel, J.-G. (1990). *Traité de correction phonétique ponctuelle* [Treatise of ad hoc phonetic correction]. Laval, Canada: Les Editions de la Faculté des Lettres, Université Laval.

Lindgaard, G. (1994). *Usability testing and system evaluation: A guide for designing useful computer systems.* London: Chapman & Hall

Mercier, G., Guyomard, M., Siroux, J., Bramoullé, A., Gourmelon, H., Guillou, A., & Lavannant, P. (2000). Courseware for Breton spelling pronunciation and intonation learning. In *Proceedings of InSTIL 2000* (pp. 145-148). Dundee, England: University of Abertay Dundee.

Myers, M. J. (2000). Voice recognition software used to learn pronunciation. In *Proceedings of InSTIL 2000* (pp. 98-101). Dundee, England: University of Abertay, Dundee.

Oxford-Hachette. (2003). *Oxford-Hachet French Dictionary on CD-ROM* (Verions 2.0). Oxford, England: Oxford University Press.

Pennington, M. C. (1996). *Phonology in English language teaching: An international approach.* London: Longman.

Protopapas, A., & Calhoun, B. (2000). Adaptive phonetic training for second-language learners. In *Proceedings of InSTIL 2000* (pp. 31-38). Edinburgh, Scotland.

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

Ralston, A., Reilly, E. D., & Hemmdinger, D. (Eds.). (2000). *Encyclopedia of computer science.* London: Nature Publishing Group.

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

Seneff, S., Wang, C., & Zhang, J. (2004). Spoken conversation interaction for language learning. In R. Delmonte, P. Delcloque, & S. Tonelli (Eds.), *Proceedings of the InSTIL/ICALL 2004 Symposium* (pp. 151-154), Venice, Italy.

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 benchmakring to advance research: A challenge to software engineering. In *Proceedings of the 25th International Conference on Software Engineering* (pp. 74-83), Portland, OR.

Skrelin, P., & Volskaya, N. (1998). Application of new technologies in the development of education programs. In S. Jager, J. Nerbonne, & A. van Essen (Eds.), *Language Teaching and Language Technology* (pp. 21-24). Lisse, The Netherlands: Swets and Zeitlinger.

Sobkowiack, W. (1998). Speech in EFL CALL. In K. Cameron (Ed.), *Multimedia CALL: Theory and practice* (pp. 23-34). Exeter, England: Elm Bank.

Sparck Jones, K., & Galliers, J. R. (1996). *Evaluating natural language processing systems: An analysis and review.* London: Springer.

Stratil, M., Burkhardt, D., Jarratt, P., & Yandle, J. (1987) Computer-aided language learning with speech synthesis: User reactions. *Programmed Learning and Educational Technology, 24*(4), 309-316.

Stratil, M., Weston, G., & Burkhardt, D. (1987). Exploration of foreign language speech synthesis. *Literary and Linguistic Computing, 2*(2), 116-119.

Talk to Me French: The Conversation Method (Version 3.5) from Auralog http://www.auralog.fr

Thomas, C., Levinson, M., & Lessard, G. (2004). Experiments in prosody for the oral generation of French. In R. Delmonte, P. Delcloque, & S. Tonelli (Eds.), *Proceedings of the InSTIL/ICALL 2004 symposium* (pp. 123-126), Venice, Italy.

Ur, P. (1984). *Teaching listening comprehension.* Cambridge, England: Cambridge University Press.

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 (Vol. 3, pp. 481-563). New York: Mouton de Gruyter.