

# *Towards Establishing a Methodology for Benchmarking Speech Synthesis for Computer-Assisted Language Learning (CALL)*

Zöe Handley

Centre for Computational Linguistics, UMIST

PO Box 88, MANCHESTER M60 1QD

[zoe.handley@postgrad.umist.ac.uk](mailto:zoe.handley@postgrad.umist.ac.uk)

## 1. Introduction

In very simple terms, speech synthesis is the process of making the computer talk. As such, speech synthesisers offer another means of providing spoken language input to the learner in CALL environments. Indeed, many potential benefits (ease of creation and editing of speech examples, generation of various kinds of modified input, generation of speech models and feedback on demand, etc.) and uses (as a reading machine, as a model which the learner can use to familiarise him/herself with segmental and suprasegmental distinctions in the Target Language (TL), a model to imitate in pronunciation training and reading exercises, feedback in pronunciation training and reading, and as the voice of a conversational partner in communication-based activities) (Hamel and Handley, to appear) of speech synthesis in CALL have been put forward. Yet, the use of speech synthesis in CALL is not widely accepted (Sobkowiak, 1998; Egan and LaRocca, 2000), and only a few applications have found their way onto the market (for example, the *Etaco* hand-held talking dictionaries, talking dictionaries for PC including *Oxford-Hachette French dictionary on CD-ROM*, and *The Longman Dictionary of American English for Macintosh*, and the CALLware *Spanish for Business Professionals*, etc.).

In my opinion, people are right to be sceptical. The use of speech synthesis in CALL is not proven. Very few evaluations of speech synthesis for and in CALL have been conducted.

One potential reason for the neglect of evaluation is the fact that it competes for time and resources which could be spent on further research and development (Hirschman and Thompson, 1996). In order to achieve a balance between development and evaluation,

*benchmarking* – an easy and efficient method of evaluating competing products – is commonly used in language engineering.

No benchmarks exist for the evaluation of speech synthesis for CALL. The goal of further research on the evaluation of speech synthesis for CALL should therefore be the development of a benchmark. The work presented in this paper constitutes the preliminary stages of the development of such a benchmark.

## 2. Evaluation

Evaluation is an integral part of the life cycle of any product (Dix et al., 1998). With respect to the product life cycle, Blasband and Paroubek (1999) identify five functions of evaluation. The first three of these which apply in chronological order are:

**Basic research evaluation** [which] tries to validate a new idea or to assess the amount of improvement it brings on older methods.

**Technology Evaluation** [which] tries to assess the performance and appropriateness of a technology for solving a problem that is well defined, simplified and abstracted.

**Usage evaluation** [which] tries to assess the usability of a technology for solving a real problem in the field (*ibid.*: 8).

As mentioned, very few ‘formal’ evaluations of speech synthesis for and in CALL have been conducted. Identification of the potential benefits speech synthesis could bring to CALL could be considered to fulfil the function of basic research evaluation. Regarding technology evaluation, the quality of the output of only one speech synthesiser has been evaluated (Stratil et al., 1987a). Yet, every speech synthesiser to be used in CALL should

undergo technology evaluation because the quality of output of different speech synthesisers differs greatly (Huang et al., 2001). As regards usage evaluation, only three speech synthesisers have been evaluated each in a different application: Stratil et al. (1987b) evaluated a TTS synthesiser for the presentation of grammar exercises in a language laboratory; Santiago-Oriola (1999) evaluated a dictation system integrating TTS synthesis; and, Hincks (2002) evaluated the use of a TTS synthesiser in conjunction with a speech editor for teaching lexical stress. Yet, not only should every application that integrates speech synthesis be evaluated, but also every task that a teacher might plan around each application (Chapelle, 2000). In summary, the technology and usage evaluations conducted to date are only partial. Benchmarks have certainly not been established. As mentioned, the use of benchmarks is preferable because evaluation is costly.

Even though CALL applications integrating speech synthesis are already available on the market, it is important to go back to technology evaluation because if it is omitted, considerable time and resources may be wasted integrating the technology into applications it is not suitable for. Also, this may impede acceptance of the technology in applications that it is suitable for. Moreover, technology evaluations enable us to identify the successes and limitations of a technology. This is important because technologies should be exploited for what they are best at (Stevens, 1989).

The aim of the work presented here is therefore to develop a methodology for benchmarking speech synthesisers at the level of technology evaluation.

### 2.1. What is Benchmarking?

*Benchmarking* consists in applying a *benchmark test* “an efficient, easily administered test, or set of tests, that can be used to express the performance of a [...] system [...] in numerical terms” (van Bezooijen and van Heuven, 1997: 497), to a system which “represents a performance level that is known to guarantee user satisfaction” (*id.*), i.e. the reference or *benchmark system*, to obtain a score or *benchmark*, against which the scores obtained

by other systems in the benchmark tests can be implicitly set off.

Before conducting an evaluation, it is necessary for evaluators to set acceptability levels which a system must obtain in order to meet the requirements of the application. Setting acceptability levels in the acoustic level evaluation of speech synthesis systems is difficult due to our limited understanding of speech perception. It is therefore common practice to compare the scores obtained by the system with those obtained by a reference condition which is known to satisfy the requirements of the application. Regarding the selection of reference conditions for use in the benchmarking of speech synthesis systems, van Bezooijen and van Heuven (1997) recommend the use of “a system of proven but still imperfect quality” (*ibid.*: 497) because “human speech will always be superior to synthetic speech, the quality of the latter will have to be expressed as a fraction, which makes it hard to compare relative differences between different types of synthetic speech” (*id.*).

### 2.3. Selection criteria

In the previous section, several criteria that should be taken into account when selecting tests for use in benchmarking came out. In this section further criteria for the selection of tests for benchmarking speech synthesis for CALL are presented.

Very generally, evaluations can be conducted from two perspectives, namely *black-box* and *glass-box*. The former treat systems as a monolith, only evaluating the final output, whereas the latter treat systems as a set of components. Glass-box evaluations therefore permit diagnosis which could be used to inform the appropriation of speech synthesis for CALL. However, glass-box evaluation is not suitable for the comparative evaluation of systems which have different internal architectures such speech synthesisers (Huang et al., 2001). The proposed benchmark will therefore treat speech synthesis systems as a black-box. Such evaluations of speech synthesisers are often referred to as *acoustic level* evaluations because they only take into account acoustic output.

The tests that can be used in such evaluations are classified along a number of

axes: *field vs. laboratory tests, automated vs. human tests, and functional vs. judgement.*

Field evaluations are conducted with real end-users in the environment in which the systems are intended for use, and laboratory evaluations are conducted under an abstract set of application-independent conditions. The second axis distinguishes between the automation of the evaluation process and the use of human subjects respectively. As regards the third axis, in functional evaluations, subjects are asked to complete a task, the results of which indicate how well a system *actually* performed with respect to a particular criterion. Whereas, in judgement evaluations, subjects are asked to rate how well *they think* a system performs with respect to a particular criterion.

When selecting between these different types of test, the features of a ‘good’ test should be taken into account. ‘Good’ tests are “*valid* – they [...] evaluate what they are really supposed to measure – and *reliable* – they [...] measure it consistently” (King, 1995: 98) – and the scores obtained from them “have *predictive validity*” (ISO, 1999: 12), that is they generalise to actual use.

Regarding validity, human evaluations should therefore be preferred over automated evaluations because automated evaluations have not been validated yet (van Bezooijen and van Heuven, 1997). And, functional tests should be preferred over judgemental tests because subjects are not reliable – you can never be sure how subjects will interpret instruction and how religiously they follow them (*ibid.*) – in particular in judgement evaluations. Where judgement evaluations are the only option, large groups of subjects should be used and inter-subjective measurements taken (*ibid.*).

Regarding field vs. laboratory evaluations, the results of field evaluations generalise better to actual use. They are, however, costly to conduct and therefore not suitable for technology level evaluation. Advantage should therefore be taken of the fact that the benchmark is application-specific and field conditions should be replicated as far as possible in the laboratory.

In addition, in order ensure predictive validity, tests should take a similar form to applications, duplicate the environment in which

applications will be used, and use subjects representative of end-users (Syrdal, 1995).

### **3. Requirements of Speech Synthesis for CALL**

As mentioned in the introduction, in addition to the first application of speech synthesis as a reading machine (Allen, 1973) in CALL applications speech synthesis is also being used to provide TL models to listen to in speech perception training, and models to imitate in pronunciation exercises, and to provide the voice of a conversational partner. CALL therefore requires different styles of speech, in particular conversational in addition to reading.

In CALL programs integrating speech synthesis, the computer takes on three main roles, ‘tutor’, ‘conversational partner’, and ‘tool’. In addition to authentic native speech, a speech synthesiser for use in CALL should therefore be able to simulate the type of speech that teachers use to address learners, i.e. Teacher Talk (TT). Features of TT are discussed in 4.1.

As regards the generation of authentic native speech,

systems available today are less than optimal, it is [therefore] important to know which aspects of a system’s performance are essential to a specific application (van Bezooijen and van Heuven, 1997).

In this respect, I believe that it is essential for speech synthesisers for use in CALL applications to be able to achieve what is expected of learners. In sections 3.2. and 3.3., the goals of pronunciation training and listening training are therefore discussed.

#### **3.1. Teacher Talk**

As mentioned, TT is the term used to describe the register of language that teachers use to address second language learners. Studies of TT have addressed a variety of different variables. The material to be synthesised when using TTS is determined by the CALL developer or the teacher, therefore only the studies of speech rate, pauses, and phonology, intonation, articulation and stress are relevant to the definition of the requirements of speech synthesis for CALL.

Regarding speech rate, studies have revealed that teachers adapt their speech rate and the duration of pauses when talking to learners (Ellis, 1994). And, results of studies of phonology etc. reveal that teachers “appear to speak more loudly and to make their speech more distinct when addressing L2 [second language] learners” (*ibid.*: 582).

A speech synthesiser for use in CALL should therefore produce ‘distinct’ speech and permit the manipulation of: speech rate, pause duration, and volume.

### 3.2. Teaching Pronunciation

Over the years there have been several different trends in the teaching of pronunciation ranging from the *Direct Method* through *Audiolingualism* to the *Communicative Approach*. These changes in approach reflect changes in the goal of pronunciation teaching. While the goal of approaches such as the direct method and audiolingualism was accuracy, the goal of the communicative approach was more realistic, to achieve a level of intelligibility which permits effective communication in the TL (Celce-Murcia et al., 1996), in other words comprehension. A speech synthesiser for use in CALL should therefore produce intelligible and comprehensible speech. While for a long time it was believed that focus on the accurate production of individual speech sounds was the most effective way to achieve intelligibility, and later the focus shifted to the accurate production of prosody, today both are considered to contribute equally to intelligibility. A speech synthesiser for use in CALL should therefore accurately produce both segments and suprasegmentals.

Returning to teaching approaches, today it is accepted that there is no one right way of teaching pronunciation; teachers draw on a range of different techniques. Of these techniques, those which involve modifying speech addressed to learners in order to make aspects of it more salient have implications for the definition of the requirements of speech synthesis for CALL. A survey of the literature on the teaching of French pronunciation reveals that the following techniques are used to teach rhythm, namely, syllable division, varying the pitch of accented syllables and varying speech rate (Lebel, 1974).

In order to be able to simulate these interventions in CALL speech synthesisers must permit the manipulation of pitch and syllable division in addition to the parameters mentioned in the previous section.

### 3.3. Teaching Listening

The goal of listening training is for the learner to be able to understand a wide range of speakers speaking in a wide range of situations. The learner therefore requires exposure to a range of different voices and speaking styles, because exposure to “variability is necessary for generalisation” (Protopapas and Calhoun, 2000; 31). Ideally, a speech synthesiser for use in CALL should therefore be able to ‘speak’ in a number of different voices and styles.

In fact, the availability of a range of different voices is also useful for pronunciation training; studies suggest that the learner’s success in acquiring pronunciation may depend on the model that they use to imitate (Scheffert and Fowler, 1999).

## 4. Selecting among Existing Tests

As mentioned in section 2, there are few existing methods for the evaluation of speech synthesis for CALL. In particular, only one technology evaluation of speech synthesis for CALL has been conducted. This evaluation, conducted by Stratil et al. (1987), aimed to determine whether the ‘quality’ of the output of the SSI 263 Spanish TTS chip was adequate for the presentation of grammar exercises in a language laboratory. Specifically, their evaluation compared the ability of both beginner and advanced learners of Spanish to repeat and transcribe Spanish sentences presented via speech synthesis, with their ability to do so for the same sentences read by a native speaker. In other words, they evaluated the intelligibility (the ease of recognition of individual sounds and words (Francis and Nusbaum, 1999)) of the output of the speech synthesiser. As an evaluation of the intelligibility of speech synthesis this paradigm overcomes the fact that the results may be affected by the subjects’ imperfect knowledge of the orthography of Spanish, which may have brought the validity of the evaluation into question, by asking beginners only to imitate the presented utterances. The

validity of the test, however, remains questionable because learners did not always write what they heard (Stratil *et al.*, 1987a:), and results are still affected by learner's imperfect perception and production of Spanish.

In summary, only one of the requirements of speech synthesis for CALL has been addressed, and the validity of that evaluation is questionable. It is therefore necessary to look outside the evaluation of speech synthesis for CALL for benchmark tests.

#### **4.1. Evaluation of Speech Synthesis: State of the Art**

As a result of both individual research teams and standards initiatives (ANSI, ESPRIT SAM, ITU-T, JEIDA and EAGLES), today many different paradigms for the evaluation of the output of speech synthesisers are available.

In section 3, it was established that speech synthesis for use in CALL should be flexible, comprehensible and intelligible. In addition, it was established that prosody played as much a role in determining intelligibility as segmental quality. A benchmark for the evaluation of speech synthesis for CALL should therefore address these factors. Paradigms have been proposed for the independent evaluation of *comprehension* (ease with messages are understood), *intelligibility* and *prosody*.

In addition, van Santen's (1993) word-pointing paradigm permits the evaluation of a number of features of the output of speech synthesisers at once. As a result, although it is a judgement evaluation, it merits consideration because it is very efficient.

##### **4.1. 1. Intelligibility**

Both judgement and functional paradigms for the evaluation of speech synthesis have been proposed. As mentioned earlier, functional evaluations are preferred as they are generally more reliable and fewer subjects are required to obtain reliable results. This discussion is therefore restricted to functional tests. Three main types of functional test have been proposed for the evaluation of the output of speech synthesisers, namely *articulation tests*, *lexical decision tests* and *word and phoneme monitoring tests*.

#### ***Articulation Tests***

Originally used for the assessment of hearing loss, articulation tests part from the assumption that intelligibility is reflected by listeners' ability to identify individual units of speech. The basic technique involves presenting subjects auditory stimuli which they are expected to identify by either selecting the appropriate orthographic transcription from a list (closed-response format) or transcribing the stimulus (open-response format). Scores are based on the number of units (phonemes, words or sentences) correctly identified. Two types of articulation test are distinguished, namely, tests which score intelligibility at the segmental level, and tests which score intelligibility at the word and/or sentence level.

##### **Segmental**

The two most well known segmental level tests are the Diagnostic Rhyme Test (or DRT) (Voiers *et al.*, 1965) and the Modified Rhyme Test (or MRT) (House *et al.*, 1965).

##### ***Diagnostic Rhyme Test***

The corpus of this test consists in 192 monosyllabic, mainly CVC words. Organised into six lists, each containing sixteen minimal pairs which differ only by a single feature in the initial consonant (e.g. *tune* vs. *dune*). Each pair of stimuli is presented to the subject visually, prior to one of the pair being presented auditorily. With an inter-stimulus interval of 3s, the test takes 15min for each system and requires 10 subjects (van Bezooijen and van Heuven, 1997).

##### ***Modified Rhyme Test***

The corpus of the MRT consists in 50 lists containing six, mainly CVC, words each which differ in either the initial or final phoneme (for example, *peel*, *reel*, *feel*, *eel*, *keel*, *heel* and *late*, *lake*, *lay*, *lace*, *lane*, *lame*). Presented in closed-response format, the subject's task is to identify which of the six orthographic stimuli were presented to them auditorily. With an inter-stimulus interval of 4s, the test takes 25min for system (van Bezooijen and van Heuven, 1997).

Both the DRT and MRT would seem ideal for use for benchmarking purposes.

Because standard orthographic responses are used, they are both easy to conduct and subjects do not require training (Pisoni, 1987). In the bargain, they are known to be reliable.

They do, however, have their limitations: corpora are fixed; only single consonant confusions are tested in stressed monosyllables appearing next to pauses; and, results are subject to ceiling effects due to the high predictability of responses.

Regarding the development of a benchmark specifically for the evaluation of speech synthesisers for use in teaching French to Anglophone learners, as mentioned above, the DRT has been adapted for the evaluation of French speech synthesisers (Peckels and Rossi, 1973). I, however, believe that it is inadequate because like the English version it tests the intelligibility of consonants in CVC syllables, yet CV syllables are significantly more common in French (Leon, 1992). Moreover, with respect to the evaluation of speech synthesis for CALL, the French DRT, as it does not test the intelligibility of vowels, does not test the majority of phonemes which pose Anglophone learners perception and production difficulties, phonemes which I believe should be particularly intelligible.

In response to the criticisms of the DRT and MRT presented above, in order to increase the number of phonemes and phonetic contexts, extensions to the DRT and MRT have been proposed, as have several entirely new test sets. In order to increase the number of confusions tested, presentation of the tests in open-response format, and the use of logatoms (i.e. nonsense words) has been suggested.

These tests however have their own limitations. The corpora of the other test sets proposed are far longer than those of the DRT and the MRT. For example, the CLuster IDentification (CLID) test (Jekosch, 1992) consists in 900 stimuli and takes 2 hours per synthesiser/reference condition (van Bezooijen and van Heuven, 1997). While tests using logatoms have the advantage that they put native subjects in the same position as learners of the language, in that they have to rely to a greater extent on the acoustic signal (Goldstein, 1983), the tests require the use of subjects trained in phonetic transcription – if orthographic

transcription is used results are open to interpretation because there is not a one-to-one correspondence between phonemes and graphemes in most languages. These tests are therefore particularly unsuitable for use in benchmark

In summary, if an existing segmental level intelligibility test is to be used for the benchmarking of speech synthesis for CALL, it will be necessary to adapt it.

### **Word/Sentence Level Tests**

The three main tests used to measure the intelligibility of speech synthesis at the word/sentence level are the Harvard Sentences Test (Egan, 1948), the Haskins Sentences Test (Nye and Gaitenby, 1974) and the Semantically Unpredictable Sentences (SUS) Test (Benoit *al.*, 1989).

#### *Harvard Sentences Test*

Presented in open-response format and scored on the basis of the number of content words correctly transcribed, the corpus of the Harvard Test (a.k.a. the Phonemically Balanced (PB) Sentences Test) consists in a fixed set of meaningful, yet not entirely predictable sentences (e.g. *A chicken leg is a rare dish.*) which are Phonemically Balanced and cover a wide range of syntactic structures.

#### *Haskins Sentences Test*

Presented and scored in the same way as the Harvard Test, the Haskins Test differs from the Harvard Test in that it is based on the use of a fixed corpus of semantically anomalous sentences – sentences which respect the grammatical constraints of the language but, in an attempt to eliminate the role of semantic and real-world knowledge, explicitly violate semantic constraints. With an inter-stimulus interval of 15s, the test takes 30min per synthesiser (van Bezooijen and van Heuven, 1997).

Like the DRT and the MRT these tests are good candidates for use in benchmarking because they are easy to conduct and extensive use of the tests has shown them to be reliable. A further benefit of both tests is that they test a wide range of confusions (single and multiple

feature consonant-consonant, vowel, etc. confusions). Consisting in a variety of sentence structures, the Harvard Sentences have the further advantage that they account for the role of prosody<sup>1</sup>. With respect to the evaluation of speech synthesis for CALL with native speakers as subjects, by eliminating the role of semantic and real-world knowledge, both tests have the advantage that they force natives to behave like learners and rely to a greater extent on the acoustic signal (Goldstein, 1983).

They do, however, have several limitations:

- Due to the unsystematic construction of material, they do not permit diagnosis (van Bezooijen and van Heuven, 1997)
- The corpora are fixed
- Because the tests are so popular, subjects may even be familiar with stimuli even if they have never participated in a test before (Lemmetty, 1999)

In response to the criticisms of the Harvard and Haskins Tests, Benoit *et al.* (1989) proposed the SUS test.

#### *Semantically Unpredictable Sentences Test*

Presented in open-response format and scored on the basis of either the number of content words, or the number of sentences correct, unlike the Harvard and Haskins tests the corpus of the multilingual SUS test is not fixed. SUS are generated by an algorithm that selects content words at random from a lexicon. Generating sentences with five different syntactic structures (declarative + adverbial *The table walked through the blue truth*, declarative *The strong way drank the day*, imperative *Never draw the house and the fact*, wh-question *How does the day love the bright word?*, declarative + relative clause *the place closed the fish that lived*), the test takes the role of prosody into account in a systematic way. 50 sentences (10 per structure) are recommended per synthesiser. With an inter-stimulus interval of 15s, the test therefore takes 15min to conduct per synthesiser (van Bezooijen and van Heuven, 1997).

With respect to the specific purposes of this study, the evaluation of speech synthesis for teaching French, as mentioned the SUS test is multilingual. In addition, a set of PB sentences has been developed for French (Combescure, 1981). While both of these tests have the advantage that they place native speakers in the same situation as second language learners, the SUS test is the most suitable for benchmarking because its corpus is not fixed.

#### **4.1.2. Comprehension**

Three functional tests have been proposed for the evaluation of the comprehension of the output of speech synthesis systems, namely *standard listening comprehension tests*, *word-gating tests*, and *sentences verification tests*.

##### ***Standard Listening Comprehension Tests***

As far as can be established, Pisoni (1987) was the first to evaluate the comprehension of synthetic speech. His approach was based on the re-use of passages from standardized adult reading comprehension tests. Specifically, subjects were presented a passage orally and then asked to answer a number of multiple-choice questions on what they had been presented. They were only allowed to listen to the passage once.

The problem with this test is, presented in closed-response format, it is subject to ceiling effects. In order to reduce ceiling effects, several suggestions have been made: the test could be presented in open response format; rather than ask subjects questions, subjects could be asked to summarise what they have heard. In these formats, the test is however more difficult to score.

Like for any test, it is not possible to use the same test set more than once with the same subjects. It is therefore necessary to be able to generate new comprehension tests. The problem is comprehension tests are difficult to design. You cannot just use reading tests because the information contained in reading comprehension test is too dense. It is therefore necessary to develop tests specifically for the purpose of testing listening comprehension. For tests to be valid, it must only be possible to answer the questions having listened to the text.

---

<sup>1</sup> The Haskins sentences, on the other hand, are all based on the same syntactic structure (*The Adjective Noun1 Verb the Noun2*).

Comprehension tests must therefore be piloted before use to verify that the questions cannot be answered on the basis of general knowledge.

### ***Word Gating Tests***

Proposed by the JEIDA initiative (Itahashi et al., 1995), word gating tests are based on the idea that subjects' comprehension of an aurally passage is reflected by their ability to complete the gaps in a written version of the same text in which has had a number of content words blanked out.

While such tests are easier to generate than comprehension tests, as all the evaluator has to do is to blank out important content words. Like standard comprehension tests, word-gating tests must be piloted to check that it is not possible to fill in the blanks correctly without having listened to the text on the basis of semantic and real-world knowledge.

### ***Sentence Verification Tests***

In the sentence verification tests proposed by Manous et al. (1985), comprehension is evaluated by measuring subjects' response accuracy and latency when asked to judge whether simple 3 and 6 word sentences presented were true (*Mud is dirty.*) or false (*Rockets move slowly.*).

While the corpora of these tests are simpler to design than standard listening comprehension tests, like lexical decision tests and phoneme/word monitoring tests, because they are based on reaction times, they are more difficult to conduct and therefore not suitable for use as benchmark tests. Moreover as all the sentences are statements, the role prosody plays in comprehension is not tested.

In summary, due to the difficulties of evaluating comprehension, its evaluation is not recommended (van Bezooijen and van Heuven, 1997). If comprehension is to be tested, the word gating is the easiest to use and develop and therefore the most suitable for the purposes of benchmarking.

### **4.1.3. Prosody**

Prosody is a complex phenomenon composed of a number of interacting variables. Functional tests have been proposed for the

evaluation of a number of these different variables including word stress, word boundaries, prosodic phrasing, sentence type and focus distribution (van Bezooijen and van Heuven, 1997). To exhaustively evaluate prosody using these functional tests is too time-consuming for the process of benchmarking. Despite preference for the use of functional tests we therefore look to judgement tests of the global quality of prosodic realisation. The two most widely used paradigms for the judgemental evaluation of prosody are the *Prosodic Form Test* (Grice et al., 1991) and the *Prosodic Function Test* (Grice et al., 1992).

In the prosodic form test, subjects are presented a range of sentences with different prosodic patterns, their task being to rate their naturalness. In the prosodic function test on the other hand subjects are presented sentences which have a range of different pragmatic function in context (e.g. Human: *I'd like to reserve a flight to Paris on Monday morning.* Synthesiser: *Are you travelling from London?*), and asked to rate their communicative appropriateness.

Although more efficient than functional evaluations of prosody, these paradigms are also time consuming and therefore not very suitable for the purposes of benchmarking. Moreover, the corpora of the tests are fixed and are only available for English (van Bezooijen and van Heuven, 1997).

### **4.1.4. Word-Pointing Paradigm**

The aim of van Santen's (1993) word-pointing paradigm is to evaluate as wide a range of contexts as possible. In order to achieve this, a frequency weighted greedy algorithm is used to find a small set of sentences which cover all the speech units to be tested. These sentences are then passed through the synthesiser and presented to learners simultaneously auditorily and orthographically. The subjects' task is to spot any errors in the auditory stimulus with respect to the text. Specifically, subjects are asked to highlight the corresponding text, indicate the nature of the error using the categories provided (*outright mispronunciation, bad letters, missing letters or words, wrong stress, wrong word emphasized, overall voice quality, choppiness, bad rhythm and other*) and

then to rate the gravity of the error on a three-point scale.

As already mentioned, the paradigm is very efficient; it enables the testing of a large number of contexts using a small test set. Into the bargain, it not only tests two features which were identified as essential requirements of speech synthesis for use in CALL, but it also tests grapheme-phoneme conversion and naturalness, features I believe it is also important for a speech synthesiser for use in CALL to get right.

It, however, relies on judgement and as a consequence necessitates the use of a large number of subjects. Into the bargain, requiring both the classification and rating of errors, it places high cognitive demands on subjects.

## **5. Proposed Benchmark**

Having analysed the suitability of existing paradigms for the evaluation of speech synthesis for the purposes of benchmarking speech synthesis for CALL, in this section a proposal for a benchmark is presented. Specifically, recommendations are given for the selection of subjects (section 5.1.), benchmark tests (section 5.2.), and reference conditions (section 5.3.).

### **5.1. Subjects**

As mentioned in section 2.3., subjects used in evaluations should be representative of the end-users of applications. In the case of CALL applications the main end-users are learners. It is however questionable whether learners are suitable subjects for the evaluation of speech synthesis for CALL. Due to the very fact that they are learners, they have an imperfect knowledge of the TL. Moreover, their linguistic ability is constantly evolving. Consequently, results of evaluations of speech synthesis using learners as subjects are likely to be inconsistent and therefore unreliable<sup>2</sup>. Rather teachers who are native speakers of the TL will be used as subjects for the purposes of

benchmarking speech synthesis for CALL because not only are they expert speakers of the TL but they are also end-users of CALL applications.

As regards the number of subjects, although a judgement paradigm is to be used, as this paradigm is only being used to support functional evaluations, it is not felt necessary to employ a large number of subjects. Rather, the use of around 10 subjects is recommended.

### **5.2. Benchmark Tests**

In section 3, it was established that the requirements of speech synthesis for CALL are: intelligibility, comprehension, the availability of different voices and styles of speaking and the ability to manipulate speech rate, pause duration, volume, pitch and syllabification.

In order to, evaluate the availability of different voices and styles and the availability of functions to permit the manipulation of speech rate etc., the use of simple checklist is proposed.

Regarding the evaluation of intelligibility and comprehension, in section 3.2., it was mentioned that prosody plays a large role in the determination of prosody of intelligibility which in turn plays a role in determining comprehension. To independently evaluate prosody in either functional or judgemental tests would make the benchmark too long. The use of intelligibility and comprehension tests which take into account the role of prosody is therefore recommended.

In addition, it is recommended that the word-pointing paradigm be used to support these functional evaluations because it will provide further data on intelligibility in a wide range of contexts and will also permit evaluators to obtain data on prosodic quality. In addition, it will also enable evaluators to obtain data on other aspects of speech which although not highlighted in the literature are important for a speech synthesiser for use in CALL to get right, such as naturalness and grapheme-phoneme conversion. And, offering an 'other' category for the classification of errors, the word-pointing paradigm also permits the evaluation of which we may not have realised are important.

---

<sup>2</sup> Indeed, the results of an experiment conducted as part of the research reported in this paper, which compared learner and native speaker performances in a Phonetically Balanced (PB) sentences test confirm this to be the case.

### **5.2.1. Intelligibility Tests**

The goal is to evaluate intelligibility taking into account the effects of both segmental and prosodic quality. Sentence level tests such as the Harvard and SUS are the best to do this however, as mentioned due to the unsystematic construction of the test sets they do not enable use to obtain diagnostics on the intelligibility of individual phonemes. In my opinion, it is important to make sure that phonemes which are known to pose learners difficulties are systematically tested, the use of a segmental level test is therefore also recommended.

#### ***Global Evaluation***

Specifically, for the global evaluation of intelligibility, the decision was taken to use the SUS test because it puts teachers in the same position as learners, has an open corpus, and takes into the account the effects of prosody on intelligibility.

The results of this test will be scored at the sentence level because as indicated this gives a better resolution of the scores.

#### ***Diagnostic Evaluation***

While, ideally the intelligibility of all phonemes would be tested, such a test is too time consuming for use in benchmarking. Minimally the intelligibility of phonemes which cause learners perception and production difficulties should be tested. Open-response rhyme tests permit the evaluation of the greatest number of confusions while remaining easy to conduct. I therefore propose the use of an open-response rhyme test using a corpus of minimal pairs extracted from Language Learning and Teaching (LL&T) corpora. In order to overcome the problem of learning effects, I propose the creation of a database of minimal pairs that can be used to test the intelligibility of phonemes that pose difficulties to learners from which two pairs should be randomly selected for each attested difficulty each time the test is run. The test will be scored on the number of words/phonemes correctly transcribed.

### **5.2.3. Comprehension Test**

As mentioned in section 5.2., a comprehension test which takes into account the

effects of prosody should be used. Both standard listening comprehension tests and word-gating tests permit this. Of these two types of test the use of word-gating paradigm using texts selected from an LL&T corpus is recommended because they are easier to develop than standard listening comprehension tests. These tests will of course have to be piloted before use to check that it is not possible to fill in the gaps without having listened to the text.

### **5.2.4. Word-Pointing Test**

Reasons for the selection of this test were already put forward in section 6.2. The question that remains is what corpus to use. Ideally, a corpus representative of the material to be synthesised in CALL applications should be used, i.e. an LL&T corpus. Use of the algorithm to select a small set of sentence with broad coverage, however, necessitates the availability of a phonetically transcribed corpus. Such a corpus is not available to us. For the time being, texts which cover a wide range of syntactic and prosodic structures will be selected by hand.

Finally, regarding the scoring of the tests, in addition to asking subjects to rate the errors they detect, these errors could be checked against a checklist of attested difficulties. Errors which correspond to attested difficulties should be penalised to a greater extent.

### **5.3. Reference conditions**

As mentioned in section 3.1., when evaluating speech synthesis, it is common practice to use reference conditions to set acceptability levels. Analysis of the requirements of speech synthesis for CALL suggests that the speech of a native teacher should be used as the reference condition when benchmarking speech synthesis for use in CALL. As mentioned in section 3.1., such a reference condition is not suitable for use in benchmarking because as native speech is of higher quality.

In the literature on the use of speech synthesis in CALL, Keller and Zellner-Keller (2000) have suggested that:

When the language competence of the system begins to outstrip that of some of the better second language users, such systems

become useful new adjunct tools (*ibid.*: 111).

While the aim should be native-like output, the speech of an advanced learner is a more appropriate reference condition for benchmarking purposes. I therefore propose to use of the speech of an advanced learner (as rated in the Canadian Language Benchmarks<sup>3</sup>, for example) as the main reference condition and minimal requirement of speech synthesis for CALL, in combination with the speech of a native teacher as the top-line (ideal) reference condition.

## 6. Conclusion

In conclusion, the aim of the study reported here was to develop a methodology for benchmarking speech synthesis for use in CALL. With this goal in mind, the requirements CALL imposes on speech synthesis have been analysed, paradigms proposed for the evaluation of speech synthesis have been assessed for their suitability for use in such a benchmark, and a preliminary proposal has been put forward.

## REFERENCES

- Allen (1973) Reading machines for the blind: the technical problems and the methods adopted for their solution. *IEEE transactions on audio and electro-acoustics*. 21 (3): 259-264
- Allen (1987) *From text to speech: the MITalk system*. Cambridge: CUP
- Balsband and Paroubek (1999). *A blueprint for a general infrastructure for natural language processing systems evaluation using semi-automatic quantitative black box approach in a multilingual environment*. ELSE.
- Benoit et al. (1989) Multilingual synthesiser assessment using SUS. In *Procs. EUROSPEECH 89*. 2: 633-636
- Celce-Murcia et al. (1996) *Teaching pronunciation*. Cambridge: CUP
- Chapelle (2000) *Computer applications in second language acquisition: foundations for teaching, testing and research*. Cambridge: CUP
- Combesure (1981) 20 listes de dix phrases phonétiquement équilibrées. *Rev. Acoustique*. 56 : 34-38
- Dix et al. (1998) *Human computer interaction*. London: Prentice Hall
- Egan (1948) Articulation testing methods. *Laryngoscope*. 58: 955-991
- Egan and LaRocca (2000). Speech recognition in language learning: a must. In *Procs. InSTiLL 2000*, Dundee: 4-9
- Ellis (1994) *The study of second language acquisition*. Oxford: OUP
- Francis and Nusbaum (1999) Evaluating the quality of synthetic speech. In Gardner-Bonneau (ed.) *Human factors and voice interactive systems*. London: Kluwer Academic Publishers.
- Goldstein (1983) Word recognition in a foreign language: a study of speech perception. *Journal of psycholinguistic research* 12(4): 417-427
- Grice et al. (1991) Assessment of intonation in text-to-speech systems - a pilot test in English and Italian. In *Procs. EUROSPEECH 91*, Genova 2: 879-882
- Grice et al. (1992) Prosodic form test. *ESPRIT SAM final report*
- Hamel and Handley (to appear) *On the use of speech synthesis in CALL*
- Hincks (2002) Speech synthesis for teaching lexical stress. *TMH-QPSR*. 44: 153-165
- Hirschman and Thompson (1996). Overview of evaluation in speech and natural language processing. In Cole et al. (eds.) *Survey of the state of the art in human language technology*. Cambridge: CUP
- House et al. (1965) Articulation testing methods: consonantal differentiation with a closed response set. *JASA*. 37 (1): 158-166
- ISO (1999) *Information technology - software product evaluation - part 1: general overview*. BS ISO/IEC 14598-1:1999
- Jekosch (1992) The CLuster IDentification Test. In *Procs. ICSLP 92*. 1: 205-208
- Keller and Zellner-Keller (2000) Speech synthesis in language learning: challenges and opportunities. In *Procs. InSTILL 2000*, Dundee: 109-116
- King (1995) The evaluation of NLP systems. In *Procs. Swan 21*, Geneva.

---

<sup>3</sup> <http://www.language.ca/home.html>

- Lebel (1974) The teaching and role of rhythm in the correction of French pronunciation.
- Lemmetty (1999) *Review of speech synthesis technology*.  
<http://www.acoustics.hut.fi/~slemmet/dippa/index.html>
- Leon (1992) *Phonétisme et prononciation du français avec des travaux pratiques d'application et leurs corrigés*. Paris: Nathan Université
- Manous et al (1985) Comprehension of natural and synthetic speech using a sentence verification task. *Research on speech perception progress report*. Indiana University. 35-57
- Morel, and Lacheret-Dujour (2001). "Kali": synthèse vocale à partir du texte. *TAL*, 42 (1), 193-222.
- Peckels and Rossi (1973) Le test de diagnostic par paires minimales. Adaptation au français du "Diagnostic Rhyme Test" de W.D. Voiers. *Rev. Acoustique* 27: 245-262
- Pisoni (1987) Some measures of intelligibility and comprehension. In. Allen et al. pp. 151-171
- Protopapas and Calhoun (2000) Adaptive phonetic training for second-language Learners. In *Procs. InSTILL 2000*, Dundee: 31-38
- Santiago-Oriola (1999) From grapheme to phoneme: diagnosis in dictation. In *Procs. ICPH'99*. San Francisco
- Scheffert and Fowler (1999) The effects of voice and visible speaker change on memory for spoken words. *Journal of memory and language*. 34: 665-685
- Sobkowiak (1998). Speech in EFL CALL. In Cameron (ed.). *Multimedia CALL: theory and practice*. Exeter: Elm Bank
- Stevens (1989) A direction for CALL: from behaviouristic to humanistic courseware. In Pennington (ed.) *Teaching languages with computers*. La Jolla, CA: Athelstan: 31-43
- Stratil et al. (1987a). Exploration of foreign language speech synthesis. *Literary and linguistic computing*. 2 (2): 116-119
- Stratil et al. (1987b). Computer-aided language-learning with speech synthesis: user Reactions. *Programmed learning and educational technology*. 24 (4): 309-316
- Syrdal (1995) Text-to-speech systems. In Syrdal et al. (eds.) *Applied speech technology*. London: CRC Press
- van Bezooijen and van Heuven (1997) Assessment of synthesis systems. In Gibbon et al. (eds.) *Handbook of standards and resources for spoken language systems*. New York: Walter de Gruyter Publishers
- van Santen (1993) Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer speech & language* 7: 49-100
- Voiers et al (1975) *Research on diagnostic evaluation of speech intelligibility*. Air force Cambridge research laboratories. Bedford, Massachusetts