Investigating the Requirements of Speech Synthesis for CALL with a View to Developing a Benchmark

Zöe Handley and Marie-Josée Hamel

Centre for Computational Linguistics, UMIST, Manchester, UK
zoe.handley@postgrad.umist.ac.uk, mjhamel@ccl.umist.ac.uk

Abstract

The evaluation of speech synthesis for CALL is particularly important because speech synthesisers developed for other purposes are being re-used. Yet, very few ‘formal’ evaluations of speech synthesis for CALL [1 - 4] have been conducted. One potential reason for the neglect of evaluation is that it is expensive in terms of both time and resources, for which it competes with further technology development [5]. In order to achieve a balance between technology development and evaluation, benchmarking, is commonly used in the evaluation of speech and language technologies. The ultimate goal of our research is the development of a benchmark for the evaluation of speech synthesis for CALL. The first stage in the development of such a benchmark is to establish the requirements of the application. This is the goal of this paper.

1. Introduction

Speech synthesis can provide spoken language input to the learner in CALL environments. Its output, or ‘voice’, can be exploited as a reading machine, a pronunciation tutor and a conversational partner. The technology has several teacher and learner-oriented advantages over the use of recordings of native speakers: ease of creation and editing of speech models [6], generation of various kinds of modified input [7], generation of speech models and feedback on demand [6], etc. Although generic applications using speech synthesis such as talking dictionaries, dictation software, and talking word processors are beginning to find their way onto the market, CALL courseware applications using speech synthesis are yet to be found. CALL developers and users are still sceptical vis-à-vis the technology. Evidence is needed to show them that it is suitable and ready for use in CALL. This could be provided through positive evaluation of the technology for use in the context. Yet, very few formal evaluations of speech synthesis for use in CALL have been conducted [1]. Evaluation is an important process in the implementation of any technology. It is however expensive in terms of time and resources, for which it competes with further technology development [5]. In order to ease the process of evaluation and consequently achieve a balance between development and evaluation, benchmarking is commonly used in language engineering. 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” [8], to a system which “represents a performance level that is known to guarantee user satisfaction” [8].

As far as it is possible to establish, 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. This paper reports on the preliminary stages in the development of such a benchmark. First, evaluations of speech synthesis that have been conducted to date are reviewed. Then, having established the goal of the benchmark, the requirements which CALL imposes on speech synthesis are presented. Finally, results of an investigation focusing on the relationship between these requirements and acceptability and appropriateness for use levels are discussed and suggestions for further research are put forward.

2. Evaluation of Speech Synthesis for CALL

The goal of our research is the development of a methodology for benchmarking speech synthesis for use in CALL applications.

The evaluation of speech synthesis for use in CALL constitutes one of a number of stages of evaluation from which CALL applications integrating speech synthesis would benefit before they are launched on the market. First, it is recommended that technology evaluation is conducted, that is that the successes and limitations of the technology are identified in order to determine whether the technology is suitable for use in a particular application [9]. Then, if the technology is deemed to be suitable and ready for use in a particular application, usage evaluation, evaluation of the technology *in* the intended application, can be conducted [9]. Usage evaluation will not be discussed further at this stage.

Regarding technology evaluation, our literature review shows that one evaluation of speech synthesis for CALL has been conducted. In this evaluation [1], the quality of the output of a Spanish Text-To-Speech synthesiser was tested to determine whether it was suitable for the presentation of grammar exercises in a language laboratory setting. More specifically, the ability of both beginner and advanced learners of Spanish to repeat and transcribe Spanish sentences presented via speech synthesis was compared with their ability to do so for the same sentences read by a native speaker. In other words, the intelligibility (the ease of recognition of individual speech sounds and words [10]) of the output of the speech synthesiser was compared with that of the native speaker. It was found that while beginners found the speech synthesis and human speech equally intelligible, the more advanced students found the speech synthesis significantly less intelligible than the human produced speech. While this evaluation addressed the intelligibility of the speech synthesiser output in a given function, as a reading machine, we feel it failed to provide us insights into the level of readiness of the technology for use in that function.

Technologies should be exploited for what they are best at. In addition to being used as a reading machine, as suggested in the above experiment, we suggest that speech synthesis can also be used as a pronunciation tutor and as a conversational partner. Further technology evaluations addressing the readiness of speech synthesis for use in these functions are therefore needed. Ultimately, a benchmark should be developed for this purpose.

3. CALL requirements

Once the purpose of the evaluation has been identified, according to the ISO standards for Software Product Evaluation [11], the next stage of the evaluation process is the analysis of the requirements of the application. One of the goals of this research is to identify the requirements imposed by CALL on speech synthesis. These are drawn from findings in Second Language Acquisition and Language Learning and Teaching research, findings specifically related to ideal conditions for SLA [12] and best practice in LL&T [13]. They suggest that in the CALL context, two aspects of the speech synthesiser would merit consideration: the quality of its output and its flexibility.

3.1 Quality of the output

As far as the quality of the output of the speech synthesiser is concerned, two more specific aspects will be of particular importance in the CALL context: the quality of the pronunciation and voice characteristics.

One central characteristic contributes to the quality of the pronunciation: its comprehensibility. Comprehensibility will therefore be taken as the minimal requirement asked from the output of the speech synthesiser in the CALL context. It has been defined as the ease with which a listener can understand a speaker's intended message [10]. Comprehensibility is determined by the adequacy, accuracy and naturalness of the realisation of the following features of speech: the quality of the pronunciation of individual phonemes, and the phrasing, rhythm and intonation of utterances. The quality of the pronunciation of individual phonemes contributes comprehensibility by rendering phonemes and words easy to identify, i.e. intelligible. Phrasing and intonation structure utterances into constituents which provide clues to information structure. Rhythm highlights word boundaries which provide cues to word identification. Intonation also conveys further meaning including the pragmatic function of the utterance and affective meaning/emotion.

Speech may be comprehensible without the aforementioned features being entirely accurate, i.e. error-free, or natural sounding, i.e. native-like. Accuracy and naturalness will therefore be considered as additional requirements imposed by CALL on the quality of the output.

Voice characteristics, such as friendliness and expressiveness will be considered additional requirements of speech synthesis for use in CALL because such voice characteristics will contribute to decreasing learner anxiety and increasing learner motivation, both factors which increase the success of SLA [18, 19].

The quality of the pronunciation depends on its level of comprehensibility, accuracy and naturalness. The requirements imposed at this level will typically be measured using perceptual experiments, as will the voice characteristics.

3.2 Flexibility

In the given CALL context, flexibility is expected with respect to register, voice, accent, pitch, duration, volume, timbre, and voice quality. The availability of a range of registers (in/formal, prepared/spontaneous, etc.), voices (male/female, old/young, etc.) and accents (standard/regional, etc.) as well as the possibility to manipulate the other features are therefore requirements imposed by the CALL context.

Depending on the function for which the speech synthesiser will be used in the CALL context, as a reading machine, a pronunciation tutor or conversional partner, the ranges of registers, voices and accents as well as the types of manipulations over pitch, duration, etc. required will vary.

Flexibility of the output depends on the availability of options and functionalities the speech synthesiser is provided with. The requirements imposed at this level will hence typically be evaluated through the use of checklists.

4. Requirements of Speech Synthesis for CALL: Investigation

The perceptual experiment presented here focuses on the quality of the pronunciation as the central requirement of speech synthesis for CALL. In particular, it addresses comprehensibility. Specifically, it compares the levels of comprehensibility found when the speech synthesiser is used as reading machine, a pronunciation tutor and a conversational partner. It also looks at the correlation between comprehensibility and acceptability levels with the objective of determining whether comprehensibility is a good indicator of acceptability for use in the different CALL contexts. Although not the goal of this experiment, ratings of acceptability also give us an indication of the suitability of the particular speech synthesiser evaluated in this experiment for use in the three different functions which it could assume within a CALL context.

The experiment reported here also addressed accuracy and naturalness, but those results have not been analysed yet. These will be reported on in a subsequent article.

4.1 Method

Thirteen out of 30 participants recruited have responded so far. They consist in a mixture of Francophone teachers and CALL researchers, with varying degrees of experience of CALL.

60 utterances produced by a research speech synthesiser were presented to participants, 20 representative of each function in CALL: as reading machine, pronunciation tutor, and conversational partner. They were selected

from existing CALL software (FreeText [14], SAFexo [15], and Tell Me More [16] respectively).

In a first pass, participants were asked to rate the comprehensibility and the acceptability of the output for the function indicated. For each context, participants were also asked to rate the appropriateness of the use of speech synthesis.

On a second pass, participants were asked to highlight any errors in the output which they believed affected its suitability for use in the function indicated. Then, for each function, they were asked to indicate the types and frequency level of errors highlighted previously and to rate the seriousness of those classes of error with respect to the use of speech synthesis in that function in CALL. Finally, participants were asked to answer a questionnaire about their familiarity with speech synthesis in general and in CALL.

4.2 Preliminary Results

Regarding the participants' experience of speech synthesis generally and in CALL. The average of the participants' ratings of their familiarity with speech synthesis on a scale of +3 (entirely familiar) to –3 (not at all familiar) was calculated for the group. This was found to be +1.92. The number of participants who had heard of CALL applications which integrate speech synthesis was also calculated. 2 out the 12 participants had heard of CALL applications that integrate speech synthesis.

Average (mean) ratings of comprehensibility and acceptability of utterances in each function were calculated for each participant. The mean ratings of comprehensibility, acceptability and overall appropriateness across participants are shown in table 1.

Table 1. Mean ratings comprehensibility (Comp.), acceptability (Accept.) and overall appropriateness (Approp.) (n =12).

	Reading	Pronunciation	Conversation
Comp.	1.21	0.91	2.07
Accept.	0.82	0.06	1.49
Approp.	0.17	-1.5	0.5

The data were analysed using the Friedman test. Significant differences were found among the ratings of comprehensibility (¯^{2}=18.67, df=11, p<0.01). Differences between the ratings of both acceptability (¯^{2}=8.17, df=11, p<0.01) and overall appropriateness (¯^{2}=6.29, df=11, p<0.01) were however not significant.

The data were also analysed using the Spearman rank correlation coefficient test to determine whether there was a correlation between ratings of comprehensibility and acceptability. Overall, a significant positive correlation was found between ratings of the comprehensibility and acceptability of utterances (*r*=0.523, N=12, p<0.05). Considering the functions of speech synthesis in CALL individually, the correlation of the ratings of the comprehensibility and the acceptability of the output was found to be most significant in the context of use as a conversational partner (*r*=0.573, N=12, p<0.05). It was also found to be significant in the context of use as a reading machine (*r*=0.534, N=12, p<0.05) but not in the context of use as a pronunciation tutor (*r*=0.505, N=12, p<0.05). Correlations between ratings of comprehensibility and appropriateness, and acceptability and appropriateness cannot be calculated because the data are not compatible.

4.3 Discussion

Ratings of comprehensibility, acceptability and overall appropriateness differ for the three different functions. Specifically, the speech was found to be most comprehensible in the context of use as a conversational partner, and least comprehensible in the context of use as a pronunciation tutor. Similarly it was found to be most acceptable and most appropriate for use as a conversational partner, and least acceptable and appropriate for use as a pronunciation tutor. While the differences in comprehensibility are significant, they are not for both acceptability and overall appropriateness.

Regarding the correlation between comprehensibility and acceptability levels, its significance was found to differ for the three different functions. It was found to be significant in the contexts of the use of speech synthesis as a reading machine and as a conversational partner, the significance being greater in the latter context, but not in the context of the use of speech synthesis as a pronunciation tutor.

These results suggest that the contribution of comprehensibility to acceptability differs in the three contexts. This supports our 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.

The correlations are not particularly high. None are greater than 0.6. This may be because other factors such as accuracy and naturalness play a significant role in determining acceptability, or because the experiment was not sufficiently contextualised for participants to make judgements of the acceptability of speech synthesis for use in these contexts - participants were left to imagine precisely how the speech synthesis might be used in each of the three functions. Contextualisation is particularly important when working with participants who are not familiar with the benefits and potential uses of speech synthesis in CALL..

5. Conclusions and Further Work

Review of the literature on SLA and LL&T suggests that CALL imposes a wide range of requirements on speech synthesis. Some of these can be evaluated simply through the use of checklists. Others can only be tested through the use of perceptual experiments. Perceptual experiments are time consuming. As mentioned, the goal of benchmarking is to be quick and efficient. A benchmark may therefore not be able to address all of these requirements. It is therefore necessary to identify which have the highest correlation with acceptability and appropriateness. The results of our experiment reveal that there is a significant correlation between ratings of comprehensibility and acceptability in the following contexts: as a reading machine and as a conversational partner. However, these correlations are not high. This suggests that other requirements may contribute more significantly to determining acceptability. Further research is necessary to investigate this. Such investigations should be more contextualised in order to facilitate/ease judgement of acceptability. The purpose of technology evaluation is to avoid wasting time and resources integrating a technology for which it is not suitable. If a higher degree of contextualisation is necessary, we should therefore be careful not to confuse contextualisation with integration. Contextualisation will therefore be a challenge for further requirements analysis As a preliminary suggestion, experiments investigating the requirements of speech synthesis for use in CALL could be contextualised through the use of mock screen shots of applications in which it might be used.

Acknowledgements

We would like to thank all our participants who have responded so far as well as those whose responses we will receive in the future.

References

[1] Stratil, M, Weston, G., Burkhardt, D. (1987). Exploration of Foreign Language Speech Synthesis. *Literary and Linguistic Computing*, 2 (2), 116-119.

[2] Stratil, M., Burkhardt, D., Jarratt, P., and Yandle, J. (1987). Computer-Aided Language Learning with Speech Synthesis: User Reactions. *Programmed Learning and Educational Technology*, 24 (4), 309-316.

[3] Santiago-Oriola, C. (1999). Vocal Synthesis in a Computerized Dictation Exercise. In *EUROSPEECH'99* (pp. 191-194). Budapest, Hungary.

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

[5] Hirschman, L. and Thompson, H. S. (1996). Overview of Evaluation in Speech and Natural Language Processing. In Battista Varile, G. and Zampolli, A. (Eds.). *Survey of the State of the Art in Human Language Technology* (pp. 409-414). Cambridge: Cambridge University Press.

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

[7] Bonneau, A., Laprie, Y. and Colotte, V. (2000). Towrads phonetic tools for speech training. In *InSTIL 2000* (pp. 77-80). Dundee: University of Abertay Dundee.

[8] van Bezooijen, R. and van Heuven, V. J. (1997). Assessment of Synthesis Systems. In Gibbon, D., Moore, R. and Winski, R. (Eds.). *Handbook of Standards and Resources for Spoken Language Systems* (Vol. 3, pp. 481-563). Berlin and New York: Mouton de Gruyter. p. 497.

[9] Paroubek, P. and Blasband, M. (1999). *ELSE Executive Summary (short version).* http://www.limsi.fr/TLP/ELSE/PreambleXwhyXwhatXrev3.htm

[10] Francis, A.L. and Nusbaum, H. C. (1999). Evaluating the Quality of Synthetic Speech. In Gardner-Bonneau, D. (Ed.). *Human Factors and Voice Interactive Systems* (pp. 63-97).

[11] ISO (1999). *Information Technology - Software Product Evaluation - Part 1: General Overview.* ISO.

[12] Chapelle, C. A. (2001). *Computer Applications in Second Language Acquisition*. Cambridge: Cambridge University Press.

[13] Kenning, M. M. and Kenning, M. J. (1990). *Computers and Language Learning: Current Theory and Practice.* Chichester: Elllis Horwood.

[14] FreeText: http://freetext6.reverso.net

[15] Hamel, M-J. (1998). Les Outils de TALN dans SAFRAN. *RECALL Journal*, 10 (1), 79-85.

[16] *Talk to Me, the Conversation Method*. Version 3.5. Auralog. http://www.auralog.com