MULTILINGUAL SYNTHESISER ASSESSMENT USING SEMANTICALLY UNPREDICTABLE SENTENCES

C. Benoît 1, A. van Erp 2, M. Grice 3, V. Hazan 3 and U. Jekosch 4

1 Institut de la Communication Parlée, Grenoble, France
2 PTT Dr Neher Laboratories, Leidschendam, the Netherlands
3 Dept of Phonetics and Linguistics, UCL, London, Great-Britain
4 Ruhr-Universität Bochum, Lehrstuhl für allg. Elektrotechnik und Akustik, Bochum, West Germany

ABSTRACT

The use of semantically unpredictable sentences was investigated as part of a standard test battery for the evaluation of synthetic speech within the ESPRIT SAM project. Such sentences were randomly generated using five syntactic structures and word lists for each syntactic category. Initial results of experiments run in English, French and German are presented.

INTRODUCTION

Within the ESPRIT "SAM" project, test material and methodologies are being developed for text-to-speech synthesiser assessment. Such standardized methods will be made available to potential users. The general goals of the SAM "Speech Output" group are to develop standard methodologies and guide-lines for test design rather than fixed material, with the aim of preventing producers from tuning their synthesisers to specific materials. A main aim is to develop material which is comparable across European languages and which would be suitable for testing all possible types of synthesiser. Material should also be designed that could be automatically generated and easily administered and scored.

Our primary concern here is to evaluate synthesiser intelligibility. Performance on sentence-length material is of particular interest as it is most representative of synthesiser use. Such material also allows for the assessment of prosody and of intelligibility across word boundaries.

Existing sentence-length material has already been used for synthesis assessment as well as for psycholinguistic and audiometric purposes. Meaningful sentence material has a number of disadvantages: the material is fixed and cannot be automatically generated and it provides contextual cues whose effect on intelligibility scores cannot readily be quantified. It is highly language-specific and does not allow for cross-linguistic comparisons. As a consequence, semantically unpredictable sentences (SUS) have been considered. Such an approach was first adopted by Miller and Isard (1) for psycholinguistic studies, and by Nye and Gaitenby (2) for synthesis assessment. Existing material is only available for English. This SUS approach has been developed further to include a set of five syntactic structures with a limited number of elements to avoid, among other things, short-term memory saturation.

SUS material has now been produced for the following languages: Dutch, English, French, German, Italian and Swedish. Intelligibility tests using SUS test material were carried out in three languages (English, German, French).

DEFINITION OF TEST MATERIAL

Two methods of categorising syntactic structures were employed: functional sentence elements which were constant across the three languages and syntactic categories whose order differed in a small number of cases (subject/verb inversion, noun/adjective inversion and position of adverbial).

Examples:
I. The table walked through the blue truth.
   La robe entre vers la science rouge.
   Der Bauch boxt mit dem wilden Dreck.

II. The strong way drank the day.
    Le verre vrai ouvre le coin.
    Ein mürbes Blatt schlürft den Rumpf.

III. Never draw the house and the fact.
     Dränge stets das Garn und den Fuss.

IV. How does the day love the bright word?
    Quand le texte pose-t-il la fille crue?
    Wann trinkt der Pelz ein grelles Kind?
V. The plane closed the fish that lived.
La chose lance le train qui pense.
Das Huhn heizt den Mann, der gräbt.

In all the five structures, nouns were singular and indicative verb forms were third person singular, simple past for English and present tense for French and German. The requirements of the verb type as regards transitivity were taken into account. Thus, transitive verbs were used exclusively in structures with a direct object, and intransitive verbs in structures without one. Minisyllabic words i.e. those words containing the smallest number of syllables within a given category in a given language, were used. This constraint was relaxed in the case of the adverb category in structure three where fewer than ten minisyllabic words could be found. Vocabulary was chosen that respected word frequency of occurrence in the language. A more detailed description of the test material can be found in van Erp and Grice (3) and Grice & Hazan (4).

TEST PROCEDURE

Tests were run in three different countries. Details of test methodology differed from site to site but the following aspects of test procedure were valid for all three studies. All tests were run in a sound-proof cubicle. Subjects were screened for normal thresholds of hearing at the beginning of the study. They listened to the SUS material via AKG 240 DF (English and German study) or EuropSonic (French study) headphones and responded to the test sentences by writing down what was heard. German and French listeners underwent a short training phase in order to acquaint themselves, at least minimally, with the synthesized speech and the anomalous aspect of the sentences.

In the English study, one synthesiser was assessed and in addition, natural speech (4 kHz bandwidth) was presented as a reference condition. The French study compared natural (degraded) speech with two diphone-based synthesizers in two different conditions: one with prosody synthesized by rule, and the second with constant pitch and phone duration ("constant prosody").The German study also assessed one synthesizer. The studies differed in the number of subjects tested and the amount of data collected per subject. Further details of the test procedures can be found in Hazan, Benoît and Jekosch (5).

RESULTS

Each study focussed on the effects of different test variables and the main findings are summarised below.

A strong effect of word frequency on sentence intelligibility was found in the English study (Fig.1). Sentences containing common words were more intelligible than sentences containing less common words. Such an effect has also been found for degraded natural speech by Howes (6). However, with a given vocabulary, different sentence lists appeared to be perceptually equivalent, as shown by a non significant difference in test scores obtained for different sentence lists at a given test session (English study).

![Intelligibility scores for sentences containing most common words (vocab.1) and less common words (vocab.2).](image)

Fig.1: Intelligibility scores for sentences containing most common words (vocab.1) and less common words (vocab.2).

The German study and a previous English pilot study (4) have both shown that SUS provide unreliable diagnostic information in terms of phoneme confusions so that this material is best used to obtain global intelligibility scores. A comparison of different scoring units (word correct vs sentence correct) was made in the French (Fig.2) and German (Fig.3) studies. The French results show a strong relationship between sentence and word scoring. These results suggest a mathematical relation between overall word intelligibility and overall sentence intelligibility. This relation, however, has yet to be quantified. In the case of word scoring, decisions have to be made as to whether all words or only content words are scored.
A strong learning effect was noted both in the English (Fig. 4) and French studies (Fig. 5) over the whole testing period. This learning effect could be due to memorisation of test words as well as to an habituation to synthetic speech. The learning effect may also have been increased because of the use of natural (even if slightly degraded) speech in clear as a reference condition. Some individual differences between subjects were noted in the English study. Some subjects showed consistently high scores across test conditions while others performed poorly at most sessions.

SUS material was compared to other test material in two of the studies. In the French study, some of the words included in the SUS sentences were used to create "semantised" sentences (eg "le vin chaud brule la bouche"). Scores obtained with these sentences were compared with intelligibility scores for SUS (Fig 5). The German study investigated the intelligibility of words in isolation vs words in unpredictable sentences (Fig.6). Scores were lower for words in isolation. This is, in part, due to the fact that isolated word tests provide no cues to the syntactic category of the word.

In both the German and English studies, only one synthesiser was assessed. In order to investigate whether the SUS material is sensitive enough to pick out differences between similar systems, two synthesis conditions differing in prosody were compared in the French study (Fig.7). Similarly, two different diphone synthesisers based on a same speaker were compared to the degraded version of natural speech by this speaker (Fig. 8). In both cases, differences between conditions were adequately detected by the SUS material.
CONCLUSION

As far as the structure of the test material is concerned, SUS do seem to fulfil the requirements set out above. Multilingual sentence-length material has been produced which can be automatically generated and which reduces contextual information. In the first evaluation of this material, tests have been found to be easy to administer, if lengthy.

Software for simultaneous testing of subjects and processing of results is at present under development within the SAM project, and may render the test administration less time-consuming. SUS has been found to be sensitive enough for the cross-comparison of similar systems. However, learning effects, due in part to the memorisation of test words and word frequency effects seem to strongly affect test results. Investigations must therefore continue into the identification and quantification of these and other factors.

ACKNOWLEDGEMENTS

This study was undertaken as part of Esprit project 1541 (Multi-lingual speech input/output: assessment, methodology and standardisation - SAM). We wish to thank all our SAM colleagues who were involved in the planning of this study.

REFERENCES


