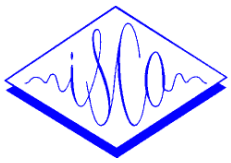

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

SEGMENTAL EVALUATION USING THE ESPRIT/SAM TEST PROCEDURES AND MONOSYLLABIC WORDS

Rolf Carlson, Björn Granström & Lennart Nord*

Dept of Speech Communication and Music Acoustics
Royal Institute of Technology
Box 70014, S-10044 Stockholm, Sweden

ABSTRACT

We have been using the preliminary version of the Esprit/SAM test procedure for synthetic speech to evaluate an experimental version of the multilingual text-to-speech system under development at our department. The proposed segmental test battery includes: a) hearing tests of the subjects. b) the familiarisation to the special type of speech synthesizer by an introductory paragraph. c) lists of CV, VC and VCV stimuli according to the phonotactic structure of the individual language. Tests on natural speech have also been performed forming a baseline for the synthesis evaluation and at the same time indicating the subjects' ability to give unambiguous orthographic response to nonsense words. We will also present data on the intelligibility of monosyllabic words drawn from the most frequent 10 000 words in Swedish.

1. INTRODUCTION

In Europe, evaluation of speech technology devises has been the objective of a joint research initiative, the Esprit/SAM project (Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation), [Fourcin, Harland, Barry & Hazan, 1989]. This report presents results on a basic segmental test proposed by SAM. This test is designed to alleviate some problems found with standard MRT or DRT tests [Carlson, Granström & Nord, 1990]. The test should also be applicable to the different European languages. The test is a nonsense word test, combining VCV, CV and VC words, where C and V denotes single consonants and vowels respectively. The test uses the extreme vowels /a/, /i/ and /u/ or the closest vowels compatible with the language structure. Consonants are the full set possible in the different positions. Since an open response, nonsense word format is used, all possible confusions are investigated. Some doubts have been expressed as to the usefulness of such a test with phonetically naive subjects. Response problems might occur due to the quite unnatural situation of listening to nonsense words. To check this and to get a base-line for the evaluation we added an initial test with the identical test material spoken by a male speaker. For similar reasons we also tested our subjects on real mono-syllabic words using the same synthesizer.

2. EXPERIMENTAL PROCEDURE

Test tapes were produced of nonsense VCV, VC and CV words. The vowels used were /a/, /i/ and /u/. Due to the phonotactic structure of Swedish the vowels were phonologically short in the VCV and VC lists (i.e. phonetically [a], [I] and [U]) but long for the CV (i.e. [ɑ:], [i:] and [u:]). In the VCV we used 18 consonants (the full set, excluding retroflex allophones, i.e /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /ŋ/, /f/, /s/, /ɧ/, /ç/, /h/, /v/, /j/, /l/ and /r/). In the CV context 17 consonants are possible (not /ŋ/). In VC words 16 consonants were used (not /h/ and /ç/). Two differently randomized lists of each structure were recorded. The lists contained all combinations, i.e. mostly nonsense words. The test lists were produced both by a male speaker and by an experimental software synthesizer. The speech was presented through headphones. Subjects were asked to respond in writing on a response form where the vowels were indicated. They were instructed to respond with a single consonant, from the phonotactically possible inventory given at the top of the response form. No training in transcription was performed, but the response inventory was explained to the subjects prior to the test. First, one natural speech list

* Names in alphabetic order

(VCV, VC or CV) was presented to the subjects. This served both as familiarization to the procedure and a baseline for the evaluation. Then the particular synthesizer was introduced to the subject by a short story. Three lists of synthetic nonsense words, one of each kind, were presented according to a rotated design. Thus, the test was run with 24 subjects and each subject heard all their lists in one session. After the nonsense word test each subject was given an additional intelligibility test using real monosyllabic words. The subjects did not have any substantial previous exposure to synthetic speech and were regarded to be phonetically naive. All subjects were native speakers of Swedish and had normal hearing as judged from their pure tone audiogram.

3. RESULTS

In the following presentation we are using our own phonetic conventions close to the Swedish orthographic conventions; the non-obvious deviations are "ng", "sj", "tj" with the IPA counterparts /ŋ/, /ɧ/, /ç/. No distinction is made for vowel quantity in the transcriptions.

In Figure 1, the over-all error rate for natural and synthetic speech is displayed. Very few confusions were observed for the human speaker. Many of the errors for the VCV and CV lists concerned /sj/, /tj/ confusions that are not possible in the VC structure. We are not making any further error analysis of the human speech due to the low error rate. The error analysis for the experimental software synthesizer reveals some definite design problems and the total error rate was in fact higher than the ones obtained with the standard hardware [Carlson et al., 1990].

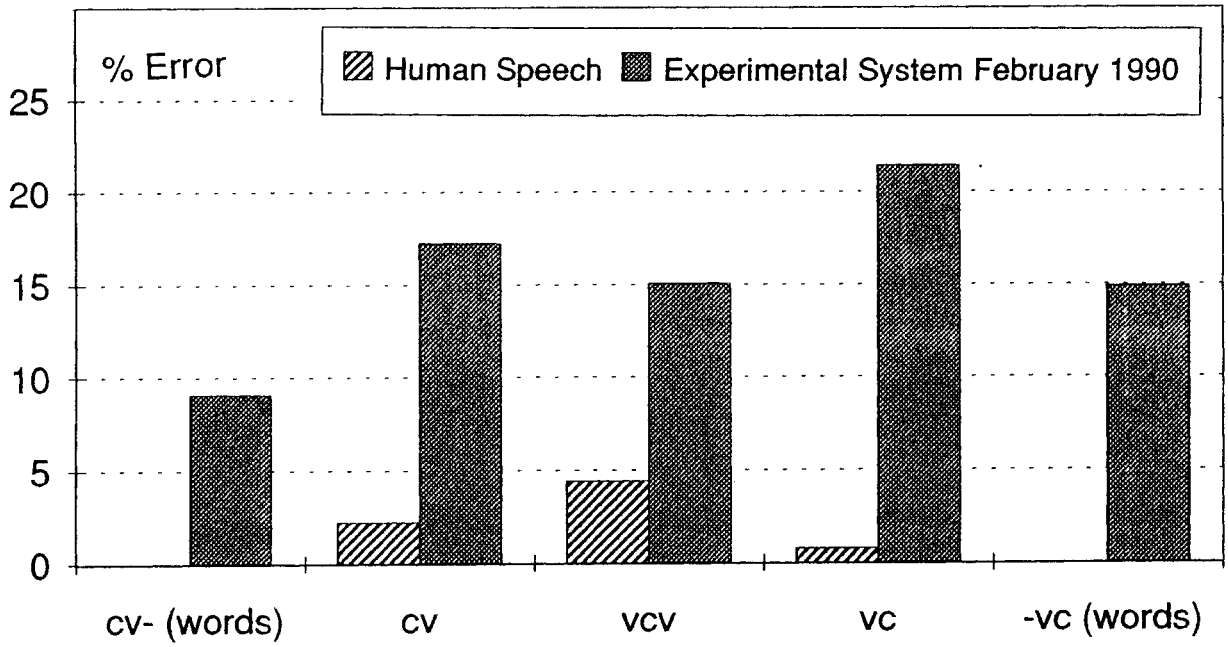

Figure 1. Over-all error rate for natural and synthetic speech. Errors in consonant recognition for the different nonsense words compared with errors in initial and final consonant clusters in real mono-syllabic words (clusters contain one to four consonants)

Due to space limitations we only present some results from the CV test as an example. Errors of the individual consonants are displayed in Figure 2, according to different vowel contexts (top). In the lower part of the figure, the data is analysed in terms of correctly perceived manner and place information. It is obvious that the confusions are strongly dependent on vowel context. Many identification problems are almost exclusive to one or two contexts as /p/ in /a/ context, /l/ in /i/ and /u/ contexts and /k/ and /g/ in /i/ context. The last confusion may be accentuated by the unusual context. Most /k/ and /g/ before /i/ will, by a phonological rule, transform to fricatives, /ç/ and /j/ respectively. Some of the confusions in this study are new compared the standard synthesizer and the results provided valuable feedback in modifying the experimental synthesizer and the rules to control it. As an example, some stop confusions could be traced back to an implementation error of the aspirative source.

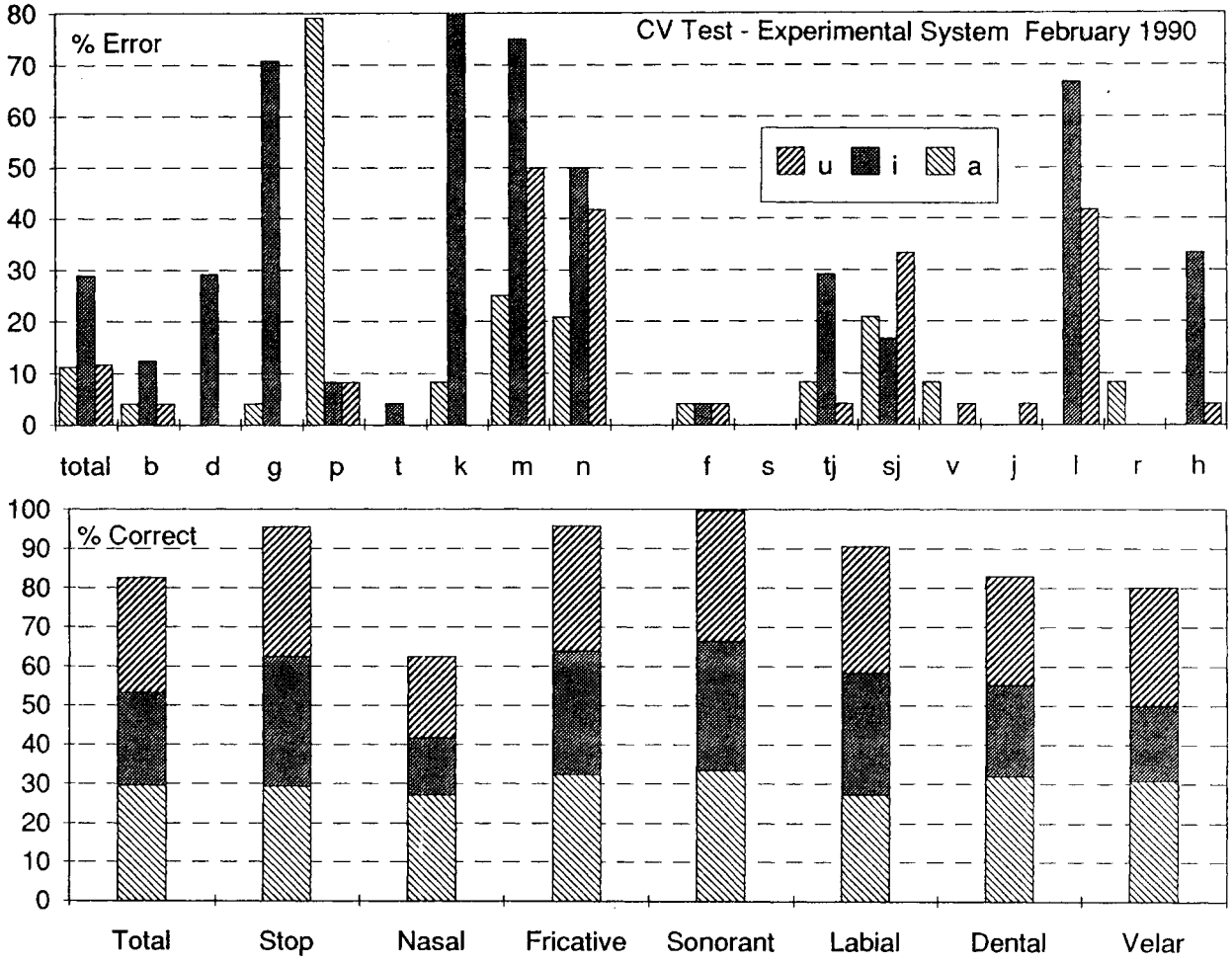

Figure 2. CV test: Errors for individual consonants according to different vowel contexts (top). Correctly perceived manner and place information (bottom)

A special kind of error concerns the fricatives /sj/ and /tj/. Most of these confusions occur between the two consonants. There are at least two explanations for the high confusion rate. There exist in Swedish two distinctly different allophones of /sj/ [ʂ,ʃ] that acoustically/articulatorily are on opposite sides of the quite similar /tj/ [ɕ] sound. The orthographic representation of these sounds are quite varied and to some extent overlapping. This can result in a response problem rather than the perceptual confusion that we have set out to evaluate. This conjecture is supported by an analysis of the few confusions in the test with human speech that showed that 23 out of the totally 31 errors involved these consonants and 13 were confusions between the two. With phonetically trained subjects the response confusions can be minimized but we feel that the value of using naive subjects dominates. A real word test would not provoke these potential problems, but could be less systematic since it relies on the somewhat arbitrary phonetic use of the lexical space.

4. MONO-SYLLABIC WORD TEST.

As mentioned above, the standard Esprit/SAM nonsense word test was complemented by an intelligibility test using real words. Common words were chosen to avoid possible gaps in the subjects' active vocabularies. 1000 mono-syllabic words were selected from the most frequent Swedish words and were randomized in 10 lists of 100 word each. Each word was used only once. Each subject listened to one word list as the final part of the test session. The result was analysed according to errors in vowels and in final and initial consonants. In Figure 1, the result is shown along with comparable results from the nonsense test. As can be seen, the error rates are substantially lower for real words. Furthermore, the consonant clusters in the real words might be partially correct, since they are scored as incorrect if not totally right. The error rates

from the nonsense test and the mono-syllabic test are not immediately comparable since in the first case phonemes are of equal probability while in the latter test frequencies represent the usage in natural language. Spiegel et al. [1988] has reported on a mono-syllabic word test for American English using equal proportions of real words and nonsense words. They found a comparable difference in intelligibility between words and nonsense words for both human and synthesized speech. Looking at single consonants, which are also represented in the nonsense test, and consonant clusters separately reveals some interesting details, Table 2.

Table 2. Results from the mono-syllabic word test.

Position	CV	C2+V	VC	VC2+
Tested types	17	31	20	88
Tested tokens	1601	681	1353	907
Errors	146	120	210	123
% errors	9.1	17.6	15.5	13.6

In the table, "C" refers to single consonants and "C2+" to consonant clusters, of two, three or four consonants, evaluated as a unit. The most intelligible position is initial single consonants. The high error rate for initial clusters could be predicted almost exactly from the error probability of single consonants if we presuppose that most of the clusters contain two consonants, (counting all consonants separately in the initial clusters gives an error rate of 9.7%). This is however a misleading view as can be seen from the corresponding data in the word final position. Phonotactic constraints are stronger in the initial positions as can be seen from the "tested types" line. These constraints should be possible to use to restrict the lexical search space in the recognition task, using real words.

A detailed analysis of the initial cluster errors are due to just a few types of misperceptions, like voicing of voiceless stops. In the Spiegel et al. study, final consonants were less intelligible than initial consonants and final clusters had considerably more errors than single consonants both for natural and synthetic speech. The initial/final difference is reproduced in our study when it comes to single consonants. However, clusters are in fact perceived more accurately than single consonants in the final position in our study, possibly due to the reduction in the lexical search space. This constraint could not come into play as strongly in the Spiegel et al. study since they used a mixed nonsense word/word test vocabulary. It is however quite promising that the results from single consonant test in the final position generalize to clusters. A detailed analysis gives valuable diagnostic information on how to improve cluster coarticulation rules, as evidenced from the large cluster error rate in initial position.

ACKNOWLEDGEMENTS

This work has been supported by grants from the Swedish National Board for Technical Development and the Swedish Telecom.

REFERENCES

Carlson, R., Granström, B. & Nord, L. (1990), "Evaluation and development of the KTH text-to-speech system on the segmental level", Proc IEEE 1990 Int. Conf. on Acoustics, Speech, and Signal Processing, (21.S6a.7), Albuquerque, New Mexico, USA.

Fourcin, A.J., Harland, G., Barry, W. and Hazan, V. (eds.) (1989), Speech input and output assessment - multilingual methods and standards, Ellis Horwood Limited, Chichester, England.

Spiegel, M., Altom, M.J., Macchi, M. & Wallace, K. (1988), "Using a monosyllabic test corpus to evaluate the intelligibility of synhesized and natural speech", Proc. of the American Voice Systems Conference, San Francisco, CA, USA.