First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SEGMENTAL INTELLIGIBILITY OF SYNTHETIC AND NATURAL SPEECH IN REAL AND NONSENSE WORDS

Rolf Carlson+, Björn Granström and Lennart Nord*
Department of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Box 70014, S-10044 Stockholm, Sweden
+presently at Laboratory for Computer Science, MIT, USA

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. An interesting question in this context is the phonemic awareness of the listeners. The Swedish fricative allophone set is a good example, where difficulties in labelling has to be studied carefully.

Results will be presented at the meeting and compared to data reported earlier. We will also present data on the intelligibility of monosyllabic words drawn from the most frequent 10 000 words in Swedish.

I. INTRODUCTION

Evaluation of speech technology devises has recently attracted considerable interest. Speech synthesis systems can be evaluated on several different levels varying from detailed studies of segmental intelligibility to global evaluations of acceptability in a real world application. The reason for the evaluation can also vary. One reason can be comparative scoring of systems before purchase; another reason can be diagnostic testing with the main objective of evaluating a change to an individual system. This requires a variety of testing methods, each with its own advantages and limitations. Pisoni and co-workers at Indiana University have reported on several different methods [1,2]. 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). The result of the first phase is available in a book [3] and the work now continues in the second phase of SAM. This report presents results on a basic segmental test proposed by SAM. In our department we have since many years conducted tests of a similar kind and some comparisons and references to that work are also given.

Our own interest in testing has mainly been for diagnostic reasons. The most widely published data on speech synthesis intelligibility are of the MRT or DRT type where real words are used with a forced choice response of 6 or 2 alternatives respectively. Many objections to these tests for speech synthesis evaluation have been raised. Our main concern is:

1. They do not test for all possible confusions, not even in an open response mode, since the test items are restricted to real words (MRT).
2. The test is too easy. Error rates are rather low for the best systems, implying that much response data needs to be collected to reach significance.
3. Test items do not occur with equal probability, implying that confusions presented as confusion matrices, for example, are hard to evaluate.
4. Consonants are only tested in positions next to pauses, which is a rather unusual context in running speech.

The basic SAM segmental test is designed to alleviate these problems and still be applicable to the different European languages. Cross-language comparisons will always be difficult, due to for example language structure differences, but the standardized procedure and test format will minimize these problems. 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. The stimulus set is small and the scoring is very simple. For a diphone system, this test would, of course, only test a small, but important, subset of the diphones. Since our approach to speech synthesis is based on phone-size units and coarticulatory rules, the testing of consonants in the context of articulatory rather extreme vowels will give us a great amount of diagnostic information and also be a valuable measure of segmental intelligibility. 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

II. SYNTHESIS SYSTEM

The KTH synthesis system used in the experiments has been under development for a long period of time [4,5]. During this work, a number of diagnostic tests have been done in order to direct the developments. Previous evaluations have used a test procedure quite similar to the one used in the present study, employing the VCV structure only [6,7]. Partial comparison with tests of the earlier versions is thus possible. The present test was performed on an experimental software synthesizer under development. The error analysis reveals some definite design problems and the total score was in fact lower than the ones obtained with the standard hardware.

* Names in alphabetic order

III. SUBJECTS

The subjects were recruited from university level electrical engineering student, attending a class in speech communication. They 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. In all 24 subjects participated in the experiment.

IV. 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 and a few meaningful words, i.e. 54 items (18 consonants * 3 vowels) in the VCV lists, 51 items (17*3) in the CV lists and 48 items (16*3) in the VC lists. The test lists were produced both by a male speaker and by an experimental software synthesizer and recorded on a Studer A 807 tape recorder. The duration between the offset of one stimulus and the onset of the next was 4.6 sec. To normalize attention a short, low-level sound preceded each stimulus by 1.3 sec. (two pulses of a /u/-like sound, not causing any masking). The test was run in a sound-treated recording booth. Two subjects were run at each occasion. The speech was presented through Sennheiser headphones (HD 250 linear, diffuse field loudness equalized). The sound was adjusted to 70 dB according to a Speech Voltmeter type SV6, manufactured by British Telecom. 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 (VCV, VC or CV) was presented to the subjects. This served both as training of the procedure and a baseline for the evaluation. Then the particular synthesizer was introduced to the subject by a short story (30 seconds of speech). The three synthetic test lists were presented according to a rotated schedule. The test was run with the 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.

V. RESULTS

In our presentation we have used our own phonetic conventions close to 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.

The data analysis was performed using the Microsoft Excel spread-sheet program. As input the individual scores and the pooled raw confusion matrices were entered.

In Figure 1, the over-all error rate for natural and synthetic speech is displayed. Each bar corresponds to approximately 400 samples for the natural speech and 1200 samples for the synthetic speech, due to the design of the experiment. As expected 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.

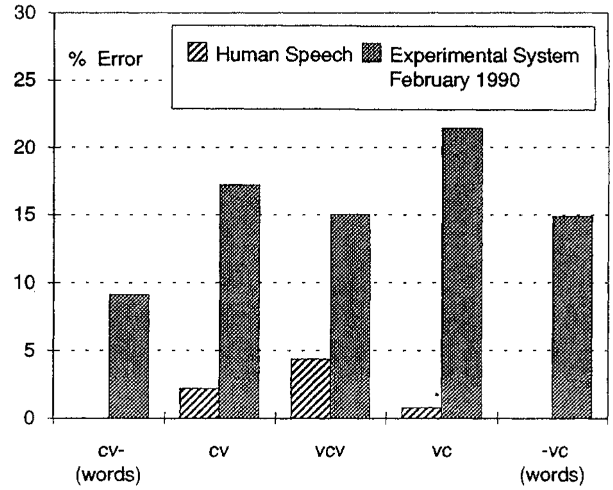

Figure 1. Over-all error rate for natural and synthetic speech in real and nonsense words.

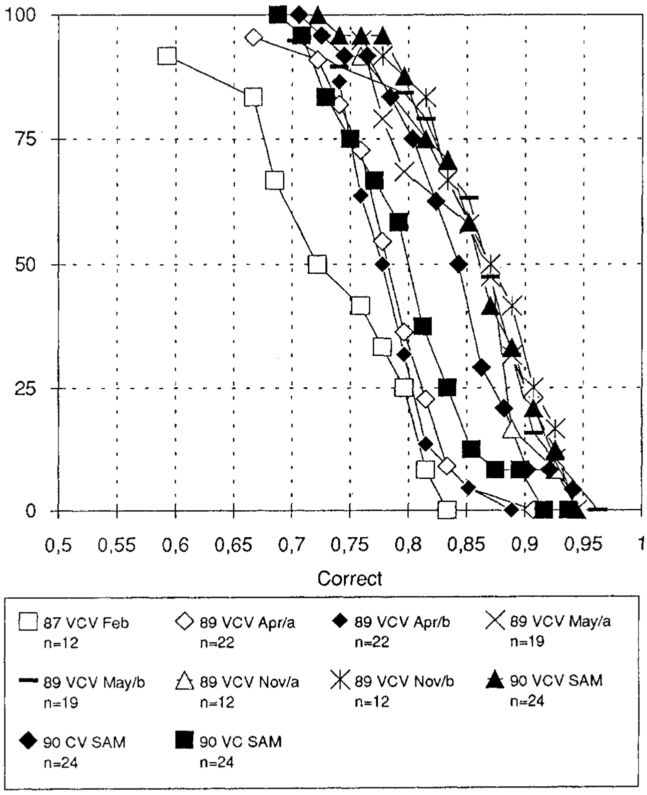

Figure 2. Distribution of individual results for different tests since 1987

	b	d	g	p	t	k	m	n	ng	f	s	tj	sj	v	j	l	r	h	???	
b	23	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	b
d	.	24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	d
g	.	.	23	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	g
p	.	.	.	5	.	5	.	.	.	.	.	1	.	.	.	.	.	13	.	p
t	.	.	.	.	24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	t
k	.	.	1	.	.	22	.	.	.	.	.	.	.	.	.	.	.	1	.	k
m	.	.	.	.	.	.	18	.	.	.	.	.	.	6	.	.	.	.	.	m
n	.	.	.	.	.	.	2	19	.	.	.	.	.	3	.	.	.	.	.	n
n.a.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ng
f	.	.	.	.	.	.	.	.	.	23	.	.	.	.	.	.	.	1	.	f
s	.	.	.	.	.	.	.	.	.	.	24	.	.	.	.	.	.	.	.	s
tj	.	.	.	.	.	.	.	.	.	.	.	22	1	.	1	.	.	.	.	tj
sj	.	.	.	.	.	.	.	.	.	.	.	4	19	.	.	.	.	1	.	sj
v	.	.	.	.	.	.	.	.	.	.	.	.	.	22	.	.	2	.	.	v
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	24	.	.	.	.	j
l	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	24	.	.	.	l
r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	22	.	.	r
h	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	24	.	h
	23	24	24	5	24	27	20	19	.	23	24	27	20	32	25	26	25	40	.	

Figure 3. Confuṣion matrix for consonants in synthetic CV words (vowel /ɑ/).

Even though VCV in average appeared to be most intelligible there are several exceptions to this among the individual subjects. The subjects appear to form a rather homogeneous group. In Figure 2 the distribution of results is displayed for several different tests since 1987. A clear general improvement since 1987 can be observed.

In Figure 3 an example of a raw confusion matrix is shown where responses are pooled across subjects but split according to vowel context, in this case the CV structure and /a/ context.

Errors of the individual consonants are displayed in Figure 4 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. The figure concerns the CV structure. 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.

A special kind of error concerns the fricatives /sj/ and /tj/. Most of the confusions occur between these two consonants. There are at least two explanations for the high confusion rate.

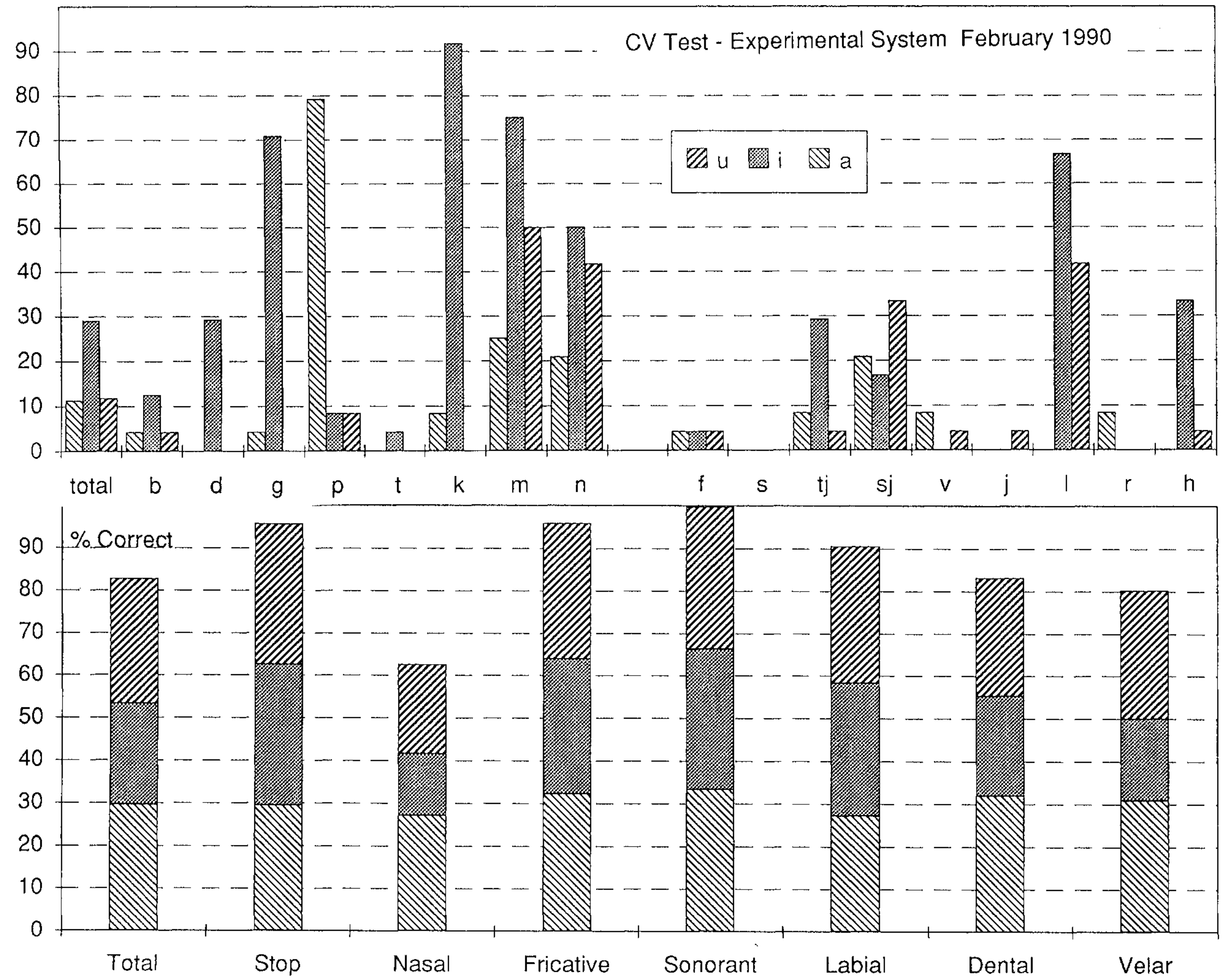

Figure 4. CV test: Errors for individual consonants according to different vowel contexts (top). Correctly perceived manner and place information (bottom).

There exist in Swedish two distinctly different allophones of /sj/ [ʂ,ɧ] 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 the lexical space.

VI. MONO-SYLLABIC WORD TEST

As mentioned above, the standard Esprit/SAM nonsense word test was complemented by a intelligibility test using real words. Common words were chosen to avoid possible gaps in the subjects' active vocabularies. 1000 monosyllabic words were selected from the most frequent Swedish words and were randomized in 10 lists of 100 words. Each word was used only once. The subjects listened to one list as the final part of the test session. The result was analysed according to errors in vowels and in final and initial consonants or consonant clusters. In Figure 1, the result is shown along with the comparable results from the nonsense test. As can be seen, the error rates are substantially lower for the 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. [8] has reported on a mono-syllabic word test for American English using equal proportions of real words and non-words. They found a comparable difference in intelligibility between their words and non-words for both human and synthesized speech.

In Table 1, some data on the word test is presented. Analyzing single consonants, which are represented in the nonsense test, and consonant clusters separately reveals some interesting details.

Table 1

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 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 Speigel et al. study since they used a mixed non-word/word test vocabulary. 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

[1] Pisoni, D.B., Nusbaum, H.C. & Greene, B. (1985): Perception of synthetic speech generated by rule", Proc. IEEE 73, No 11., pp. 1665-1676.

[2] Logan, J.S., Greene, B.G. & Pisoni, D.B. (1989): "Segmental intelligibility of synthetic speech produced by rule", J. Acoust. Soc. Am., 86 (2), pp. 566-581.

[3] 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.

[4] Carlson, R., Granström, B. & Hunnicutt, S. (1982): "A multi-language text-to-speech module", Proc. ICASSP 82, Vol. 3, Paris, pp. 1604-1607.

[5] Carlson, R., Granström, B. & Hunnicutt, S. (1990): "Multilingual text-to-speech development and applications", in A.W. Ainsworth (ed), Advances in speech, hearing and language processing, JAI Press, London

[6] 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.

[7] Barber, S., Carlson, R., Cosi, P., Di Benedetto, M.G., Granström, B. & Vagges, K. (1989): "A rule based Italian text-to-speech system", Proc. of EUROSPEECH 89, Paris.

[8] 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.