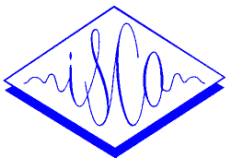

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

IS % OVERALL ERROR RATE A VALID MEASURE OF SPEECH SYNTHESISER AND NATURAL SPEECH PERFORMANCE AT THE SEGMENTAL LEVEL?

Mikael Goldstein and Ove Till

Telia Research AB, Human Factors and Services Research
136 80 Haninge, Stockholm, Sweden

ABSTRACT

Intelligibility of 18 Swedish consonants embedded in 3 symmetrical vowel contexts (a_a, i_i and o_o) was assessed using the VCV test procedure SOAP v3.0 (ESPRIT PROJECT 2589 (SAM)) implemented on a PC computer. 24 subjects with normal hearing assessed the 54 different VCV combinations.

Two synthesizer systems, the KTH and the INFOVOX system were assessed. Natural speech was used as a baseline condition.

Although the % overall error rate for Natural speech was very low (5.56%) as compared to the KTH (8.72%) and INFOVOX (12.27%) synthesis systems, it was located to three VCV words that obtained very high error rates: otjo (54%), atja (50%) and oho (34%). For the VCV word oho, Natural speech yielded a significantly higher % error rate than that obtained for each of the two synthesis systems.

Differences between two correlated proportions (% correct) obtained for each VCV combination and system assessed by the same group of subjects were tested using the new program module DPROP.EXE, which uses the result files generated by the SOAP software as input files. The program tests for differences between all VCV combinations generated by two different systems that have been assessed by the same group of subjects. The testing (at the 2%-level, two-tailed t-test) of VCV words between KTH-Natural speech yielded 4 significant (*) differences (+ongo*, +ingi*, +ini* and -oho*), between INFOVOX-Natural speech 9 significant differences (+ovo*, +ingi*, +ongo*, +omo*, +ibi*, +opo*, +olo*, +ama* and -oho*) and between INFOVOX-KTH synthesis 3 significant differences (+ovo*, +opo* and +ibi*). In order to be significant, the % error difference had to be of the order of 30%.

The use of the % overall error rate as a valid diagnostic synthesiser performance measure is discussed. A measure that treats all insignificant VCV differences the same way as significant differences, by adding them together into an % overall error rate. This measure is compared to significance testing of individual VCV words, as well as the use of Natural speech as a ´true´ baseline.

10.21437/ICSLP.1992-133

I. INTRODUCTION

This report presents results from the SAM segmental VCV test (SOAP v3.0) [1] for the Swedish language. The test is performed on two rule-based synthesisers generating the Swedish language and on Natural speech. The first system is an experimental software synthesiser under development at the Royal Institute of Technology (KTH), Stockholm [2][3]. The second system is a rule based commercially available synthesiser implemented on a PC board, manufactured by the Swedish company Infovox (IVX), owned by Swedish Telecom AB.

The VCV segmental intelligibility test is implemented on the SAM output workstation (PC-386) and is completely automatic. Eighteen Swedish consonants are presented visually on the monitor screen immediately after an aural VCV stimulus presentation, and each consonant is surrounded by a black frame. The subject´s (S´s) task is to identify which test consonant he heard aurally through the headphones, by clicking on the corresponding consonant presented on the screen, using the mouse.

In order to establish a baseline for the evaluation, the VCV test material was recorded for Natural speech (NAT) as well, using a male speaker. In this way a "true" reference condition was established, against which % error scores obtained for the synthesised conditions could be compared.

The ability to discriminate between VCV intelligibility scores obtained for various synthesisers is discussed. Especially the conclusions drawn from a statistical point of view. When conducting VCV segmental intelligibility tests, within the ESPRIT/SAM project, usually no statistical approach is applied to the data, in order to verify if the obtained % differences are significant or not [3][4]. A remedy is suggested, whereby significance testing of differences between correlated proportions is proposed. Various implications that this statistical approach puts forward, are discussed.

II. METHODOLOGY

2.1. VCV material

The test is a VCV (vowel-consonant-vowel) segmental intelligibility test where the consonants appear in a symmetric vowel context, consisting of the vowels /a_a/, /i_i/ and /o_o/.

Eighteen Swedish consonants and consonant clusters were used in the VCV test: /b, d, g, p, t, k, m, n, ng, f, s, tj, sj, h, v, l, r, and j/. Each random list thus consisted of 54 VCV test items (18 consonants x 3 vowel contexts). Each synthesiser condition obtained a randomization order of its own. Each S was exposed to three different conditions: VCV words generated by Natural speech (NAT) male speaker, VCV words generated by the KTH experimental speech synthesis under development and VCV words generated by the IVX speech synthesis.

2.2 Subjects

Twenty-four (n=24) Ss, working at the Swedish Telecom premises took part in the experiment (mean age=30 years, sd=8,5 years). They were screened using a Bekesy audiometer. A maximum deviation of 20 dB from perfect hearing for frequencies between 0.25-8 kHz was allowed.

2.3 Testing conditions

One S at the time, was seated in a sound-proof booth. The segmental VCV words were delivered binaurally through headphones (Sennheiser HD250 linear, diffuse field loudness equalized). Written instructions prior to the VCV testing were given to each S. The Ss were also instructed, that the configuration of the eighteen consonants appearing visually on the screen (six consonants per row), was presented in alphabetical order.

Examples of words containing the consonant clusters /sj/, /tj/ and /ng/ were given. After having read through the instructions and familiarised with the mouse, each S had a training session consisting of 18 VCV words presented in Natural speech. The result from this training session was not recorded. After the training session, each S was first presented with 54 VCV words generated by Natural speech. After that, the two synthesiser systems (KTH and IVX synthesis) were presented according to a randomised block design, in order to balance out any experimental position effects. Each condition had its own randomisation order. Each synthesised speech condition was preceded by a short story generated by synthesised speech (30 s of synthesised speech), in order for the S to familiarise himself with the synthetic speech condition.

The S had approximately 5 s time to recognise and mark (by moving the mouse to the appropriate consonant frame and click) which consonant he had heard. If no answer was given within this 5 s period of time, a "time-out" error occurred and the next VCV combination on the randomisation list was presented aurally.

2.4 Active speech level

In order to adjust the 'three' systems to produce approximately equal loudness levels, the method of measuring the active speech level of running speech was utilized [5].The active speech level was measured using the digital speech volt meter SV6. The long-term

active speech level of the short story generated by the two synthesiser systems was set to approximately 73 dB SPL. For Natural speech, no introductory story being available, the active speech level of the first ten VCV words was set to approximately 73 dB SPL. Comparing the active speech level of the first ten VCV words for both conditions of synthetic speech gave a SPL reading of the same magnitude.

2.5 Statistical treatment of significance of the difference between two correlated proportions

No significance testing has been made in earlier studies to establish if a certain percentile increase or decrease is significant or not. If system KTH receives a 5% higher correct rate than system IVX for a certain VCV word, does that imply that KTH synthesis is better for that VCV word, or is this difference only due to chance? In order to test if a % difference is significant we have to apply statistical analysis tools to obtain a firm base for our judgments.

The significance of the difference between two correlated proportions or percentages based on the same sample of Ss [6] is a statistical test computed by McNemar [7] well suited for this purpose. Since the same S assesses the same VCV word x generated by both (all three) systems, there might exist a correlation between the two responses. One S may "pass" VCV word x generated by system KTH and also "pass" VCV word x generated by system IVX. A second S may "pass" VCV x generated by system KTH but "fail" when listening to the same word generated by system IVX. A third S "fails" VCV word x generated by both systems. The paired observations may be tabulated in a 2x2 contingency table. The proportion passing (% correct) VCV word x generated by system KTH and IVX is denoted pKTH and pIVX. We wish to test the difference between pKTH and pIVX for the VCV word x.

Thus each VCV word (18 x 3=54) can be statistically tested to see, if the difference between proportion correct responses for system KTH and IVX is significant. In this way, a more firm judgment can be made regarding differences between systems.

We have to realise, that if there are only 24 observations for each VCV word, each observation when converted to percentages is approximately 4%. A difference of around 8-10% between a VCV word generated by two systems only implies that two more Ss actually managed to assess the consonant correctly. Thus it seems difficult to believe that even a 10-20% difference would actually yield a significant result. What we will get by adopting statistical significance testing is a tool that converts percentages into meaningful units.

By including 2x2 contingency tables for all VCV words and test variables for significance testing for all 54 VCV words between systems (KTH-NAT, IVX-NAT and IVX-KTH) in the SOAP Output assessment scoring software as an additional software module, a new diagnostic tool was implemented. The statistical test also supplies the user with % correctly answered VCV words for each system at the individual S level.

2.6 Data Analysis

The data computations were made using the SOAP scoring module. The automatic scoring software produces confusion matrices, articulation scores for the 18 consonants and an environmental matrix based on vowel context. The new run file DPROP.EXE (source code DPROP.C) computes the difference between the two correlated proportions for each of the 3 x 18=54 VCV combinations. The DPROP.EXE program computes the differences between KTH-NAT, IVX-NAT and IVX-KTH synthesis. The SOAP result files containing the individual S's preprocessed scoring files with the extension "___.slf" are used as input files for the DPROP.EXE program. A high significance level (*, p<0.02, two-tailed t-test) is chosen, due to mass-significance problems. For n=24 Ss, this implies a significance level of t=2.492.

III. RESULTS

In Fig. 1, the over-all % error rates for Natural and synthetic speech are displayed. As can be seen, NAT obtained approximately 5.6% overall error rate, while KTH obtained approximately 8.7% and IVX synthetic speech obtained slightly above 12% overall error rate. The % error rate for IVX is lower than that obtained in an earlier testing (15% error rate) [4]. The error rate for Natural speech is of around the same magnitude as obtained before.

In Fig. 2, the % error rates are shown for NAT, KTH and IVX synthesis respectively, according to the different vowel contexts. Although the over-all % error rate for VCV NAT is very low (5,6%), it is located to a few consonants that alone yield very high error rates. Thus, consonants most vulnerable to generate high error rates for NAT are /tj/ (40% error rate), /sj/ (21% error rate) and /h/ (18% error rate).

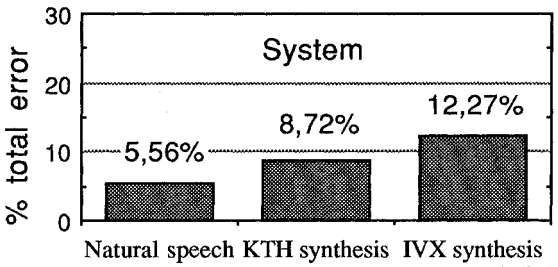

Fig. 1 % total error for NAT, KTH and IVX

The vowel contexts for NAT that are especially vulnerable for the consonants /tj/, /sj/ and /h/ to appear in yield still higher error rates. Error rates up to 54% are present. The NAT consonant clusters /tj/ and /sj/ and consonant /h/ that appear in the following vowel contexts yield especially high error rates: **otjo** (54%), **atja** (50%), **oho** (33%), **isji** (25%), **asja** (21%), **osjo** (17%), **ihi** (17%) and **itji** (17%).

The KTH error rates are highest for the consonants /ng/, /tj/ and /n/ (36%, 33% and 21%). The consonant clusters /ng/, /tj/ and the consonant /n/ that appear in the following VCV contexts yield especially high error rates: **ongo** and **ingi** (54%, respectively), **otjo, itji** and **atja** (between 29%-36%) and **ini** (42%).

The IVX error rates are highest for the consonants /ng/ (33%), /tj/ (34,7%) and /b/ (26.4%). The consonant clusters /ng/ and /tj/ and consonants /b, p v and m/ that appear in the following VCV contexts yield especially high error rates: **ingi** (62%), **ongo** (36%), **otjo** (42%), **atja** (36%), **ibi, opo** and **ovo** (42%, respectively), **ama** and **omo** (29%, respectively).

Since the condition NAT is supposed to form a yardstick, against which all synthesised results should be compared, the % error rates obtained for Natural speech is subtracted from the error rates obtained for each synthesiser system for each VCV word.

Fig. 3 shows the % error difference between synthesised speech (KTH and IVX) and NAT across all VCV combinations.

The % error difference can be either positive or negative. If negative (-%), it implies that the synthesised speech generates lower error rates than Natural speech (NAT). If positive (+%), it implies that synthesised speech generates higher error rates than the NAT condition.

The % error differences for the VCV words **ongo, ingi** and **ini** are much higher for KTH synthesized speech than for NAT (% error difference +50%, +50% and +44% higher, respectively).

However, for the consonant /h/ and the consonant cluster /tj/ appearing in the vowel context **o_o**, the % error differences are much lower for KTH synthetic speech than for the NAT condition! (-29% and -26% lower than NAT). Thus the intelligibility score for KTH synthesis is here higher than for NAT. How is this possible? The reason why the KTH synthesis obtains better score than NAT is not because the error rates of the VCV words oho and otjo are much lower than for NAT (otjo has 28% error rate for KTH syntesis), but because Natural speech itself generates such high error rates for these two words (see Fig. 2).

The following 4 VCV differences between KTH-NAT were significant (* at the 2%-level, two-tailed t-test); **+ongo***, **+ingi***, **+ini*** and **-oho***, thus indicating that a % error difference in the vicinity of 30% is necessary to obtain a significant result.

The following 9 VCV differences between IVX-NAT were significant; **+ovo***, **+ingi***, **+ongo***, **+omo***, **+ibi***, **+opo***, **+olo***, **+ama*** and **-oho***. For the VCV word oho, the same line of reasoning applies as above. The % error difference in favour of IVX-NAT is -34%. A % error difference in the vicinity of 30% is necessary to obtain a significant result.

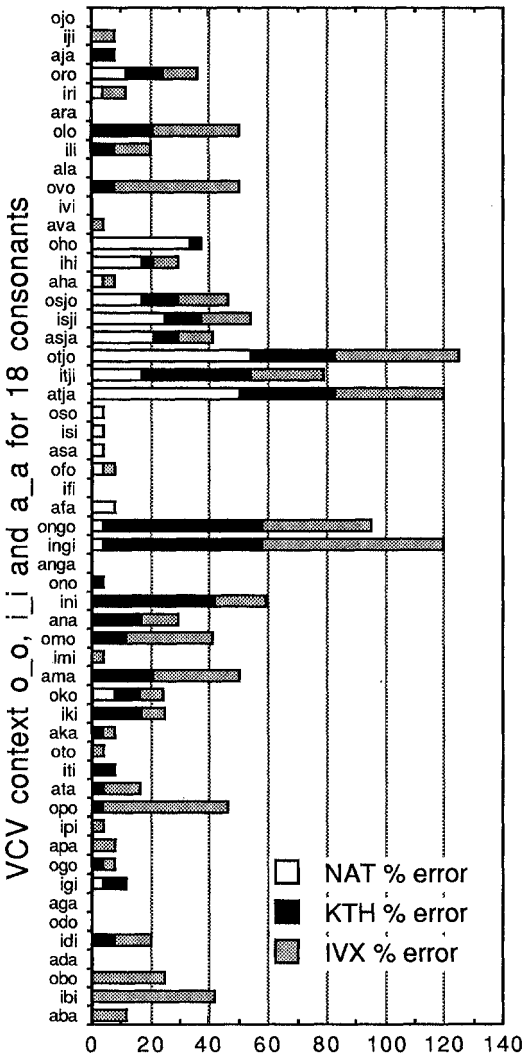

Fig. 2 % error rate for NAT, KTH and IVX for each VCV word

In Fig. 4, the % error difference between IVX-KTH are depicted. As can be seen, the % error differences are significantly higher for IVX than for KTH for the following 3 VCV words; **+ovo***, **+opo*** and **+ibi***, thus indicating a % error difference of the order of 30% to obtain a significant result.

IV. DISCUSSION

By using NAT as a baseline, several questions must be put forward: If % error rates in the vicinity of 50% are encountered for a certain VCV Natural speech context, how well can Natural speech serve as a baseline or yardstick against which evaluation of synthetic speech is made? What are the reasons for Natural speech to perform so poorly?

Even if the over-all error rate turns out to be very low for Natural speech, it is of interest to note that the high error rates only occur at certain instances. Usually, the highest error confusion rates for human Swedish (natural) speech are obtained for the fricatives /sj/ and /tj/, but no error analysis is usually made due to the low over-all error rate [4]. The NAT consonant cluster /tj/ was mistaken for /sj/, yielding an error rate of 36%, while the reverse confusion obtained 21% error. These confusion rates are higher than those obtained for the same VCV words generated by the two synthesisers. The result obtained in the Carlson et al experiment is looked upon as a "response problem" rather than as a "perceptual confusion" problem [4] since around 2/3 of all the errors encountered for human speech involved these consonants and 40% of the errors referred to confusions between the two. However, if the problem is of the response type, then similar error rates would be obtained for all systems regarding these VCV combinations? In fact, for the VCV words otjo and atja, both synthesisers seem to have difficulties. For the VCV word oho however, the two synthesisers yield very low error rates.

The problem encountered here is one of interpretability. If natural speech performs badly, is this a good yardstick to use for evaluating synthetic speech? If this is a typical outcome for the Swedish language as such, is difficult to forecast. More data gathered for various other languages must be examined in order to rule out language-specific causes.

However, the articulation of the natural speech baseline words must be undertaken a closer look. Is there not enough articulation when pronouncing the words in NAT condition? If this is the case, perhaps several speakers must be examined in order to find the

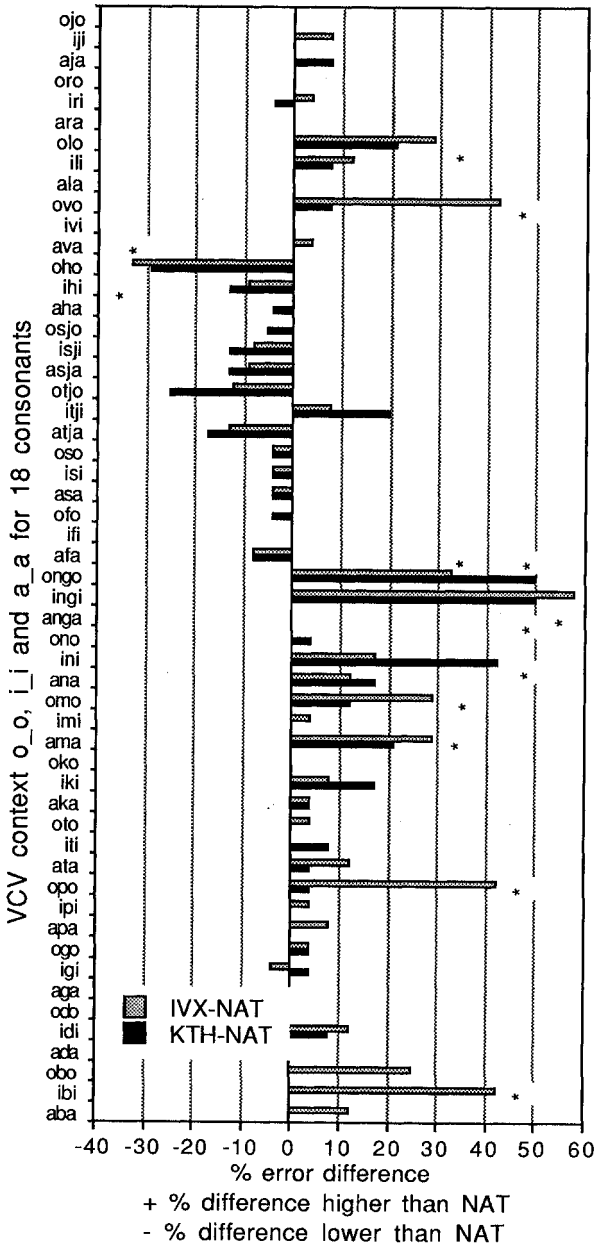

Fig. 3 % error difference compared to NAT

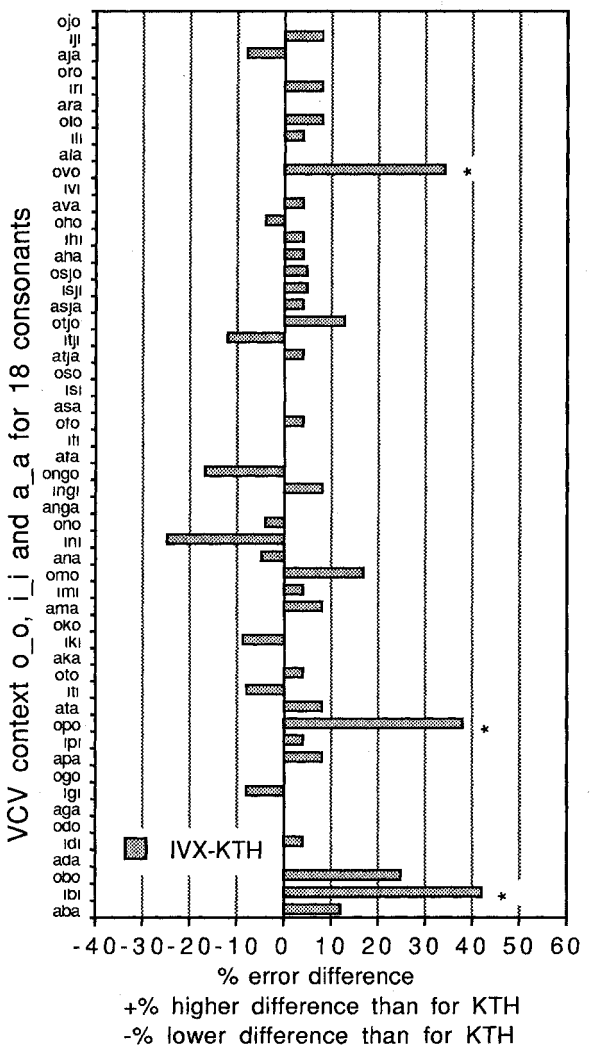

Fig. 4 % error difference IVX-KTH

speaker that generates the minimal error rate? This brings us back to the speaker variability problem. Perhaps the pronunciation of the VCV words in natural speech should be overarticulated, to secure low error rates? Or is it a fact that synthetic speech is in itself overarticulated in order to maximise intelligibility instead of naturalness? Only by establishing a true baseline can we evaluate the quality of synthesised speech.

Since the proper level of analysis is at the VCV level relative to some kind of base-line level (e.g. Natural speech), and not at the % over-all level, one should be vary careful when evaluating over-all % error rates. Usually high error rates occur for certain vowel contexts. To establish if a phoneme occuring in a certain context is evaluated significantly different for two systems is of interest. By counting all the insignificant differences (% error differences below 30%) into an over-all score, and then using this as a total unit of comparison between various systems, interpretations that lack validity may occur. As can be seen from the results, very few significant differences at the 2% level were encountered. For the KTH-NAT comparison, 4 significant VCV combinations were encountered. For the IVX-NAT comparison, 9 significant VCV combinations were encountered and for the KTH-IVX comparison, 3 significant VCV combinations were obtained. Only significant differences should be regarded.

Acknowledgements

This work was done within the framework of ESPRIT/SAM Project No. 2589. We wish to thank Assistant Professor Bertil Lyberg, Telia Research AB, for valuable hints. We also wish to thank Ritch Schulman, Swedish Telecom and Roger Lindell, Royal Institute of Technology for valuable technical support.

References

[1] P. Howard-Jones. "SOAP, Speech Output Assessment Package, Version 4.0". SAM UCL-042, 28 Feb 1992.

[2] R. Carlson and B. Granström. "A phonetically oriented programming language for rule description of speech". In: *Speech Communication*, (Ed.) by G. Fant. Stockholm: Almqvist & Wiksell 1975.

[3] R. Carlson, B. Granström and L. Nord. "Results from the SAM segmental test for synthetic and natural speech in Swedish (VCV, CV and VC)". ESPRIT PROJECT No. 2589 (SAM), internal report, Sweden, Stockholm, 1990.

[4] R. Carlson, B. Granström and L. Nord. "Segmental evaluation using the ESPRIT/SAM test procedures and mono-syllabic words". In: *Talking Machines: Theories, Models and Applications* (Eds.) by G. Bailly, G. C. Benoit. Elsevier-North-Holland Publishers 1991.

[5] *CCITT. TELEPHONE TRANSMISSION QUALITY.* Rec. P.56. Objective measurement of active speech level. IXth plenary session, Blue book, Melbourne, 14-25 November, 1988.

[6] G.A. Ferguson. *Statistical Analysis in Psychology and Education*, Chapter 12.3: Significance of the difference between two correlated proportions. New York: McGraw-Hill 1966.

[7] Q. McNemar. "Note on the sampling error of the differences between correlated proportions or percentages", *Psychometrica*, vol. 12, pp. 153-157, 1947.