VOWEL DYNAMICS IN A TEXT-TO-SPEECH SYSTEM - SOME CONSIDERATIONS.

Rolf Carlson and Lennart Nord*

Department of Speech Communication and Music Acoustics
KTH, Stockholm, Sweden

*Names in alphabetic order

ABSTRACT

The purpose of the present study is to increase the naturalness of the rule synthesis developed at the Department, especially positional variants of phonemes. We have in this study chosen the Swedish short vowel /e/, as this vowel is highly variable and can show a great degree of reduction.

A statistical analysis of the second formant value as a function of the phoneme context (CVC) gave as result that the variability expressed as standard deviation was reduced by 38% on the training material and 24% on the test material.

Analysis of the spectra of /e/ vowels showed that there was a tendency of acoustically different allophones to appear, that is, different spectral envelopes, due to context and position.

Keywords: analysis, synthesis, reduction, vowels

1. INTRODUCTION

In this study we focus on the dynamics of segments in reduced positions. From a phonetic point of view reduced segments display a variety of acoustic shapes that are partly predictable. Our aim is to be able to map vowel formant movements in realisations differing in stress, duration, consonantal frame etc. These are some of the reasons for the mechanical speech quality and the impression of over-articulated pronunciation of synthetic speech. Earlier work on Swedish vowel reduction as well as material extracted from our Swedish speech data bank indicate the importance of exploring the predictability of acoustic variability.

The variability for a single formant, depending on context and stress level might approach as much as a few hundred Hertz. This is not surprising regarding the flexibility of the signal, the "speech code", is difficult to describe and to formulate the rules governing this change is part of the challenge. In an earlier study, Nord 1986, it was found that for the Swedish short /e/ F2 varied between 1470 and 1660 Hz in a test where stress and position were varied.

Regarding the analysis of consonants we have earlier published some results on Swedish sonorants, especially regarding the interaction between a rule description and an expanded inventory of allophones (Carlson, Nord, 1991). The main goal with the present project is to get a better understanding of the acoustics of human speech and to use this knowledge in improving the quality of synthetic speech (Carlson et al. 1991). In this paper we report on the analysis of the Swedish short vowel /e/, as this vowel is highly variable and reduced in many instances.

Studies on the dynamics of vowels have dealt with a number of languages, such as English (Stevens and House, 1963), Dutch (Koopmans-van Beinum, 1980) and Swedish (Lindblom, 1963; Nord, 1975, 1986). Results have shown formant deviations due to stress, position, duration etc.

There are evidently language differences. In English, the /e/ vowel is described as being changed by a laxing rule that will turn the /e/ into a schwa in certain positions. There is no such rule in Swedish, a language that does not change the vowel colour to the same degree as the English language does.

2. CORPUS

We are currently developing a large single-speaker database of read speech. The corpus is taken from Swedish text corpora such as novels and newspaper material. The goal is to have about 2000 sentences phonemically (and phonetically) labelled. Two parts of this material are used in this study, see Table 1. Corpus 1 consists of eleven short text passages. This corpus has already been used in an evaluation study on speech synthesis (Carlson et al., 1992; Neovius and Raghavendra, 1993). Corpus 2 consists of corpus 1 plus 223 unrelated sentences from newspaper text. The recording was made in an unechoic chamber using a DAT recorder and the data was subjected to standard processing.

Table 1. Presentation of corpus 1 and 2.

Corpus	Number of Sentences	Mean Sentence Duration (s)	Number of /e/ vowels
1	190	4.05 (sd 2.48)	386
2	413	4.33 (sd 2.52)	813

3. ANALYSIS

We have performed several different kinds of acoustic analysis of the corpus. The speech data was segmented and labelled, partly using an automatic method (Blomberg and Carlson, 1993). A software by Carlson and Glass based on "analysis-by-synthesis" (Carlson and Glass, 1992) was used to generate spectral envelopes and also to extract formant frequencies for the different realisations of /e/. A spectral representation was created by analysing the middle third of the vowel with FFT-analysis. In the case of short vowels at least 20 ms was used for the analysis.

A gradient-descent procedure was used to minimise the error between the input and a synthetic spectral representation of the vowel. The error metric consisted of a weighted Euclidean distance emphasising spectral peaks. A fixed step-size was used during the search which effectively quantized the parameter space. The resulting spectra could then be synthesized simply by adding a perturbation vector, rather than by being completely regenerated from the parameters themselves. The analysis-by-synthesis procedure also made it possible to estimate the source amplitude of the vowels.

The first two formants were manually corrected after the automatic analysis in corpus 1, while the remaining vowels in corpus 2 only have automatically derived values. The first formant in nasalized vowels was associated to the higher peak in frequency if two peaks could be observed. Most of the following discussion is based on corpus 1. The lower peak is a result of the nasal coupling that introduces a low pole-zero pair.

4. SENTENCE POSITION EFFECTS

In Figure 1, the amplitude level of the vowel is plotted as a function of its position. There is a trend of a level decrease as the vowel is found later in the sentence, probably due to the general prosodic pattern of the sentences. The final reduction of amplitude is well known but it is interesting to note the speaker´s habit to have a higher effort in the beginning of the utterance.

It was also noted in the material, although not shown here, that the relative position is correlated with segmental duration. Early in the sentence vowels can rarely be of long duration, while later in a sentence short and long durations occur.

Figure 2 displays another kind of reduction. The F2-value of the vowel preceding /n/ is displayed as a function of utterance position. A regression line is added to the figure, and has a slope of about 150 Hz over the whole sentence. The regression coefficient is about .30 indicating a weak but possible correlation. Most of the /en/ sequences constitute unstressed endings in Swedish which might make them more sensitive to reduction compared to vowels in stressed position. An analysis of all vowels in all contexts showed a correlation of .24 with about the same slope. This means that the F2 value drops 150 Hz over an utterance. However, more research is needed to support such a claim.

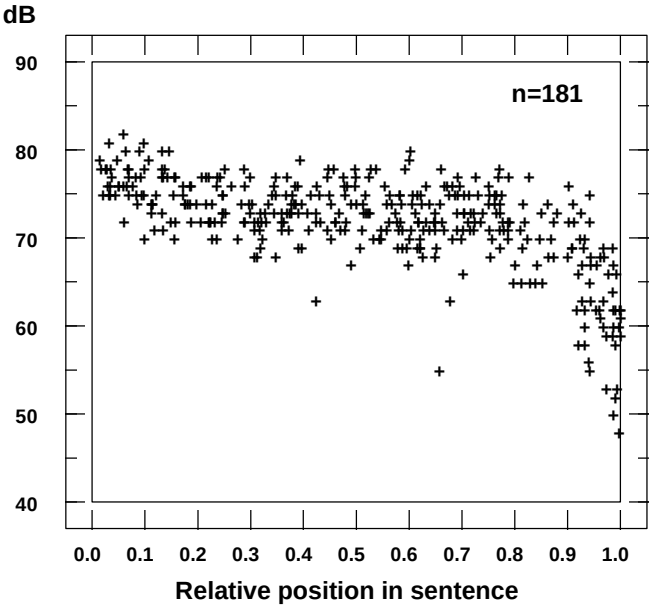

Figure 1. Amplitude level of the vowel /e/ as a function of its relative position in the sentence.

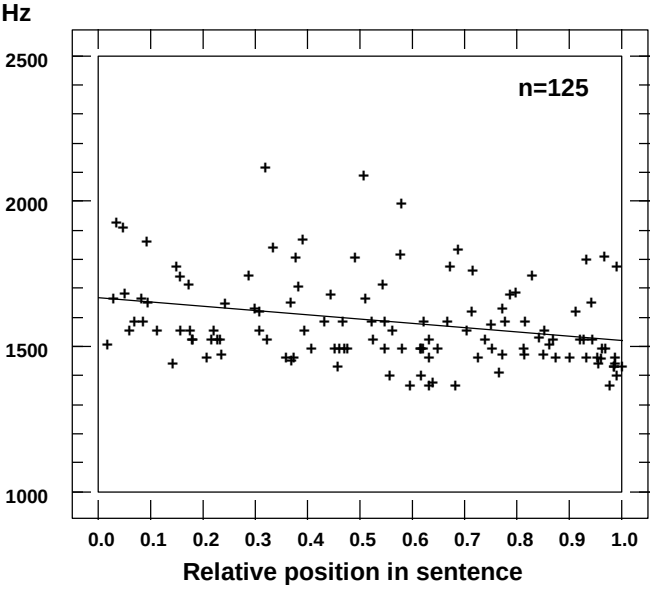

Figure 2. Second formant frequency of the vowel /e/ before /n/ as a function of its relative position in the sentence.

5. CONTEXT EFFECTS

Several aspects of contextual influence on the /e/ vowel were studied. Most of the analysis dealt with the F2-value as a function of segmental context. In a study by Slater and Hawkins (1992) locus equations for velar stops were discussed. Despite the good correlation in a specific context the equation changed depending on many other factors. A rule based text-to-speech system has to contain rules that covers all possible contexts and still has a restricted number of rules. Figure 3 gives an example of how the second formant in the sequence /re/ is dependent on vowel duration and final

context. The labels in the figure indicate the following segment. It can be seen how the utterance final position (marked by "." in Figure 3) extends the duration, how the nasal pulls down the second formant and how a velar context pushes the second formant up.

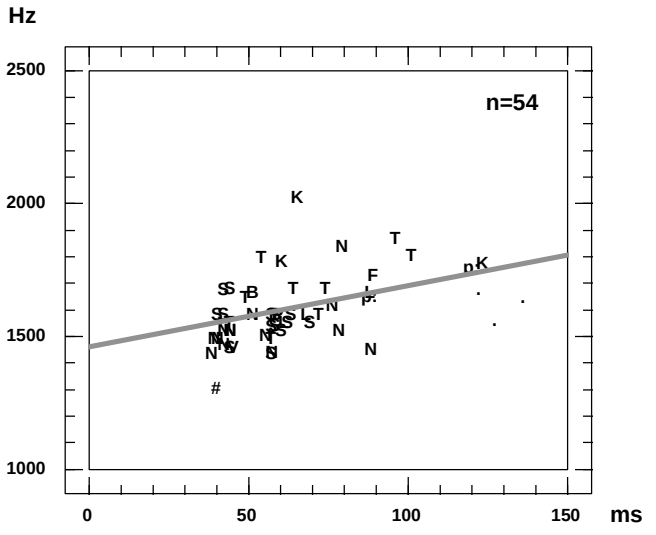

Figure 3. F2 value of the vowel /e/ following /r/, as a function of its duration.

To study this behaviour in a more general way we applied a technique earlier explored in the context of speech recognition, Phillips et al. 1991. We initially measured the mean and standard deviation for the vowel /e/ in all contexts and position. The task was then set to reduce the standard deviation as much as possible by applying automatically derived rules. These rules can then be thought of as starting points for rules in a text-to-speech system. The following assumptions were made:

The F2-value is:

dependent on the position in the sentence.
dependent on its duration.
dependent on the preceding segment
dependent on the following segment.

All these dependencies can be described as:

F2 = a+b*d

where **a** and **b** are unknown

d is the vowel position or its duration.

The context and sentence dependent **a** and **b** were calculated in an iterative process. For corpus 1 the standard deviation was with this method reduced from 167 Hz to 103 Hz, a reduction of 38%. In the next experiment the same procedure described above was repeated but each of the eleven passage was used as a test corpus while the other ten were used as training corpus. The standard deviation was reduced 24.3% taken as a mean of the test material.

6. DISCUSSION

6.1 Spectral analysis - allophones

The spectral distribution of the /e/ vowels is displayed in the form of spectral envelope histograms, Figure 4. The spectral energy distribution is thus given in 10% steps for each frequency and the corresponding points are connected forming spectral envelopes. The resulting curves do not necessarily correspond to observed spectra, but give a statistical view of the distribution. The reduction of vowel colour or the willingness to coarticulate with context can clearly be seen.

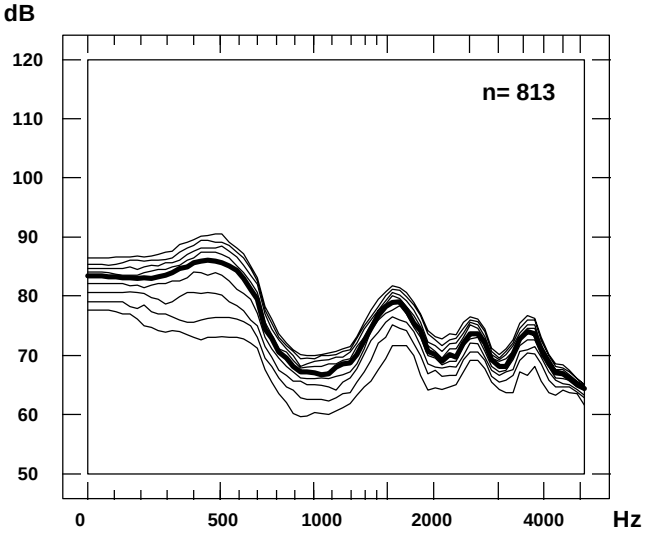

Figure 4. Spectral envelope histograms stressed and unstressed /e/. in all contexts. A line is drawn for each frequency at every 10% of the distribution. The thick line corresponds to 50%.

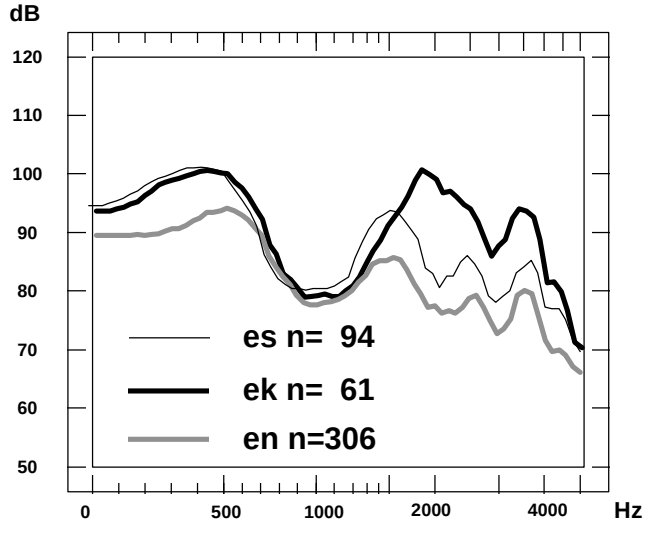

Figure 5. Spectral envelopes (50% line) for the vowel /e/ in different contexts.

Should the variability be accounted for by rules or by a set of allophones? We have decided to use a number of extra

allophones and also to adjust time and frequency values by rules.

Some velar consonants will raise the F2 of the /e/, such as /k,g,j/ and some consonants lower F2 toward a more neutral value, /m, n, *l*/. This would call for different acoustic allophones in addition to a rule description. Compare the spectral envelopes in Figure 5.

Table 2. Formant values for a selected number of /e/ allophones.

context	F1	sd	F2	sd	F3	sd
(all)	529	93	1647	167	2501	210
es	486	50	1612	131	2523	223
ek	492	59	1893	117	2391	157
en	565	57	1585	150	2512	227

6.2 Articulation

A point that has not been discussed here is an explanation in terms of articulation. Especially features like lip rounding, jaw opening and velar coupling, to name a few, sometimes have far-reaching consequences for a string of phonemes. For example, if a rounded vowel is stressed it may cast a "rounding" shadow over a number of preceding and following phonemes, and vowels as well as consonants may get a lowering of the formants for several hundred milliseconds.

7. CONCLUSION

Acoustically different allophones of the short Swedish /e/ vowel have been derived from a speech data bank with semi-automatic labelling and analysis. We have found acoustic forms of the /e/ vowel based on our work with the data bank at the Department. An important part of the that serves as the bases for the synthesis development.

ACKNOWLEDGEMENT

This work has been supported by grants from The Swedish National Language Technology Program.

REFERENCES

Blomberg, M. and Carlson, R. (1993): "Labelling of speech given its text representation." This conference.

Carlson, R. and Glass, J. (1992): "Vowel classification based on analysis-by-synthesis," pp. 575-578 in (J.J. Ohala, T.M. Nearey, B.L. Derwing, M.M. Hodge, & G.E. Wiebe, eds.) ICSLP 92 Proceedings, Vol. 1, University of Alberta, Canada.

Carlson, R., Granström, B. and Hunnicutt, S. (1991): "Multilingual text-to-speech development and applications," A. W. Ainsworth (Ed.), Advances in speech, hearing and language processing, JAI Press, London, UK.

Carlson, R., Granström, B., Neovius, L., and Nord, L. (1992): "The "listening speed" paradigm for synthesis evaluation", Chalmers Technical Report No 10, Department of Information Theory, Chalmers University of Technology, pp. 63-66.

Carlson, R. and Nord, L. (1991): "Positional variants of some Swedish sonorants in an analysis-synthesis scheme", Journal of Phonetics, Vol. 19, pp. 49-60.

Koopmans-van Beinum, F.J. (1980): "Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions", Doct. thesis, University of Amsterdam.

Lindblom, B. (1963): "Spectrographic study of vowel reduction", J.Acoust Soc.Am. 35, pp. 1773-1781.

Nord, L. (1975): "Vowel reduction - centralization or contextual assimilation?", pp. 149-154 in (G. Fant, ed.): Speech Communication, Vol. 2, Almqvist & Wiksell Int., Stockholm.

Nord, L. (1986): "Acoustic Studies of Vowel Reduction in Swedish", *Speech Transmission Laboratory Quart. Progress & Status Report, Dept of Speech Comm & Music Acoustics,. STL-QPSR* 4/1986, pp. 19-36, (KTH) Stockholm.

Philips, M., Glass, J. and Zue, V. (1991): "Automatic learning of lexical representations for sub-word unit based speech recognition systems," Proc. European Conference on Speech Communication and Technology.

Slater, A. and Hawkins, S. (1992): "Effects of stress and vowel context on velar stops in British English," Proc. ICSLP92, Banff, Canada, pp. 57-60.

Stevens, K.N. and House, A.S. (1963): "Perturbations of vowel articulations by consonantal context: an acoustical study", J. Speech Hearing Res. 6, pp. 111-128.