AUTOMATIC DERIVATION OF PHONETIC RULES BY AN ITERATED NORMALISATION PROCEDURE

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

ABSTRACT

In this paper an iterative normalisation procedure to automatically derive phonetic rules from a labelled speech corpus is described. It is assumed that the acoustic influence of coarticulatory constraints can be superimposed to model natural spectral variation. The algorithm proves to be promising when used to analyse the effect of phonetic context, stress and duration of Swedish front vowels on F1 and F2.

INTRODUCTION

Phonetic spectral variability in the realisation of a phoneme is due to numerous factors such as context, stress, speaker, speaking style, etc. Analyses in studies of coarticulation have traditionally dealt with syllables or words in a strictly controlled context. However the interaction between different factors in connected speech which influence segmental quality is very complex. In an attempt to describe essential coarticulatory phenomena we propose a data-driven method applied to a labelled speech corpus. The description is given in terms of a set of allophones or phonetic rules adjusting the spectral parameters of the phone to be modelled [1][2].

In this paper we describe a step by step normalisation procedure to automatically derive phonetic rules. The rules are easily interpreted and can be applied directly in the KTH text-to-speech system [3]. Thus, we combine the strengths of data driven and knowledge based techniques. The aim is to both produce more natural-sounding synthetic speech and also to gain deeper knowledge about speech production and perception.

In the current experiment we address the problem of modelling how F1 and F2 of Swedish front vowels are influenced by phonemic context, lexical stress and position. The variations along the speaker and style dimensions have been reduced by analysing read speech of one speaker.

METHOD

The speech material

The speech material consists of 11 short stories read by one male speaker. Formant frequencies of 2944 Swedish front vowels were manually measured. See Table 1 for the vowel distribution. This material has been used in several other investigations, e.g. [1][4]. Carlson & Nord [2] have also used the corpus to study context dependencies for the short vowel /e/.

Table 1. Number of analysed phonemes.

a	872	ø:	25
ʉ	83	ø	33
i:	184	e:	164
ɪ	355	e	647
y:	28	æ:	157
ʏ	49	æ	54
œ:	63	ɛ:	28
œ	37	ɛ	165

Derivation of rules

A data sample in the analysis consists of a prediction vector, *X*, and a response vector *Y*. The aim is to correctly predict *Y*=[F1, F2], of a front vowel given *X* which contains information about the vowel's duration, lexical stress and phonemic context.

The phonemic context is defined by the identity of the target phoneme itself, the three preceding and the three following phones. Each sample also includes information about whether it is word final or word initial.

The algorithm is based on the assumption that the acoustic realisation of a phoneme can be modelled by superimposing the influence of the most important predictor variables.

A superpositional model has proven to be reasonably reliable despite statistically significant predictor variable interaction [5]. Context normalisation techniques have also been applied with success in automatic speech recognition, e.g. [6].

The samples are subjected to binary questions to find the group of samples that minimises the acoustic spread of the entire data set when those samples have been normalised and replaced. The amount of spread is evaluated by means of the function S,

$$S(y) = \sum_{i=1}^{2} \frac{1}{N}\sum_{j=1}^{N}\left(y_{i,j} - \bar{y}_i\right)^2$$

where $y_{i,j}$ and $\bar{y}_i$ are sample number *j* and the mean of the *i*:th formant frequency respectively. *N* is the total number of samples. All frequencies are calculated on the technical mel-scale. Hence, S is basically the sum of the formant frequency variances in mel.

A significant advantage of the replacement procedure is that all data are available for analysis in every iteration.

A categorical variable is a variable taking on unordered values, A question on such a variable can be of the type *"Is the phone immediately following the target a nasal?"* That is, phonetically meaningful features are used to form questions as well as single phoneme identities. Ideally, all phoneme combinations should be used to form questions. However, this task becomes unfeasible as the number of combinations, *n*, is given by $n=2^m$, where *m* is the number of phonemes. Currently 42 features are used apart from the single phoneme identities. A typical question on a continuous variable is *"Is the duration of the target phoneme < 100 ms?"* Questions are made on all unique target phoneme durations occurring in the speech corpus.

Samples responding positively to the question are normalised on the mel-scale towards the grand mean. The normalisation term that is added to the sample is the difference between the mean frequency of all samples and the mean frequency of the selected samples. One normalisation term is used for each formant frequency. The question which minimises the variance of the entire data set, in combination with the corresponding normalisation terms, is chosen to specify a rule.

A cross validation procedure is used to determine how many rules can be used without loss of predictive power for unseen data. Thus, for V-fold cross validation, (N/V)*(V-1) parts of the material is used for training and the remaining part is used for testing. The material is permuted so that each sample is used both for training and testing. The test score is calculated using the function S. The value of S, when applied in testing, is expected to decrease with increasing training until a critical point where the effect of overtraining will become noticeable and the variance will increase again. In the last step all data are used to generate rules. The cross validation result indicates the maximum appropriate number of rules that can be used without loss of generality.

In the experiments described below five-fold cross validation was used and no rule applying to less than ten samples was accepted. Moreover, all standard deviations are calculated on the mel-scale.

RESULTS

The overlap is considerable in the F1-F2 vowel space. In the first experiment, we employ the phonetic label of the target phoneme as a feature. Thus the predictive power of the phonetic labels can be compared to that of other features. All front vowels were analysed simultaneously and normalised towards a single front vowel prototype. F1 and F2 of this vowel, the grand mean, equal 514 and 1599 Hz respectively. The algorithm was iterated to generate 200 rules.

The normalised value of S as a function of the number of rules is plotted in Figure 1. The solid line indicates the mean cross validation score and the dashed line represents the result of the training on the entire data set.

The cross validation score indicates that no further improvement will be gained using more than about 50 rules. At this point 50% of the standard deviation for F1 and 52% for F2 is explained. As expected, the most significant rules concern the target phoneme itself in terms of features. In fact, one third of the first fifty rules are of this type. The second most important factor influencing F1 and F2 is velar context. Rules number five and six concern front vowels in the immediate context of velar phones.

The distribution of rules based on the right and left context is quite symmetric. The exception is the far context, three phones away, in which the right context seems to be somewhat more important than the left context.

Only one rule among the first fifty, concerns duration or stress. This is quite natural since these aspects influence the target samples differently depending on their phonemic identity. Therefore, the question set was expanded to include composite questions such as *'Is the target /a/ AND stressed?'* Apart from simple phoneme identities, 16 additional features were used for the targets implying a dramatic increase in computational load.

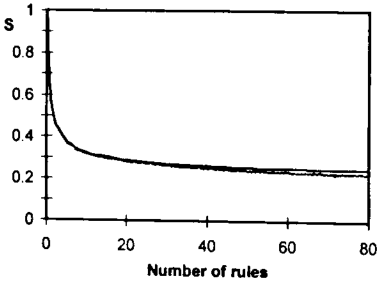

Figure 1. The value of the spread, S, as a function of the number of rules. The solid line indicates the mean cross validation score. The dotted (lower) line indicates the score from the training of the entire material.

The standard deviation of F2 is reduced from 162 to 72 mel using the first 50 composite rules. This explains more than 55% of the standard deviation. The standard deviation of F1 is practically unchanged. Thus, the introduction of composite questions yield only a slight improvement. However, the improvement seems to become more important when more rules are used. The cross validation score is one percent lower in the composite case when 50 rules are used.

A separate set of rules was generated for /a/, the most frequently occurring front vowel, to illustrate the power of the method more clearly and to extend the analysis to some extent without increasing the computational complexity.

The cross validation score indicates that some thirty rules suffice to model the major contextual influence on /a/. Degeneration occurs when more than fifty rules are used. The most significant factor is, again, velar context followed by lexical stress. The following two rules describe the coarticulatory influence of bilabials and nasals in the close vicinity of the target vowel. Vowel features are also important: rule number five and six consider /a/ coarticulated with other front and low vowels. The first 30 rules explain 27% and 39% of the standard deviation for F1 and F2 of /a/ respectively. This corresponds to a decrease from 48 to 35 mel in F1 and from 107 to 65 mel in F2. More rules are based on the left context than on the right among the top thirty rules. This means that the left context has stronger predictive power than the right. Moreover, the stronger explanatory power of the left context mainly concerns F2. It is unclear whether this has any implications for reasoning about carry-over vs. anticipatory coarticulation before the phonetic distribution of the context is analysed more thoroughly.

Since only one speaker, reading text passages, is analysed we expect the articulatory effort and speaking style to be about the same throughout the speech material. It is reasonable to believe that the duration of the target phoneme will have a systematic effect on the formant frequency values [7]. Therefore, when the best rule has been found a duration-dependent normalisation adjustment is introduced to refine the analysis. The formant frequency displacement is assumed to increase linearly with a logarithmic decrease in segment duration.

Figure 2 shows an example of the relation between the second formant frequency of /a/ tokens following immediately after a velar segment and the logarithmic value of the duration.

There was a decrease in the standard deviation of both F1 and F2 on the training of the entire data set. The mean cross validation score indicates a small improvement compared to the normalisation independent of segment duration.

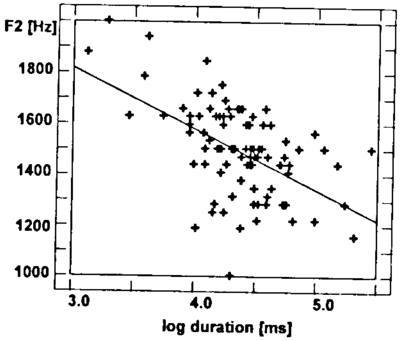

Figure 2. F2 of /a/ following a velar segment plotted vs. the logarithmic value of the duration. r = -0.54.

DISCUSSION

In this paper we have proposed a method for automatic derivation of phonetic rules from a labelled speech corpus. In the first experiment the phonetic labels assigned to the target vowel proved to be powerful formant frequency predictors as expected. However, all front vowel identities were not more important than the context. This might change if F3 is taken into account as well.

The cross validation score plotted in Figure 1 implies that the method is robust. The curve does not turn upwards, indicating over-fitting to test data, not even after 200 iterations. The cross validation of the /a/ rules displays a slight deterioration if more than 60 rules are used. The robustness is probably due to the fact that the overall decrease in variance depends on more than the magnitude of a coarticulatory effect. Just as important is the number of samples influenced. The rules were used to predict formant frequencies given the predictor vectors of the training material for /a/. The result showed that there is a systematic tendency of underestimating high formant frequencies and overestimating low frequencies. That is, the formant frequency displacement seems to be underestimated on an average. One plausible explanation for this is that the normalisation terms are based on differences in mean values that are biased by other coarticulatory effects. We conclude that the algorithm proposed in this paper is robust and provides easily interpretable results that potentially can be used to enhance the quality of synthetic speech.

ACKNOWLEDGEMENTS

I would like to thank Rolf Carlson for initiating this work and for great moral and practical support. This work has been supported by grants from The Swedish National Language Technology Programme.

REFERENCES

[1] Högberg, J (1994), A phonetic investigation using binary regression trees. In: Papers from the Eighth Swedish Phonetics Conference, Lund, Sweden.

[2] Carlson R & Nord L (1993), Vowel dynamics in a text to speech system - some considerations. In: Proc. of Eurospeech 93, 1911-1914.

[3] Carlson R, Granström B & Hunnicutt S (1991). Multilingual text-to-speech development and applications. In: (Ainsworth AW, ed.), *Advances in speech, hearing and language processing*, London: JAI Press, UK.

[4] Neovius L & Raghavendra P (1993), Comprehension of KTH text-to-speech with 'listening speed' paradigm. In: Proc. of Eurospeech 93, 1687-1690.

[5] Broad DJ & Fertig RH (1970) Formant-frequency trajectories in selected CVC-syllable nuclei. In: J.Acoust SocAm, Vol 46, 1572-1582.

[6] Philips M, Glass J & Zue V (1991) Automatic learning of lexical representaions for sub-word unit based speech recognition systems. In: Proc European Conf. on Speech Comm. and Technology.

[7] Lindblom B (1963) Spectrographic study of vowel reduction. In: J.Acoust SocAm Vol 35, 1773-1781.