A Phonetic Investigation Using Binary Regression Trees

Jesper Högberg
Dept of Speech Communication and Music Acoustics, KTH,
Box 70014, S-100 44 Stockholm, Sweden.

ABSTRACT

This paper is concerned with the analysis of Swedish front vowels in different contexts. The analysis is made using binary regression trees. The values of the first two formant frequencies are analysed as a function of context and duration. The trees provide both quantitatively and qualitatively interpretable representations.

INTRODUCTION

The realisation of a phoneme display a considerable amount of variation even if we consider only one speaker. A large part of this variation is due to the fact that neighbouring phonemes coarticulate with the target phoneme. However, to study these variations by hand is a tedious work. Classification and regression trees have proven to provide good support for the researcher analysing large amounts of data. The method is described in full detail in *"Classification And Regression Trees"* (Breiman, Friedman, Olshen and Stone 1984) and has recently gained attention within the speech community. It has been used in several studies aiming at refining the context dependent variants of phonemes both in time and frequency, that is allophones. The results have been applied to speech recognition, speech synthesis, etc., e.g. (Bahl, Souza, Gopalakrishnan, Nahamoo and Picheney 1991, Downey, Russel, Nowell, Bijl, Galloway and Ponting 1993, Wang, Campbell, Iwahashi and Sagisaka 1993).

This paper is concerned with the analysis of the first and second formant frequency of Swedish front vowels using binary regression trees.

METHOD

Speech data and analysis

The material consists of 11 short stories read aloud by a single male speaker. The corpus has previously been used in an investigation of speech intelligibility (Neovius and Raghavendra 1993). The recording was made in an unechoic room. The corpus has also been used in a study of the short vowel /e/ (Carlson and Nord 1993). The formant frequencies were manually measured for the 2947 Swedish front vowels analysed in this study. Their distribution is shown in Table 1.

Table 1. *The table shows the number of phonemes analysed.*

a	ʉ	ɪ:	ɪ	ʏ:	ʏ	œ:	œ	ø:	ø	e:	e	æ	æ:	ɛ:	ɛ
868	83	184	356	28	49	64	37	25	32	164	648	161	56	28	164

Construction of the regression tree

A data sample in the regression analysis consists of a prediction vector $\boldsymbol{X}$ and a response variable vector $\boldsymbol{Y}$. The components of $\boldsymbol{X}$, x_i can be either continuous or categorical variables, the latter assuming unordered values ,c_i , representing categories in C, where C = $\{c_1, c_2, ..., c_N\}$.

In this study one continuous and seven categorical predictor variables are considered. $\boldsymbol{X} = [p_{-3}, p_{-2}, p_{-1}, p_0, p_{+1}, p_{+2}, p_{+3}, dur_0]$ where p_i is the identity of the phoneme in position i relative to the target phoneme. Thus, p_0 and dur_0 are the identity and the duration of the target phoneme itself. The response vector contains the first and second formant frequency of the target phoneme, $\boldsymbol{Y} = [F1_0, F2_0]$.

The starting point in a regression tree is the entire data set which constitutes the root node. The goal is to split the root node into two subnodes maximising a purity function I for some node t.

$$I(t) = R(t) - R(t_r) - R(t_l)$$

Where t_r and t_l are the right and left subnodes of t.

$$R(t) = \sum_{i=1}^{2} \frac{1}{\bar{y}_i} \frac{N_t}{N_{tot}} \frac{1}{N_t} \sum_{j=1}^{j=N_t} \left(y_{i,j} - \bar{y}_i \right)^2$$

Where N_{tot} is the total number of samples in the data set and N_t is the number of samples in node t. Thus, R(t) is a weighted measure of the response variance in node t.

The split is a binary question of the type; "Is $x \leq x_0$?" where x_0 is a real number if x is a continuous variable or "Is x in S?" if x is categorical. S can be any subset of C. Ideally, the best split is found by exhaustively evaluating all possible questions on $\boldsymbol{X}$. However, the number of subsets in C = $\{c_1, c_2, ..., c_N\}$ is 2^N. With N=63 the computation cost is far too high. Therefore a question set was specified containing, apart from phoneme identities, phonetically interesting features such as place and manner of articulation. The vowel feature system was designed according to Elert (1989), dividing the vowel height into three regions, high, midhigh and low. The exception is the vowel /ʉ/ which is also included in the material.

The tree is grown recursively, splitting the current terminal nodes into new nodes until some stop criterion is met.

The size of the tree is a crucial issue. A very large tree with few training samples in each terminal node is likely to reflect the detailed properties of the training material and will have poor predictive power for unseen data. One method to cope with this problem is proposed in (Gelfand, Ravishankar and Delp 1991). The data set is divided into two subsets L1 and L2. A large tree is grown on L1. The samples of L2 are used to evaluate the tree and all bad branches are cut off. Now L2 is used to grow the tree from the new terminal nodes and the obtained tree is evaluated on L1 and pruned again. Thus L1 and L2 are alternated as test and training materials. The algorithm guarantees convergence and the iteration stops when two subsequent optimally pruned trees contain the same number of nodes.

The current implementation deviates from the method proposed in (Gelfand *et al.* 1991) in one aspect. In this investigation we assign a node terminal status if the best split provide any of the two subnodes with less than N_{min} samples.

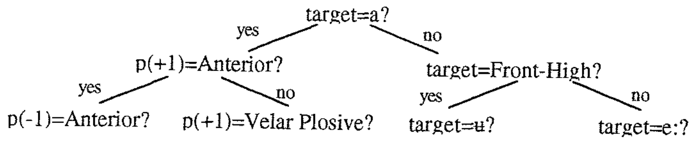

Figure 1. *The top of the tree. N_{min} =10.*

RESULTS

Figure 1 shows the percentage of the standard deviation explained in trees generated with the parameter N_{min} set to different values. It appears that the lower N_{min} is the better the prediction. However, the curves in Figure 1 also reflect the complexity of the trees. N_{min} = 2, 4 and 10 result in trees with 203, 99 and 49 terminal nodes respectively. The very large trees give very good predictions for most samples but also contain some terminal nodes with large standard deviations. Therefore we can expect a large tree to behave well for most data but occasionally predict some samples badly. The coherence between the curves in Figure 1 indicates that F1 and F2 have been properly weighted in the objective function R(t).

Let us consider the tree generated with N_{min} set to 10 in more detail. The terminal nodes of that tree comprise 60 samples on an average. Figure 2 shows the top of the tree. The most significant split on the initial data isolates the vowel /a/ at once. Generally the nature of the target phoneme itself is of great importance as anticipated. Following the rightmost branch the next split is "Is target = Front-Rounded?". The tree contains one branch of mixed clusters for /i/, /ɪ/, /y/ and /ʏ/ and one branch for /ɸ/, /ɸ:/, /œ:/ and /œ/. The rightmost part of the tree terminates in nodes comprising samples of /e/, /ɛ/, /ɛ:/, /æ:/ and /æ/.

Splits made on nasal and velar context are important. Nasal context pushes F1 up for front-high and mid-high vowels. Conversely velar context brings F1 down and forces F2 up. These results agree well with the findings of Carlson and Nord (1993). Figure 3 shows the number of the splits made on the predictor variables. As can be seen there is a skew to the right. The distribution is not the same for the whole assembly of trees but the same tendency is found when N_{min} is in the interval 8-20. The tree of N_{min} = 4 contain relatively more splits on the far context. A plausible explanation is that as N_{min} assumes smaller values more specific splits are allowed that capture variations due to more uncommon contexts. The duration of the target phoneme is a powerful predictor for trees of all sizes. These splits are readily interpreted in terms of reduction.

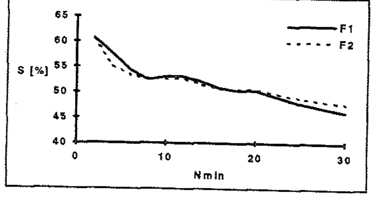

Figure 2. *The percentage explained, S, in the standard deviation as a function of N_{min} .*

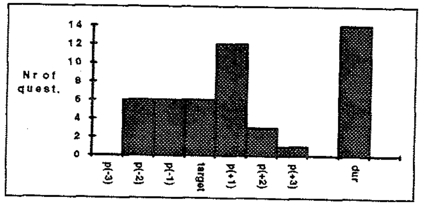

Figure 3. *The number of splits made on the predictors.*

The subnode containing the shorter segments predicts formant values approaching those of an ideal neutral vowel of a typical male speaker. That is F1 approaches 500 Hz and F2 1500 Hz. The sole exception to this rule is the velar context in which a reduced segment will have a high F2 and a somewhat lower F1.

DISCUSSION AND CONCLUSIONS

The problem of determining the right-sized tree remains. However, it is reasonable to assume that a tree holding about 30 samples per terminal node will make a reliable predictor. This figure corresponds to the tree grown with N_{min} = 4, explaining 53.2 % of the standard deviation in F1 and 52.7% in F2. These rates could probably be improved by dealing with some of the problems inherent in the current implementation. Firstly, the stop criterion is an important matter that should be overlooked. Secondly the question set should be extended to account for frequent contexts modelled not only by simple questions but also by more complex ones.

Word boundaries have been given phoneme status in this investigation which has probably had a non-negligible impact on the results. Further on, the balance in the two data sets matters a great deal, interchanging L1 and L2 results in different trees.

We conclude that regression trees can provide a powerful tool when investigating a large corpus of speech data. They provide both a qualitatively and a quantitatively meaningful description of coarticulated phonemes. This investigation will be the starting point of a more extensive study, addressing the problems mentioned above.

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

Breiman, L., Friedman, J.H, Olshen, R.A., Stone, C. J. (1984): "*Classification and regression trees.*", Belmont, CA: Wadsworth.

Bahl, L.R, Souza, P.V., Gopalakrishnan P.S., Nahamoo D. Picheney, M.A. (1991): "Context dependent modeling of phones in continuous speech recognition using decision trees.", Proc. of DARPA Speech and natural language workshop, 264-269.

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

Downey, S., Russel, M., Nowell, P., Bijl, D., Galloway, K., Ponting, K. (1993): "Experiments in vocabulary independent speech recognition using phoneme decision trees.", Proc. of Eurospeech 93, 1575-1578.

Elert C-C. (1989): "*Allmän och svensk fonetik.*", Norstedts Förlag, Stockholm, 32.

Gelfand, S. B., Ravishankar, C.S, Delp E.J. (1991): "An iterative growing and pruning algorithm for classification tree design.", IEEE Trans. on PAMI, Vol 13, No 2, 163-174.

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

Wang, W. J., Campbell, W.N, Iwahashi, N., Sagisaka, Y. (1993): "Tree-based unit selection for English speech synthesis.", Proc. of ICASSP Vol. 2, 191-194.