BROAD PHONETIC CLASS BASED SPEAKER MODELLING

A. Satriawan and J.B. Millar

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
The Australian National University

ABSTRACT - This paper discusses the results of speaker modelling experiments based on broad phonetic classes of a speaker using different hidden Markov model structure and testing on different durational classes of broad phonetic class test segments. The results show that for fricatives and plosives the full covariance normal models are much better than other models, independent of segment duration, but especially for medium and long segments. For nasals, vowels, and approximants, left-to-right models are preferable. However, for short segments, left-to-right with skip models are the best for all broad phonetic classes.

INTRODUCTION

A recent approach to speaker modelling is to use phoneme based speaker models. This approach is based on evidence that certain phonemes have higher inter-speaker variability than others which makes them particularly useful as good speaker discriminators. However individual phonemes are not easy to recognise automatically with very high accuracy when the speaker is unknown. State-of-the-art figures for such recognition in optimal conditions are of the order of 80% averaged across all phonemes. Further, specific instances of individual phonemes can be very limited within short segments of running speech. In this paper we investigate the use of the broad phonetic classes (BPC) as the unit for speaker modeling. These classes are fricatives, plosives, approximants, nasals, and vowels, each treated as a single broad class. BPCs are easier to recognise than phonemes and they occur more frequently in running speech. BPC-based speaker models may therefore be more easily trained if their added internal phonetic complexity can be accommodated.

There are already several works (e.g. Eatock and Mason, 1994 ; Savic and Gupta, 1990) that investigate the discrimination ability of the BPCs. However, they are mainly done in terms of verification which is affected by many decision choices, like threshold and reference template selection. Besides, their results are only implicitly based on BPC, through the spectral clustering properties of ergodic hidden Markov models (HMM) and vector quantisation (VQ) which does not necessarily classify all frames of a given BPC segment entirely into a specific BPC. Instead the frames of BPC segments may be classified into several different BPC-like classes. Therefore, the results obtained do not really measure the speaker discriminating performance of individual BPC segments. They only reflect the performance improvement when speech utterances variations can be localised into *BPC-like* classes. Also, their results are mostly reported as aggregated results, obtained through concatenation of the BPC classes instead of using individual BPCs, making it even more difficult to assess the BPC's performance. Only Eatock and Mason (1994) give the results of each BPC although also based on a verification paradigm. This approach is satisfactory when the objective is for speaker verification where the test utterance is of short duration (of order more than 1.0 second). However if a very short duration is needed, of order less than 1.0 second, a better understanding of specific phoneme class is important.

In this paper a speaker identification paradigm was used to evaluate the performance of five BPC classes mentioned above and the use of different model structure to model these five classes. We also investigated the effect using several classes of test segment durations on the performances of the BPCs and the models. The second section describes the method and the models used. The results of the five different BPCs and the different models are given in the third section. The optimal type of model for different durational and phonetic classes is given in the fourth section.

METHOD

Our experiment used data from an Australian version of the SCRIBE corpus which had previously been phonetically labeled manually by phoneticians at Macquarie University. Independent training and testing data from five male speakers were used to evaluate five BPCs : vowels, nasals, approximants, fricatives and plosives. These training and test data were obtained from four sets of 50 sentences making a total

of 200 sentences for each speaker. These two pairs of sets were used as the two replications for the experiment.

Each BPC of each speaker was modeled using four model architectures : Gaussian mixture HMMs having single state (VQ), ergodic HMM (ERG), left-to-right (LTR), and LTR with skips (SLTR) configurations. We used five types of ergodic models : E2s1, E3s1, E4s1, E2s3, and E2s4, where the first letter denotes ergodic model, the first number denotes the number of states, and the number after the small 's' denotes the number of mixture components for each state. For LTR models, we used 16 types of models : L2s11, L3s111, ..., L3s333, where the first letter denotes LTR, the first number denotes the number of states, and the numbers after the small 's' denotes the number of mixture components in each state, starting from the initial to the final state. For SLTR, we use 7 types of models with the same notational convention as for the LTR with 'L' is replaced by 'S'. For VQ we used 7 types : V2, V4, V8, V16, V32, and the single mixture component (Normal distribution) V1, and V1f, where the number denotes the number of mixture components, and the letter 'f' in V1f denotes that a full covariance matrix is used. We shall refer to the components of the mixture of Gaussian output probability of a state simply as components or components of a state and to the number of all components of all states of a model as total components if the context is clear. All models used Gaussian components with diagonal covariance matrix except the V1f model. The choice of model types was based on the acoustic phonetic consideration of the BPCs and on the possibility of testing the effect of a different number of states with the same total number of mixture components and testing the effect of a different total number of components with the same number of states.

The acoustic features comprised 22 mel-scale cepstral coefficients plus energy. They were obtained from a digitised speech with sampling rate of 20000 samples/s by using Hamming windows of size 25.6 ms, shifted each 10 ms and pre-emphasised.

The experiments were done using a speaker identification paradigm. In each replication, each model was trained using the HTK-Toolkit on the BPC segments of the training set and the performance of each model was tested using the testing set for that replication. We had two set of experiments. In the first experiment we used all BPC segments from the test set. This was to test the effect of different model structure to model each BPC, and to see how well each BPC performs in an identification test. In the second experiment, the BPC segments from the test set were divided into three durational classes : SHORT, MED, and LONG based on whether the segment's duration was more than one standard deviation less than the mean logarithmic duration, within one standard deviation from the mean, and more than one standard deviation larger than the mean respectively. The performance was then obtained based on these three durational classes. This was to test the effect of durational classes on the performance of the different BPCs and different speaker model structure in an identification test. In both experiments, we used normalised non-error count score, or simply referred to as score, and its reliability score, deviation of the score from its estimated expected value, as our performance criteria. The data were analysed using a repeated measure experimental design where the different models were regarded as the repeated treatments.

COMPARISON AMONG DIFFERENT MODEL STRUCTURES

The result of the first experiment is shown in Figure 1. It is given as the average of the non-error scores from the two replications. From an analysis of variance, averaging across all model types within a structure, there was no significant effect between different model structure on the average non-error count score. However, within each structure, the different model types significantly affected the non-error score ($p < 0.012$) for all BPCc but nasals.

Fricatives. In general, for fricative models with the same number of total components, the LTR models with many states and with components evenly distributed across the states and the LTR models with more components in the initial and in the medial than in the final states, gave better results than other models. The latter models were usually slightly more reliable. The models with more components in the initial and in the final states than in the medial state gave worst discrimination performance. Comparing across different total components, the four total component models gave the best results followed by the fewer than three total components and the larger five, eight, and nine total components. Similar to LTR, within the same total components, the SLTR models with many states, on average, performed better than those models with fewer states. For mixture of Gaussian models (VQ), V1f stands out as the best model. V1, however, gave the worst result. Comparing across different model structure the normal

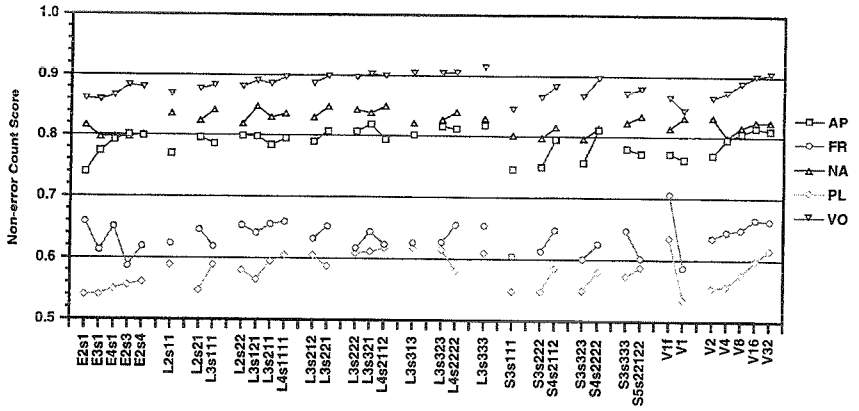

Figure 1: Average non-error count score for overall test data divided into approximants (AP), fricatives (FR), nasals (NA), plosives (PL), and vowels (VO).

model V1f gave the best result with a score of 0.706 followed by E2s1 model and the LTR models with four total components L4s1111 with scores of 0.659.

Plosives. Similar to fricatives, LTR models with many states for plosives also gave good results but different from fricatives, the models with fewer components in the medial state than in the initial state performed better than the other models. This pattern also showed up in the SLTR models except for the eight component models. In the VQ group, similar to fricatives, the normal model with full covariance matrix V1f stands out as the best model for plosives with a score of 0.635. In general, ergodic models were not good models for plosives even the one with many components E2s4 which gave the best result for replication 1 but on average still performed worse than the simplest LTR model L2s11.

Comparing among different total components, models with more than six total components gave good results. For models with fewer total components, only models with four and five total components with many states or with more components in initial and final than in the medial states gave comparable results, Comparing across different model structures, in general LTR models gave good performance. For other structures, only the normal model with full covariance matrix could outperform the LTR models which also gave good reliability. In general, ergodic models and SLTR models were not good models for plosives.

Approximants. For ERG and VQ structures, models with many total components (more than four) were better than those with fewer components achieving scores greater than 0.799. For LTR, models with more components in the initial and medial states than in the final state gave better results than the other models. Comparing across different model structure, the LTR models with six and eight total components gave comparable results with VQ models having more than eight total components. The best model for approximants was L3s333 with a score of 0.818.

Nasals. The results and trend of the results of nasals across different replications were quite varied. This is one of the reasons for the non-significant effect of different models for nasals. In average, however, using LTR and SLTR, and for the same total components, the many-states models performed better than the fewer-states models. In other words, the models with fewer components in each state are better than the ones with more components for each state. Of these models, only the nine component SLTR gave good stable results with scores of 0.823 and 0.834 for S3s333 and S5s22122 respectively.

For VQ and ERG models, however, the simple V1, V2, and E2s1 models gave better average results than the other models although these results were not stable across replications. Only models with

many components V16 and V32 gave consistently high results in both replications with average scores of 0.825.

Vowels. For LTR and SLTR, the models with many components in the initial or with many components in the initial and medial states always gave better results than the ones with many components in the initial and finals but a few components in the medials. For the same total components, the ones with the many states (up to four states) gave better results than the ones with fewer states. For VQ models, the models with many components gave better results, i.e. in V16 and V32 with scores of 0.901 and 0.905 respectively. For models with less than 16 mixture components, the normal model with full covariance matrix (V1f) gave a good result with a score of 0.866. For the ergodic models, the four state model with a single component per state gave a better result than the one having two states with a single component per state (E4s1 is better than E2s1).

However, in general the models with many components in each state gave good results which were more stable across replications. For example, L3s333, E2s4 and L4s2222 could achieve scores of more than 0.880. The best model for vowels was L3s333 with a score of 0.915.

COMPARISON AMONG TEST SEGMENT DURATIONS

The results for the second experiment are given in Figures 2 to 4. They are based on the average of the two replications. Note that in the SHORT duration class, some or all of the LTR models could not be used because of the short duration of the test segments and the strict state order of the left-to-right HMM models.

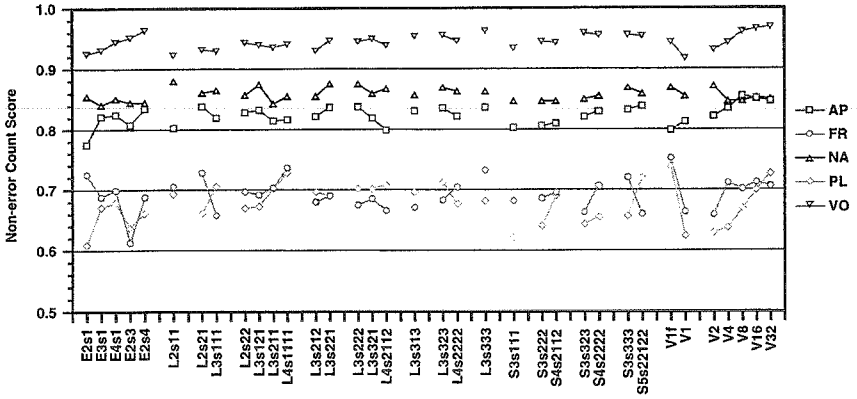

Figure 2: Average non-error count score for long duration test data divided into approximants (AP), fricatives (FR), nasals (NA), plosives (PL), and vowels (VO).

Fricatives. Although, in the duration independent experiment, fricatives could only reach a score of 0.706 using V1f, using different models for different durational class they could improve their performance. For long duration fricatives, using V1f they could reach 0.752. Using V1f and for medium duration they could reach 0.716. For short duration, however, fricatives could not be used. The best score was only 0.535, using V1f. Comparing across different model structure, V1f outperformed the other models for long and medium durations.

Plosives. In the duration independent experiment, the plosives average score was only 0.635, lower than fricatives which could reach 0.706 for the same model V1f. However examining their durational results, for short duration, plosives could be better than fricatives. Using model L2s22 or L2s21 with scores of 0.588 and 0.570 respectively they could outperform the best of fricatives (V1f and V16 with scores of 0.532 and 0.535 respectively). For long segments, plosives were also comparable to fricatives

using V1f. For medium duration, however, plosives were much worse than fricatives. While fricatives could reach 0.716, plosives could only achieve 0.621.

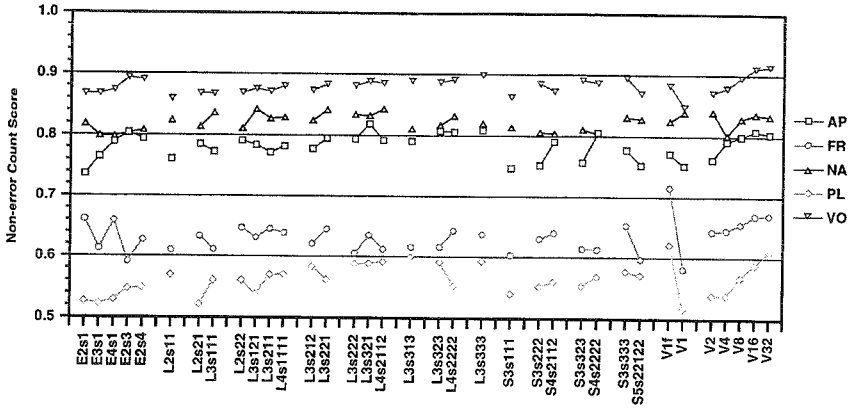

Figure 3: Average non-error count score for medium duration test data divided into approximants (AP), fricatives (FR), nasals (NA), plosives (PL), and vowels (VO).

Approximants. Using duration information, only a small fraction of LTR models could be used for short duration. In fact only skipped models could be used for short approximants. Of these skipped models, V16 and V32 with scores of 0.775 and 0.789 respectively gave the best results. However some simpler models V1 and E3s1, with scores of 0.775 and 0.764, came close to V16 and V32. For medium duration, model L3s321 gave the best result with a score of 0.819. Also in this durational class, models with a few components, usually fewer than four components, gave poor results.

Vowels. In general, long vowels gave very good performance with a score of 0.969. Although the VQ models with many total components, V16 and V36, gave the best results, the eight and nine total component models came close to them. L3s323, S3s333, S4s2222, E2s4, and E2s3 gave scores of more than 0.952. This suggests that introduction of structure into the model gave better modelling ability. Long vowels gave much higher score than the duration independent score which was only 0.915. For medium vowels, VQ models with many components, V16 and V32, scored 0.911 and 0.914 respectively, clearly better than the best of the other models, L3s333 with a score of 0.899.

Nasal. For short duration test segments, nasals in general outperformed vowels, especially when LTR models with a few components and a few states or SLTR with more components in the initial and final states than in the medial (i.e S4s2222, S4s2112, S5s22122) were used. S4s2112 and S5s22122 could achieve more than 0.845 compared to vowels using S4s2222 which was only 0.792. For long duration nasals, models as simple as L2s11 could achieve a score of 0.880 which was comparable to V2 or V1f models.

DISCUSSION AND SUMMARY

The good results of LTR models with more components in the initial and the medial than in the final states in approximants suggest that the coarticulation effect can be modelled better using these models than the other models. Moreover these models can accommodate the slowly changing spectrum of approximants. On the other hand, for nasals, where a simple L2s11 can give a good result, this indicates that mainly only the initial and the final spectra are important to differentiate speakers.

The good results of V1f, compared to other models, for fricatives and plosives, indicates the high inter-frame dependency of the spectra of these signals. Note that the V1f model is much more complicated in terms of the number of parameters to be estimated than the other models with less than five total

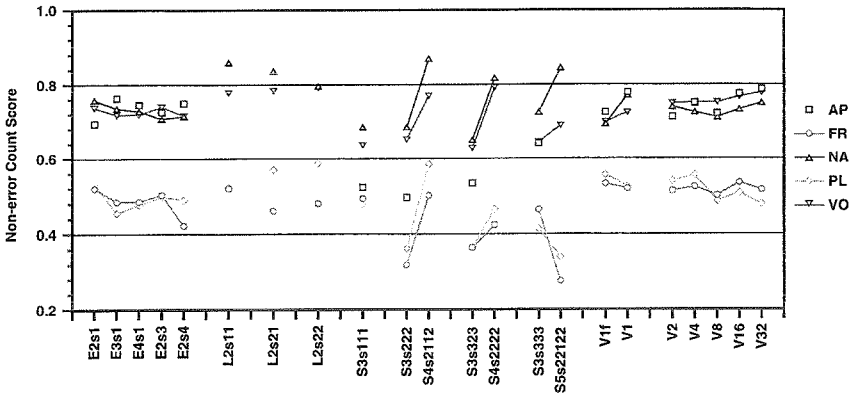

Figure 4: Average non-error count score for short duration test data divided into approximants (AP), fricatives (FR), nasals (NA), plosives (PL), and vowels (VO).

components. Only when using many total components is the complexity of the V1f models approached.

In summary, the results indicated that the use of LTR models could improve the non-error score up to 10% compared to the use of ERG and SLTR models. This improvement was much larger for the nasals and plosives compared to that for the other BPCs. There was also a tendency for LTR models which have more components in their initial state to perform better than those with the same components in all states. This was especially evident for the approximants. This suggests that the timing of spectral change was important, and that onset coarticulation was also important for these BPCs.

The test data, divided into three durational classes within each BPC, gave very heterogeneous performance with respect to duration for vowels, fricatives, and plosives. For example fricatives longer than 0.126s, with a score of 0.752, could approach the performance of vowels shorter than 0.035s with a score of 0.792. The vowels longer than 0.129s however gave a score of 0.969. Approximants and nasals resulted in more homogeneous best performance across different durational classes ranging from 0.787 to 0.855 for approximants and from 0.842 to 0.880 for nasals.

CONCLUSION

The implication of the result described in this paper is that to have a good speaker recognition performance in very short periods of running speech based on BPC, a specific model structure operating on different durationally classified BPC segments is needed.

ACKNOWLEDGEMENT

The provision of phonetically labelled data by the Speech Hearing and Language Research Centre, Macquarie University, is gratefully acknowledged.

REFERENCES

Eatock, J.P. and Mason, J.S. (1994) *A quantitative assessment of the relative speaker discriminating properties of phonemes*, Proc. ICASSP 1994, pp I-133 -- I-136.

Savic, M. and Gupta, S.K. (1990) *A Variable parameter speaker verification system based on hidden Markov modelling*, Proc. ICASSP 1990, pp 281 -- 284.