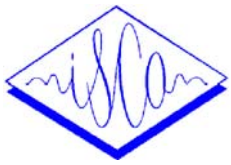

5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

JOINT RECOGNITION AND SEGMENTATION USING PHONETICALLY DERIVED FEATURES AND A HYBRID PHONEME MODEL.

Naomi Harte, Saeed Vaseghi, Ben Milner

The Queen's University of Belfast, N.Ireland.

*British Telecom Research Laboratories, England

E-Mail: N.Harte@ee.qub.ac.uk

ABSTRACT.

This paper encompasses the approaches of segmental modelling and the use of dynamic features in addressing the constraints of the IID assumption in standard HMM. Phonetic features are introduced which capture the transitional dynamics across a phoneme unit via a DCT transformation of a variable length segment. Alongside this, the use of a hybrid phoneme model is proposed. Classification experiments demonstrate the potential of these features and this model to match the performance of standard HMM. The extension of these features to full recognition is explored and details of a novel recognition framework presented alongside preliminary results. Lattice rescoring based on these models and features is also explored. This reduces the set of segmentations considered allowing a more detailed exploration of the nature of the model and features and the challenges in using the proposed recognition strategy.

1. INTRODUCTION.

The HMM framework is now widely accepted as the foundation of successful speech recognition systems. This strongly established statistical framework, supported by the elegance of Baum Welch Re-estimation and the Viterbi algorithm for the training and recognition phases, has become the preferred foundation for further work on improving the performance of speech recognition systems. In using HMM to model the temporal evolution of speech feature vectors, it is assumed that successive vectors are Independently and Identically Distributed (IID) within a state of a HMM. That is to say that the feature vectors are assumed to be randomly distributed about a fixed mean and that there is no correlation between successive vectors. By nature of the speech process and the feature extraction process it is clear that successive feature vectors will in fact be highly correlated and there is useful information for recognition in the evolution of this spectral time trajectory. The IID assumption is thus recognised as an obstacle in achieving higher performance in HMM based speech recognition. Attempts can be made to lessen the effect of this assumption by incorporating temporal information into the actual feature vectors, e.g. the use of dynamic cepstrum, and by extending conventional HMM. Another product of these research efforts has been the development of segmental models. The essence of segmental modelling lies in the association of states with feature vector sequences rather than with individual frames, in order to overcome the IID assumption. A comprehensive review of approaches to segmental modelling can be found in [5] where the authors present a unified view of the approaches. One advantage of segmental modelling is the potential for the use of segmental features which are not apparent at a frame level. Difficulties arise in extending segmental features beyond the realms of classification, one problem being that the computational demand of recognition commonly outreaches the potential benefits to be gained. This paper investigates alternative features as segmental features in conjunction with a framework to extend the use of these features to recognition.

2. SEGMENTAL PHONETIC FEATURES.

Features are conventionally extracted on a frame by frame basis with heavy overlap between successive fixed length windows. The current work instead proposes that phonetic features can be calculated over the duration of a phoneme in order to capture the transitional dynamics and the correlation within that segment. The success of cepstral-time matrices has been demonstrated before in isolated recognition tasks [5]. This work presents the use of closely related segmental phonetic features for continuous speech recognition. For a given unit of speech, identified as a phoneme unit or phonetic segment, and of length T vectors, the phonetic features for that segment can be derived as

$$Y = A_T X \tag{1}$$

where $X = [x_t, ..., x_{t+T-1}]$ is the segment and A_T is a transformation dependent on the segment length T. In this work, A_T is the T length DCT and the phonetic features Y are hence derived via a DCT on the stacked cepstral vectors X as

$$c(n,m) = \frac{1}{T}\sum_{k=0}^{T-1} c_k(n) \cdot Cos\left(\frac{(2k+1)m\pi}{2T}\right) \tag{2}$$

where $c_k(n)$ is the nth coefficient of the kth cepstral vector in the segment of staked MFCC vectors. The $\frac{1}{T}$ factor accounts for the variable length of the segment. These phonetic features thus yield a fixed length representation of a phoneme irrespective of the original frame length of the segment. Alongside the use of these features, a novel phoneme model is presented which uses a hybrid representation of a phoneme.

3. HYBRID PHONEME MODEL.

The current work also proposes the use of a novel acoustic model for modelling phonemes for phoneme based continuous speech recognition. It is a hybrid phonetic model in that it allows both the use of conventional features (here MFCC) and phonetically derived features. This model aims to exploit the success of conventional HMM but to also supplement the information used in recognition with information derived directly from a segment hypothesised to represent a phoneme unit.

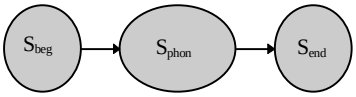

Figure 1: Phonetic Model Topology.

The phonetic model as shown in Figure 1 is similar to the standard monophone HMM in that it has three distinct states and a left-to-right nature. There are however no transition probabilities associated with this model. The model as a whole aims to model a segment which corresponds closely to a phonetic event. The beginning and end state model the conventional cepstrum feature frames flagging the segment. The middle or phonetic state is dedicated to modelling the phonetic features derived across the duration of that segment. The distribution of features in all states are modelled using conventional mixture gaussian densities.

4. CLASSIFICATION.

Preliminary experiments were performed in classification where the phoneme boundaries are known. This was to assess the initial performance of the features and phonetic model. When the boundaries are given, the classification of a segment is performed as identifying the phoneme α which maximises the likelihood of the segment as

$$\hat{\alpha} = \arg\max_{\alpha} P\left(x_1 \middle| s_{\lambda_\alpha, s_{beg}}\right) P\left(Y \middle| s_{\lambda_\alpha, s_{ph}}\right) P\left(x_T \middle| s_{\lambda_\alpha, s_{end}}\right) P\left(T \middle| \lambda_\alpha\right) \quad (3)$$

Boundary information was taken from the transcriptions provided with the TIMIT database.

5. RECOGNITION.

The extension of the phonetic features and hybrid model to recognition involves some key challenges, many encountered by other researchers in the use of segmental style features. The main challenges are :

- Boundaries of phonetic segments are unknown and must be hypothesised at some level.
- The set of possible segmentations grows exponentially becoming computationally unmanageable.
- Features are no longer time synchronous and hence Viterbi based recognition cannot be used.

Previous approaches to recognition using segmental features have included a dynamic programming approach [2], more advanced split and merge algorithms [4] and the segmentation-first approach [2] [6] This work proposes a new strategy to recognition based on the use of phonetic features and the current hybrid model. It involves the hypothesis of successive segments to expand a segmentation network but employs strategic pruning to maintain the network at a manageable size and avoid the exponential growth mentioned above.

5.1. Joint Segmentation and Pruning Strategy.

A potential segment boundary can be considered to establish a node in the segment network. Given a particular node or start time of a segment, the possible duration and corresponding identity of a phoneme segment starting at that time is considered. With a set of N phoneme models with duration varying between a global minimum and maximum duration, $\tau_{\min}$ and $\tau_{\max}$ respectively, this could amount to $N\left(\tau_{\max} - \tau_{\min}\right)$ segments considered for any one node. The number of evaluation is immediately reduced by recognising that the different phonemes will vary in minimum and maximum duration and these statistics can be established from training data. Computation is further reduced by first hypothesising each phoneme at three or four durations between it's minimum and mean duration to eliminate candidates which are unlikely to show a good match at other durations. In the current experiments 10 of 39 phoneme candidates were eliminated from subsequent evaluation for a given segment with no perceivable affect on performance.

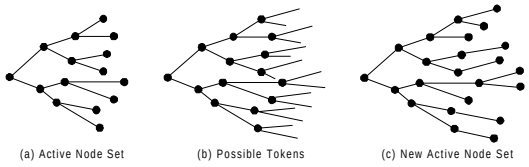

Figure 2: Emergence of Segmentation Network.

In this way, a small set of most likely segments given that start time can be established as branches which establish new possible nodes in the network. The process is demonstrated graphically in Figure 2. The initial active node set in (a) corresponds to possible start times for the nth segment and the process described above leads to the establishment of a new active node set of possible start times for the n+1th segment. Clearly, to prevent the network expanding excessively, not all possible segments can be kept in the network. Computationally it is more efficient with this method to explore all the possibilities for the nth segment in parallel as shown. Pruning is enabled so that as in this example initially only the two most likely branches from each node are preserved. Then the most likely branches for all hypothesised start times from that segment are globally examined and the least likely paths deactivated to maintain a maximum number of nodes for any

segment. Merge detection is also important and when merges occur the path with the worse score is deactivated. Recognition is then complete when all paths terminate in the final frame of a test sentence. Comparison of paths with different numbers of segments is difficult as the paths with the least number of segments will clearly have better log likelihood scores generating a bias towards deletions, a reported problem in [3]. This work seeks to examine other bases for comparison of paths such as average score per segment.

5.2 Lattice Rescoring.

Since much of the computation involved in the recognition strategy is in the hypothesis of segmentations, this issue can be alleviated in some way by attaining a prior set of most likely segmentations for a sentence. Such approaches have been used before in [2] and [6] where the first stage of the systems is to establish a single or a set of segmentations. In a effort to better understand the behaviour of the models under investigation, experiments were carried out where a set of segmentations was established before evaluation of the segmentations with the phonetic features and hybrid models. HMMs were used to output lattices in the course of Viterbi-based recognition. This was done using HTK where the size of the lattice could be controlled by changing the number of tokens preserved in a state in the token passing paradigm implementation of the Viterbi algorithm.

It is vital that in using such a method that the segmentation set yields a distribution of segmentations close to the actual segmentation on the sentence. Ideally the distributions would show many possible segment boundaries close to the actual boundaries with significantly fewer segment boundaries detected within phonemes. This would maximise the potential of the phonetic features in identifying the correct segmentations and phoneme sequence.

6. EXPERIMENTS.

All experiments reported were performed using the TIMIT database. The full training data was used to build up duration statistics of minimum and maximum duration for each phoneme and to establish a gamma distribution for the duration term where necessary. Monophone models were trained using segmented data from the labelling transcriptions provided with the corpus. The beginning and end states of the phonetic model were trained with MFCC frames including first and second order derivatives calculated over a number of adjacent frames in the conventional manner and hence were influenced by neighbouring segments. The classification experiments were performed on the complete test set while subsequent experiments except where otherwise indicated were performed using the core test set as recommended in the TIMIT corpus documentation.

6.1 Classification Experiments

Initial experiments in classification demonstrate the potential of the phonetic features and the hybrid models. The original MFCC feature vectors were derived at a frame rate of 1.5ms with a window length of 15ms and also at a frame rate of 2.5ms with a window length of 20ms. Phonetic features for a segment of frames $[x_1 \ldots x_T]$ were derived over frames $[x_2 \ldots x_{T-1}]$ while the beginning and end frame were modelled by the distribution in the beginning and end states respectively. Columns zero through two were preserved as phonetic features after the DCT transformation of a segment. Table 1 shows the results obtained with percentage correct classification given.

MFCC data	15 Mix.	24 Mix	28 Mix	36 Mix.
1.5ms/15ms	64.08	65.07	65.24	65.46
2.5ms/20ms	65.79	67.03	67.40	67.49

Table 1: Classification Experiments using Hybrid Phonetic Model.

These results demonstrate the ability of the phonetic features to match the performance of standard HMM with first and second order derivatives which yield 66.40% classification on the same data.

6.2 Lattice Rescoring Experiments:

The models trained on the 1.5ms MFCC data were used in the lattice rescoring experiments. Again column zero to two were preserved as phonetic features. A lattice for each test sentence was output from HMMs trained on MFCC data with a standard frame rate of 10ms and a window length of 25ms. The use of lattices generated from preserving 5, and 10 tokens in each original HMM state was investigated. Results are reported in Table 2.

Tokens	%Classification.	%Del	%Sub	%Ins
5	46.72%	12.31	40.97	29.29
10	43.26%	17.71	39.13	22.70

Table 2: Lattice Rescoring Experiments.

For these experiments it was found that the inclusion or exclusion of the duration term in (3) had little affect on performance. Further experiments were performed to assess whether the phonetic model could discriminate between the correct and alternative segmentations when the actual segmentation was forced into the lattice and completely evaluated. This resulted in no affect on the output. The fact that the correct segmentation did not yield a better score than alternative segmentations demonstrates that a more detailed understanding of the nature of these models is necessary. The level of deletions, substitutions and insertions is included in Table 2. Both the insertion and deletion rate are high indicating the difficulty in locating accurate boundaries using the segmental features. On average transcriptions tended to have approximately 15% more segments than the correct segmentation indicated.

The issue of how to optimally combine scores from the different segments in competing segmentations needs to be fully addressed. A segment score could be weighted by the number of frames in the segment. This naturally would lead to a tendency towards insertions however which is already high. Sentences could then be compared on either a total likelihood or likelihood per segments basis but initial experiments with these models have confirmed that the level of insertions is too great in both cases. It would therefore appear that there is some middle ground to be found between this and the pronounced tendency towards deletions found in recognition experiments reported in the next section. More extensive research has been undertaken into optimally combining the elements which make up the complete segment score and how to weight scores in such a manner as to improve the discriminative ability when comparing paths.

6.3 Full Recognition Experiments.

Experiments were carried out using the recognition strategy outlined where successive segments were hypothesised for each test sentence and the pruning strategies employed. Initial performance figures are indicated in Table 4. The maximum number of paths maintained was varied between 10 and 20 paths with little affect on the results. The MFCC frames were again extracted at a 1.5ms rate.

Active Paths	%Recognition	%Del	%Sub	%Ins
10	31.12	29.62	39.25	9.73
20	31.04	28.45	41.60	8.98

Table 4: Recognition Performance.

There was a pronounced tendency towards deletions at almost 30% in both experiments. Paths were compared on an average score per phoneme basis but there was still an underlying tendency for longer segments to score better than shorter segments. It can be seen that increasing the number of paths had no affect. It is felt that problems with the model are the over-riding factor and first must be addressed before other factors such as number of active paths and branches from each node will become more significant. The evolution of the segment network during recognition has been closely studied. In instances where a node corresponds closely to an actual phoneme boundary the model performs well and tokens with close to correct duration and correct identity will be propagated. This sort of behaviour is demonstrated by Figure 3 which shows the emergence of the different likelihood terms when an occurrence of the phoneme ch is hypothesised beginning from the actual start time of the phoneme. The phoneme is of duration of approximately 50 frames and this can be clearly picked out from the likelihood terms. Improvements in recognition are anticipated when the task of rescoring lattices is fully resolved.

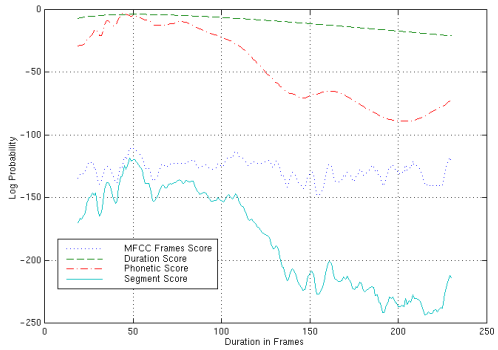

Figure 3: Emergence of Segment Score for Hypothesized Occurrence of ch, from Actual Phoneme Start Time

7. CONCLUSIONS.

Initial experiments in recognition have reflected some of the difficulties encountered by others in using segmental features for recognition tasks. The lattice rescoring experiments have given a greater insight into the nature of the problems of deletions in full recognition and insertions in lattice rescoring and the balance of probability terms in light of the modelling assumptions been made. The continuation of lattice experiments is seen as presenting the best opportunity for improvement. Once performance is improved at this level, the performance of the presented recognition strategy can be fully assessed. One direction for research and experimentation is the use of discriminative methods in the segmental framework to optimally combine likelihood terms. The work is also being extended to include context modeling to model correlation across segments aswell as within the phonetic unit.

8. REFERENCES.

1. Digalakis, V., Ostendorf, M., Rohlinek, J., *Fast Algorithms for Phone Classification and Recognition Using Segment-Based Methods* , IEEE Trans S.P. Vol.40, No.12, 1992.

2. Glass, J., Chang, J., McCandless, M., *A Probabilistic Framework for Feature-Based Speech Recognition*, Proc ICSLP, pp2277-2280, 1996.

3. Ostendorf, M., Roukos, S., A Stochastic Segment Model for Phoneme Based Continuous Speech Recognition. IEEE Trans. Acoust. Speech, Signal Processing, pp 1857-1869, Vol.37, No.12, Dec.1989.

4. Ostendorf, M., Digalakis, V., Kimball, O., *From HMMs to Segment Models: A Unified View of Stochastic Modelling for Speech Recognition*. IEEE Trans ASSP, vol. ASSP-37, no.12, pp1857-1869

5. Vaseghi, S., Conner, P, Milner, B., *Speech Modelling Using Cepstral-Time Feature Matrices in Hidden Markov Models*, Proc IEE-I Vol.140, No.5, pp317-320, Oct. 1993.

6. Zue, V., Glass, J., Phillips, M., Seneff, S., *Acoustic Segmentation and Phonetic Classification in the SUMMIT system*. IEEE ICASSP. pp389-392, 1989.