A NOVEL MODEL FOR PHONEME RECOGNITION USING PHONETICALLY DERIVED FEATURES

Naomi Harte, Saeed Vaseghi, Paul McCourt
School of Electrical Engineering and Computer Science,
The Queen's University of Belfast, N.Ireland
Tel: +44 1232 274275; fax: +44 1232 667023
e-mail: n.harte@ee.qub.ac.uk

ABSTRACT

This paper presents work on the use of segmental modelling and phonetic features for phoneme based speech recognition. The motivation for the work is to lessen the effects of the IID assumption in HMM based recognition. The use of phonetic features which are derived across the duration of a phonetic segment is discussed. In conjunction with the use of these features, a hybrid phoneme model is introduced. In a classification task on the TIMIT database, these features are capable of outperforming standard HMM. The extension of the work to recognition is presented in detail. The challenges are identified and a novel algorithm presented for recognition based on phonetic features and the hybrid phoneme model. The approach is built around a segmentation hypothesis approach employing pruning at a number of levels.

1 INTRODUCTION

HMMs have now been firmly established as the most widely used and successful acoustic model for speech recognition. The extent of their popularity stems largely from the existence of efficient mechanisms in Baum Welch Re-estimation, the Forward Backward Algorithm, and Viterbi Decoding for both training and recognition. This strongly established statistical framework has thus become the preferred foundation for further work on improving the performance of speech recognition systems. The successive states of a HMM model the temporal evolution of a set of observations derived from the original speech wave. However, within each state the observation vectors are assumed to be independent and identically distributed (IID). Thus the probability of an observation is only dependent on the current state and not on any previous observation, though it is clear that successive vectors will in fact be highly correlated. The IID assumption is clearly violated and therefore an obstacle in achieving higher performance in HMM based speech recognition. Attempts can be made to lessen the effect of this IID assumption by incorporating temporal information into the actual feature vectors and by extending conventional HMM models.

The use of dynamic coefficients is a well established approach to extending conventional feature vectors to include temporal information. Much research has been directed towards explicit methods of including temporal correlation in the actual HMM framework to overcome the IID assumption. In [8] and [9] the probability of a feature vector is conditioned on the previous frame. Work by Ming et al [7] has extended the inter-frame dependence model to account for the correlation dependency structure both forwards and backwards in time. An alternative approach at a model level to overcome the weaknesses of the IID assumption has been the use of segmental HMMs. The essence of segmental HMM lies in the association of states with feature vector sequences rather than with individual frames. In [2] the segmental HMM has an underlying semi-Markov process which models speech at a segment level with separate models then used at a state level to model the so called extra-segmental and intra segmental sources of variability. Both a static and linear segmental model have been explored. In [3], the stochastic segment model is used to model variable duration phonemes by using a fixed length representation of a variable length sequence of frames.

2 PHONETIC FEATURES

Conventionally, features are extracted on a frame by frame basis with heavy overlap between successive windows. The current work instead proposes that phonetic features be calculated over the duration of a phoneme in order to capture the transitional dynamics in that segment. The use of such features was first proposed in [1]. For a given unit of speech of length T frames, identified as a phonetic segment, the phonetic features for that segment are 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. Here, A_T is the variable length T by N DCT used to decode the transitional dynamics across the duration of the phonetic event. N denotes the number of features in any vector x_t. Phonetic features thus yield a fixed length representation of a phoneme irrespective of the original frame length of the segment and are derived via a DCT on the stacked cepstral vectors as

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

The $\frac{1}{T}$ normalising factor is introduced to take into account the variable length of the segment. The phoneme feature matrix is then made up of M of the T columns of the transformed matrix with N features in each column. With the phonetic features derived from the original cepstral coefficients as described, the phonetic features are a compressed representation, having a measure of the average of each coefficient and the rate of change at different rates over the

duration of the segment. A mixture Gaussian density is used to model the phonetic features distribution. This is closely related to the use of cepstral-time matrices with a variable length transform matrix. The use of cepstral-time matrices has proven highly successful in isolated word recognition [2].

3 PHONETIC MODEL

The most widely employed HMM for modelling monophones in speech recognition, is the 3-state left to right model with self loops permitted. A new hybrid model is proposed where 3 states are used, but the beginning and end states are used to model the transitions between phones, and the middle state is intended to model the phonetic features derived across the phoneme. One potential advantage of this is that co-articulation effects at the beginning and end of a phoneme, where the actual distinction between successive phones may be fuzzy, are not considered in deriving the segmental phonetic features. With this possible effect removed, the phonetic state is dedicated to modelling the phonetic event proper.

Figure 1: Proposed Phoneme Model.

3.1 Computation of Probabilities

The question arises, given an observation sequence $X = [x_1 ... x_T]$ and a set of K phoneme models $\lambda^K = [\lambda_1 ... \lambda_K]$, how to efficiently compute $P\left(X|\lambda^K\right)$ and identify the sequence of phonemes generating that sequence. The score for a segment, when employing the three state phonetic model, becomes

$$P(X|\lambda) = P(x_1|s_b)P(Y|s_p)P(x_T|s_e)P(T|\lambda) \qquad (3)$$

for a particular phoneme model λ, where Y is the transformed representation of $[x_2 ... x_{T-1}]$. The quantity $P(T|\lambda)$ denotes the probability of duration T for that phoneme. It is clear that there are distinct elements contributing to the overall score: the contribution of the beginning and end frames, the phonetic features score and the duration term. This presents the possibility of a hierarchical approach to the evaluation of a segment, discussed in later sections.

4 RECOGNITION AND THE PHONETIC MODEL

4.1 Classification: Known Segmentation

The phonetic model can be readily employed where the phoneme segment boundaries are given. Classification involves simply transforming the segment to yield the phonetic features and identifying the phoneme as

$$\widehat{\alpha} = \arg\max_{\alpha} P(x_1|s_{b,\lambda_\alpha})P(Y|s_{p,\lambda_\alpha})P(x_T|s_{e,\lambda_\alpha})P(T|\lambda_\alpha) \qquad (4)$$

This enables a direct comparison of the performance of the phonetic features compared to standard cepstrum using first and second order derivatives. Depending on the number of columns of the cepstral time matrix preserved and the number of frames in the phonetic segment, a significant data compression can be achieved. For instance, for a typical segment of 100ms, a typical cepstrum representation of this may be at a 10ms frame rate with 39 features in each frame. The phonetic representation of this segment may be between just 26 and 78 features for the entire segment.

4.2 Recognition: Unknown Segmentation.

The preliminary difficulty in the use of the proposed phonetic model for recognition is that the phonetic segment boundaries needed for the derivation of the phonetic features are unknown. Thus any recognition strategy must involve, at some level, the hypotheses of possible phonetic segments lengths and subsequent evaluation of the soundness of the hypothesis. Previous approaches to the task of recognition for segmental models have employed dynamic programming techniques [3], split and merge algorithm [4] and a segmentation-first strategy[5]. In the context of the phonetic model many of the challenges are similar:

- Segment boundaries are not available.
- For a hypothesised segmentation, feature frames are no longer time synchronous and thus Viterbi decoding cannot be employed.
- The number of possible segmentations grows exponentially becoming computationally unmanageable.

The computational requirements of an exhaustive evaluation of segmentations must be avoided. The proposed method is based on hypothesising segments to expand a possible network while employing an intelligent pruning strategy to control the potential exponential growth of the resultant network.

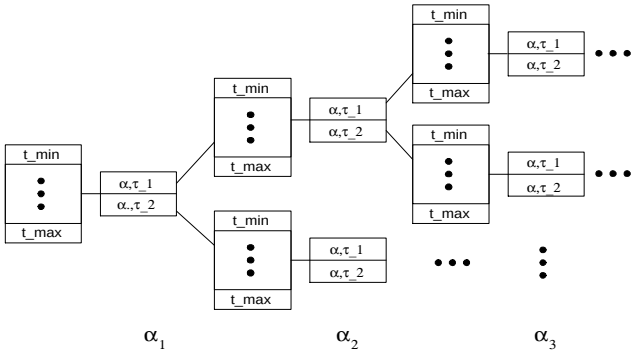

Figure 2: Segmentation Network.

The set of segmentations for a test sentence can be arrived at via two distinct routes: sequentially or in parallel as can be seen from Figure 2 where $\alpha_1, \alpha_2, \alpha_3$ represent the successive phonemes for the sentence. The duration (denoted τ inthe diagram) and identity (denoted α) of each segment is hypothesised with the duration varying between a global maximum and minimum $t_{max}$and t_{min}. In seeking a pruning regime to reduce the overall number of segments evaluated, it becomes clear that a hybrid method encompassing elements of both a parallel and a sequential evaluation of segments is

suitable. Upon enabling pruning, only a controlled number of candidates and durations survive for a segment. In the example shown, only the best two candidates are preserved. In this way the number of branches growing at each point is significantly reduced. This is demonstrated in Figure 2 where a segment is examined and a decision made to the most likely duration and identities of that phoneme. The successive segments are hypothesised based on the possibilities for the previous segment. Any hypothesised segment start time can be considered as establishing a node in a segment network. Phonemes of particular duration establish branches or tokens in the network possibly leading to new nodes. Hence possible paths through the network emerge. This is shown at the node and branch level in Figure 3. The

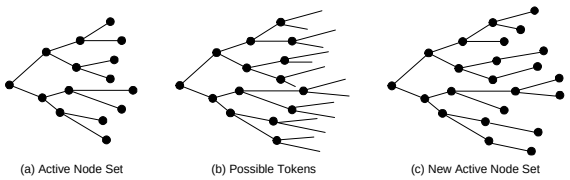

Figure 3: Emerging Network

basis of the approach is to employ pruning at two distinct levels: the first when hypothesising branches and new nodes, and the second through deactivating emerging paths which are unlikely when compared to other paths.

4.2.1 Branch Level Pruning

Any segment in the network will have a number of possible start times or nodes which have been previously established as the most likely end points of the segmentation of the sentence into $n-1$ segments, Figure 3(a). The duration and identity of segment n, given this set of possible start times must be subsequently hypothesised. To reduce computation, it is undesirable to calculate the likelihood of all phonemes of all possible duration to identify the most likely candidates. Instead a hierarchical strategy can be employed alongside using indicators of match for the phonemes.

For a given node, a pointer of the potential soundness of match of the models to the data is taken. This can be done on the basis that some of the candidates will produce a bad match to the data regardless of segment length. From the duration statistics gathered on the TIMIT database, it is possible to model the duration of each phoneme with a gammadistribution varying about an average duration. By taking the start time and hypothesising each phoneme to have a possible set of three to five different durations somewhere between their minimum and average duration, an initial estimate of how well the data matches the model is possible. For this first evaluation, durations greater than the average for each phoneme are not considered. This initial evaluation is used then to eliminate models from further evaluation that demonstrate a poor match to the data while preserving candidates which are more likely to demonstrate good matches at different durations.

At this point, the number of models being considered has been significantly reduced, and the best duration for each of the surviving candidates can be examined in finer detail. For each candidate, the duration which maximises the likelihood for that phoneme is found by allowing the duration to vary between the minimum and maximum duration for that candidate. To further reduce computation, this need not be examined at a resolution of each possible duration as each frame may represent as little as 1ms. In this way a number of most likely branches for that node are arrived at. When all nodes or possible start times for a segment are examined the situation as in Figure 3 (b) arises. When the number of possible new nodes exceed a predefined maximum number of nodes and thus paths which can be kept active simultaneously, the second form of pruning, node level pruning, is enabled. This maintains the network at a manageable size. In the example shown, the number of possible new nodes is sixteen but the maximum number of parallel paths is set at eight. Thus the performance of each path is examined to determine the paths to preserve.

4.2.2 Node Level Pruning

The aim of pruning at this level is to deactivate paths in this growing tree which are highly unlikely when compared to more likely emerging paths. In conventional HMM recognition, the Viterbi decoding is often implemented via a token passing paradigm whereby pruning out unlikely paths is straightforward as it involves investigating all paths up to the current time instance, and deactivating those which have fallen a threshold below the best or average path and are unlikely to recover in subsequent evaluation. The paths can be compared on the basis that they represent a route to a common end point. If all hypotheses for a segment are worked out in parallel as shown, the number of segments while the same are not representative of an equal number of frames. Further, by transforming a variable length segment into a fixed length representation, the dynamic range is effected: the dynamic range for the phonetic model becomes constant for a segment which is of variable length. This contributes to a difficulty in comparing paths. One solution to this is to compare alternative paths on an average score per phone basis rather than absolute likelihood and to prune out unlikely paths based on this measure. This should also help prevent a bias towards segmentations containing fewer segments of longer duration, a problem reported in [2] and [3]. Another issue is the possibility of re-merging paths. Quite often, distinct paths will re-merge at a later stage. Thus before deactivating unlikely paths, merges must be identified as it would be undesirable to keep two copies of one path to the detriment of exploring another possible path through the network.

In this way, the identity and duration of each segment in turn is hypothesised until each active path terminates in the final frame of the test sentence. Again paths may be compared on an average score per phoneme basis rather than the overall score for a segmentation, allowing segmentations comprising of a different number of hypothesised phonemes to be compared.

In the early stages of the segmentation, it is important to keep as many alternative paths active as possible as it is here that the effect of eliminating a good segmentation which has scored poorly would have the most serious consequences.

Thus particularly for the first segment, it would be preferable to keep more than the demonstrated two best branches per node. Also for many databases such as the TIMIT database, network attributes such as the fact that each sentence begins and ends with silence can be exploited.

5 EXPERIMENTAL RESULTS

Baseline experiments in classification were performed on the 39-phoneme set from the TIMIT database. Classification experiments were performed on the full training and test sets as recommended in the TIMIT corpusdocumentation. 39 features were used, comprising 13 mel-frequency cepstral coefficients with first and second order derivatives. A frame rate of 5ms andwindow length of 25.6ms were employed. Each phoneme was modelled with acontext independent 3 state left to right HMM with 12 mixtures per state. The classification rate was 66.40%.

Experiments were performed to investigate the effect of increasing the number of phonetic features retained for a segment. As described in earlier sections, when deriving the phonetic features, M of the T columns of the feature matrix are preserved. Classification experiments were carried out where an increasing number of columns were presrved. Typical results are shown in Table 1. These experiments were performed using MFCC features derived at a frame rate of 2.5ms. The results are shown for window lengths of between 12ms and 25ms. Only the phonetic features were used for classification rather than the hybrid model. For classification, the contribution from the begining and end state would be constant for a constant frame and window length and thus Table 1 demonstrates the discriminative ability of the phonetic features alone.

Columns Preserved:	0-1	0-2	0-3	0-4
12ms, 15 mixtures	57.95%	61.79%	62.76%	62.51%
12ms, 24 mixtures	58.61%	62.68%	62.76%	62.51%
20ms, 15 mixtures	58.09%	61.86%	62.29%	61.98%
20ms, 24 mixtures	58.09%	62.94%	63.35%	63.37%
25ms, 15 mixtures	58.18%	61.49%	61.96%	61.63%
25ms, 24 mixtures	58.77%	62.83%	63.44%	63.04%

Table 1: Effect of using Increasing Numbers of Phonetic Features.

It is interesting to note that the performance starts to decrease when more than four columns are preserved (beyond column 3). A longer window length gave better performance but there was little difference in the results at 20ms and 25ms. The best performance was achieved using three and four columns.

No.Mixtures	%Classification
15	65.79
24	67.03
28	67.40

Table 2: Performance of Phonetic Model for Classification.

The performance of the phonetic model was investigated using a frame rate of 2.5ms and a window length of 20ms for the original MFCC features. Columns 0-2 were preserved as phonetic features. Table 2 shows the classification performance of the model. The figure of 67.40% demonstrates that the phonetic features and model are capable of matching the performance of standard HMM of 66.40% using first and second order time derivatives. Experiments are on-going to investigate the performance of the recognition strategy presented. The effects of varying the number of active paths maintained, the number of tokens preserved for any node in the network and the decision mechanism for pruning paths are being examined in detail.

6 CONCLUSIONS

The use of phonetic features and a hybrid phonetic model has been proposed. It has been demonstrated that the phonetic features and models are readily employed for classification where segment boundaries are given. Experimental results demonstrate that these features and models can give improved performance over standard HMM in a classification task on the TIMIT database. A novel approach to recognition has been presented and the difficulty of preventing exponetial growth of a segmentation network has been addressed via a multi-level pruning approach. The method encompasses both the approaches of segmental modelling and dynamic features to overcome the IID assumption in standard HMM. Further experiments are on-going in assessing the performance of the recognition algorithm.

7 REFERENCES

1. S.Vaseghi, N.Harte, B.Milner, "Multi-Resolution Phonetic Segmental Features and Models for HMM-Based Speech Recognition", Proc. IEEE ICASSP,Vol.2, pp.1263-1266
2. W.J.Holmes and M.J.Russell,"Linear Dynamic Segmental HMMs: Variability Representation and Training Procedure", Proc. IEEE ICASSP, Munich, Vol.2, pp1399-1402, 1997.
3. M.Ostendorf and S.Roukos, "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. 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.
5. Zue,V., Glass,J., Phillips,M., Seneff, S., Acoustic Segmentation and Phonetic Classification in the SUMMIT system. IEEE ICASSP. Pp389-392, 1989.
6. B.Milner, "Inclusion of Temporal Information into Features for Speech Recognition", Proc. Int, Conf. On Spoken Language Processing, Vol.1, pp256-259, 1996
7. J.Ming and F.J.Smith, "Modelling of the Interframe Dependence in an HMM using Conditional Gaussian Mixtures", Computer Speech and Language, 10, pp.229-247, 1996
8. K.K.Paliwal, "The Use of Temporal Correlation Between Successive Frames in a Hidden Markov Model based Speech Recogniser", Proc IEEE ICASSP.pp.209-212, 1993.
9. C.J.Wellekens, "Explicit Correlation In Hidden Markov Models for Speech Recognition", Proc. IEEE ICASSP, pp.384-387, 1987.