ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# COMBINED TEMPORAL AND SPECTRAL MULTI-RESOLUTION PHONETIC MODELLING*

*Paul McCourt  Naomi Harte  Saeed Vaseghi*

School of Electrical & Electronic Engineering
Queens University Belfast, United Kingdom
pm.mccourt,s.vaseghi@ee.qub.ac.uk
http://www.ee.qub.ac.uk/dsp/research/speech/

## ABSTRACT

Incorporating discriminative strengths from alternative acoustic models is an important topic of recent increasing interest. Multi-resolution sub-band models and a novel phonetic segmental model independently achieve improvements on HMMs with standard MFCCs of 70.21% and 70.63% respectively from a baseline TIMIT classification score of 66.4%. Discriminatively trained weighted combination of the log likelihood scores from these acoustic modelling strategies is shown to successfully extend the performance to 72.6%.

## 1. INTRODUCTION

The combination of alternative acoustic, language and even visual modelling strategies for speech recognition is receiving increasing interest in an effort to exploit the differing class discriminative strengths of particular models into a global optimum [1,2]. Multi-resolution sub-band features and models [3] demonstrate that important cues for phonetic discrimination exist in localised spectral correlates that are not captured by full band cepstra. Segmental phonetic features extracted over the duration of a phoneme and incorporated into a novel hybrid acoustic phonetic model presented here are also demonstrated to give performance improvements over conventional HMM. An aspect of the discussion on human recognition [4] which stimulated recent interest in sub-band based models is the possibility that features extracted at multiple temporal resolutions may play an important role in human recognition. This is explored within the proposed segmental model by the inclusion of sub-segmental phonetic features. Despite overall similar performances, different acoustic modelling strategies can reveal quite different class discrimination characteristics. Incorporating discriminative strengths from alternative acoustic models is hence a topic of recent developing interest. This can take the form of some kind of voting scheme based on the transcriptions generated by different acoustic models [5]. The model combination approach used here is through direct linear weighting of the log probability scores of the alternative modelling strategies. Discriminative training of a class-specific linear weight set according to the Minimum Classification Error (MCE) criterion [6] is demonstrated [3] to increase the performance of the multi-resolution sub-band acoustic models. Incorporating discriminatively weighted combination of the segmental phonetic model is shown here to also successfully extend TIMIT classification performance

## 2. MUTI-RESOLUTION SPECTRAL FEATURES AND MODELS

For standard MFCC features, cepstral analsyis is performed on the mel-spaced filterbank log energy vector $\mathbf{E}$ of each short-time analysis frame, as expressed by the linear transformation

$$\mathbf{X} = \mathbf{A}\mathbf{E} \qquad (1)$$

where $\mathbf{A}$ typically represents the DCT basis functions. The log energy vector $\mathbf{E}$ can be split into $N$ sub-vectors $\mathbf{E} = \left[\mathbf{E}_1^{\mathrm{T}} \quad \cdots \quad \mathbf{E}_b^{\mathrm{T}} \quad \cdots \quad \mathbf{E}_N^{\mathrm{T}}\right]^{\mathrm{T}}$ (where T indicates matrix transpose) such that each sub-vector $\mathbf{E}_b^{\mathrm{T}}$ effectively represents a grouped bandwidth of log energies. Separate cepstral analysis using appropriately dimensioned DCT transforms $\mathbf{A}_b$ yields a new feature vector $\mathbf{X}^r$ (r simply identifies the resolution index for a given sub-band decomposition) created from the set of sub-band ceptsral vectors, thus:

$$\left[\mathbf{X}_1^{\mathrm{T}} \quad \cdots \quad \mathbf{X}_N^{\mathrm{T}}\right]^{\mathrm{T}} = \left[(\mathbf{A}_1\mathbf{E}_1)^{\mathrm{T}} \quad \cdots \quad (\mathbf{A}_N\mathbf{E}_N)^{\mathrm{T}}\right]^{\mathrm{T}} \qquad (2)$$

* This work is supported by EPSRC grant GR/L60463

This is illustrated by Figure 1 for the case of two sub-bands. This analysis is based on the conjecture that important cues for discrimination exist in the local spectral correlates that may not be captured by the full band cepstral analysis. A multi-resolution analysis simply involves performing the above at several decomposition levels, using a different number of sub-bands at each resolution.

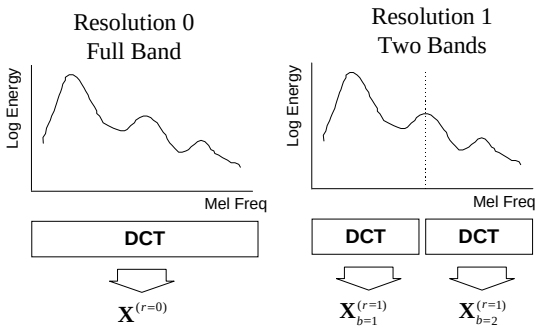


Fig 1. Sub-band Cepstral Feature Extraction

The multi-resolution features can be concatenated or the model decomposition can match the spectral decompsition by independent stream modelling, the timing of the combination of stream scores depending on whether the task is classification or recognition.

## 3. PHONETIC TEMPORAL MODELLING

### 3.1 Phonetic Segmental Features

For a given unit of speech of length $T$ vectors, identified as a phonetic unit, the phonetic features for that segment are calculated as

$$\mathbf{Y} = \mathbf{A}_T \mathbf{X} \qquad (3)$$

where $\mathbf{X} = [\mathbf{x}_t, \ldots, \mathbf{x}_{t+T-1}]$ is the segment and $\mathbf{A}_T$ is a transformation dependent on the segment length $T$. $\mathbf{A}_T$ is the $T$ length DCT used to decode the transitional dynamics across the duration of the phonetic event. $\mathbf{Y}$ hence denotes the phonetic features for that segment and is derived via a DCT on the stacked cepstral vectors $\mathbf{X}$ as

$$c(n,m) = \frac{1}{T}\sum_{k=0}^{T-1} c_n(k) \cdot Cos\left(\frac{(2k+1)m\pi}{2T}\right) \qquad (4)$$

where $c_n(k)$ is the $n$th coefficient of the $k$th cepstral vector in the segment of $T$ conventional MFCC vectors. This is illustrated in Fig 2.

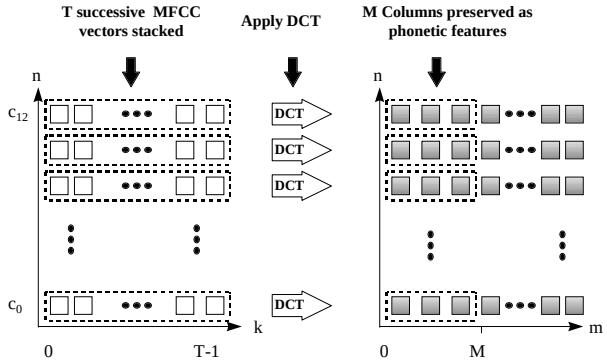


Fig 2. Transformation Across A Segment of MFCC Vectors to Yield Phonetic Features

The first $M$ columns of the matrix are preserved as phonetic features for the complete segment such that a fixed length representation is yielded from variable length sequences.

### 3.2 Segment based Phonetic Model

The proposed phonetic model in Fig 3 is similar to the standard monophone HMM. There are however no transition probabilities associated between states.


Fig 3. Phonetic Model Topology

The model as a whole is used to model a segment which corresponds closely to a phonetic event. The beginning and end state model the conventional cepstrum feature frames bounding the segment. The middle or phonetic state is dedicated to modelling the segmental phonetic features derived across the duration of that segment. Given a segment $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ with known boundaries, the likelihood for that segment when using the full phonetic model $\lambda_\alpha$ is expressed as

$$P(\mathbf{X}|\lambda_\alpha) = P(\mathbf{x}_1|s_b) \cdot P\left(Y(2,T-1)\middle|s_{ph}\right) \cdot P(\mathbf{x}_T|s_e) \cdot P(T|\lambda_\alpha) \qquad (5)$$

where $Y(2,T\text{-}1)$ are the transformed phonetic features derived across the vectors $x_2 \ldots x_{T-1}$ with the individual scores for the beginning frame, the phonetic feature columns and the end frame conditioned on the relevant states. The factor $P(T|\lambda_\alpha)$ is a measure of the probability of that segment having a duration of $T$ frames and is calculated from a

gamma distribution derived from the duration statistics. The dynamic range of scores from the phonetic feature terms is fixed for segments of variable duration. In standard HMMs, the log likelihood score for a segment is accumulated across successive frames. With the phonetic model only a segment score is calculated

### 3.3 Sub-Segmental Phonetic Features

Multi-resolution temporal feature extraction as defined here mirrors that applied in the spectral domain i.e. segmental features are to be extracted within localised time segments. Our initial sub-segment decomposition and feature extraction experiments are with equal length sub-segments. Sub-segmentation could be hypothesised by forced state alignment with well-trained HMMs, the sub-sgmental features then effectively capturing state trajectory dynamics. This however necessarily implies joint decoding with HMMs to extract representative sub-segment features during recognition.

## 4. MODEL COMBINATION

Let $\lambda_j^{(rb)}$ denote independent phoneme models for each band *b* and resolution *r*, and $\lambda_j^{Seg}$ an independent hybrid segmental model, for a phoneme $j$ Given the definitions

$$B_j^{(rb)}(\mathbf{X}^{(rb)}) = \log P(\mathbf{X}^{(rb)} | \lambda_j^{(rb)}) \qquad (6a)$$

$$B_j^{Seg}(\mathbf{X}) = \log P(\mathbf{X} | \lambda_j^{Seg}) \qquad (6b)$$

the log-likelihood score of the segment belonging to class $j$ is

$$g_j(\mathbf{X}) = \left[ \sum_{r=1}^{R} \sum_{b=1}^{B} \omega_j^{(rb)} B_j^{(rb)}(\mathbf{X}^{(rb)}) \right] + \omega_j^{(Seg)} B_j^{Seg}(\mathbf{X}) \quad (7)$$

Minimum Classification Error (MCE) training [6] can be used to train the weights $\omega_j^{(rb)}$ and $\omega_j^{(Seg)}$. The misclassification measure $d_k(\mathbf{X})$ for an observation known to belong to class *k* is given by

$$\begin{aligned} d_k(\mathbf{X}) &= -g_k(\mathbf{X}) + \max_{j \neq k} g_j(\mathbf{X}) \\ &= -g_k(\mathbf{X}) + g_\eta(\mathbf{X}) \end{aligned} \qquad (8)$$

where $\eta$ represents the model with the nearest score ie. the most confusable class. A smoothed continuous loss function is defined as a sigmoidal function of $d_k(\mathbf{X})$

$$\Gamma_k(\mathbf{X}) = \frac{1}{1 + e^{-\gamma d_k(\mathbf{X})}} \qquad (9)$$

with "good" classification tending towards zero and incorrect tending toward one. $\gamma$ controls the slope of the sigmoid function. Generalised Probabilistic Descent (GPD) (or stochastic descent) token-by-token training implies the gradient of the local class-specific loss function drives the parameter updates [6]. Thus the weight update equation is

$$\omega_k^{n+1} = \omega^n - \varepsilon \frac{\partial \Gamma_k(\mathbf{X})}{\partial \omega_k^n} \qquad (10)$$

where $\omega_k^n$ is a model weight after the i[th] iteration, $\partial \Gamma_k(\mathbf{X}) / \omega_k$ is the gradient of the local loss function and $\varepsilon$ is a small positive learning constant. The gradient function is expanded according to the chain rule of calculus. The update equations for the n+1th iteration are summarised below:

$$\omega_j^{(rb)^{n+1}} = \omega_j^{(rb)^n} - \varepsilon(\Gamma_j(\mathbf{X})[\Gamma_j(\mathbf{X}) - 1]) B_j^{(rb)}(\mathbf{X}^{(rb)})$$

$$\omega_\eta^{(rb)^{n+1}} = \omega_\eta^{(rb)^n} + \varepsilon(\Gamma_j(\mathbf{X})[\Gamma_j(\mathbf{X}) - 1]) B_\eta^{(rb)}(\mathbf{X}^{(rb)})$$

$$\omega_j^{(Seg)^{n+1}} = \omega_j^{(Seg)^n} - \varepsilon(\Gamma_j(\mathbf{X})[\Gamma_j(\mathbf{X}) - 1]) B_j^{Seg}(X)$$

$$\omega_\eta^{(Seg)^{n+1}} = \omega_\eta^{(Seg)^n} + \varepsilon(\Gamma_j(\mathbf{X})[\Gamma_j(\mathbf{X}) - 1]) B_\eta^{Seg}(X)$$

## 5 EXPERIMENTAL RESULTS

### 5.1 Phonetic Segmental Model

| Mix/Duration | Col 0_3 | Col 0_4 | Col 0_5 |
|---|---|---|---|
| 36mix/no duration | 69.29% | 69.34% | 68.97% |
| 36mix/duration | 70.31% | 70.41% | 70.12% |
| 40mix/no duration | 69.22% | 68.94% | 69.19% |
| 40mix/duration | 70.45% | 69.98% | 70.06% |
| 44mix/no duration | 69.18% | 68.83% | 69.04% |
| 44mix/duration | 70.44% | 70.19% | 70.21% |
| 48mix/no duration | 69.27% | 68.45% | 68.84% |
| 48mix/duration | 70.63% | 69.98% | 69.88% |

Table 1. Phonetic Model TIMIT Classification

Results are presented for the core test set of the TIMIT database. MFCC vectors from which the phonetic features are calculated are extracted with a window-length of 15ms at frame rate of 1.5 ms in order to have sufficient data from which to consistently extract *M* columns. Table 1 below shows results for different numbers of columns retained (0_3 means the first four columns Fig 2). The re-

sults clearly show that at most 3 columns should be retained. A consistent performance improvement is shown when the duration modelling term equation (7) is included. The best result of 70.63% gives a satisfying improvement on a baseline score of 66.4% using HMMs trained on standard MFCCs with delta and acceleration coefficients.

### 5.2 Multi-Resolution Temporal Features

Results presented in Table 2 are for the male only portion of the TIMIT training and test set, and are for a single state model trained purely on the phonetic features with the first three columns. Whilst performance initialy fell if the zero'th column from each of three equal length sub-segmental feature matrices where retained (denoted Col_02_0), a significant gain in performace is achieved if columns one and two across the three sub-segments are concatenated to the segmental features. Good spectral-time trajectory modelling is thus achieved at both the phonetic level and sub-phonetic level

| Mix | Col_02 | Col_02_0 | Col_02_01 | Col_02_02 |
|---|---|---|---|---|
| 1 | 52.78 | 53.63 | 55.82 | 55.58 |
| 15 | 61.64 | 58.10 | 62.42 | 63.48 |
| 24 | 61.40 | 58.48 | 62.70 | 64.09 |
| 28 | 62.22 | 58.81 | 62.83 | 63.92 |
| 36 | 61.87 | 58.98 | 63.53 | 64.48 |

Table 2 Sub-Segmental Feature Concatenation

### 5.3 Model Combination

The model combination experiments reported Table 3 are for the full TIMIT test set. In the case of the full-band and sub-band models the cepstral features stated were supplemented by delta and delta-delta coefficients in training 20 mixture CDHMMs. Linear combination of the classification scores from these models demonstrates improvement over full-band alone. Inclusion of scores for the segmental model, of 36 mixtures and trained on three columns only, gives additional improvement. Discriminatively trained weight re-combination (as denoted by the * last row) improves classification to 72.16%. This is a significant improvement over the baseline result of 66.4 % and clearly demonstrates that the combined models scores extend the discriminative performance achieved by separate models. Approaches to segmental model combination for continuous recognition are currently being investigated.

| Bandwidth (kHz) | Cepstral Analysis | Classification (%) |
|---|---|---|
| 0-7.9 | (13) | 66.40 |
| 0-2 | (7) | 59.31 |
| 2-7.9 | (7) | 45.87 |
| Segmental | (39) | 66.91 |
| 0-7.9, 0.2, 2-7.9 | (13)+(7,7) | 69.77 |
| 0-7.9, 0.2, 2-7.9, Seg | (13)+(7,7)+(39) | 71.20 |
| 0-7.9, 0.2, 2-7.9, Seg* | (13)+(7,7)+(39) | 72.16 |

Table 3. Model Combination Results

## 6. CONCLUSIONS

The hybrid phonetic model, based on boundary MFCC vectors and fixed dimension across-time segmental features extracted from variable length segments of MFCC vectors, is shown to improve significantly on HMM for TIMIT phoneme classification. Multi-resolution temporal features in the form of sub-segmental transform trajectory modelling is also demonstrated to yield notable performance advantages. Direct linear combination of log likelihood scores from the segmental model and multi-resolution sub-band models using weights derived by MCE training is shown to improve discriminative potential by extending TIMIT classifciation result beyond independent modelling performances.

## 7. REFERENCES


[1] P. Beyerlin, "Discriminative Model Combination", p. 166-169, Proc. ICASSP-98

[2] G Potaminos & H. Graf, "Discriminative Training of HMM Stream Exponents for Audio-Visual Speech recognition", p. 3733-3736, Proc. IEEE ICASSP-98

[3] P. McMahon, P. McCourt, S. Vaseghi, "Discriminative Weighting of Multi-Resolution Sub-Band Cepstral Features for Speech Recognition", Proc. ICSLP-98

[4] J. Allen, "How Do Humans Process and Recognise Speech?", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, 1994, pp. 567-577

[5] Fiscus, J., " A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)", Proc. IEEE ASRU Workshop, pp 347-352, 1997

[6] B. Juang, W. Chou, C. Lee, "Minimum Classification Error Rate Methods for Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 3, May 1997