# DISCRIMINATIVE SPECTRAL-TEMPORAL MULTI-RESOLUTION FEATURES FOR SPEECH RECOGNITION

*P. McMahon N. Harte S. Vaseghi P. McCourt*

School of Electrical Engineering , The Queens University of Belfast
E-mail: (p.mcmahon, n.harte, s.vaseghi, pm.mccourt) @ ee.qub.ac.uk

## ABSTRACT

Multi-resolution features, which are based on the premise that there may be more cues for phonetic discrimination in a given sub-band than in another, have been shown to outperform the standard MFCC feature set for both classification and recognition tasks on the TIMIT database [5]. This paper presents an investigation into possible strategies to extend these ideas from the spectral domain into both the spectral and temporal domains. Experimental work on the integration of segmental models, which are better at capturing the longer term phonetic correlation of a phonetic unit, into the discriminative multi-resolution framework is presented. Results are presented which show that including this supplementary temporal information offers an improvement performance for the phoneme classification task over the standard multi-resolution MFCC feature set with time derivatives appended. Possible strategies for the extension of theses techniques into the area of continuous speech recognition are discussed.

## 1. INTRODUCTION

The multi-resolution framework which seeks to exploit discriminative cues in the spectral domain by supplementing cepstral features derived from the full frequency band-width with those obtained from smaller sub-bands, has been shown to outperform the standard MFCC feature set for both classification and recognition tasks on the TIMIT database [5]. The multi-resolution framework is based on the following premises:

- The Human Auditory System (HAS) relies on features derived from numerous overlapping sub-band filters [1],
- It appears that HAS uses an across-time muti-level processing scheme, i.e the combination of localised detailed information from a relatively small window in time with longer term temporal information
- Through a discriminative optimisation process we can model the potential for phonetic discrimination across these dependent feature processing channels

This idea that the inclusion of supplementary information provides potential for phonetic discrimination though some discriminative optimisation process, can be extended from the spectral domain into the temporal domain in an attempt to model this *across-time processing scheme.* While the conventional three state HMM offers a powerful statistical model, the assumption that within each state the observation vectors are independent and identically distributed (IID), is clearly violated by the high degree of correlation between successive vectors. The use of dynamic coefficients is a well established approach to extending conventional feature vectors to include temporal information and lessen the effect of the IID assumption. This paper presents some preliminary investigation which seeks to apply the multi-resolution model to the temporal domain by supplementing first and second order regression coefficients taken over a small number of frames, with coefficients taken over an increasingly large number of frames, thereby extending the degree of frame dependence to account for the high degree of correlation between successive vectors.

An alternative approach to overcome the weaknesses of the IID assumption is the use of segmental HMMs, which have the reverse properties of a standard HMM in that they are better at modelling the longer term phonetic correlation but miss out on localised sub-phonetic cues. This paper explores the discriminative combination of both multi-resolution and segmental models in an attempt to combine the capabilities of these different feature sets.

## 2. MULTI-RESOLUTION SPECTRAL TEMPORAL FEATURES

Let $E = [E_1, E_2 \ldots E_T]$ be a sequence of log mel-filter bank energy vectors. Cepstral features are derived from a linear transformation of

$$X_t = AE_t \tag{1}$$

$A$ is conventionally the DCT, but it can be a general discriminative feature transform [7]. Multi-resolution feature vectors are a set of feature transformations such as

$$X_t = \left[A_0 E_t, \left(A_{11} E_{t11}, A_{12} E_{t12}\right), \left(A_{21} E_{t21}, A_{22} E_{t22}, A_{23} E_{t23}, A_{24} E_{t24},\right) \ldots\right]^T \tag{2}$$

$A_1 E_t$, yields the cepstral features over the whole bandwidth, $\left(A_{12} E_{t12}, A_{22} E_{t22}\right)$ yield cepstral features over, the lower half and the upper half sub-bands, and $\left(A_{14} E_{t14}, A_{24} E_{t24}, A_{34} E_{t34}, A_{44} E_{t44}\right)$ yield the features over four sub-band quadrants and so on [5]

## 2.1 Sub-Phonetic and Segmental Models

In a conventional three state HMM, the states model sub-phonetic segments of speech corresponding to the beginning, the middle and the end of each phonetic unit. In contrast segmental models capture the spectral-temporal features of a phone over its duration. A three state HMM is parameterised by a multiple-mixture Gaussian density and a Markovian state transition probability. The segmental model is rather similar to a one state HMM, with the difference that segmental modelling involves an estimate of the beginning and end of each phone that is necessary for time normalisation.

The current work uses closely related segmental phonetic features for phoneme classification. 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

$$Y = A_T X \tag{3}$$

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

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

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 used which uses a hybrid representation of a phoneme. Previous work on these features and models have shown that the features can match the performance of standard cepstrum with first and second order derivatives for a classification task on the TIMIT database.

## 2.2 Multi-Resolution Regression Coefficients

The performance of a speech recognition system can be greatly enhanced by adding time derivatives to the basic feature vector, and is now a well established approach in an attempt to include temporal information within conventional feature vectors. First order regression (delta) coefficients are calculated using the following formula

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{5}$$

where $d_t$ is a delta coefficient at time t computed in terms of the corresponding static coefficients $c_{t+\theta}$ to $c_{t-\theta}$. The value of $\Theta$ dictates the window size, or the number of preceding and succeeding vectors used to calculate the coefficients. The same formula is applied to the delta coefficients to obtain second order regression (acceleration) coefficients. By supplementing coefficients calculated with a small window size with coefficients based on a longer window size we can effectively apply the idea of including several levels of supplementary temporal information in the same manner as was applied to the multi-resolution framework in the frequency domain, based on the premise that there may be more potential for phonetic discrimination in one set of regression coefficients than another.

# 3. DISCRIMINATIVE COMBINATION

Discriminative weighting of individual multi-resolution models has been shown to outperform conventional MFCC features for the phoneme classification task on the TIMIT database [5], while segmental features have demonstrated the ability to match the performance of standard HMM with first and second order derivatives [6]. The inclusion of segmental features and models into the discriminative multi-resolution framework is therefore a natural progression in the effort to extend the multi-resolution model from the just the spectral domain to both the spectral and temporal domains.

A given segment $X$ of length T vectors will have a number of multi-resolution subband cepstral feature vectors $X^{(rb)}\{r = 1\ldots R, b = 1\ldots B_r\}$ where $r$ identifies the resolution level and $b$ the sub-band index within that resolution (for $r = 1$ indicating the full band $B_r = 1$). It will also have a fixed length segmental representation $Y$ of the phoneme derived via the process outlined in **2.1.**

Let $M_j^{(rb)}$ denote independent phoneme models for each band b and resolution r, and $S_j$ an independent hybrid segmental model for each phoneme $j$. Then the combined log likelihood for class $j$ can be given as

$$\begin{aligned}\log p\left(X \middle| M_j, S_j\right) = &\left[\sum_{r=1}^{R}\sum_{b=1}^{B_r} \omega_j^{(rb)} \log p(\mathbf{X}^{(rb)} \middle| M_j^{(rb)})\right] \\ &+ \omega_j^{(S)}\left[\log p(\mathbf{X}_1|S_j) \log p(\mathbf{Y}|S_j) \log p(\mathbf{X}_T|S_j)\right]\end{aligned} \tag{6}$$

The weights $\omega_j^{(rb)}$ should ideally reflect the discriminative potential or confidence of each multi-resolution sub-band for a particular class, and the weights $\omega_j^{(S)}$ the degree of confidence of the hybrid segmental model

In keeping with this principle, discriminative training of the weights $\omega_j^{(rb)}$ and $\omega_j^{(S)}$ is proposed, using a minimum classification error (MCE) criterion.

The following notation is used where

$$B_j^{(rb)}(\mathbf{X}^{(rb)}) = \log p(\mathbf{X}^{(rb)} \middle| M_j^{(rb)}) \tag{7a}$$

$$S_j(\mathbf{X}_1, \mathbf{Y}, \mathbf{X}_T) = \log p(\mathbf{X}_1|S_j) \log p(\mathbf{Y}|S_j) \log p(\mathbf{X}_T|S_j) \quad (7b)$$

(7a) describes the partial recognition score for a multi-resolution sub-band vector sequence $X^{(rb)}$ given a sub-band model, and (7b) the score for a segmental model. Using this the log-likelihood score of the segment belonging to class $j$ can be defined as

$$g_j(\mathbf{X}) = \left[ \sum_{r=1}^{R} \sum_{b=1}^{B} \omega_j^{(rb)} B_j^{(rb)}(\mathbf{X}^{(rb)}) \right] + \omega_j^{(S)} S_j(X_1, Y, X_T) \quad (8)$$

Let a misclassification measure $d_k(\mathbf{X})$ for a training segment belonging to class $k$ be given by

$$\begin{aligned} d_k(\mathbf{X}) &= -g_k(\mathbf{X}) + \max_{j \neq k} g_j(\mathbf{X}) \\ &= -g_k(\mathbf{X}) + g_\eta(\mathbf{X}) \end{aligned} \quad (9)$$

where $\eta$ represents the model with the nearest score ie. the most confusable class. A loss function can be defined [7] as a sigmoidal function of $d_k(\mathbf{X})$

$$\Gamma_k(\mathbf{X}) = \frac{1}{1 + e^{-d_k(\mathbf{X})}} \quad (10)$$

The loss function is minimised for each training vector by adaptively adjusting the sub-band and segmental model weights, according to

$$\omega^{i+1} = \omega^i - \varepsilon \frac{\partial \Gamma(\mathbf{X})}{\partial \omega^i} \quad (11)$$

where $\omega^i$ is the parameter value after the i$^{th}$ iteration, $\partial \Gamma(X) / \omega^i$ is the gradient of the loss function and $\varepsilon$ is a small positive learning constant. The weight update equations can be derived for $X$ belonging to class $k$ and $\eta$ being the most confusable class. The equations are presented below in equations 12a to 12d.

$$\omega_k^{(rb),i+1} = \omega_k^{(rb),i} + \varepsilon(\Gamma_k(\mathbf{X})[\Gamma_k(\mathbf{X}) - 1]) B_k^{(rb)}(\mathbf{X}^{(rb)}) \quad (12a)$$

$$\omega_\eta^{(rb),i+1} = \omega_\eta^{(rb),i} + \varepsilon(\Gamma_k(\mathbf{X})[\Gamma_k(\mathbf{X}) - 1]) B_\eta^{(rb)}(\mathbf{X}^{(rb)}) \quad (12b)$$

$$\omega_k^{(S),i+1} = \omega_k^{(S),i} + \varepsilon(\Gamma_k(\mathbf{X})[\Gamma_k(\mathbf{X}) - 1]) S_k(X_{1,}, Y, X_T) \quad (12c)$$

$$\omega_\eta^{(S),i+1} = \omega_\eta^{(S),i} + \varepsilon(\Gamma_k(\mathbf{X})[\Gamma_k(\mathbf{X}) - 1]) S_\eta(X_1, Y, X_T) \quad (12d)$$

## 4. EXTENSIONS TO CONTINUOUS SPEECH RECOGNITION

The method of discriminative combination of partial scores described in the previous section is readily applied to classification tasks. An extension to recognition involves more challenges. Results have been reported on extending multi-resolution features to recognition using state dependant weights for each model whereby each frame assigned to a particular model in Viterbi based recognition has a weight applied to the likelihood score for that frame [5]. To extend recognition to incorporate segmental features is inherently problematic. The transformation of variable length segments to a fixed length representation can potentially lead to excessive computation. This is because segment boundaries must be hypothesised if not established through some pre-processing stage. Thus the Viterbi Algorithm where the recognition lattice can be elegantly extended on a frame by frame basis is not applicable. Approaches to recognition using the phonetic model and features are explored more thoroughly in [6]. This work seeks to explore methods to introduce segmental models into a framework using segmental and multi-resolution features.

The recombination of different recognisers is complicated by difference in the alignment of transcriptions and it is not a case of simply deciding which class is correct for a particular segment. ROVER (Recogniser Output Voting Error Reduction) is a system developed at NIST to produce a composite system where outputs from many recognition systems are available. There are two stages to the system. The transcriptions from two or more systems are combined in a word transition network using a modified version of the dynamic programming alignment protocol traditionally used by NIST to evaluate ASR systems. Each branching point in the network is then evaluated with a voting scheme and the best scoring word chosen and output in the new transcription.

It would be possible to employ ROVER in this way to combine decisions from recognition based on phonetic models and features with standard HMM models. However, this would not be the optimal way of combining segmental models with standard HMMs. The transcriptions from the segmental models have a much lower accuracy and overall would not be suitable for a voting scheme which relies on a certain amount of similitude across transcriptions. This would not properly exploit the complimentary strengths and weaknesses of segmental modelling, multi-resolution features and HMM models. It would be more advantageous to devise a rescoring scheme that given a reduced set of possible segmentations through a lattice output from HMM models used the phonetic features and models and multi-resolution features to reassess the output of the HMM and truly combine the abilities of the different modelling paradigms.

Previous experimentation has shown that the highest level of accuracy achieved has been with the combination of multi-resolution features from a number of bands. These features and models could be used to output a lattice for each sentence to enable the revaluation of the n-best hypotheses for a sentence. The lattice consists of a series of links and nodes which correspond to definite phonemes of associated log likelihood. The score for each link can be adjusted by extracting the segmental features for each link and calculating the likelihood for the appropriate model. Discriminative weights could be used to determine how the likelihood scores should be combined. The paths are then reassessed to obtain the best path.

# 5. EXPERIMENTAL RESULTS

## 5.1 Discriminative Combination

Experiments were performed to assess the potential of discriminatively combining the multi-resolution and segmental feature sets using 39 context-independent 20 mixture HMM models for each multi-resolution sub-band and 36 mixtures for the phonetic models. The full TIMIT training and test sets were used throughout, with the exception that classified phonemes of less than five frames were excluded from the experiments. Previous experimentation [7] showed that while supplementing the full band cepstra with either 2 or 4 sub-bands gave improved results, use of both resolution levels was seen to yield no further advantage. For the purpose of these experiments therefore, three multi-resolution bands are used - a full band supplemented with two sub-bands.

| Bandwidth (kHz) | Cepstral Analysis | Classification (%) |
|---|---|---|
| 0-7.9 | (13) | 68.88 |
| 0-2 | (7) | 59.31 |
| 2-7.9 | (7) | 45.87 |
| Segmental | (39) | 65.46 |
| 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 (1)** 'Multi-Resolution/Segmental Combination'

Table (1) shows the individual results for each of the sub-band, segmental, multi-resolution features and then the performance when the feature sets are linearly combined. Earlier work has shown that multi-resolution recombination is shown to yield an increase when each band is given equal weightings, but the inclusion of segmental models into the framework improves the classification result further. Discriminatively training recombination weights according to the process outlined in **3.** for both multi-resolution and segmental models extends this improvement (indicated in the above table by an asterix beside the sub-band boundary). This shows that the assumption that segmental models can be integrated into the multi-solution framework, and can contribute discriminative information which is not present within the MFCC feature set to hold true.

# 6.0 CONCLUSIONS

Multi-resolution features strive to utilise and combine the discriminative features of speech in both time and frequency at both localised sub-phonetic levels and across longer length segmental phonetic levels. The premise that cues for phonetic discrimination may exist in one part of the spectral domain but not another may be applied to the temporal domain, leading to the conjecture that multi-resolution temporal features and models may be able to exploit discriminative cues in localised regions of the temporal domain. Linearly weighted inclusion of segmental models into the multi-resolution framework offers an improvement in performance for the phoneme classification task on the TIMIT database, highlighting the potential for the inclusion of supplementary temporal information. Future work will explore the issue of discriminative multi-resolution segmental transforms which seek to exploit discriminative cues in both the time and frequency domains, as well as experimentation with various levels of regression coefficients in an attempt to model pertinent temporal information.

# 7.0 REFERENCES


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

[2] S. Tibrewala & H. Hermansky, "Sub-band Based Recognition of Noisy Speech", Proc. ICASSP-97, Vol. 2, pp. 1255-1258

[3] H. Bourlard, S. Dupont, H. Hermansky, N. Morgan, "Towards Sub-Band based Speech Recognition", Proc. EUSIPCO-96, pp. 1579-1582

[4] R. Chengalvarayan & L. Deng, "HMM-Based Speech Recognition Using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features", IEEE Trans. on Speech and Audio Processing, Vol. 5, No. 3, 1997, pp. 243-256

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

[6] N. Harte, S. Vaseghi, B. Milner, "Joint Recognition and Segmentation using Phonetically Derived Features and Hybrid Phoneme Model", Proc. ICSLP-98

[7] P. McCourt, S. Vaseghi, N. Harte, "Discriminatively Weighted Multi-Resolution Cepstra for Phonetic Recognition", Proc. ICASSP-98, Vol. 1, pp. 577 - 580

[8] 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

[9] B. Juang, S. Katagiri, "Discriminative Learning for Minimum Error Classification", IEEE Transactions on Signal Processing, Vol. 40, No. 12, December 1992