DISCRIMINATIVE WEIGHTING OF MULTI-RESOLUTION SUB-BAND CEPSTRAL FEATURES FOR SPEECH RECOGNITION

Philip McMahon *Paul McCourt* *Saeed Vaseghi*
The Queens University of Belfast, N. Ireland
E-mail : (p.mcmahon, pm.mccourt, s.vaseghi)@ee.qub.ac.uk

ABSTRACT

This paper explores possible strategies for the recombination of independent multi-resolution sub-band based recognisers. The multi-resolution approach is based on the premise that additional cues for phonetic discrimination may exist in the spectral correlates of a particular sub-band, but not in another. Weights are derived via discriminative training using the 'Minimum Classification Error' (MCE) criterion on log-likelihood scores. Using this criterion the weights for correct and competing classes are adjusted in opposite directions, thus conveying the sense of enforcing separation of confusable classes. Discriminative re-combination is shown to provide significant increases for both phone classification and continuous recognition tasks on the TIMIT database. Weighted recombination of independent multi-resolution sub-band models is also shown to provide robustness improvements in broadband noise.

1. INTRODUCTION

In recent years there have been a number of papers on sub-band based speech recognition [3, 4], inspired by Allen's paper reviewing the earlier work of Fletcher [1]. The central conclusion of this work is the proposition that the human auditory system relies on the recognition of independent spectral-temporal features. The multi-resolution approach extends the purely sub-band approach by supplementing, rather than substituting, conventional full-band MFCC's with more detailed sub-band cepstral features. Experimentation with multi-resolution cepstral features based on concatenation of cepstral vectors from a number of sub-bands, outperforms conventional MFCC features for the continuous phoneme recognition task on the TIMIT database [2].

As an extension to this earlier work, a new approach presented and evaluated in this paper is to combine the log-likelihood scores of independent sub-band acoustic models. As a result of such an approach a number of issues need to be addressed, including possible strategies for the recombination or merging of the individual sub-band/model recogniser scores. While non-linear recombination has been explored via the use of a 'Multi Layer Perceptron' (MLP) [3], the approach adopted here is the principle of linearly weighting confidence for each sub-band recogniser according to its discriminative potential, based on the premise that additional cues for phonetic discrimination may exist in the spectral correlates of a particular sub-band, but not in another. The recombination weights should ideally reflect the contribution of each sub-band for discrimination of a particular class. In keeping with this principle, weights are derived via discriminative training using the 'Minimum Classification Error' (MCE) criterion on log-likelihood scores. Using the MCE criterion, weights for the correct and competing classes are adjusted in opposite directions, thus conveying the sense of enforcing separation of confusable classes

This principle of deriving recombination weightings can be extended to derive state-based weights for each multi-resolution sub-band hidden markov model. This is to acknowledge that the spectral information across sub-bands will be different for the states of a multi-resolution HMM .

Extending the multi-resolution cepstral decomposition from the feature space to the model space also gives the possibility of improving robustness in noisy conditions by exploiting variations of sub-band SNR to weight the reliability or confidence of the partial information from each recogniser. We present some results which demonstrate this advantage in white noise.

2. MULTI-RESOLUTION SUBBAND FEATURES

Let $\boldsymbol{E} = [\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_T]$ be a sequence of log mel-filter bank energy vectors. Cepstral features are derived from a linear transformation of

$$\mathbf{X}_t = \mathbf{AE}_t \qquad (1)$$

$\boldsymbol{A}$ is conventionally the DCT, but it can be a general discriminative feature transform [5]. Multi-resolution feature vectors are a set of feature transformations such as

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

$\boldsymbol{A}_1 \boldsymbol{E}_t$, yields the cepstral features over the whole bandwidth, ($\boldsymbol{A}_{12} \boldsymbol{E}_{t12}$,$\boldsymbol{A}_{22} \boldsymbol{E}_{t22}$) yield cepstral features over, the lower half and the upper half subbands, and ($\boldsymbol{A}_{14} \boldsymbol{E}_{t14}$,$\boldsymbol{A}_{24} \boldsymbol{E}_{t24}$, $\boldsymbol{A}_{34} \boldsymbol{E}_{34}$,$\boldsymbol{A}_{44} \mathrm{E}_{t44}$) yield the features over four subband quadrants and so on.

3. MODEL RECOMBINATION STRATEGIES

3.1 Discriminative Class Based Weightings

Consider the multi-resolution subband cepstral feature vectors $\mathbf{X}^{(rb)}$ {r=1,..,R; b=1,..,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). If we associate independent models $M_j^{(rb)}$ for each band b within resolution r, the combined log likelihood for class j can be given as

$$\log p(\mathbf{X}|M_j) = \sum_{r=1}^{R}\sum_{b=1}^{B_r} \omega_j^{(rb)} \log p(\mathbf{X}^{(rb)}|M_j^{(rb)}) \tag{3}$$

The multi-resolution sub-band weights $\omega_j^{(rb)}$ should ideally reflect the discriminative potential or confidence of each sub-band for a particular class. Fully independent models $M_j^{(rb)}$ will have separate state transition probability matrices. However for our initial experiments the state transition probabilities are effectively tied for the sub-band models of each phoneme. In keeping with this principle we propose to perform discriminative training of the weights $\omega_j^{(rb)}$ using a minimum classification error (MCE) criterion eg. [5]. Let

$$B_j^{(rb)}(\mathbf{X}^{(rb)}) = \log p(\mathbf{X}^{(rb)}|M_j^{(rb)}) \tag{4}$$

describe the partial recognition score for a sub-band vector $\mathbf{X}^{(rb)}$ given a sub-band model, we define the log-likelihood score of the full parameter vector belonging to class j as

$$g_j(\mathbf{X}) = \sum_{r=1}^{R}\sum_{b=1}^{B_r} \omega_j^{(rb)} B_j^{(rb)}(\mathbf{X}^{(rb)}) \tag{5}$$

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

$$\begin{aligned} d_k(\mathbf{X}) &= -g_k(\mathbf{X}) + \max_{j\neq k} g_j(\mathbf{X}) \\ &= -g_k(\mathbf{X}) + g_\eta(\mathbf{X}) \end{aligned} \tag{6}$$

where η represents the model with the nearest score ie. the most confusable class. A loss function can be defined [5] as a sigmoidal function of $d_k(\mathbf{X})$

$$\Gamma_k(\mathbf{X}) = \frac{1}{1+e^{-d_k(\mathbf{X})}} \tag{7}$$

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

$$\omega^{n+1} = \omega^n - \varepsilon \frac{\partial\Gamma(X)}{\partial\omega^n} \tag{8}$$

where ω^n is the parameter value after the nth iteration, $\partial\Gamma(X)/\omega^n$ is the gradient of the loss function and ε is a smalll positive learning constant. For the sake of brevity the weight update equations are quoted without direct derivation of the gradient function as follows, for **X** belonging to class k and η being the most confusable class.

$$\omega_k^{(rb),n+1} = \omega_k^{(rb),n} - \varepsilon(\Gamma_k(\mathbf{X})[\Gamma_k(\mathbf{X})-1])B_k^{(rb)}(\mathbf{X}^{(rb)}) \tag{9a}$$

$$\omega_\eta^{(rb),n+1} = \omega_\eta^{(rb),n} + \varepsilon(\Gamma_k(\mathbf{X})[\Gamma_k(\mathbf{X})-1])B_\eta^{(rb)}(\mathbf{X}^{(rb)}) \tag{9b}$$

3.2 Discriminative State Based Weightings

The issue of at which segmental level to recombine log-likelihood scores is one which has as yet proved inconclusive [3]. The recombination criterion at a model level outlined above is advantageous for phoneme classification, but does not address the issues raised with regards to continuous speech recognition.

The MCE criterion outlined above can be extended to derive state based weightings for each sub-band model, which in turn can be applied as state based stream weights, thus allowing the discriminative multi-resolution weights to be used within the standard HMM framework for continuous speech recognition. Given the sequence of multi-resolution feature vectors

$$\mathbf{X}^{(rb)} = [\mathbf{X}_1^{(rb)}, \mathbf{X}_2^{(rb)}, \ldots, \mathbf{X}_T^{(rb)}] \tag{10}$$

the optimal state sequence for a sub-band model j is defined as

$$\Theta_j^{(rb)} = [\theta_1, \theta_2, \theta_3, \ldots, \theta_T] \tag{11}$$

such that θ_t represents the state associated with feature vector $\mathbf{X}_t^{(rb)}$. Without taking into account the state-to-state transition probabilities, the partial classification score for $\mathbf{X}^{(rb)}$, given a sub-band model j and state segmentation $\Theta_j^{(rb)}$ is then

$$B_j^{(rb)}(\mathbf{X}^{(rb)}, \Theta_j^{(rb)}) = \sum_{t=1}^{T} \omega_{j,\theta_t}^{(rb)} \log B_{j,\theta_t}^{(rb)}(\mathbf{X}_t^{(rb)}) \tag{12}$$

where $\omega_{j,i}^{(rb)}$ represents the linear weight associated with class j for sub-band b (of resolution r) and state i. If $T_{i,j}$ is defined to represent the set of time indices such that the state association of the feature vector $\mathbf{X}_t^{(rb)}$ belongs to state i, and where N represent the number of states in the model k, i.e.

$$T_{i,j} = \{t | \theta_t = i\} \quad 2 \le i \le N-1, \; 1 \le t \le T \qquad (13)$$

The weight update equations of (9a) and (9b) are refined to

$$\omega_{k,i}^{(rb),n+1} = \omega_{k,i}^{(rb),n} - \varepsilon(\Gamma_k[\Gamma_k - 1]) \sum_{t \in T_{i,k}} \log \; B_{k,t}^{(rb)}(\mathbf{X}_t^{(rb)}) \quad (14a)$$

$$\omega_{\eta,i}^{(rb),n+1} = \omega_{\eta,i}^{(rb),n} + \varepsilon(\Gamma_k[\Gamma_k - 1]) \sum_{t \in T_{i,\eta}} \log \; B_{\eta,t}^{(rb)}(\mathbf{X}_t^{(rb)}) \quad (14b)$$

3.3. SNR Weighted Recombination

An advantage of splitting the spectral information into sub-bands is that variations in sub-band SNR may be exploited for improved recognition in noisy conditions. Thus by weighting the confidence in each multi-resolution sub-band stream according to its SNR, the influence of low SNR information can be reduced with a corresponding shift to reliance on partial recognition from higher SNR regions of the spectrum. Thus equation (3) can be refined to

$$\log p(\mathbf{X} | M_j) = \sum_{r=1}^{R} \sum_{b=1}^{B_r} \omega_j^{(rb)} (SNR_j^{(rb)}) \log p(\mathbf{X}^{(rb)} | M_j^{(rb)}) \quad (15)$$

where $\omega_j^{(rb)}(SNR_j^{(rb)})$ specifies the sub-band weighting to be a function of the local SNR (for band b in resolution level r) for model j. We initially have experimented with Weiner -type weightings of the following form

$$\omega_j^{(rb)} = \frac{S_j^{(rb)}}{S_j^{(rb)} + N^{(rb)}} \qquad (16)$$

$S_j^{(rb)}$ defines the signal power in sub-band b of resolution r for the phoneme class j. This value is obtained experimentally by averaging the energy within each sub-band over all occurrences of each particular phoneme across the TIMIT database. $N^{(rb)}$ specifies the noise energy within a sub-band. As the spectral characteristics within each state of a phonetic HMM are different a refinement to the weighting functions would be to make them, not only model dependent, but also state dependent.

4. EVALUATIONS

The performance of the multi-resolution cepstral feature set was tested on the TIMIT speech database using 39 context-independent 12 mixture HMM models for each multi-resolution sub-band. The full TIMIT training and test sets were used throughout, with the exception that classified phonemes of less than three frames were excluded from the experiments. Previous experimentation [2] 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.

4.1 State-Independent Weighting of Multi-Resolution Sub-bands

Table (1) gives results for phoneme classification using independent multi-resolution sub-band models, along with the sub-band boundaries implemented. The fourth row indicates the linear recombination of all three multi-resolution bands. The rightmost column of the table shows the results of discriminatively weighting each sub-band class with twelve epochs of MCE criterion based via the strategy outlined in Section 3.1. The column headed 'Cepstral Analysis' indicates the number of mel-filtered cepstral coefficients extracted from each band before delta and acceleration coefficients are appended. One point of interest is that the classification results improve even for sub-band only discrimination, showing the potential for between band discrimination. Multi-resolution sub-band recombination is shown to yield an increase when each band is given equal weightings, and when combination is based upon discriminatively trained weights, the final improvement in classification score is significantly above the original full-band MFCC classification score.

Bandwidth (kHz)	Cepstral Analysis	Equal Weightings	MCE Derived Weightings
0-7.9	(13)	65.62	67.6
0-2	(7)	56.72	58.4
2-7.9	(7)	44.05	45.3
0-7.9, 0-2, 2-7.9	(13)+(7,7)	67.04	70.0

Table (1) 'TIMIT State-Independent Classification Results'.

Bandwidth (kHz)	Cepstral Analysis	Recognition (%)
0-7.9	(13)	68.8
0-2,2-7.9	(7,7)	69.9
0-7.9,0-2,2-7.9	(13)+(7,7)	70.6

Table (2) Concatenated Cepstra Recognition Results

4.2 State Dependent Weighting of Multi-Resolution Sub-bands

Table (2) shows the results for models trained using concatenated multi-resolution cepstral feature vectors, with Table (3) giving results when weighting independent multi-resolution sub-band model states for continuous phoneme recognition. State based weights were trained on classified data as outlined in Section 3.2 using twelve epochs of discriminative training. As demonstrated by Table (2), there is some improvement in performance using sub-band cepstral features alone, compared to the full bandwidth cepstra, with further improvement in recognition performance when the multi-resolution features sets are employed.

Bandwidth (kHz)	Cepstral Analysis	Recognition (%)
0-7.9,0-2,2-7.9	(13)+(7,7)	70.21
0-7.9*,0-2,2-7.9	(13)+(7,7)	72.9
0-7.9*,0-2*,2-7.9	(13)+(7,7)	73.0
0-7.9*,0-2*,2-7.9*	(13)+(7,7)	72.4

Table (3) Weighted Independent Stream Recognition Results

Cepstral Analysis	Concatenated Features	Equally Weighted Streams	Wiener-Weighted Streams
(13)	37.1	-	-
(7,7)	35.9	34.7	40.0
(13)+(7,7)	37.7	36.6	44.4
(5,5,4,2)	34.6	33.5	36.5
(13)+(5,5,4,2)	36.0	36.0	44.2

Table (4) Recognition Results in Noise

The first result of Table(3) is produced when all the model-dependent state weights are set equal to unity. For the remaining results, an asterix beside the sub-band boundary indicates that discriminatively trained stream weights were applied to that multi-resolution sub-band whilst the other sub-bands were left unweighted. One point of interest is that while weighting one or two bands offers an improvement in performance, weighting all three bands offers no further improvement, attributable to the fact that the sub-band weights are trained independently of each other and may be at times conflicting. Nonetheless the state-based model weighting achieves a further significant performance improvement beyond that produced by using straightforward concatenated multi-resolution feature vectors over full-band MFCCs.

4.3 Independent Stream Weighting in Noise

We have experimented using the fixed Weiner type weightings (16) for recombination of independent multi-resolution sub-band streams according to (15) for continuous recognition. The results are for performance in white noise with a signal to noise ratio of 15dB.

In obtaining values for the stream weights of each phonetic HMM, the sub-band signal powers for each monophone were averaged over their occurrences across the full TIMIT training set. Whilst the weights are model-dependent, for these initial experiments they are initially the same for each state within a model. The results quoted are based on the core TIMIT test set. The results indicate the benefits of separate sub-band model weighting over concatenated multi-resolution feature vectors in improving on the MFCC score. The advantage of a multi-resolution as opposed to a purely sub-band approach is also seen. The number of sub-bands used does not change the weighted multi-resolution result significantly. The application of state-based weights is still to be explored along with the possibility of time-varying weights for non-stationary noise.

5. CONCLUSIONS

Multi-resolution sub-band cepstral features strive to exploit discriminative cues in localised regions of the spectral domain by combining HMM models trained on full band-width cepstral features with HMM models trained on cepstral features derived from several levels of sub-band decomposition. Linear weighted recombination of these independent classifiers is shown to outperform conventional MFCCs for phoneme classification on the TIMIT database. The recombination weights should ideally reflect the contribution of each sub-band for discrimination/recognition of a particular class. In keeping with this principle, weights are derived via discriminative training using the 'Minimum Classification Error' (MCE) criterion on log-likelihood scores at both a class and state level. Discriminatively weighted recombination yields a further improvement for TIMIT phoneme classification, while state-dependent stream weighting offers a similar improvement for continuous phoneme recognition. By exploiting the sub-band variations in signal to noise ratio for linearly weighted recombination of the log likelihood probabilities improved phoneme recognition performance in broadband noise is also obtained. This is an advantage over a purely sub-band approach using non linear recombination which is robust only to narrow band noise.

Recent papers [5] have shown that using discriminative methods such as the MCE criterion to derive optimal linear transforms is superior in performance to using a global fixed linear transform such as the DCT. This, coupled with the extension of the multi-resolution model from the spectral domain into both the spectral and temporal domains, beginning with the discriminatively weighted inclusion of segmental models into the multi-resolution framework, will provide the direction of research for the near future.

6. 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] P. McCourt, S. Vaseghi, N. Harte , "Multi-Resolution Cepstral Features for Phoneme Recognition Across Speech Sub-bands ", Proc. ICASSP-98, Vol.1, pp.577-580.

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

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

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