MULTI-RESOLUTION CEPSTRAL FEATURES FOR PHONEME RECOGNITION ACROSS SPEECH SUB-BANDS

Paul McCourt *Saeed Vaseghi* *Naomi Harte*

School of Electrical Engineering and Computer Science, The Queens University of Belfast, N. Ireland
E-mail : (s.vaseghi, pm.mccourt,n.harte)@ee.qub.ac.uk

ABSTRACT

Multi-resolution sub-band cepstral features strive to exploit discriminative cues in localised regions of the spectral domain by supplementing the full bandwith cepstral features with sub-band cepstral features derived from several levels of sub-band decomposition. Mult-iresolution feature vectors, formed by concatenation of the subband cepstral features into an extended feature vector, are shown to yield better performance than conventional MFCCs for phoneme recognition on the TIMIT database. Possible strategies for the recombination of partial recognition scores from independent multi-resoltuion sub-band models are explored. By exploiting the sub-band variations in signal to noise ratio for linearly weighted recombination of the log likelihood probabilities we obtained improved phoneme recognition performance in broadband noise compared to MFCC features. This is an advantage over a purely sub-band approach using non linear recombination which is robust only to narrow band noise.

1 INTRODUCTION

The choice of any acoustic feature set for speech recognition is motivated by its potential for increased class seperability, the success of which is ultimately reflected in a reduction in recognition error rate. In recent work we presented promising results using multi-resolution feature sets in both time and frequency. The structure of the multi-resolution spectral feature set [1] was to supplement the familiar Mel-filterbank cepstral coefficients (MFCC) taken over the full spectral bandwidth, with cepstral analysis of the mel-filterbank log energies grouped into sub-bands eg. a low band (0-2000Hz) and a high band (2000-7900Hz). This is based on the conjecture that important additional cues for phonetic discrimination may exist in the local spectral correlates that are not captured by the full band cepstral analysis. The level of sub-band decomposition and subsequent cepstral analysis can be increased such that features may be selected from a pyramid or hierarchy of resolution levels.

In recent years there has been a number of papers on sub-band based recognition [2,3,4,5]. These have been primarily inspired by Allens paper [6] summarising important conclusions on experiments conducted by Fletcher, during the early part of this century, on the nature of Human Speech Recognition (HSR). The central conclusion of this work is the proposition that the human auditory system relies on the recognition of independent spectral-temporal features, merged at some higher processing level into recognition of basic phonemes, and subsequent syllables, words etc. Results reported on the merging of partial recognition scores from independent HMM based sub-band recognisers using an MLP [2] concluded that, while no increase in basic recognition was achieved, the system is more robust in narrowband noise conditions.

While a similarity exists with the purely sub-band based approaches of [2] and [3], there are important distinctions and extensions underlying the motivation for multi-resolution features. The first is that the local spectral or sub-band information supplements, rather than substitutes, full band-width discriminative information. One of the design choices for sub-band based schemes is the number of bands and the exact sub-band boundary decomposition. In an optimal sense the sub-band boundaries and hence cepstral analysis lengths should be class specific. No model formulation or method of training yet exists however where the varying transform lengths for sub-band feature extraction can be expressed as a trainable parameter. A possible advantage of the multiresolution feature set is that the inclusion of different resolutions of sub-band decomposition in effect relaxes the restriction of using a single fixed sub-band decomposition.

Our initial presentation and experimentation with multi-resolution cepstral features is based on concatenation of cepstral features from each multi-resolution sub-band to form a single extended feature vector. An alternative to this approach is to combine the likelihood scores of *independent* resolution and sub-band acoustic models. The recombination must in some way reflect discriminative confidence between the individual sub-band and resolution models, within and between each phone class. MLP-based recombination [2] attempts to create re-combination discriminative functions for each phoneme model over all other phonetic sub-band model scores. Here we have experimented with weighted combination of the log likelihood scores of the multi-resolution sub-band models for each phoneme class, based on the principle of weighting confidence in each sub-band recogniser according to its discriminative potential and its SNR. Experiments with phoeneme dependent weighting of the sub-band and resolution log probability scores to improve robustness in broadband noise in particular are described. Further refinements include employing state-dependent weighting and possible weight adaptation based on time-varying sub-band SNR of speech.

0-7803-4428-6/98 $10.00 © 1998 IEEE

2 MULTI-RESOLUTION SUBBAND FEATURES

The multi-resolution cepstra combines sets of features offering different trade-offs between the spectral resolution, the variance and the class separability. 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

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

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

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

$\boldsymbol{A}_1\boldsymbol{E}_t$, 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 $(\boldsymbol{A}_{14}\boldsymbol{E}_{t14}, \boldsymbol{A}_{24}\boldsymbol{E}_{t24}, \boldsymbol{A}_{34}\boldsymbol{E}_{34}, \boldsymbol{A}_{44}\mathbf{E}_{t44})$ yield the features over four subband quadrants and so on. Fig(1) illustrates the first three basis functions in each sub-band for these resolution levels.

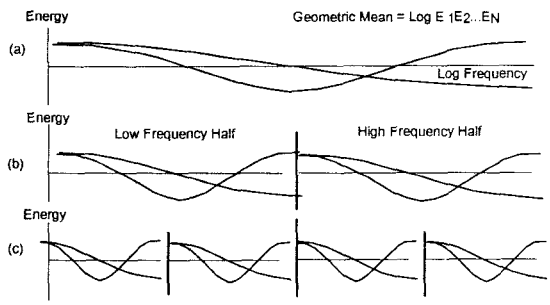


Figure 1 - DCT basis functions in a 3-level multi-resolution cepstral analysis.

3 PARTIAL RECOGNITION RECOMBINATION

Whilst our initial presentation and experimentation is based on straightforward concatenation of the cepstral coefficients from each multi-resolution sub-band to form a single feature vector, the recombination or merging of scores from independent models for each sub-band stream may permit further enhancements to performance, particularly in noisy operating conditions. It is a straightforward step to extend the principle of merging individual sub-band recogniser scores to multi-resolution sub-band recombination. In [2] trained MLPs were employed to perform non-linear merging to model re-combination functions for each phoneme classifier over all class sub-band model scores. Here we have experimented with linear weighted merging of the log likelihood scores of the multi-resolution models within each classification class.

Consider the multi-resolution subband cepstral feature vector $\boldsymbol{X}$ split into separate stream vectors $\boldsymbol{X}^{(rb)}\{r=1,..,R;\ b=1,..,B_r\}$ 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 $M_l^{(rb)}$ for each band b within resolution r, the combined log likelihood for class l can be given as

$$\log p(\mathbf{X}|M_l) = \sum_{r=1}^{R} \sum_{b=1}^{B_r} \omega_l^{(rb)} \log p(\mathbf{X}^{(rb)} | M_l^{(rb)}) \qquad (3)$$

The multi-resolution sub-band weights $\omega_l^{(rb)}$ should ideally reflect the discriminative potential or confidence of each subband for a particular class. Fully independent models $M_l^{(rb)}$ 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. As previously reported [2] the potential benefits of relaxing the temporal synchrony of spectral transitions within sub-bands could not be confirmed, so the use of tied state transitions is hence a reasonable approximation to adopt.

In [2,3] some discussion was devoted to the question of which temporal resolution was most appropriate at which to perform recombination. ie. merging scores at a segmental level such as phones or syllables. For continuous speech recognition it is necessary to merge scores at the "frame" level ie. for each incoming acoustic vector. Choosing an appropriate increased temporal resolution or segmental level for recombination would be difficult to assess optimally due to the variations in phoneme durations.

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

$$\log p(\mathbf{X}|M_l) = \sum_{r=1}^{R} \sum_{b=1}^{B_r} \omega_l^{(rb)}(SNR_l^{(rb)}) \log p(\mathbf{X}^{(rb)} | M_l^{(rb)}) \qquad (4)$$

where $\omega_l^{(rb)}(SNR_l^{(rb)})$ specifies the sub-band weighting to be a function of the local SNR (for band b in resolution level r) for model l. This recombination strategy is illustrated Fig(2). A possible function for weight variation could be

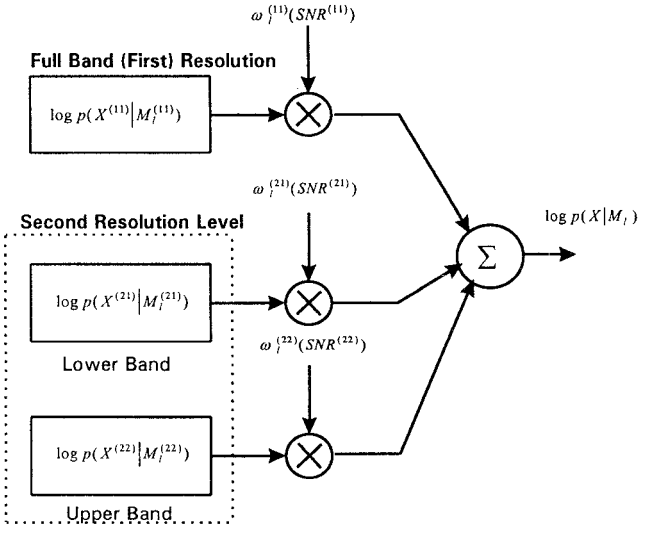
$$\omega_l^{(rb)}(SNR_{dB}^{(rb)}) = \frac{1}{1 + e^{-\alpha\, SNR_{dB}^{(rb)}}} \qquad (5)$$

In this case, α is also necessarily model dependent, in order to reflect the SNR range of the particular phonetic sub-band.

Another approach which avoids the need for making a further parameter choice is to use what is effectively a Weiner-type weighting of the multi-resolution sub-band streams. This can be expressed by

$$\omega_l^{(rb)} = \frac{SNR_l^{(rb)}}{SNR_l^{(rb)} + 1} \qquad (6)$$

or alternatively by equation (7)



Fig(2) Linear Recombination of Independent Sub Band Scores

$$\omega_l^{(rb)} = \frac{S_l^{(rb)}}{S_l^{(rb)} + N^{(rb)}} \qquad (7)$$

$S_l^{(rb)}$ defines the signal power in sub-band b of resolution r for the phoneme class l. 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. Thus, for the weighting function equation (7), the weight for state j in model l becomes

$$\omega_{jl}^{(rb)} = \frac{S_{jl}^{(rb)}}{S_{jl}^{(rb)} + N_j^{(rb)}} \qquad (8)$$

A possible estimation of state-dependent sub-band energies could be obtained by inverse DCT of the cepstra from trained phonetic HMM states, followed by summation of mel-filterbank energies over a sub-band group. In the above representation, the multi-resolution sub-band weightings are model-dependent and fixed. However in non-stationary noise environments it may be better to weight confidence on the features extracted within sub-bands on the time-varying or signal-dependent SNR. These weights would in effect be used to modify the underlying confidence in the partial recognition for a particular phonetic sub-band model conditioned on the time-varying signal and noise spectra. Thus if the function $\omega^{(rb)}(SNR(\mathbf{X})^{(rb)})$ indicates that the weight ω depends on the use the SNR in band b or resolution r as a function of the parameter vector $\mathbf{X}$, the recombination expression (3) becomes

$$\log p(\mathbf{X}|M_l) = \sum_{r=1}^{R}\sum_{b=1}^{B_r} \omega^{(rb)}(SNR(\mathbf{X})^{(rb)}) \log p(\mathbf{X}^{(rb)}|M_l^{(rb)}) \qquad (9)$$

4 EXPERIMENTAL RESULTS

4.1 Multi-Resolution Cepstra on TIMIT Database

The performance of the multi-resolution cepstral feature set was tested on the TIMIT continuous speech database using 39 context-independent HMM models. The full TIMIT training and test sets were used throughout. Table(1) gives results for purely sub-band based feature sets, along with the subband boundaries implemented. Table(2) gives the recognition performance for the multiresolution sets, where for example, the set (13)+(7,7) indicates that 13 cepstral coefficients are taken from a full-band analysis (ie. the first resolution), and 7 cepstral coefficients taken from the lower and 7 from the upper band of the 2nd resolution level. C_0 is included for each sub-band cepstra, and the results represent the coefficients supplemented in all cases by the delta and delta-delta trajectory coefficients. A mel-spaced filterbank implementation, rather than DFT, was used for the initial log spectral feature extraction. The results given are for 12 mixture HMMs.

Number of Bands	Bandwidths (kHz)	Cepstral Analysis	Recognition(%)
1	0-7.9	(13)	68.8
2	0-2,2-7.9	(7,7)	69.9
4	0-0.7,0.7-2 2-4,4-7.9	(5,5,3,2)	69.6

Table(1) TIMIT Subband Recognition Results

MultiResolution Cepstral Analysis	Recognition(%)
(13)+(7,7)	70.6
(8)+(4,4)	67.9
(13)+(5,5,3,2)	70.6
(13)+(7,7)+(5,5,3,2)	70.5

Table(2) TIMIT MultiResolution Recognition Results

All the experiments were carried out using the HTK toolkit. (The baseline recognition score using HTKs own MFCC features is 69%).

The results from Table(1) show some improvement in performance using subband cepstral features alone, compared to the full bandwidth cepstra. Table(2) however indicates further improvement in recognition performance when the multi-

resolution features sets are employed. Supplementing the full band cepstra with either 2 or 4 sub-band cepstra gives similar results. Use of both resolution levels is seen to yield no further advantage however. A reduction in the number of cepstral coefficients retained from each band, as indicated by the multi-resolution feature set (8)+(4,4), leads to a decrease in performance.

4.2 Independent Stream Weighting in Noise

Table (3) summarises some initial experiments using the fixed Weiner type weightings (7) for recombination of independent mulit-resolution sub-band streams according to (3) for continuous speech recognition. The results are for performance in white noise with a signal to noise ratio (pertaining over the full TIMIT training set) 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. The sub-band decompositions are the same as defined Table(1)

MultiResolution Cepstral Analysis	Recognition(%)
(13)	36.46
(7,7)	38.97
(13)+(7,7)	43.65

Table(3) Recognition Performance in 15dB white noise

The baseline recognition rate is low given that the 0^{th} ceptsral coefficient is retained in the feature vector. The performance is however increased with the use of two sub-bands. The best result obtained thus far is from use of the weighted multi-resolution sub-band models, with a reduction in error rate of 7.5%. These results compare favourably with those reported [2], where no improvements in performance using sub-band merging under white noise were obtained over using conventional features.

5 FUTURE WORK

Future work will explore more thoroughly the use of state-dependent weighting in a range of broadband noise conditions. Discriminative training of the stream weights for baseline recognition is also to be investigated, based on a minimum classification error (MCE) criterion. The Wiener-type weightings for use in noise would in some manner modify these baseline confidence measures. The manner of this modification is an issue for investigation but in the simplest sense would be simply the product of the baseline and noise dependent weights. Discriminative training of sub-band state-dependent linear transforms may also yield further improvements in performance [7]. The results did not provide a clear conclusion on the optimum number of sub-bands. The use of two sub-band decomposition resolutions was also seen to give no benefit in recognition performance. It may be that full exploitation of sub-band feature extraction for extending basic recognition performance can be achieved through using model-dependent sub-band boundaries. This however requires a new framework for training which incorporates some form of probabilistic decision making for optimum linear transform lengths. Continued examination of the multi-resolution temporal or segmental features introduced [1] is also currently being further explored. The re-convergence of the multi-resolution sub-band decomposition with the segmental feature set is nonetheless envisaged for future experimentation.

6 CONCLUSIONS

Multi-resolution cepstral features supplement cepstral analysis over the full-band mel filterbank log energies with cepstral analysis of grouped sub-banded energies. The multi-resolution cepstral features are demonstrated to improve monophone recognition on the TIMIT database compared to single resolution MFCCs. By exploiting the spectral variations in signal to noise ratio in terms of weighting the log likelihood scores from independent multiresoluion and sub-band streams we also achieve a more significant improvement in performance under white noise. This approach shows promise for increased robustness and future work will focus on employing discriminative training of stream weights and by extending SNR based weightings to state and signal dependent versions.

REFERENCES

[1] S.Vaseghi, N.Harte, B. Milner, "Multi-Resolution Phonetic/Segmental Features and Models for HMM-Based Speech Recognition", Proc. ICASSP-97, Vol. 2, pp.1263-1266

[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, "Subband-Based Speech Recognition", Proc. ICASSP-97, Vol 2, pp. 1251-1254

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

[5] H. Hermansky, M. Pavel, S. Tribrewala, " Towards ASR on Partally Corupted Speech", Proc. ICSLP-96, pp.462-465

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

[7] 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