MULTI-RESOLUTION PHONETIC/SEGMENTAL FEATURES AND MODELS FOR HMM-BASED SPEECH RECOGNITION

Saeed Vaseghi *Naomi Harte* *Ben Milner**

Queen's University of Belfast, N. Ireland.
E-mails : (S.Vaseghi, N.Harte)@ee.qub.ac.uk

*British Telecom Research Laboratories, UK.
Ben@saltfarm.bt.co.uk

ABSTRACT

This paper explores the modelling of phonetic segments of speech with multi-resolution spectral/time correlates. For spectral representation a set of multi-resolution cepstral features are proposed. Cepstral features obtained from a DCT of the log energy-spectrum over the full voice-bandwidth (100-4000 Hz) are combined with higher resolution features obtained from the DCT of upper subband (say 100-2100) and lower subband (2100-4000) halves. This approach can be extended to several levels of different resolutions.

For representation of the temporal structure of speech segments or phonetic units, the conventional cepstral and dynamic cepstral features representing speech at the sub-phonetic levels, are supplemented by a set of phonetic features that describe the trajectory of speech over the duration of a phonetic unit. A conditional probability model for phonetic and sub-phonetic features is considered. Experiments demonstrate that the inclusion of the segmental features result in about 10% decrease in error rates.

1 INTRODUCTION

In speech recognition, as in any pattern classification, the choice of signal features has a substantial influence on the separability of different classes of sounds and hence on the word recognition error rate. This paper explores the use of multi-resolution spectral/time correlates for speech recognition.

Most speech recognition systems employ mel filterbank cepstral coefficients (MFCCs), first difference (delta) MFCCs, and second difference (delta-delta) MFCCs for speech representation. A close variant of the cepstral features, with good performance, is the cepstral-time features derived from a 2-D DCT of stacked log-spectral vectors [1]. Alternative features based on higher order spectra, wavelets, and auditory models have been proposed but so far have proved less successful. Cepstral features are derived from a linear transformation (usually DCT) of the logarithm of the output energies of a set of mel-scaled spectral channels. Some optimisation and improvement in the feature extraction process can be obtained by using a discriminative optimisation criterion to estimate the parameters of the linear transformation and even the centre frequencies and the bandwidths of the channels.

This paper investigates the effects of using multi-resolution features for speech recognition. The current practice of using a single time window of about 30 ms, and splitting the signal within the window to about 20-25 frequency channels, is a best compromise between the requirements for time and frequency resolutions and low variance. An alternative is to employ features obtained at several different levels of spectral resolutions, and at sub-phonetic and phonetic durations.

A novel contribution of this paper is the introduction of a set of multi-resolution cepstral features. The multi-resolution cepstral idea is based on the hypothesis that for many speech sounds the localised features of the time-frequency trajectory of speech can provide crucial clues for classification. Hence it is desirable that in addition to the conventional cepstral features, a set of features that describe the more localised features of the log-spectral energy are also used.

The multi-resolution concept is extended to the time domain through defining sub-phonetic and phonetic speech features. To model speech at a segmental [1-4] or phonetic level, appropriate features together with statistical models for the inclusion of these features need to be defined. In [2] a segmental model is proposed in which the mean feature vector of each segment is taken as the segment feature. Since a segment models a section of the time-varying trajectory of speech in time and frequency, it is expected that dynamic features modelling the trajectory of each segment are a more appropriate feature set.

In this paper speech features are derived at what are effectively three different time resolutions. These include two sub-phonetic time windows; one is the cepstral feature vectors sampled with a time resolution of 5-10ms , and the second is the short-time dynamic features averaged over the duration of at least 3 vectors i.e. 15-30 ms. The other set of features are phonetic features which span a time window of the order of the duration of a phonetic unit which can be much above that of the conventional time windows. A drawback of using long windows is that the time window would often contain signals from different adjacent phones. To avoid this problem phonetic segment boundaries are required. Hence one method for using the phonetic features is in the second pass of the decoding algorithm using the ML segmentation boundaries provided by the first pass. Two issues explored in this paper are : (a) the choice of phonetic and sub-phonetic features, and (b) the statistical models for combination of phonetic and sub-phonetic features.

2 MULTI-RESOLUTION SUBBAND CEPSTRAL FEATURES

In this section we propose a set of multi-resolution cepstral features. The motivation is to explore new spectral correlates that may provide more separable features for speech classification. Let the vector sequence $\mathcal{E}=[\mathcal{E}_1, \mathcal{E}_2,..., \mathcal{E}_L]$ denote a sequence of L P-dimensional mel-spaced log-filter bank energy vectors. The N-dimensional feature vectors are extracted from the P dimensional log-spectral energy features using a $P \times N$ transformation as

$$X_t = A\mathcal{E}_t \qquad (1)$$

Conventionally A is the DCT matrix. Alternatives to DCT include the Karhunen-Loeve transform (KLT), and the linear discriminative analysis (LDA) transform. A relatively recent method is HMM-based state dependent transformation of log-spectral energy features [6].

The transformation A in eq(1) yields a set of features that are averaged over the entire speech bandwidth. An alternative is to combine the cepstral features extracted from the whole signal bandwidth, with those extracted at subbands, for example from say the upper and the lower half bands. Hence the multi-resolution cepstrum proposed here can be expressed as a combination of a set of linear transformations as

$$X_t = [A_0\mathcal{E}_t, (A_{12}\mathcal{E}_{t12}, A_{22}\mathcal{E}_{t22}), (A_{14}\mathcal{E}_{t14}, A_{24}\mathcal{E}_{t24}, A_{34}\mathcal{E}_{t34}, A_{44}\mathcal{E}_{t44}), \ldots]^T \qquad (2)$$

where $A_0\mathcal{E}_t$, yields the cepstral features over the entire banwidth, $(A_{12}\mathcal{E}_{t12}\ ,A_{22}\mathcal{E}_{t22})$ yields cepstral features over, the lower half and the upper half subbands, and $(A_{14}E_{t14}\ ,A_{24}\mathcal{E}_{t24}, A_{34}\mathcal{E}_{t34}\ ,A_{44}\mathcal{E}_{t44})$ yields the features over four subband quadrants as in Figure(1). The subscript notation ij refers to the i^{th} band given that the spectrum has been divided into j subbands.

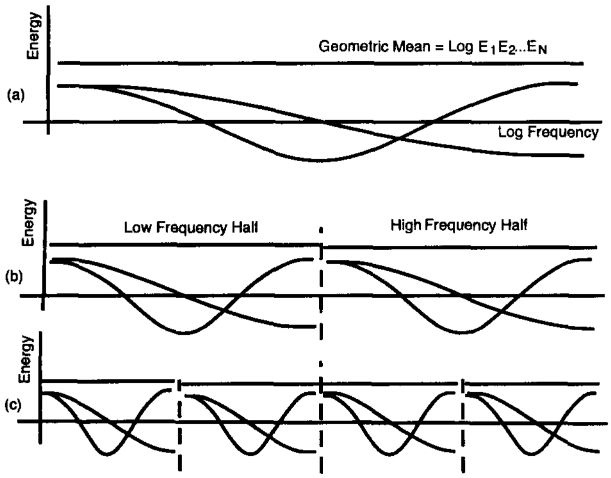

Figure 1 - The first three DCT basis functions in a 3-level multiresolution cepstral analysis.

Figure(1) is an illustration of the multi-resolution feature extraction process where the voice band is progressively divided first into 2 subbands and then into 4 subbands respectively. The figure also illustrates the DCT basis functions for each subband in eq(1). The first cepstrum coefficient in Figure(1.a) taken over the entire spectrum gives a measure of the slope of the spectrum, alternatively it can be interpreted as the log ratio of the weighted geometric mean of the energy in the lower half-band to that in the upper half-band. The first cepstral coefficients from the two subbands in figure (1.b) give more detailed measures of the slopes of the spectra in the upper half-band, and the lower half-band respectively. In figure(1.c) the multiresolution pyramid is extended down to a third level. The higher resolution features give a set of correlates that can be used as an alternative to the higher coefficients in the conventional cepstrum.

3 PHONETIC/SEGMENTAL FEATURES

Continuous speech recognition systems model the acoustic speech signals as a string of elementary phonetic symbols. Each phone, or its triphone variant, is typically modelled by a three-states HMM. As each phone is a relatively short speech segment with a distinct spectral-time composition, the use of a segmental model is appropriate for the phonetic units. A segmental model avoids the finite state quantisation involved in the HMMs, and models the segments as a continuous signal process. However, there are two main drawbacks associated with a segmental model when compared to HMMs : (1) HMMs can easily model the variation in speech duration through the self-loop state transitions, in contrast for a segmental model the segment length has to be normalised and therefore some knowledge of the segment boundaries are required, (2) In continuous speech recognition, HMMs can model the various probable boundaries of phonemes, whereas segmental models need to estimate the boundaries.

A solution to this problem is a two stage method of combination of HMMs and segmental models. In the first stage, using conventional features, speech is decoded to obtain the N-most likely candidates and their most likely phonetic boundaries. In the second stage the estimates of the phonetic boundaries are used to extract phonetic features, which will then be employed as an additional set of features in a revaluation of the probability score for each phone. Let $X(t,T)=[x(t), ..., x(t+T-1)]$, denote a feature vector sequence of length T. The variables t and T can be estimated from an ML estimation of the underlying phonetic boundaries in a first pass of the Viterbi decoding of speech features. The phonetic features can be formulated as

$$X_{\mathcal{P}} = A\ B\ X \tag{3}$$

where B is a TxT_n resampling matrix for normalisation of the phone duration from T samples to a constant preseleced value of T_n, and A is a transformation matrix for extracting a set of features. Phonetic features normally span a sequence of several cepstral vectors, and hence dynamic features that describe the temporal trajectory of the cepstral vectors in a phoneme may be considered as appropriate features for speech representation at the phonetic or segmental level. A convenient choice for A is the DCT. Alternatively the matrix A can be obtained using a linear discrimination analysis (LDA), or from a discriminative training of HMMs [7].

3.1 SEGMENTAL PHONETIC HMMS

For an HMMs $\mathcal{M}_k$ the log-likelihood of a sequence of T feature vectors, along a state sequence s_k, is given as

$$\log f(X|\mathcal{M}_k,s)=\sum_{t=0}^{T-1}\log a_{s_{t-1}s_t}+\sum_{t=0}^{T-1}\log f(x(\mathrm{t})|\mathcal{M}_k,s_t) \tag{4}$$

Eq(4) assumes that within each state the features are independent and identically distributed. The correlation across the states are modelled through the Markovian state transitions parameters $a_{s_{t-1}s_t}$. In this section conditional probability models are used to capture the dependencies between cepstral vectors and phonetic or segmental features. Conditional probability models have been used in HMMs for inclusion of the correlation of successive speech cepstral vectors [7].

If we also include a phonetic feature $\boldsymbol{\chi}_{\mathcal{P}}$ in the model then the log likelihood becomes

$$\log f(X,\chi_{\mathcal{P}}|\mathcal{M}_k,s)=\sum_{t=0}^{T-1}\log a_{s_{t-1}s_t}+\sum_{t=0}^{T-1}\log f(x(\mathrm{t})|\mathcal{M}_k,\chi_{\mathcal{P}},s_t)+T\log f(\chi_{\mathcal{P}}) \tag{5}$$

For a multi-variate Gaussian density the conditional mean vector and the covariance matrix are given by

$$\begin{aligned}\mu_{(x|\chi_{\mathcal{P}})}&=\mathcal{E}[x|\chi_{\mathcal{P}}]\\&=\mu_x+\Sigma_{x\chi_{\mathcal{P}}}\Sigma^{-1}_{\chi_{\mathcal{P}}\chi_{\mathcal{P}}}(\chi_{\mathcal{P}}-\mu_{\chi_{\mathcal{P}}})\end{aligned} \tag{6}$$

$$\Sigma_{(x|\chi_{\mathcal{P}})}=\Sigma_{\chi_{\mathcal{P}}\chi_{\mathcal{P}}}-\Sigma_{x\chi_{\mathcal{P}}}\Sigma^{-1}_{\chi_{\mathcal{P}}\chi_{\mathcal{P}}}\Sigma_{\chi_{\mathcal{P}}x} \tag{7}$$

The phonetic features can be used as additional features in a second pass of speech decoder.

4 EXPERIMENTAL RESULTS

An important aspect of speech processing is to capture the nonstationary character of the signals. The concept of a nonstationary spectrum may be more appropriately associated with the time-variations of the output of a digital filter-bank because unlike the Fourier transform, the input to a filterbank is not assumed to be short-time stationary. The feature extraction system used in this paper is based on a mel-scaled filter bank shown in figure(2). The spectral features are obtained from averaging the output of the filter bank over a time window of about 10-20 ms. Successive overlapping frames of log spectral energy are stacked to form a spectral-time matrix of a duration of 32 ms. A duration of this order is considered to represent speech at a sub-phonetic level. Each log spectral-time matrix is converted via a 2-DCT to a cepstral-time matrix. Phonetic features are derived as the trajectory of sub-phonetic features over the duration of a phone.

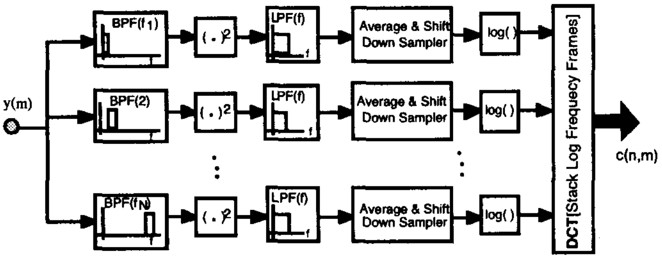

Figure 2 - A filter bank implementation of cepstral-time feature extraction.

The following experiments are based on two databases; an isolated digit data base, and a continuous data base of 1000 talkers recorded over the telephone trunk network. The continuous-speech data base is modelled using phonetic units. The speech was sampled at 8 kHz, and the subband processor of figure(2) was used for feature extraction. The filter consists of 20 mel-scaled channels. Table-1 shows the experimental results for the isolate-digit data base. Table-2 shows the experimental results for the continuous speech database. The segmental features are compared with those for the conventional cepstral plus delta and delta-delta features. The table shows the followings; (row 1) when 12 conventional cepstral coefficients extracted over the full voice band of 0.1-4 kHz are used, (row 2) multi-resolution cepstrum using 6 conventional features, plus 3 features from each of the lower and upper half subbands, and (row 3) multi-resolution cepstrum using 4 conventional cepstral features plus 4 from each of the lower and upper subbands. From table-1, it is encouraging that multi-resolution cepstral features compete well with the more conventional cepstral coefficients.

Table-1

Subbands (kHz)	Coeff Pattern	c+dc+ddc	Segmental Features
0.1-4	12	95.5	97.6
(0.1-4), (0.1-2, 2-4)	6 + (3 , 3)	96.2	97.6
(0.1-4), (0.1-2, 2-4)	4 + (4 , 4)	96.3	97.8

Table-2

Subbands (kHz)	Coeff Pattern	c+dc+ddc	Segmental Features
0.1-4	12	66.8	73.2
(0.1-4), (0.1-2, 2-4)	6 + (3 , 3)	67.1	73.7
(0.1-4), (0.1-2, 2-4)	4 + (4 , 4)	67.3	73.9

For the continuous database, table-2 clearly demonstrates that the use of long term segmental features provides substantial improvement in accuracy. Table-3, like table-2, presents some results for the continuous speech data base. The main difference is that the multi-resolution approach is extended to 3 levels as in fig(1).

Table-3

Subbands (kHz)	Coeff Pattern	c+dc+ddc	Segmental Features
0.1-4	16	67.2	73.6
(0.1-4), (0.1-2, 2-4), (0.1-1,1-2, 2-3, 3-4)	4 +(4,4)+(1,1,1,1)	67.6	73.8
(0.1-4), (0.1-2, 2-4), (0.1-1,1-2, 2-3, 3-4)	4 +(2,2)+(2,2,2,2)	67.9	73.97

CONCLUSION

Further improvements in the performance of speech recognition systems is likely to result from advances in the modelling of phonemes and speech features. The work described in this paper explored a new approach to speech feature extraction. In proposing a multi-resolution approach the aim is to obtain a new set of correlates in time and frequency for improved speech recognition. Experiments demonstrated that the multi-resolution cepstrum compares well with the more conventional method, and that the use of segmental features results in a significant improvement in speech recognition. To obtain the full potential benefits of phonetic features the feature extraction and the training/decoding processes need to be well integrated.

Acknowledgement The authors acknowledge the support of British Telecom Laboratories, Martlesham UK.

REFERENCES

[1] Milner B., "Inclusion of Temporal Information into Features for Speech Recognition", Int. Conf. on Spoken Language Processing, ICLSP 96, Pages 256-259.

[2] Ostendorf M., Rouskos, "A Stochastic Segment Model for Phoneme Based Continuous Speech Recognition", IEEE Trans. ASSP, Vol 37, No 12. Pages 1857-1869, 1989.

[3] Gales M., Young S., "Segmental Hidden Markov Models", Eurospeech93, Pages 1579-1582, (1993)

[4] Bacchiani M, Ostendorf M., Sagisaka Y., Paiwal K., "Design of a Speech Recognition System Based on Acoustically Derived Segmental Units", ICASSP-96 443-446.

[5] Holmes W. J., Russell M. J., "Speech Recognition Using a Linear Dynamic Segmental HMM", Proc. Eurospeech-95, Pages 1611-1614 (1995)

[6] C. Rathinavelu, L. Deng, "HMM-Based Speech Recognition Using State-Dependent Linear Transform on Mel-Warped DFT Features", ICASSP-96, pages 9-12.

[7] F.J Smith, J. Ming, P. O'Boyle, A.D. Irvine," A Hidden Markov Model with Optimise Interframe Dependence", ICASSP95 Pages 209-212.