HIERARCHICAL SUBBAND LINEAR PREDICTIVE CEPSTRAL (HSLPC) FEATURES FOR HMM-BASED SPEECH RECOGNITION

Rathinavelu Chengalvarayan
Speech Processing Group,
Lucent Speech Solutions Department
Lucent Technologies, Naperville, IL 60566, USA
Email: rathi@lucent.com

ABSTRACT

In this paper, a new approach for linear prediction (LP) analysis is explored, where predictor can be computed from a mel-warped subband-based autocorrelation functions obtained from the power spectrum. For spectral representation a set of multi-resolution cepstral features are proposed. The general idea is to divide up the full frequency-band into several subbands, perform the IDFT on the mel power spectrum for each subband, followed by Durbin's algorithm and the standard conversion from LP to cepstral coefficients. This approach can be extended to several levels of different resolutions. Muti-resolution feature vectors, formed by concatenation of the subband cepstral features into an extended feature vector, are shown to yield better performance than the conventional mel-warped LPCCs over the full voice-bandwidth for connected digit recognition task.

1. INTRODUCTION

The structure of a typical continuous speech recognizer consists of a frontend feature analysis stage followed by a statistical pattern classifier. The feature vector, interface between these two, should ideally contain all the information of the speech signal relevant to subsequent classification, be insensitive to irrelevant variations due to changes in the acoustic environments, and at the same time have a low dimensionality in order to minimize the computational demands of the classifier [3]. Several types of feature vectors have been proposed [10]. Most speech recognizers have traditionally utilized cepstral parameters derived from an LP analysis due to the advantages that LP provides in terms of generating a *smooth* spectrum, free of pitch harmonics, and its ability to model spectral peaks reasonably well. Mel-based cepstral parameters, on the other hand, take advantage of the perception properties of the human auditory system by sampling the spectrum at mel-scale intervals. Logically, combining the merits of both LP analysis and mel-filter bank analysis should, in theory, produce an improved set of cepstral features.

This can be performed in several ways. For example, one could compute the log magnitude spectrum of the LP parameters and then warp the frequencies to correspond to the mel-scale. Previous studies have reported encouraging speech recognition results when warping the LP spectrum by a bilinear transformation prior to computing the cepstrum, as opposed to not using the warping [11]. Several other frequency warping techniques have been proposed, for example in [13] a mel-like spectral warping method through all-pass filtering in the time domain is proposed. Another approach is to apply mel-filter bank analysis on the signal followed by LP analysis to give what will be refered to as mel-lpc features [11]. The computation of the mel-lpc features is similar in some sense to PLP coefficients [4]. Both techniques apply a mel filter bank prior to LP analysis. However, the mel-lpc uses a higher order LP analysis with no perceptual weighting or amplitude compression. The basic ideas of all the above techniques are to try to *perceptually* model the spectrum of the speech signal, which lead to an improved speech quality and provide more efficient representation of the spectrum for speech analysis, synthesis and recognition.

In recent years there has been a number of papers on subband-based feature extraction techniques [1, 7, 9, 15]. Recent theoretical and empirical results have shown that auto-regressive spectral estimation from subbands is more robust and more efficient than full-band auto-regressive spectral estimation [12]. In this work, a new approach for prediction analysis is proposed, where predictor can be computed from a bunch of mel-warped

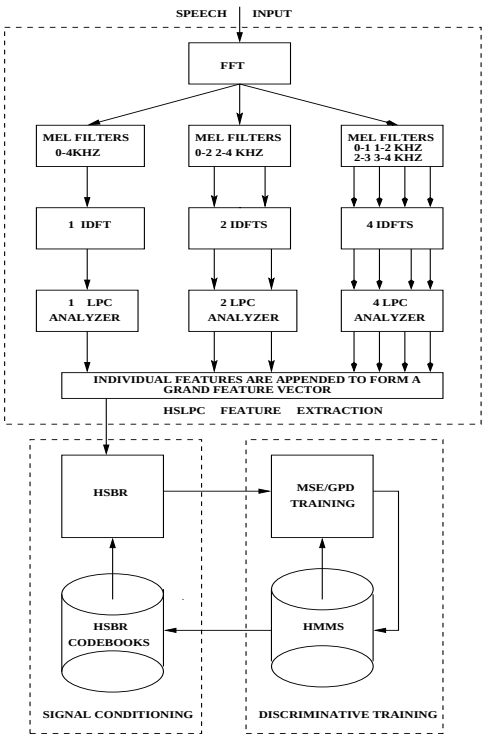

Figure 1. Block diagram of hierarchical subband linear predictive speech analysis with signal conditioned minimum string error rate training.

subband-based autocorrelation functions obtained from the power spectrum. Further, the level of subband decomposition and subsequent cepstral analysis can be increased such that features may be selected from a pyramid resolution levels. In this study, an extended feature vector is formed based on concatenation of LP cepstral features from each multi-resolution subband, defining a large dimensional space on which the statistical parameters are estimated. We restrict our presentation to only the recognizer based on hidden Markov model (HMM) approach using continuous density mixtures to characterize the states of the HMM. A relative advantage of multi-resolution feature set is that the inclusion of different resolutions of subband decomposition in effect relaxes the restriction of using a single fixed subband decomposition. One of the design choices for subband-based schemes is the number of bands and the exact subband boundary decomposition.

2. HSLPC CEPSTRAL FEATURES

In this section we explore a set of hierarchical subband-based linear predictive cepstral (HSLPC) features. The motivation is to explore new spectral correlates that may provide more separable features for classification. Figure 1 shows the overall process of computing the hierarchical mel-lpc features from a frame of speech. The steps in the process are as follows:

- Mel-filter bank analysis: This includes preemphasis, blocking speech into frames, frame windowing, Fourier transformation and mel-filter bank analysis. The center frequencies of the filters are spaced equally on a linear scale from 100 to 1000 Hz and equally on a logarithmic scale above 1000 Hz. Above 1000 Hz, each center frequency is 1.1 times the center frequncy of the previous filter. Each filter's magnitude frequency response has a triangular shape that is equal to unity at the center frequency and linearly decreasing to zero at the center frequencies of the two adjacent filters. The spectrum for each frame is passed through a set of M triangular mel-filter banks, where M is set to 24 in this study.
- Autocorrelation analysis: The IDFT is applied to the smoothed power spectrum (without the log operation) to yield Q autocorrelation coefficients, where Q is set to 10 for level 1. For level 2, Q is set to 8 for lower half and upper half subbands (0-2 KHz and 2-4 KHz). For level 3, Q is set to 6 for each subband quadrants (0-1 KHz, 1-2 KHz, 2-3 KHz and 3-4 KHz) and so on. The Figure 1 illustrates the sequence of operations in each subband for these resolution levels.
- Cepstral analysis: Each set of autocorrelation coefficients is converted first to LP coefficients, using Durbin's recursion algorithm, and then to cepstral parameters using the standard LP to cepstrum recursion. This process is repeated for each level and for each subband, until we arrive at the required number of cepstral features from all the levels. Then the multi-level subband features are concatenated to form a single extended feature vector. The final dimension of the cepstral vector is set to 12 in this study. We explored three types of feature sets:
  - (12,0,0) represents 12 features from level 1.
  - (0,6,6) indicates 12 features from level 2 (6 features from lower subband and 6 features from upper subband).
  - (6,3,3) represents 6 features from level 1 and six features from level 2 (3 features from lower subband and 3 features from upper subband).

For each frame of speech, the input feature vector is extended beyond the 12 HSLPC features (and energy) to include the first and second order derivatives. In total, a 39-dimensional feature vector is used for each frame [6].

3. SPEECH DATA

The connected digit (CD) database used in this study is a good challenge for speech recognizers because of its diversity. It is a compilation of databases collected during several independent data collection efforts, field trials, and live service deployments. The CD database contains the English digits *one* through *nine*, *zero* and *oh*. It ranges in scope from one where talkers read prepared lists of digit strings to one where the customers actually use an recognition system to access information about their credit card accounts. The data were collected over wireline network channels using a variety of telephone handsets. Digit string lengths range from 1 to 16 digits. The CD database is divided into two sets: training and testing. The training set includes both *read* and *spontaneous* digit input from a variety of network channels, microphones and dialect regions. The testing set is designed to have data strings from both matched and mismatched environmental conditions. All recordings in the training and testing set are valid digit strings, totaling 7461 and 13114 strings for training and testing, respectively [14].

4. SIGNAL CONDITIONED HMM RECOGNIZER

Following feature analysis, each feature vector is passed to the recognizer which models each word in the vocabulary by a set of left-to-right continuous mixture density HMM using context-dependent head-body-tail models [6]. Since the signal has been recorded under various telephone conditions and with different transducer equipment, each HSLPC feature vector was further processed using the hierarchical signal bias removal (HSBR) method in order to reduce the effect of channel distortion [11]. Each word in the vocabulary is divided into a head, a body, and a tail segment. To model inter-word coarticulation, each word consists of one body with multiple heads and multiple tails depending on the preceding and following contexts. In this paper, we model all possible inter-word coarticulation, resulting in a total of 276 context-dependent sub-word models. Both the head and tail models are represented with 3 states, while the body models are represented with 4 states, each having 4 mixture components. Silence is modeled with a single state model having 32 mixture components. This configuration results in a total of 276 models, 837 states and 3376 mixture components. Training included updating all the parameters of the model, namely, means, variances, and mixture gains using maximum-likelihood estimation (MLE) followed by three epochs of minimum string error (MSE) training to further refine the estimate of the parameters [5, 6, 8]. The HSBR codebook of size four is extracted from the mean vectors of HMMs, and each training utterance is signal conditioned by applying HSBR prior to being used in MSE training [2]. The number of competing string models was set to four and the step length was set to one during the model training phase. The length of the input digit strings are assumed to be unknown during both training and testing [14].

Feature Vector Type	Training Scheme	
	ML Training	MSE Training
$HSLPC_{12,0,0}$	78.38%	90.69%
$HSLPC_{0,6,6}$	79.06%	91.06%
$HSLPC_{6,3,3}$	81.65%	92.10%

Table 1. String accuracy rate for an unknown-length grammar-based connected digit recognition task using the ML and MSE training methods as a function of HSLPC feature type.

5. EXPERIMENTAL RESULTS

Several sets of experiments were run to evaluate the connected digit recognizers using three types of HMMs ($HSLPC_{12,0,0}$, $HSLPC_{0,6,6}$ and $HSLPC_{6,3,3}$) and two types of training (ML and MSE). The overall performance of the recognizers, organized as the string accuracy as a function of the feature type is summarized in Table 1. For example, the set $HSLPC_{6,3,3}$ indicates that 6 mel-lpc features are taken from the first resolution, and 3 mel-lpc features are taken from the lower and 3 from the upper band of the second resolution level. The normalized frame energy is included along with the multi-resolution features, and the results represent the features supplemented in all cases by the delta and delta-delta trajectory features. Table 1 illustrates four important results. First, the MSE training is superior to the MLE training and the MSE-based recognizer achieves an average of 55% string error rate reduction, uniformly across all types of speech models, over the MLE-based recognizer. Second, some improve-

ment in performance using subband cepstral features alone ($HSLPC_{0,6,6}$), compared to the full bandwidth cepstra $HSLPC_{12,0,0}$ is also observed. Thirdly, further improvement in recognition performance is obtained when the multi-resolution feature sets are employed as shown in third row of Table 1. Finally, the best result obtained thus far is from use of the features from both resolution levels ($HSLPC_{6,3,3}$), with a reduction in error rate of 15% when compared with the first resolution feature set alone ($HSLPC_{12,0,0}$). From Table 1, it is encouraging that the multi-resolution mel-lpc features are demonstrated to improve recognition on the telephone connected digit database compared to single resolution mel-lpc features. These results compare unfavorably with those reported in [7], where use of both resolution levels is seen to yield no further advantage.

6. CONCLUSIONS

We have addressed the problem of using multi-resolution subband LP cepstral features for speech recognition. The original contribution of this paper is the introduction of a set of hierarchical subband-based LP cepstral features as speech recognition features. We described the multi-resolution LP cepstral extraction technique and formulated extended feature vectors by concatenation of the hierarchical subband LP cepstral features. Experimental results on connected digit recognition task demonstrated a 15% string error rate reduction by using the extended feature set as compared to conventional mel-warped LP cepstral features over the full voice-bandwidth. This suggests that the important additional cues for speech discrimination may exist in the local spectral correlates that are not captured by the full band LP cepstral analysis and the inclusion of this new multi-level of feature parameters further enhances the recognizer performance.

REFERENCES

[1] H. Bourlard and S. Dupont, "Subband-Based Speech Recognition", *Proc. ICASSP*, 1997, pp. 1251-1254.

[2] W. Chou, M.G. Rahim and E. Buhrke, "Signal Conditioned Minimum Error Rate Training", *Proc. EUROSPEECH*, 1995, pp.495-498.

[3] T. Eisele, R. Haeb-Umbach and D. Langmann, "A Comparative Study of Linear Feature Transformation Techniques for Automatic Speech Recognition", *Proc. ICSLP*, 1996, pp. 252-255.

[4] H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of Speech", *Journal of Acoustical Society of America*, Vol. 87, No. 4, 1990, pp. 1738-1752.

[5] B.H. Juang and L.R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Transactions on Acoustics Speech and Signal Processing*, Vol. 38, No. 9, 1990, pp. 1639-1641.

[6] B.H. Juang, W. Chou and C.H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No.3, pp. 257-265, 1997.

[7] P. McCourt, S. Vaseghi and N. Harte, "Multi-Resolution Cepstral Features for Phoneme Recognition Across Speech Subbands", *Proc. ICASSP*, 1998, pp. 557-560.

[8] E. McDermott and S. Katagiri, "String-Level MCE for Continuous Phoneme Recognition", *Proc. EUROSPEECH*, 1997, pp. 123-126.

[9] S. Okawa, E. Bocchieri and A. Potamianos, "Multi-Band Speech Recognition in Noisy Environments", *Proc. ICASSP*, 1998, pp. 641-644.

[10] J.W. Picone, "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, Vol. 81, No. 9, 1993, pp.1215-1247.

[11] M. Rahim and B.H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", *IEEE Transactions on SPeech and Audio Processing*, Vol. 4, No. 1, 1996, pp. 19-30.

[12] S. Rao and W.A. Pearlman, "Analysis of Linear Prediction, Coding and Spectral Estimation from Subbands", *IEEE Transactions on Information Theory*, Vol. 42, 1996, pp. 1160-1178.

[13] H.W. Strube, "Linear Prediction on a Warped Frequency Scale", *Journal of Acoustical Society of America*, Vol. 68, No.4, 1980, pp. 1071-1076.

[14] D.L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features", *Proc. ICASSP*, 1998, pp. 21-24.

[15] S. Tibrewala and H. Hermansky, "Subband Based Recognition of Noisy Speech", *Proc. ICASSP*, 1997, pp. 1255-1258.