ROBUST SPECTRO-TEMPORAL FEATURES BASED ON AUTOREGRESSIVE MODELS OF HILBERT ENVELOPES

Sriram Ganapathy[1], *Samuel Thomas*[1], *Hynek Hermansky*[1,2]

[1]Department of Electrical and Computer Engineering
[2]Human Language Technology Center of Excellence
Johns Hopkins University, USA
{ganapathy,samuel,hynek}@jhu.edu

ABSTRACT

In this paper, we present a robust spectro-temporal feature extraction technique using autoregressive models (AR) of sub-band Hilbert envelopes. AR models of Hilbert envelopes are derived using frequency domain linear prediction (FDLP). From the sub-band Hilbert envelopes, spectral features are derived by integrating these envelopes in short-term frames and the temporal features are formed by converting these envelopes into modulation frequency components. The spectral and temporal feature streams are then combined at the phoneme posterior level and are used as the input features for a recognition system. For the proposed features, robustness is achieved by using novel techniques of noise compensation and gain normalization. Phoneme recognition experiments on telephone speech in the HTIMIT database show significant performance improvements for the proposed features when compared to other robust feature techniques (average relative reduction of 10.6 % in phoneme error rate). In addition to the overall phoneme recognition rates, the performance with broad phonetic classes is also reported.

***Index Terms*—** Frequency domain linear prediction (FDLP), Hilbert Envelopes, Robust spectro-temporal features, Phoneme recognition.

1. INTRODUCTION

Conventional speech analysis techniques estimate the spectral content of relatively short (about 10-20 ms) segments of the signal (short-term spectrum). Each estimated vector of spectral energies represents a sample of the underlying dynamic process in production of speech at a given time-frame. Most of the information contained in these acoustic features relate to formants which provide important cues for recognition of basic speech units. Stacking such estimates of the short-term spectra in time provides a two-dimensional (time-frequency) representation of speech that forms the basis for most speech features (for example [1]).

An alternate way to describe a speech signal is that of a summation of a number of amplitude modulated narrow frequency sub-bands. In this view, every frequency band can be considered to consist of a carrier signal (fine structure) and a time-varying envelope [2]. One can directly estimate trajectories of spectral energies in the individual frequency sub-bands, each estimated vector then representing the underlying dynamic process in a given sub-band. Such estimates, stacked in frequency, also form a two-dimensional representation of speech (for example [3]).

For human phoneme recognition, it has been shown that modulation information of 0-16 Hz is important [4]. The interaction between the spectral and temporal modulations has been conducted by varying the number of sub-bands in input speech analysis [5]. Here, it is shown that a combination of spectral and temporal information is important for human perception of phonemes.

In our previous work [6], we had proposed a feature extraction technique that combines short-term spectral features and the temporal modulation features for the task of phoneme recognition. Specifically, speech signals in frequency sub-bands are analyzed over long temporal segments using the Frequency Domain Linear Prediction (FDLP) to estimate the Hilbert envelopes. The short-term spectral features are derived by integrating the sub-band Hilbert envelopes in short analysis windows and the temporal modulation features are obtained by the application of cosine transform on the compressed (static and adaptive compression) long term sub-band Hilbert envelopes [6].

In this paper, we propose a noise compensation technique and a gain normalization technique for FDLP in deriving spectro-temporal features. For noise compensation, an estimate of the noise envelope is derived from the input noisy speech signal in each sub-band. This estimate is subtracted from the noisy sub-band envelope before the application of linear prediction in frequency domain. Once the FDLP envelopes are estimated, we apply a gain normalization procedure on the FDLP envelopes which tries to alleviate convolutive distortions in speech. The application of these techniques improves the robustness of the proposed features in mismatched train/test conditions.

Experiments are performed on phoneme recognition task in HTIMIT database [8] (which contains telephone channel recordings of TIMIT data) using the models trained in clean TIMIT dataset. We use a hybrid Hidden Markov Model - Artificial Neural Network (HMM-ANN) phoneme recognition system [9]. The proposed features provide considerable improvements in phoneme recognition accuracies for this task.

The rest of the paper is organized as follows. In Sec. 2, the FDLP technique for deriving sub-band envelopes is described. The conversion of these sub-band envelopes into spectral and temporal features is explained in Sec. 3. Experiments with the proposed features for phoneme recognition task are reported in Sec. 4. In Sec. 5, we conclude with a discussion of the proposed features.

2. FREQUENCY DOMAIN LINEAR PREDICTION

The Hilbert envelope, which is the squared magnitude of the analytic signal, represents the instantaneous energy of a signal in the time domain. A discrete time analytic signal, as defined in [11], can be obtained by forcing the causality of the discrete Fourier trans-

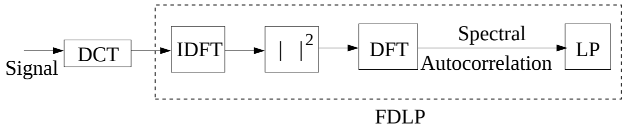

Fig. 1. Block schematic for the frequency domain linear prediction (FDLP).

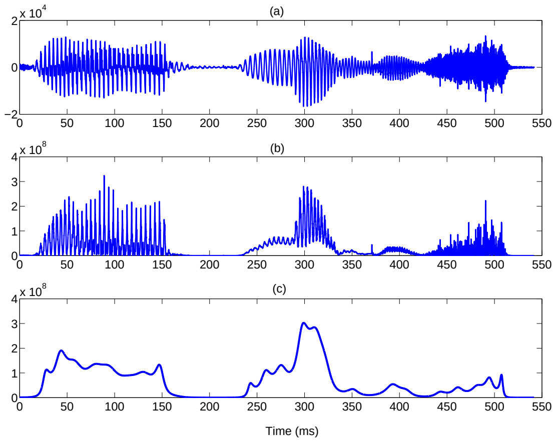

Fig. 2. Illustration of the all-pole modelling property of FDLP. (a) a portion of the speech signal, (b) its Hilbert envelope computed using DFT [11], and (c) all pole model obtained using FDLP.

form (DFT) and by ensuring the orthogonality of real and imaginary parts. Mathematically, it can be shown that the autocorrelation of discrete cosine transform (DCT) of the input signal and its discrete time Hilbert envelope are Fourier transform pairs [7]. This means that the application of linear prediction on the cosine transform of the signal yields an AR model of the Hilbert envelope of the signal. Thus, a parametric model for the Hilbert envelopes can be obtained using FDLP [2, 7].

Fig. 1 shows the block schematic for the implementation of FDLP technique. Long segments of the input signal (of the order of 1000 ms) are transformed into frequency domain using DCT. The inverse DFT (IDFT) of the DCT coefficients represents the discrete time analytic signal [7]. Spectral autocorrelations are derived by the application of DFT on the squared magnitude of analytic signal. These autocorrelations are used for linear prediction (similar to the application of TDLP using time domain autocorrelations [10]). In deriving spectro-temporal features for phoneme recognition, FDLP is applied on the critical bands of the input speech signal.

Fig. 2 shows the AR modelling property of FDLP. It shows (a) a portion of speech signal, (b) its Hilbert envelope computed using the Fourier transform technique [11] and (c) an all pole approximation for the Hilbert Envelope using FDLP.

2.1. Noise Compensation

When additive noise is present in speech signal, the FDLP envelope is modified in such a way its dynamic range is reduced. The effect of noise is more pronounced in the valleys of the FDLP envelopes, where the mismatch between clean and noisy speech is significant. When features are derived from the uncompensated FDLP envelopes, the performance of the phoneme recognition system degrades significantly in noisy conditions.

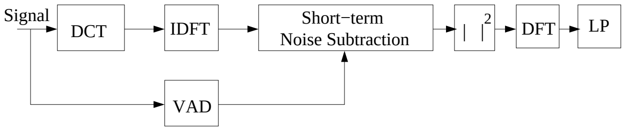

Fig. 3. Noise compensation in frequency domain linear prediction.

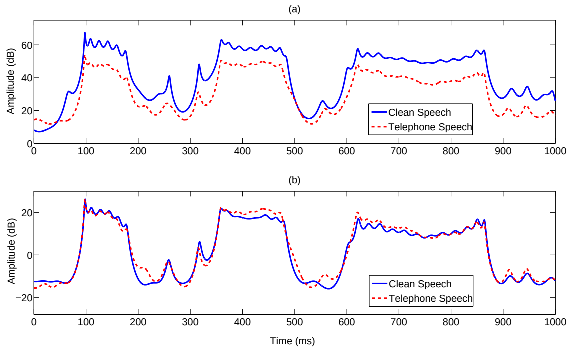

Fig. 4. Log-FDLP envelopes for a sub-band of clean speech and telephone speech (a) without gain normalization and noise compensation, (b) with gain normalization and noise compensation.

We apply noise compensation technique on FDLP as shown in Fig. 3. A voice activity detector (VAD) operates on the input speech signal to indicate the presence of non-speech frames. The VAD is implemented using the same technique proposed in [12]. The VAD output is a flag indicating the speech/non-speech decision for every short-term frame of speech (with a length of 25 ms and a shift of 10 ms).

As mentioned before, long segments of the input speech signal are transformed to DCT domain. The discrete time analytic signal is obtained as the magnitude IDFT of the DCT signal. If noise signal is additive in signal domain, it continues to be additive in the analytic signal domain. Hence, the effect of noise can be compensated by short-term noise subtraction on the analytic signal. This is achieved in two steps. In the first step, we window the analytic signal into short-term segments (of length 25 ms with a shift of 10 ms). The next step is to subtract an estimate of the short-term noise component derived from non-speech segments (in the initial portion of speech utterances).

2.2. Gain Normalization

The Hilbert envelope and the spectral autocorrelation function form Fourier transform pairs [7]. When speech is corrupted by convolutive distortions, the sub-band Hilbert envelopes can be assumed to be a convolution of the sub-band Hilbert envelope of the clean speech with the sub-band Hilbert envelope of the noise [13]. This means that the spectral autocorrelation function of convolutive noise can be approximated as the multiplication of spectral autocorrelation function of the clean speech with that of the convolutive noise. Typically, for long segments of input signal, the spectral autocorrelation function in frequency sub-bands can be assumed to be slowly varying compared to that of the speech signal. Thus, normalizing the gain of the sub-band FDLP envelopes suppresses the multiplicative effect present in the spectral autocorrelation function of noise [13].

When speech signal is passed through a telephone channel, the output signal can be modelled as a combination of back-ground additive noise and a convolutive noise in the channel. In such conditions,

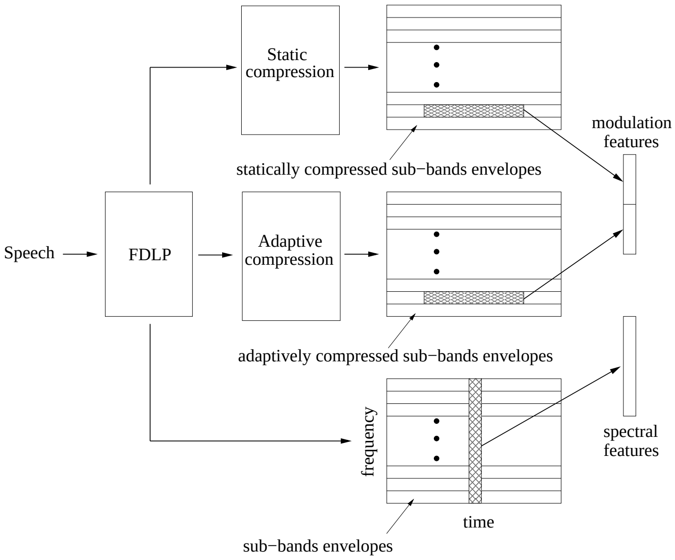

Fig. 5. Schematic of the joint short-term spectral and temporal modulation feature extraction technique

the combination of noise compensation and gain normalization provides considerable robustness to the FDLP technique. This is illustrated in Fig. 4, where we plot the log FDLP envelopes for clean and telephone speech in two conditions, (a) without gain normalization and noise compensation and (b) with gain normalization and noise compensation. This figure shows that the application of these techniques reduces the mismatch between the FDLP envelopes extracted from clean and noisy speech. In the next section, we use the gain normalized and noise compensated FDLP envelopes for extracting spectro-temporal features from input speech.

3. SPECTRO-TEMPORAL FEATURE EXTRACTION

3.1. Short-term Spectral Features

As mentioned before, Hilbert envelope represents the instantaneous energy of a signal in the time domain. Since integration of signal energy is identical in time and frequency domain, the sub-band Hilbert envelopes can equivalently be used for obtaining the sub-band energy based short-term spectral features. This is achieved by integrating the sub-band temporal envelopes in short term frames (of the order of 25 ms with a shift of 10 ms). These short term sub-band energies are then converted into 13 cepstral features along with their first and second derivatives (similar to 39 dimensional PLP features [1]). Each frame of these short-term spectral features is used with a context of 9 frames for training a phoneme posterior probability estimator [14].

3.2. Long-term Modulation Features

The long-term sub-band envelopes from the FDLP form a compact representation of the temporal dynamics over long regions of the speech signal. The sub-band FDLP envelopes are compressed using a static compression scheme which is a logarithmic function and a dynamic compression scheme [15]. The dynamic compression is realized by an adaptation circuit consisting of five consecutive nonlinear adaptation loops [15]. Each of these loops consists of a divider and a low-pass filter with time constants ranging from 5 ms to 500 ms. The input signal is divided by the output signal of the low-pass filter in each adaptation loop. The compressed temporal envelopes are divided into 200 ms segments with a shift of 10 ms. Discrete Cosine Transform (DCT) is applied on the static and the adaptive segments to yield the static and the adaptive modulation spectrum respectively. We use 14 modulation frequency components from each cosine transform, yielding modulation spectrum in the $0 - 35$ Hz region with a resolution of 2.5 Hz. The static and adaptive modulation features for each sub-band are stacked together to obtain modulation features for each sub-band and fed to the posterior probability estimator.

Table 1. Phoneme Recognition Accuracies (%) in clean speech and telephone speech (average performance for 9 channel conditions).

Feat	Clean	Tel
PLP	65.4	34.3
ETSI	64.0	47.7
FDLP-S	63.5	52.2
MRASTA	62.8	48.0
FDLP-M	62.3	55.4
PLP+MRASTA	67.5	47.5
ETSI+MRASTA	66.8	52.9
FDLP-S+FDLP-M	66.7	57.9

We combine the short-term spectral and modulation frequency features at the phoneme posterior level using the Dempster Shafer (DS) theory of evidence [16].

4. EXPERIMENTS AND RESULTS

The proposed features are used for a phoneme recognition task on the HTIMIT database [8]. We use a phoneme recognition system based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [9]. The system is trained on clean speech using the TIMIT database downsampled to 8 kHz. The training data consists of 3000 utterances from 375 speakers, cross-validation data set consists of 696 utterances from 87 speakers and the test data set consists of 1344 utterances from 168 speakers. The TIMIT database, which is hand-labeled using 61 labels is mapped to the standard set of 39 phonemes [14]. For phoneme recognition experiments in telephone channel, speech data collected from 9 telephone sets in the HTIMIT database [8] are used, which introduce a variety of channel distortions in the test signal. Each of these telephone channels consist of 842 test utterances, which also have clean recordings in the TIMIT test set. The system is trained only on the original TIMIT data, representing clean speech without the distortions introduced by the communication channel but tested on the clean TIMIT test set as well as the HTIMIT degraded speech.

Table 1 shows the results for phoneme recognition accuracies for various feature extraction techniques in clean and telephone speech. In the base-line experiments, the proposed features are compared with other feature extraction techniques on the same task - the PLP features with a 9 frame context [14] and Advanced-ETSI (noise-robust) distributed speech recognition front-end [12] with a 9 frame context which are similar to short-term spectral features derived using FDLP (FDLP-S) and MRASTA features [17] which are similar to modulation features derived from FDLP (FDLP-M). We combine the short-term spectral and modulation frequency features [6] using the DS theory of evidence to obtain three more feature sets - PLP features with MRASTA features (PLP+MRASTA), ETSI features with MRASTA features (ETSI+MRASTA) and FDLP-S fea-

Table 2. Recognition Accuracies (%) of broad phonetic classes obtained from confusion matrix analysis on TIMIT database

Class	PLP + MRASTA	ETSI + MRASTA	FDLP-S + FDLP-M
Vowel	87.6	87.5	88.8
Plosive	82.2	81.5	80.9
Fricative	81.8	81.2	79.7
Semi Vowel	75.2	75.5	74.4
Nasal	85.2	84.0	83.6
Avg.	82.4	81.9	81.5

tures with FDLP-M features (FDLP-S+FDLP-M).

The FDLP-S features provide comparable results as the ETSI features in clean conditions whereas it shows good robustness in telephone channel conditions. The modulation features (FDLP-M) result in significant improvements for phoneme recognition rate in telephone speech compared to the MRASTA features. The joint short-term spectral and modulation features yield robust phoneme recognition compared to the baseline systems. We obtain a relative improvement of 10.6 % in telephone speech phoneme recognition.

5. DISCUSSION AND CONCLUSION

The previous section showed that the proposed feature extraction provides promising phoneme recognition performance on HTIMIT database. Here, we analyze the improvements in terms of decompositions into broad phoneme classes using phoneme confusion matrices. Table. 2 and Table. 3 show the recognition accuracies of broad phoneme classes for the proposed feature extraction technique along with baseline systems for clean and telephone speech respectively. For clean conditions, the proposed features (FDLP-S+FDLP-M) provide recognition accuracies that are competent with other techniques for all the phoneme classes. For telephone speech, the FDLP-S features provide significant robustness for fricatives and vowels (which is due to modelling property of the signal peaks in FDLP) whereas the FDLP-M features provide good robustness for plosives (where the fine temporal fluctuations like onsets and offsets carry the important phoneme classification information). Hence, the combination of these feature streams results in considerable improvement in performance for most of the broad phonetic classes.

In summary, we have proposed a robust spectro-temporal feature extraction scheme for ASR. Sub-band Hilbert envelopes, estimated using FDLP, are processed to derive both short-term spectral and temporal modulation features. Gain normalization and noise compensation techniques add robustness to FDLP envelopes. These features provide considerable improvements for phoneme recognition tasks in noisy conditions.

6. REFERENCES

[1] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis of Speech,” *J. Acoust. Soc. Am.*, vol. 87(4), pp. 1738-1752, 1990.

[2] R. Kumerasan and A. Rao, “Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications”, *Journal of Acoustical Society of America*, Vol. 105 (3), Mar. 1999, pp. 1912-1924.

Table 3. Recognition Accuracies (%) of broad phonetic classes obtained from confusion matrix analysis on TIMIT database

Class	PLP + MRASTA	ETSI + MRASTA	FDLP-S + FDLP-M
Vowel	83.7	85.7	87.9
Plosive	60.0	68.4	73.0
Fricative	71.3	73.5	78.2
Semi Vowel	70.7	74.7	72.0
Nasal	69.8	72.1	80.1
Avg.	71.1	74.9	78.2

[3] B.E.D. Kingsbury, N. Morgan and S. Greenberg, “Robust speech recognition using the modulation spectrogram”, *Speech Comm.*, Vol. 25 (1-3), pp. 117-132, 1998.

[4] R. Drullman, J.M. Festen and R. Plomp,“Effect of Reducing Slow Temporal Modulations on Speech Reception”, *J. Acoust. Soc. Am.*, Vol. 95(5), pp. 2670-2680, 1994.

[5] L. Xu and Y. Zhang, “Spectral and temporal cues for phoneme recognition in noise,” J. Acoust. Soc. Am., Vol. 122(3), pp. 1758-1764.

[6] S. Thomas, S. Ganapathy and H. Hermansky, “Phoneme recognition using spectral envelope and modulation frequency features”, *Proc. of ICASSP*, 2009.

[7] M. Athineos and D.P.W. Ellis, “Autoregressive modelling of temporal envelopes”,*IEEE Trans. Speech and Audio Proc.*, Vol. 55, pp. 5237-5245, 2007.

[8] D.A. Reynolds, “HTIMIT and LLHDB: speech corpora for the study of hand set transducer effects,” in *Proc. ICASSP*, 1997, pp. 1535-1538.

[9] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, Boston, 1994.

[10] J. Makhoul, “Linear Prediction: A Tutorial Review”,in *Proc. of the IEEE*, Vol 63(4), pp. 561-580, 1975.

[11] L.S. Marple, “Computing the Discrete-Time Analytic Signal via FFT”, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 47, 1999, pp. 2600-2603.

[12] “ETSI ES 202 050 v1.1.1 STQ; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms”, 2002.

[13] S. Thomas, S. Ganapathy and H. Hermansky, “Recognition of Reverberant Speech Using Frequency Domain Linear Prediction,” IEEE Sig. Proc. Let., Vol. 15, pp. 681-684.

[14] J. Pinto, B. Yegnanarayana, H. Hermansky and M.M. Doss, “Exploiting Contextual Information for Improved Phoneme Recognition”, *Proc. of INTERSPEECH*, 2007, pp. 1817-1820.

[15] J. Tchorz and B. Kollmeier, “A model of auditory perception as front end for automatic speech recognition”, *J. Acoust. Soc. Am.*, Vol. 106(4), 1999, pp. 2040-2050. 0

[16] F. Valente and H. Hermansky, “Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence,” in *Proc. of ICASSP*, 2007, pp. 1129-1132.

[17] H. Hermansky and P. Fousek, “Multi-resolution RASTA filtering for TANDEM-based ASR”, *Proc. of INTERSPEECH*, 2005, pp. 361-364.