IDIAP RESEARCH REPORT

Hilbert Envelope Based Spectro-Temporal Features for Phoneme Recognition in Telephone Speech

Samuel Thomas [a b] Sriram Ganapathy [a b]
Hynek Hermansky [a b]

IDIAP–RR 08-18

June 2008

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-18

Hilbert Envelope Based Spectro-Temporal Features for Phoneme Recognition in Telephone Speech

Samuel Thomas Sriram Ganapathy Hynek Hermansky

June 2008

Abstract. In this paper, we present a spectro-temporal feature extraction technique using sub-band Hilbert envelopes of relatively long segments of speech signal. Hilbert envelopes of the sub-bands are estimated using Frequency Domain Linear Prediction (FDLP). Spectral features are derived by integrating the sub-band Hilbert envelopes in short-term frames and the temporal features are formed by converting the FDLP envelopes into modulation frequency components. These are then combined at the phoneme posterior level and are used as the input features for a phoneme recognition system. In order to improve the robustness of the proposed features to telephone speech, the sub-band temporal envelopes are gain normalized prior to feature extraction. 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 11% in phoneme error rate).

n this paper, we present a spectro-temporal feature extraction technique using sub-band Hilbert envelopes of relatively long segments of speech signal. Hilbert envelopes of the sub-bands are estimated using Frequency Domain Linear Prediction (FDLP). Spectral features are derived by integrating the sub-band Hilbert envelopes in short-term frames and the temporal features are formed by converting the FDLP envelopes into modulation frequency components. These are then combined at the phoneme posterior level and are used as the input features for a phoneme recognition system. In order to improve the robustness of the proposed features to telephone speech, the sub-band temporal envelopes are gain normalized prior to feature extraction. 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 11% in phoneme error rate).

Index Terms: Frequency Domain Linear Prediction, Spectro-Temporal Features, Telephone Speech, Phoneme Recognition.

1 Introduction

Traditionally, acoustic features for Automatic Speech Recognition (ASR) systems are extracted by applying Bark or Mel scale integrators on power spectral estimates in short analysis windows ($10-30$ ms) of the speech signal. Typical examples of such features are the Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) [2]. The signal dynamics is represented by a sequence of short-term feature vectors with each vector forming a sample of the underlying process.

On the other hand, it has been shown that important information for speech perception lies in the $1-16$ Hz range of the modulation frequencies [3]. In order to exploit the information at these modulation frequencies, relatively long segments of speech signal need to be analyzed. For example, an explicit incorporation of the information about these long-term speech dynamics have been proposed in [4].

In this paper, we propose a feature extraction technique that combines long temporal features with the short-term spectral features for the task of phoneme recognition in telephone channels. These spectral and temporal features are derived from sub-band Hilbert Envelopes of relatively long segments of speech and are combined in the phoneme posterior level to form a joint spectro-temporal feature set. These posterior features are used in a hybrid Hidden Markov Model - Artificial Neural Network (HMM-ANN) phoneme recognition system [5].

For estimating the Hilbert envelopes of the sub-band signal, we use the technique of linear prediction in spectral domain. Frequency Domain Linear Prediction (FDLP) was originally proposed for temporal noise shaping in audio coding [6] and was used for ASR feature extraction in [7]. By applying linear prediction on the Discrete Cosine Transform (DCT) of the signal, the FDLP technique provides an efficient way for auto-regressive (AR) modeling of the Hilbert Envelope of a signal.

When the proposed spectro-temporal features are used for phoneme recognition task in telephone channels, there is degradation in performance due to the linear filtering from the handset microphone and the telephone channel [8]. Several techniques have been proposed in the literature to compensate the telephone channel noise [9, 10]. For the proposed spectro-temporal features based on FDLP, the artifacts introduced by the telephone channel affect the gain of the sub-band temporal envelopes. In order to reduce the mismatch between the features trained from clean speech and telephone speech, we propose a technique that normalizes the gain of the auto-regressive models in frequency sub-bands prior to the spectro-temporal feature extraction. Experiments on a phoneme recognition task in the HTIMIT database [11] show significant performance improvements for the proposed features compared to the other robust feature extraction techniques.

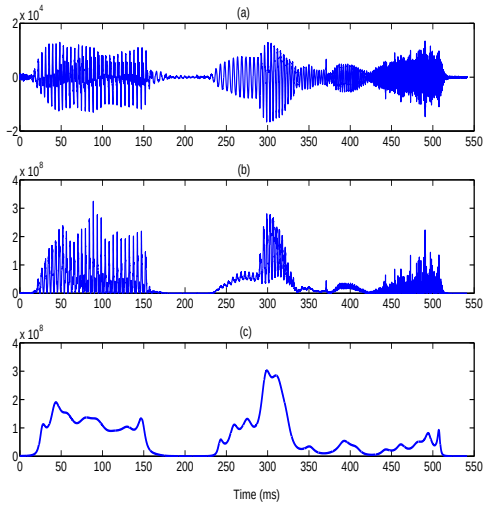

Figure 1: *Illustration of the all-pole modelling property of FDLP. (a) a portion of the speech signal, (b) its Hilbert envelope (c) all pole model obtained using FDLP*

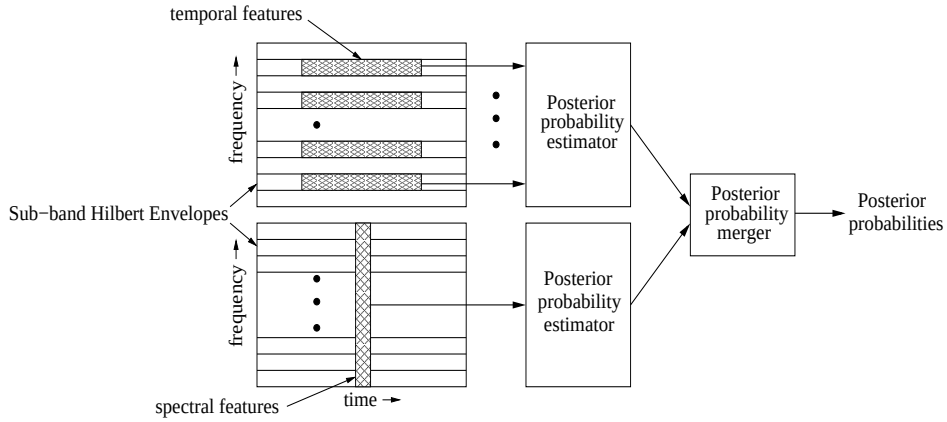

Figure 2: *Schematic of the joint spectro-temporal features for posterior based ASR*

2 Frequency Domain Linear Prediction

Typically, Auto-Regressive (AR) models have been used in speech/audio applications for representing the envelope of the power spectrum of the signal (Time Domain Linear Prediction (TDLP) [12]). This paper utilizes AR models for obtaining smoothed, minimum phase, parametric models for temporal rather than spectral envelopes. Since we apply the LP technique to exploit the redundancies in the frequency domain, this approach is called Frequency Domain Linear Prediction (FDLP). For the FDLP technique, the squared magnitude response of the all-pole filter approximates the Hilbert envelope of the signal (in a manner similar to the approximation of the power spectrum of the signal using TDLP [12]). By using analysis windows of the order of hundreds of milliseconds, the technique automatically decides the distribution of the poles to best model the temporal envelope. Fig. 1 shows an illustration of the AR modelling property of FDLP. It shows (a) a portion of speech signal of 500 ms duration, (b) its Hilbert envelope computed using the Fourier transform technique [13] and (c)

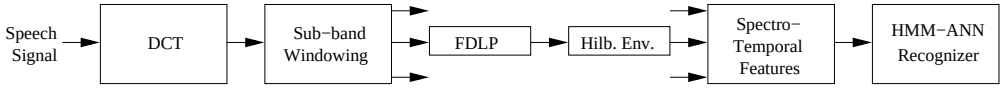

Figure 3: *Schematic of the FDLP based speech recognition system*

an all pole approximation (of order 50) for the Hilbert Envelope using FDLP.

Modelling long temporal envelopes of speech in sub-bands using FDLP provides some important advantages:

- Fine time-dependent resolution provides information about transient events in time like stop bursts.
- Long-term summarization of power in spectral bands presents the ability to capture complete description of the linguistic units lasting more than 10 ms.

We implement the FDLP technique in two parts - first, speech signal is decomposed into frequency sub-bands by windowing the DCT. Linear prediction is then performed on the DCT coefficients to obtain a parametric model of the Hilbert envelope.

3 Spectro-Temporal Features from sub-band Hilbert Envelopes

The whole set of sub-band temporal envelopes forms a two dimensional (time-frequency) representation of the input signal energy. Conventional ASR feature extraction techniques derive short-term spectral features by integrating the estimate of power spectrum of the signal in sub-bands for short segments of the signal. Similarly, we extract spectral features from the sub-band Hilbert envelopes by integrating them in short-term frames (of the order of 25 ms with a shift of 10 ms). These are then converted to short-term cepstral features in a manner similar to the PLP feature extraction technique [2]. As the features are derived from short temporal segments, they capture spectral details in the order of 10 ms. A context of 9 frames is used for the spectral features in training the ANN for spectral features [14].

Since the long-term sub-band FDLP envelopes form a compact representation of the temporal dynamics over long regions of the speech signal, they can also be used to extract temporal features for ASR. In a manner similar to the conversion of the magnitude spectrum of a signal into cepstral coefficients, the temporal envelopes can also be converted into modulation frequency components [7]. Since the phoneme recognition system requires features at a frame rate of 10 ms, the modulation frequency components for the current frame in each sub-band are obtained along with contextual information of neighboring frames (in a manner similar to the setup in [4]). From phoneme recognition experiments, a context of 250 ms for the temporal features gave the best phoneme recognition performance [14].

The spectral and temporal features are combined using the Dempster Shafer (DS) theory of evidence [15] to form a joint spectro-temporal posterior feature set [14]. Fig. 2 shows the schematic of the proposed combination of spectral and temporal features. These features are used for in a phoneme recognition as shown in Fig. 3.

4 Gain Normalization of Hilbert Envelopes for Telephone Speech

The Hilbert envelope and the spectral autocorrelation function form Fourier transform pairs [6]. For the telephone speech, 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 telephone channel impulse response. This means that the spectral autocorrelation function of telephone speech can be approximated as the multiplication of spectral autocorrelation function of the clean speech with that of the channel impulse response. Typically, for long segments of telephone channel impulse responses,

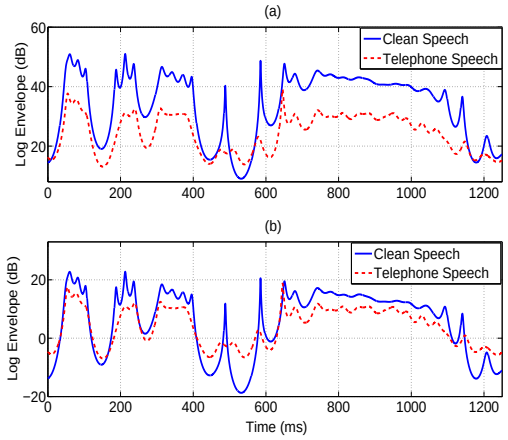

Figure 4: *FDLP envelopes for clean and telephone speech for second sub-band (a) without gain normalization (b) with gain normalization.*

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 telephone speech. For example, Fig. 4 provides an illustration of the effect of gain normalization on the sub-band FDLP envelopes for clean and for telephone speech. Gain normalization of the sub-band FDLP envelopes is achieved by setting the gain of the inverse FDLP filter to unity. We use the gain normalized FDLP envelopes for extracting spectro-temporal features from telephone speech.

5 Experiments and Results

We apply the proposed features and techniques in a phoneme recognition task using the HTIMIT database [11]. The HTIMIT corpus is a re-recording of a subset of the TIMIT corpus through different telephone handsets. The phoneme recognition system is based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [5].

The ANN estimates the posterior probability of phonemes given the acoustic evidence $P(q_t = i|x_t)$, where q_t denotes the phoneme index at frame t, x_t denotes the feature vector. The relation between the posterior probability $P(q_t = i|x_t)$ and the likelihood $P(x_t|q_t = i)$ is given by the Bayes rule,

$$\frac{p(x_t|q_t = i)}{p(x_t)} = \frac{P(q_t = i|x_t)}{P(q_t = i)}. \tag{1}$$

The scaled likelihood in an HMM state is given by Eq. 1, where we assume equal prior probability $P(q_t = i)$ for each phoneme $i = 1, 2...39$. The state transition matrix is fixed with equal probabilities for self and next state transitions. Viterbi algorithm is applied to decode the phoneme sequence. Fig. 3 shows the block schematic for the phoneme recognition system.

The phoneme recognition 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 [16].

The ANN is trained using the standard back propagation algorithm with cross entropy as the training error. The learning rate and stopping criterion are controlled by the frame classification rate on the cross validation data. In our system, the MLP consists of 1000 hidden neurons, and 39

Table 1: Phoneme Recognition Accuracies (%) for different feature extraction techniques on clean speech and telephone speech *cb1*

Feature	Accuracy(*clean*)	Accuracy(*cb1*)
PLP-9	64.9	42.5
RASTA-PLP-9	61.2	50.7
ETSI-9	63.9	55.1
MRASTA	63.1	52.0
MRASTA-PLP-9	67.6	53.7
FDLP-G	68.1	47.7
FDLP-GN	67.1	57.2

output neurons (with soft max nonlinearity) representing the phoneme classes. For the decoding step, all phonemes are considered equally probable (no language model). The performance of phoneme recognition is measured in terms of phoneme accuracy.

For phoneme recognition experiments in telephone channel, speech data collected from 9 telephone sets in the HTIMIT database are used (corresponding to the following transducer names - *cb1, cb2, cb3, cb4, el1, el2, el3, el4, pt1*). For each of these telephone channels, 842 test utterances, also having clean recordings in the TIMIT test set, are used. Features extracted for the telephone speech were tested on the models trained with clean TIMIT training set.

The results for the proposed FDLP technique are compared with those obtained for several other robust feature extraction techniques namely RASTA [10], Multi-resolution RASTA (MRASTA) [4], and the ETSI advanced (noise-robust) distributed speech recognition front-end [17]. The first set of experiments compare the performance of these feature extraction techniques for the clean test conditions in TIMIT database. The results of this experiment are shown in the first column of Table 1. The conventional PLP feature extraction used with a context of 9 frames [16] is denoted as PLP-9. RASTA-PLP-9 features use 9 frame context of the PLP features extracted by applying the RASTA filtering [10]. The ETSI-9 corresponds to 9 frame context of the features generated by the ETSI front-end. In a manner similar to the proposed spectro-temporal features, we also combine the posterior probabilities for the MRASTA features with PLP-9 [18]. For the proposed FDLP based technique, we investigate the effect of gain normalization of the Hilbert Envelopes for the spectro-temporal feature extraction. The FDLP-G features are derived from temporal envelopes without the gain normalization, whereas the gain normalized temporal envelopes are used for deriving the FDLP-GN features. Table 1 also shows the phoneme recognition results for one of telephone sets (*cb*1) in HTIMIT database. It can be seen that the spectro-temporal features extracted from temporal envelopes without removing the gain (FDLP-G) perform better than other features on clean speech. Without much degradation in performance for clean speech, the FDLP-GN features provide significant improvements for the telephone speech.

For all the telephone channel sets in the HTIMIT database, Table 2 shows the phoneme recognition performance using the feature extraction techniques that provided the best results in Table 1 along with PLP-9 baseline. The proposed features, on the average, provide a relative error improvement of 11% over the other feature extraction techniques considered.

6 Conclusions

We have proposed a novel method of extracting spectro-temporal features for phoneme recognition. For this purpose, Hilbert envelopes of frequency sub-bands are modelled using Frequency Domain Linear Prediction. Further, for the task of phoneme recognition in telephone speech, the proposed technique uses features extracted from gain normalized temporal envelopes to alleviate the artifacts introduced by the channel. The proposed features provide noticeable improvements over other robust

Table 2: Phoneme Recognition Accuracies (%) for different feature extraction techniques on telephone speech

Set	PLP-9	MRASTA-PLP-9	ETSI-9	FDLP-GN
cb1	42.5	53.7	55.1	57.5
cb2	46.0	57.2	57.3	61.9
cb3	23.3	36.8	39.5	40.6
cb4	32.5	43.3	39.7	47.8
el1	43.4	56.5	55.5	61.6
el2	27.4	44.8	46.8	55.9
el3	38.3	47.0	46.8	50.9
el4	31.9	47.5	41.9	54.9
pt1	24.5	43.1	46.6	50.0
Avg.	**34.4**	**47.8**	**47.7**	**53.5**

feature extraction techniques for phoneme recognition tasks in the HTIMIT database. The results are promising and encourage us to experiment on other tasks with different test conditions.

References

[1] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", in *IEEE Trans. on Acoustics, Speech and Signal Processing* Vol. 28, pp. 357-366, 1980.

[2] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, 1990.

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

[4] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. INTERSPEECH*, Lisbon, Portugal, pp. 361-364, 2005.

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

[6] J. Herre and J.D Johnston, "Enhancing the Performance of Perceptual Audio Coders by using Temporal Noise Shaping (TNS)," in *Proc. of 101st AES Conv.*, Los Angeles, USA, pp. 1-24, 1996.

[7] M. Athineos, H. Hermansky and D.P.W Ellis, "LP-TRAPS: Linear Predictive Temporal Patterns," in *Proc. of Interspeech*, Jeju Island, Korea, pp. 1154-1157, 2004.

[8] H. Hermansky, N. Morgan, A. Bayya and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech," in *Proc. Eurospeech*, Genova, Italy, 1991, pp. 1367-1370.

[9] J.D. Veth and L. Boves, "Channel normalization techniques for automatic speech recognition over the telephone," *Speech Communications*, vol. 25, pp. 149-164, Aug. 1998.

[10] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 578-589, Oct 1994.

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

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

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

[14] —, "Spectro-Temporal Features for Automatic Speech Recognition using Linear Prediction in Spectral Domain," to appear in *Proc. of EUSIPCO*, Lausanne, Switzerland, Aug 2008.

[15] F. Valente and H. Hermansky, "Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence," in *Proc. of ICASSP*, Hawaii, U.S.A, pp. 1129-1132, 2007.

[16] J. Pinto, B. Yegnanarayana, H. Hermansky and M. M. Doss, "Exploiting Contextual Information for Improved Phoneme Recognition", in *Proc. of Interspeech*, Antwerp, Belgium, pp. 1817-1820, 2007.

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

[18] S. R. M. Prasanna, B. Yegnanarayana, J. Pinto, and H. Hermansky, "Analysis of Confusion Matrix to Combine Evidence for Phoneme Recognition," *IDIAP Research Report (IDIAP-RR-07-27)*, July 2007.