Robust Phoneme Recognition Using High Resolution Temporal Envelopes

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

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.)
[2]Human Language Center of Excellence (HLTCOE), Johns Hopkins University, Baltimore, MD, USA.)
ganapathy@jhu.edu, hynek@jhu.edu

Abstract

Frequency domain linear prediction (FDLP) is a technique for auto-regressive (AR) modeling of Hilbert envelopes of the signal. The model is derived by the application of linear prediction on the discrete cosine transform (DCT) of the signal. In this paper, we propose modifications of the basic FDLP approach for deriving high resolution envelopes. We determine various factors which affect temporal resolution in FDLP such as the location of the input peaks within the analysis segment, type of window applied in the DCT of the signal, and order of the FDLP model. This analysis enables us to improve the resolution of temporal envelopes derived from FDLP. The features extracted from high resolution envelopes outperform MFCC features in noisy phoneme recognition experiments (relative improvements of 10 %) and phoneme recognition in conversational telephone speech (relative improvements of 5 %).

Index Terms: Frequency Domain Linear Prediction, Resolution Analysis, Feature Extraction, Phoneme Recognition

1. Introduction

Conventionally, time domain linear prediction (TDLP [1]) is used for AR modeling of power spectrum. Various modifications in the model estimation can yield different variants of TDLP [2]. The effect of the TDLP window shape and the method for computing the AR coefficients were well studied in [3]. TDLP is still widely used in speech coding and speech feature extraction (e.g. perceptual linear prediction (PLP) [4].

Frequency domain linear prediction (FDLP) analysis approximates the Hilbert envelope of the signal by its auto-regressive model [5, 6]. The sub-band envelopes estimated using FDLP have been applied for feature extraction in speech recognition [6, 7]. When speech is corrupted by noise, the low-energy regions of the speech signal are modified significantly by noise. A robust feature extraction scheme should aim to focus only on the high energy regions of the signal. This can be achieved by the AR modeling procedure appearing in FDLP. There is also an addition demand of representing the high energy regions in the noisy signal with good resolution so that the mis-match between the features derived from clean and noisy conditions is reduced. In the FDLP framework, this corresponds to estimation of temporal peaks with high resolution. While one would agree that the higher the FDLP model order, the better its temporal resolution, the effect of various factors on the resolution was not to our knowledge systematically studied. In particular, the questions such as whether the temporal resolution is constant over the whole time interval being analyzed, the effect of various types of windows on the DCT sequence, and the influence of various ways of deriving the auto-regressive models, are still open.

In order to analyze the resolution properties of FDLP, we need to define an objective measure of resolution. We propose to define the temporal resolution of FDLP by using synthetic signals with two closely spaced peaks. The separation between the peaks is varied and the minimum separation between the two peaks in the input for which the output of the AR model has two peaks is determined (critical time-span). Then, the resolution is computed as the inverse of the critical time-span. We show the resolution is a function of the relative location of the peaks within the analysis window, model order, type of LP method as well as the type of window function used for the analysis. The results reveal that the temporal resolution is significantly better in the central part of the analysis segment than it is at its boundaries. The analysis suggests several modifications of FDLP which can improve its temporal resolution like symmetric padding of the signal at the boundaries of the analysis window, suitable window functions and the use of least-squares linear prediction technique.

When speech is corrupted by noise, temporal envelopes estimated from noisy speech do not match those obtained from clean training conditions. Using the techniques proposed in this paper, we show that the high resolution estimation of the envelopes can reduce this mis-match. In phoneme recognition experiments, the proposed high resolution FDLP features provide significant improvements in additive noise as well as matched conditions of conversational telephone speech (CTS).

The rest of the paper is organized as follows. Sec. 2 describes the FDLP framework for AR modeling of Hilbert envelopes. Here, we provide a simple derivation for the relation between the auto-correlation of DCT and the analytic signal. The temporal resolution analysis of FDLP is provided in Sec. 3. Phoneme recognition experiments using FDLP features is described in Sec. 4, followed by a summary in Sec. 5.

2. AR Model of Hilbert Envelopes

The fundamental relation in TDLP is that the auto-correlation of a signal and its power spectrum form a Fourier transform pair. In a dual manner, we show that the auto-correlation of DCT sequence and the Hilbert envelope (squared magnitude of analytic signal (AS)) are related by the Fourier transform. This would mean that the application of linear prediction on the DCT sequence provides an AR model of the Hilbert envelopes [5, 6] (similar to the application of linear prediction in time domain sequence to obtain the power spectrum [3]).

Let $x_a[n]$ denote the AS of a discrete sequence $x[n]$ for $n = 0, \ldots, N-1$. We assume that the discrete-time sequence $x[n]$ has a zero-mean property in time and frequency domains, i.e., $x[0] = 0$ and $X[0] = 0$. In a discrete-time case, the spectrum

of AS ($X_a[k]$) can be defined [8] as

$$X_a[k] = \begin{cases} 2X[k] & \text{for } 0 \leq k \leq \frac{N}{2} \\ 0 & \text{for } \frac{N}{2}+1 \leq k \leq N \end{cases} \quad (1)$$

The type-I odd DCT $y[k]$ of a signal for $k = 0, \ldots, N-1$ is defined as

$$y[k] = 4 \sum_{n=0}^{N-1} c_{n,k} x[n] \cos\left(\frac{2\pi nk}{M}\right) \quad (2)$$

where the constants $c_{n,k} = 1$ for $n, k > 0$ and $c_{n,k} = \frac{1}{2}$ for $n, k = 0$ and $c_{n,k} = \frac{1}{\sqrt{2}}$ for the values of n, k, where only one of the index is 0 and $M = 2N - 1$. The DCT defined by Eq. 2 is a scaled version of the original orthogonal DCT with a factor of $2\sqrt{M}$.

We also define the even-symmetrized version $q[n]$ of the input signal,

$$q[n] = \begin{cases} x[n] & \text{for } n = 0, .., N-1 \\ x[M-n] & \text{for } n = N, \ldots, M-1 \end{cases} \quad (3)$$

A important property of $q[n]$ is that it has a real spectrum given by,

$$Q[k] = 2 \sum_{n=0}^{N-1} x[n] \cos\left(\frac{2\pi nk}{M}\right) \quad (4)$$

For signals with the zero-mean property in time and frequency domains, and using Eq. 2, 4, we get,

$$y[k] = 2Q[k] \quad (5)$$

for $k = 0, \ldots, N-1$. Let $\hat{y}$ denote the zero-padded DCT with $\hat{y}[k] = y[k]$ for $k = 0, \ldots, N-1$ and $\hat{y}[k] = 0$ for $k = N, \ldots, M-1$. From the definition of Fourier transform of the analytic signal in Eq. 1, and using the definition of the even symmetric signal in Eq. 3, we find that,

$$Q_a[k] = \hat{y}[k] \quad (6)$$

for $k = 0, \ldots, M-1$. This says that the AS spectrum of the even-symmetric signal is equal to the zero-padded DCT signal. In other words, the inverse DFT of the zero-padded DCT signal is the even-symmetric AS. Now, the auto-correlation of the DCT sequence is defined as,

$$r_y[\tau] = \frac{1}{N} \sum_{k=|\tau|}^{N-1} y[k] y[k-|\tau|] \quad (7)$$

Using Eq. 6 and Eq. 8, it can be shown that,

$$r_y[\tau] = \frac{1}{N} \sum_{n=0}^{M-1} |q_a[n]|^2 e^{-j\frac{2\pi n\tau}{M}} \quad (8)$$

i.e., the auto-correlation of the DCT signal and the squared magnitude (Hilbert envelope) of the even-symmetric AS are Fourier transform pairs. Thus, we can deduce that the linear prediction of DCT components results in AR model of the Hilbert envelope of the even-symmetrized signal.

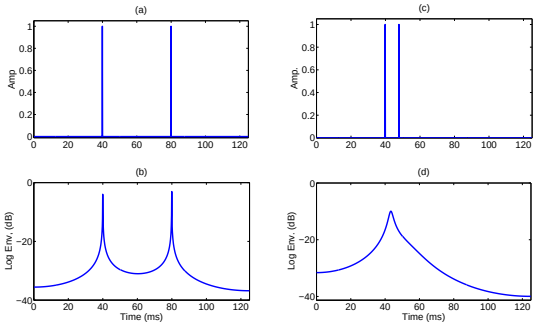

Figure 1: Plot of 125 ms of input signal in time domain (a), (c) and the corresponding log FDLP envelopes (b), (d).

3. Temporal Resolution in FDLP

In this section, we analyze the temporal resolution in FDLP models using signals with distinct temporal peaks (impulses). We use artificial signals for this analysis and compute FDLP models on the full-band DCT signal (as opposed to sub-band FDLP models used in speech feature extraction discussed in Sec. 4). The main factors considered here are the type of the DCT window, relative position of the temporal peak within the analysis window, model order for FDLP and type of LP method used (auto-correlation LP versus least squares LP). Before we discuss the resolution properties of FDLP, we propose an objective method to determine temporal resolution.

3.1. Defining the Temporal Resolution

We generate a signal with two peaks as shown in Fig. 1(a). The FDLP envelope of this signal (Fig. 1(c)) is computed by the application of linear prediction on DCT components. As seen in Fig. 1(a),(c), if the input signal has peaks which are far enough, two distinct peaks emerge in the FDLP envelope. As the spacing between the input peaks is decreased (Fig. 1(b)), the resulting peaks in the FDLP envelope start merging (Fig. 1(d)). The time interval between the two peaks in the input signal below which the resulting peaks in the FDLP envelope merge to form a single peak is referred to as the critical time-span. We define the resolution as the inverse of the critical time-span. In order to determine the resolution of the FDLP model, we use a peak picking mechanism on the log FDLP envelope.

In the discussions that follow, the input signal has two distinct peaks and the interval between the two peaks is varied. The FDLP envelope for this signal is input to the peak picking algorithm and the critical time-span is used to calculate the resolution.

3.2. Effect of Various Factors on Resolution

We analyze the effect of various factors on the temporal resolution, namely 1) the method of computing the linear prediction coefficients, 2) different types of window on the DCT signal, and 3) the FDLP model order. The main aspect of interest is the variation of the resolution as a function of the location of the first peak within the analysis window (Fig. 2) for a 125 ms signal (1000 samples at 8 kHz).

As shown in Fig. 2, we find that the resolution is not uniform within the analysis window and it is relatively poor at the boundaries of the analysis window. Fig. 2 (a) shows that the resolution can be improved by least-squares linear prediction

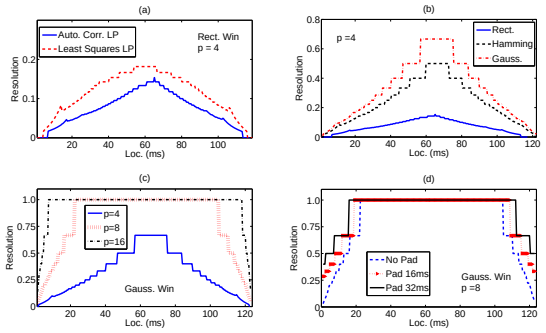

Figure 2: Normalized resolution in FDLP as function of the location of the first peak for a 125 ms long signal. (a) Two LP methods, (b) Various DCT windows, (c) FDLP model order and (d) symmetric padding at the boundaries.

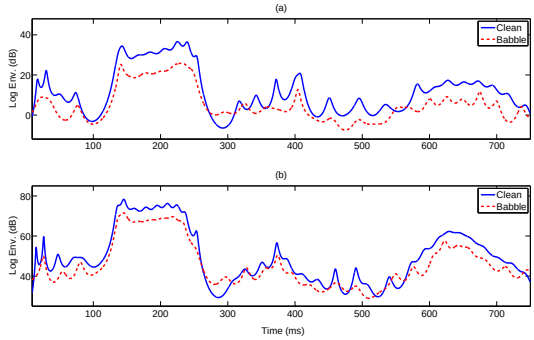

Figure 3: Log FDLP envelopes from clean and noisy (babble at 10 dB) sub-band speech. (a) Low resolution envelopes and (b) High resolution envelopes.

method replacing the standard auto-correlation method. The main drawback of the least-squares method is that the resulting AR model may be unstable (the roots of the AR polynomial lying outside the unit-circle). However, as observed in TDLP studies [3, 2], this can be partially alleviated when the number of samples N is significantly larger than model order p. Fig. 2 (b) shows that the Gaussian window in the DCT domain provides good temporal resolution among various window types considered here. An increase in the model order also improves the resolution as shown in Fig. 2 (c). However, this is not valid for noisy speech, where we found that increasing the model order beyond a limit tends to degrade the system performance as the model starts fitting the noisy regions.

In Fig. 2 (d), we provide one possible solution for improving the resolution at the boundaries of the analysis window. This is done by symmetric padding of the signal at the beginning and end of the analysis window. Once the FDLP envelope is derived, the portion of the envelope in the padded regions can be ignored. This eliminates the lower resolution parts of the FDLP model and improves the temporal resolution within the region of interest. We find that about 32 ms of padding provides good resolution at the boundaries.

In order to illustrate the effect of improved resolution in clean and noisy speech signals, the FDLP envelopes are estimated from sub-band (700-1100Hz) DCT components for clean

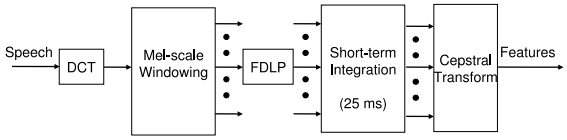

Figure 4: Feature extraction using sub-band FDLP models

Cond.	MFCC	FDLP-LR	FDLP-HR
Clean	31.5	31.1	30.9
0 dB	84.8	84.0	78.2
5 dB	77.6	77.9	70.8
10 dB	69.0	68.9	61.9
15 dB	59.1	58.8	52.2
20 dB	49.1	48.6	43.4
Avg.	67.9	67.6	38.7
CTS	46.9	46.7	44.4

Table 1: Phoneme error rates (PER) (%) in clean and noisy condition (Avg. performance over four noises.)

speech and noisy speech (babble noise at 10 dB). Fig. 3 (a) and (b) shows the plot of the envelopes without and with the modifications developed for higher resolution. As seen in this figure, estimating high resolution envelopes from noisy speech reduces the mismatch between clean and noisy conditions without making any assumptions about the noise.

4. Experiments and Results

We use a phoneme recognition system based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [9] trained on clean speech using the TIMIT database (16 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 [10].

For noisy phoneme recognition experiments, we create noisy version of the test data with additive noise (Babble, Restaurant, Ex-hall and Subway) at various SNRs (0,5,10,15,20 dB). The ANN models are trained using the clean speech data and they are tested with noisy and clean versions of the test data. The baseline features are MFCC features [11] with a 9 frame context [10] forming an ANN input vector of dimension 351.

The feature extraction scheme using FDLP is shown in Fig. 4. Long segments of the input signal (2000 ms segments) are analyzed using DCT. Gaussian windows that vary in width and position according to mel perceptual frequency scale are applied on DCT and linear prediction is performed on the windowed DCT components to obtain the FDLP envelopes on frequency sub-bands. These envelopes are integrated in 25 ms frames with a shift of 10 ms to obtain sub-band energy representation. The application of logarithm and DCT across sub-bands provides cepstral features. We derive delta and acceleration features and use a 9 frame context on the FDLP features to yield 351 features at the input of ANN. In order to illustrate the usefulness of improved temporal resolution in FDLP, we compare the performance of the FDLP model proposed in this

	Clean Speech		Noisy Speech	
Class	MFCC	FDLP-HR	MFCC	FDLP-HR
Vowel	13.4	12.9	14.2	16.1
Plosive	17.3	18.0	91.2	87.4
Semi-Vowel	24.9	25.2	64.9	62.6
Fricative	15.6	15.3	23.0	20.8
Nasal	17.0	17.0	51.4	42.0

Table 2: Broad class phoneme recognition error rate (%) in clean and noisy condition (babble at 10 dB SNR).

work and its earlier implementation [7] (without the proposed modifications). The old implementation uses auto-correlation method of LP (75 poles per second per sub-band) and is denoted as FDLP-Low-Res (FDLP-LR). For the proposed features, we obtain high resolution FDLP envelopes using the parameters detailed in Sec. 3, namely the application of least-squares LP method, Gaussian mel-spaced DCT windows, symmetric padding at the boundaries and a higher model order (100 poles per second per sub-band). These features are denoted as FDLP-High-Res (FDLP-HR).

The results for various phoneme recognition experiments are shown in Table. 1. In these experiments, the FDLP-LR features perform similar to the baseline MFCC features in clean and noisy conditions. The FDLP-HR features provide significant improvements in noisy conditions (average relative improvements of about 10 % over the baseline). These improvements are consistent across all SNR levels from 0-20 dB. The results show that an improved resolution in the sub-band FDLP envelope estimation translates to improvements in phoneme recognition performance.

We also perform phoneme recognition experiments in matched telephone channel conditions using large amounts of conversational telephone speech data (CTS) data [12]. The training data consists of 100 hours of speech, cross-validation data set consists of 30 hours of speech and the test data consists of 10 hours of speech. It is labeled using 45 phonemes (44 speech classes and 1 silence class) obtained by force aligning the word transcriptions to the previously trained HMM-GMM models [12]. Here, the ANN consists of 5000 hidden neurons, and 45 output neurons (with soft max non-linearity) representing the phoneme classes. The results of these phoneme recognition experiments are reported in last row of Table 1. In these experiments, the FDLP-HR features provide noticeable improvements (relative improvements of 5 %).

In order to obtain more insight into the observed improvements, we show the broad phonetic class error rate in Table 2. In clean conditions, the performance for various phoneme classes are similar for proposed front-end and MFCC features. In the noisy case (babble noise at 10 dB), the FDLP-HR provides noticeable improvements for plosives and nasals, where the fine temporal representation is important. This table also shows that noise has adverse effects on certain phoneme classes like semi-vowels, nasals and plosives as opposed to vowels and fricatives.

The improvements reported here are obtained without any assumptions about the noise or distortions. For instance, spectral subtraction and gain normalization techniques can be applied with the FDLP feature to improve the performance in noise [13]. Similarly, other noise compensation techniques can be applied along with the baseline MFCC features to improve the performance. In future, we plan to apply some of these techniques in addition to the high resolution envelopes reported here.

5. Summary

We have analyzed the temporal resolution properties in FDLP envelope. Our analyzes show that the resolution is not uniform across the temporal segment and a higher resolution is obtained at the center of the window. In order to improve the temporal resolution of FDLP, we suggest several techniques like the use of Gaussian DCT window, least-squares method of LP estimation and symmetric padding at the boundaries. These methods improve the resolution of the FDLP envelopes in clean and noisy conditions. Phoneme recognition experiments using noisy speech show noticeable improvements with higher resolution FDLP models.

6. Acknowledgements

This research was funded by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015.

7. References

[1] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *JASA*, vol. 47, pp. 637–655, 1970.

[2] D. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood," *Proc. of the IEEE*, vol. 70, pp. 975–989, 1982.

[3] J. Makhoul, "Linear prediction: A tutorial review," *Proc. of the IEEE*, vol. 63, pp. 561–580, 1975.

[4] H. Hermansky, "Perceptual linear predictive analysis of speech," *JASA*, vol. 87, pp. 1738–1752, 1990.

[5] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *JASA*, vol. 105, p. 1912, 1999.

[6] M. Athineos and D. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Tran. on Sig. Proc.*, vol. 55, pp. 5237–5245, 2007.

[7] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Sig. Proc. Lett.*, vol. 15, pp. 681–684, 2008.

[8] L. Marple Jr, "Computing the discrete-time analytic signal via fft," *IEEE Tran. on Sig. Proc.*, vol. 47, pp. 2600–2603, 1999.

[9] H. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer, 1994.

[10] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai-Doss, "Exploiting contextual information for improved phoneme recognition," in *IEEE ICASSP 2008*. IEEE, 2008, pp. 4449–4452.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Tran. on Acous., Speech Sig. Proc.*, vol. 28, pp. 357–366, 1980.

[12] T. Hain and L. e. a. Burget, "The development of the ami system for the transcription of speech in meetings," *Machine Learning for Multimodal Interaction*, pp. 344–356, 2006.

[13] S. Ganapathy, S. Thomas, and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum," *JASA*, vol. 128, pp. 3769–3780, 2010.