Temporal resolution analysis in frequency domain linear prediction

Sriram Ganapathy[1] and Hynek Hermansky[2]

[1] *IBM T.J Watson Research Center, Yorktown Heights, NY, USA.*

[2] *Dept. of ECE, Johns Hopkins University, Baltimore, USA.*

Email: ganapath@us.ibm.com, hynek@jhu.edu

Running title: Resolution analysis in frequency domain prediction

Abstract

Frequency domain linear prediction (FDLP) is a technique for auto-regressive (AR) modeling of Hilbert envelopes. In this letter, the resolution properties of the FDLP model are investigated using synthetic signals with impulses immersed in noise. The effect of various factors are studied which affect the temporal resolution and this analysis suggests ways to improve the resolution of the FDLP envelopes in noisy speech. The high resolution FDLP envelopes are used to derive robust features for phoneme recognition in noisy and reverberant speech. In these experiments, the FDLP features derived from high resolution envelopes provide significant improvements.

© "2012" Acoustical Society of America

PACS numbers: 43.72.Ne, 43.72.Ar

I. Introduction

Frequency domain linear prediction (FDLP) analysis approximates the Hilbert envelope of the signal by its auto-regressive model [Athineos and Ellis, 2007; Kumerasan and Rao, 1999]. In this letter, we analyze the temporal resolution properties of FDLP using an objective measure. More than a century ago, Lord Rayleigh [Rayleigh, 1880] offered one possible definition for objective measurement of resolution: "As the power of a telescope is measured by the closeness of the double stars which it can resolve, so the power of a spectroscope ought to be measured by the closeness of the closest double lines in the spectrum which it is competent to resolve." Along these lines, we propose to define the temporal resolution of FDLP.

In our prior work, we analyzed some factors which affect the resolution properties of FDLP in clean conditions [Ganapathy and Hermansky, 2012]. However, some of the factors useful for improving the resolution in a clean signal do not show significant benefits in the presence of noise. In this study, we show additional factors which improve the resolution in noisy signals.

In order to investigate the resolution properties in noise, the pilot signals used in our study are corrupted with white noise at 20 dB. The separation between the impulses is varied and the maximum separation between the two peaks in the input for which the output of the AR model has a single peak is determined (critical time-span). Then, the resolution is computed as the inverse of the critical time-span. In this letter, we show that 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.

In the past, it has been shown that time domain linear prediction (TDLP) can be modified to estimate a transformed spectral envelope [Hermansky, 1983]. In the FDLP framework, this can be incorporated by deriving the spectral autocorrelations from transformed Hilbert envelopes. In this paper, we show that the transformed LP method applied to FDLP significantly improves the temporal resolution.

When speech is corrupted by noise or reverberation, temporal envelopes estimated from noisy speech do not match those obtained from clean training conditions. Using the factors derived from the resolution analysis of noisy signals, we show that the high resolution estimation of the envelopes can reduce this mis-match. Phoneme recognition in noisy speech continues to be a challenging task because of the mis-match between the acoustic feature derived from clean training conditions and noisy test conditions. In this paper, we show that the use of the proposed techniques to improve the resolution of sub-band envelopes derived from noisy speech can lead to a robust feature extraction technique. For phoneme recognition of noisy speech, we develop a feature extraction scheme which uses high resolution (HR) envelopes from FDLP. In these phoneme recognition experiments, the FDLP-HR features provide significant improvements in all mis-matched conditions like additive noise, reverberation and telephone channel noise.

The rest of the paper is organized as follows. The temporal resolution analysis of FDLP is provided in Sec. II. Phoneme recognition experiments using FDLP features is described in Sec. III, followed by a summary in Sec. IV.

II. Temporal Resolution in FDLP

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, the auto-correlation of DCT sequence and the Hilbert envelope (squared magnitude of analytic signal (AS)) are related by the Fourier transform [Athineos and Ellis, 2007; Kumerasan and Rao, 1999]. Let $y[k]$ denote the DCT sequence of a discrete input signal $x[n]$ for $n = 0, \;..\;, N-1$. The auto-correlation of the DCT sequence is defined as,

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

Let $q[n]$ denote the even-symmetrized and $q_a[n]$ denote the corresponding analytic signal defined using Fourier transform [Marple, 1999]. It can be shown that [Athineos and Ellis, 2007; Ganapathy and Hermansky, 2012],

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

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.

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. III). The main factors considered here are the type of the DCT window, relative position of the temporal peak within the analysis window 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.

We generate a pilot signal with two peaks and artificially add white noise at 20 dB. The use of additive white noise helps in understanding the factors which improve the resolution in noisy signals. The FDLP envelope of the pilot signal is computed by the application of linear prediction on DCT components. As the spacing between the input peaks is decreased, the resulting peaks in the FDLP envelope merge. The time interval between the two peaks for 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.

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) transform domain LP estimation. 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. 1) for a 125 ms signal (1000 samples at 8 kHz).

As shown in Fig. 1, 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. 1 (a) shows that the Gaussian window in the DCT domain provides moderate improvements in temporal resolution. This is because the spectrum of a smooth analysis window like the Gaussian window has the advantage of lower leakage in its side-lobes compared to a rectangular window. Fig. 1 (b) shows that the resolution can be improved by a least-squares linear prediction method replacing the standard auto-correlation method. The main advantage of the least-squares LP method is that it can accurately estimate the peaks for small sample size N as it based on unbiased auto-correlation estimates [Makhoul, 1975].

Fig. 1 (c) shows that application of transform domain LP [Hermansky, 1983], provides significant improvements in resolution. This is attributed to the fact that the peaks are enhanced in the transformed Hilbert envelope and this enables an efficient estimation of the peaks in the FDLP model.

In Fig. 1 (d), we provide one possible solution for improving the resolution at the boundaries of the analysis window at the expense of a moderate degradation in resolution at the center. 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 method improves the lower resolution parts of the FDLP model at the boundaries. We find that about 32 ms of padding provides good resolution at the boundaries.

In order to illustrate the effect of proposed techniques for speech signals, we derive FDLP envelopes for clean speech, noisy speech (babble noise at 10 dB), artificially reverberated speech ($T60 = 400ms$) and telephone speech (Electret Mic.). In Fig. 2, we compare the FDLP envelopes using the proposed modifications for high-resolution estimation (FDLP-HR) with the previous low-resolution implementation (FDLP-LR). As seen in this plot, the FDLP-HR estimation procedure reduces the mis-match between clean and noisy conditions consistently for additive noise, reverberation and telephone channel speech without making any assumptions about the type of distortion. When features are derived from FDLP-HR envelopes, the characteristics of features obtained from noisy speech match better with those obtained from clean training conditions resulting in robust phoneme recognition.

III. Experiments and Results

We use a phoneme recognition system based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [Bourlard and Morgan, 1994] trained on clean speech using the TIMIT database ($8, 16$ 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 is hand-labeled using the standard set of 39 phonemes [Pinto et al., 2007].

For noisy phoneme recognition experiments, we create 1) noisy versions of the test data with additive noise (Babble, Restaurant, Ex-hall and Subway) at various SNRs (0,5,10,15,20 dB) and 2) reverberated versions of the test data using different impulse responses with reverberation time ranging from 200 to 600 ms [ICSI Meeting Task]. We also experiment with natural telephone speech derived from HTIMIT database [Reynolds, 1997] which has a re-recording of the TIMIT data using various types of telephone channel involving Carbon microphone, Electret microphone and cordless phone. The ANN models are trained using the clean speech data at $16kHz$ and they are tested with additive noise and clean versions of the test data. For experiments with reverberated speech and telephone speech, we use the ANN models trained using clean speech data at 8 kHz. The baseline features are MFCC features [Davis and Mermelstein, 1980] with a 9 frame context [Pinto et al., 2007] forming an ANN input vector of dimension 351.

The feature extraction scheme using FDLP is shown in Fig. 3 [Thomas et al., 2008]. Long segments of the input signal (2000 ms segments) are analyzed using DCT. Gaussian windows that vary in width and position according to the mel perceptual frequency scale are applied on DCT and linear prediction is performed on the windowed DCT components to obtain the FDLP envelopes on sub-bands. We use transform domain LP estimation with an expansion factor $r = 1.5$. The use of expansion factors beyond 1.5 caused modification of feature characteristics resulting in performance degradation.

The FDLP envelopes are integrated in 25 ms frames with a shift of 10 ms. The application of logarithm on the FDLP envelopes followed by a 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 for ANN.

We compare the performance of the FDLP model proposed in this work with its previous implementation [Thomas et al., 2008] (without the proposed modifications). The old implementation uses an auto-correlation method of LP (75 poles per second per sub-band) without symmetric padding and is denoted as FDLP-low-resolution (FDLP-LR). For the proposed features, we obtain high resolution FDLP envelopes using the parameters detailed in Sec. II, namely the application of least-squares LP method, Gaussian mel-spaced DCT windows, symmetric padding at the boundaries and transform domain LP estimation with $r = 1.5$. We use a slightly higher model order (100 poles per second per sub-band). These features are denoted as FDLP-high-resolution (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 8%, 12% and 14% over the baseline for additive noise conditions, reverberation and telephone channel noise respectively). These improvements are consistent across all SNR levels from 0-20 dB, various types of room impulse responses and telephone microphone types. These results show that an improved resolution in the sub-band FDLP envelope estimation translates to improvements in phoneme recognition performance.

In order to investigate the source of improvements for FDLP-HR features, we report the accuracy of individual broad phonetic classes in Table 2. In all the noisy conditions considered here, the FDLP-HR features provide significant improvements for plosives and nasals. As the plosives and nasals are characterized by temporal dynamics (transient behavior and envelope valleys respectively), high resolution estimation of FDLP envelopes benefits the recognition of these classes. For reverberation conditions, significant improvements are also observed for vowels and fricatives.

IV. Summary

We have analyzed the temporal resolution properties in FDLP envelope. 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, application of transform domain LP 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 high resolution FDLP models.

Acknowledgments

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

References and links

Athineos, M., and Ellis, D.P.W. (**2007**). "Autoregressive modelling of temporal envelopes," IEEE Trans. on Signal Proc., Vol. 55, pp. 5237-5245.

Bourlard, H. and Morgan, N. (**1994**). "Connectionist speech recognition - A hybrid approach", Kluwer Academic Publishers.

Davis, S. and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoust. Speech and Signal Proc., Vol. 28, pp. 357-366.

Ganapathy, S., Hermansky, H. (**2012**). "Robust Phoneme Recognition Using High-Resolution Temporal Envelopes," Proc. of INTERSPEECH.

Hermansky, H. (**1983**). "Analysis and synthesis of speech based on spectral transform linear predictive method," Proc. of ICASSP, Vol. 8, pp. 777-780.

Kumerasan, R. and Rao, A. (**1999**). "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," J. Acoust. Soc. Am., Vol. 105(3), pp. 1912-1924.

Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. of IEEE, Vol. 63, pp. 561-580.

Marple, Jr. L., (**1999**). "Computing the discrete-time analytic signal via FFT," IEEE Trans. on Signal Proc., Vol. 47, pp. 2600-2603.

Pinto, J., Yegnanarayana, B., Hermansky, H. and Doss, M.M. (**2007**). "Exploiting contextual information for improved phoneme recognition," Proc. of INTERSPEECH, pp. 1817-1820.

Rayleigh, L. (**1880**). "Scientific papers, Vol. 1," Proc. Lond. Math. Soc., Vol. 11, pp. 57-80.

Reynolds, D.A. (**1997**). "HTIMIT and LLHDB: speech corpora for the study of hand set transducer effects," Proc. ICASSP, pp. 1535-1538.

"The ICSI Meeting Recorder Project," <http://www.icsi.berkeley.edu/Speech/mr> (date last viewed 8/18/2012).

Thomas, S., Ganapathy, S. and Hermansky, H. (**2008**). "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction," IEEE Sig. Proc. Let., Vol. 15, pp. 681-684.

Table 1. Phoneme recognition accuracy (%) in clean, additive noise (Avg. performance over four noises.), reverberation (Avg. performance over two impulse responses) and telephone speech.

Cond.	MFCC	FDLP-LR	FDLP-HR
Clean (16 kHz)	68.5	68.9	67.8
Additive Noise			
0 dB	15.2	16.0	20.4
5 dB	22.4	22.2	27.5
10 dB	31.0	31.1	36.4
15 dB	40.9	41.2	46.4
20 dB	50.9	51.4	55.6
Avg.	32.1	32.4	37.3
Clean (8 kHz)	66.3	66.4	65.9
Reverberation			
Revb. (200ms)	29.2	32.2	45.1
Revb. (400ms)	26.4	28.8	34.0
Revb. (600ms)	21.5	21.6	24.2
Avg.	25.7	27.5	34.4
Telephone Speech			
Carbon Mic.	33.6	31.6	41.6
Electret Mic.	32.8	30.6	43.3
Cordless	24.7	24.0	34.9
Avg.	30.4	28.7	39.9

Table 2. Broad class phoneme recognition accuracy (%) in clean, additive noisy condition (babble at 10 dB), reverberation (400 ms) and telephone speech (Electret Mic).

Feat.	Vowel	Plosive	Semi-Vowel	Fricative	Nasal
Additive Noise					
MFCC	85.8	8.8	35.1	77.0	48.6
FDLP-LR	83.8	11.2	25.0	86.0	49.2
FDLP-HR	83.4	13.9	36.5	77.9	58.2
Reverberation					
MFCC	68.6	4.5	50.1	21.8	9.0
FDLP-LR	68.2	11.0	48.4	19.9	12.2
FDLP-HR	75.6	16.9	44.5	41.4	21.2
Telephone Speech					
MFCC	83.0	62.1	64.3	65.5	52.2
FDLP-LR	82.2	63.9	62.1	60.6	51.0
FDLP-HR	84.0	67.9	67.8	73.4	63.1

List of figures

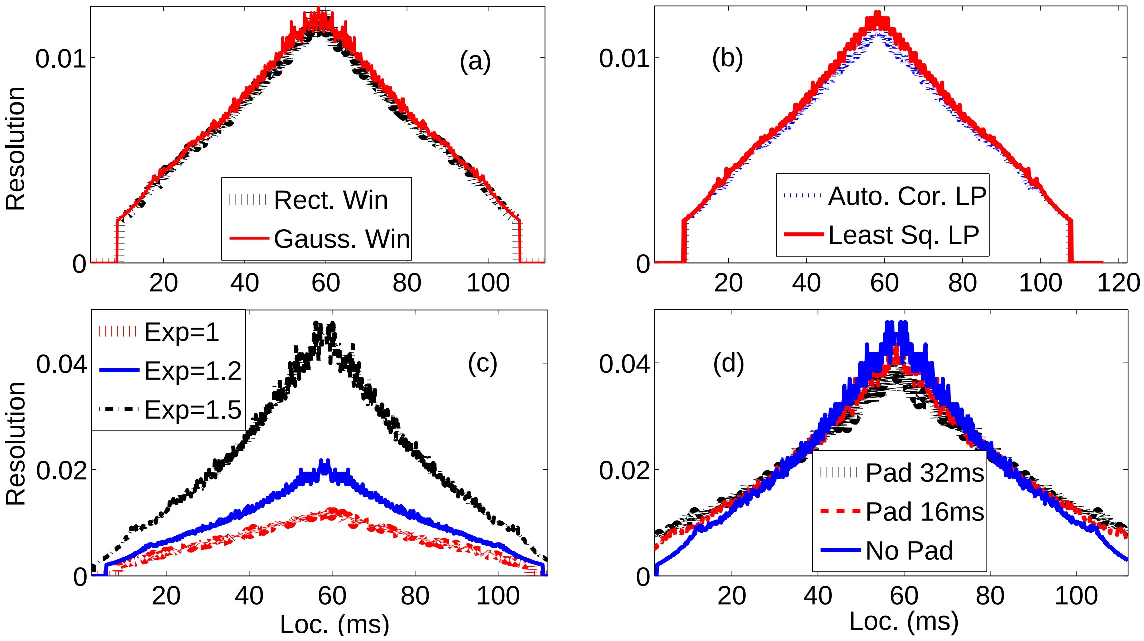
(a)
Resolution
0.01
0
20
40
60
80
100
Rect. Win
Gauss. Win
(b)
0.01
0
20
40
60
80
100
120
Auto. Cor. LP
Least Sq. LP
(c)
Resolution
0.04
0.02
0
20
40
60
80
100
Loc. (ms)
Exp=1
Exp=1.2
Exp=1.5
(d)
0.04
0.02
0
20
40
60
80
100
Loc. (ms)
Pad 32ms
Pad 16ms
No Pad

FIG. 1.

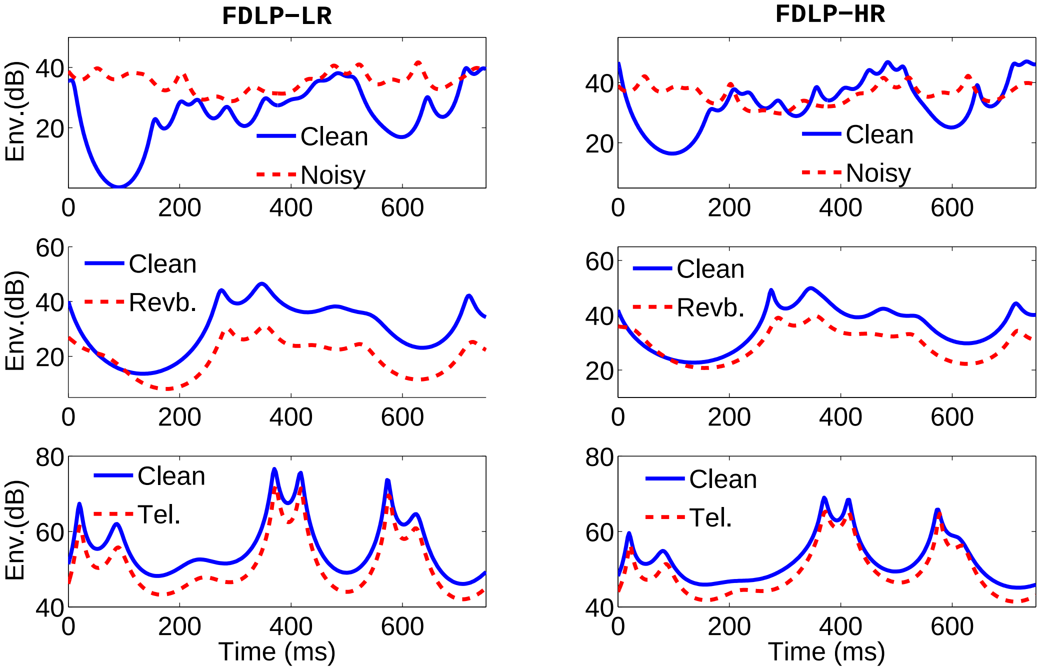
FDLP−LR
FDLP−HR
Env.(dB)
Clean
Noisy
Revb.
Tel.
0
200
400
600
20
40
60
80
Time (ms)

FIG. 2.

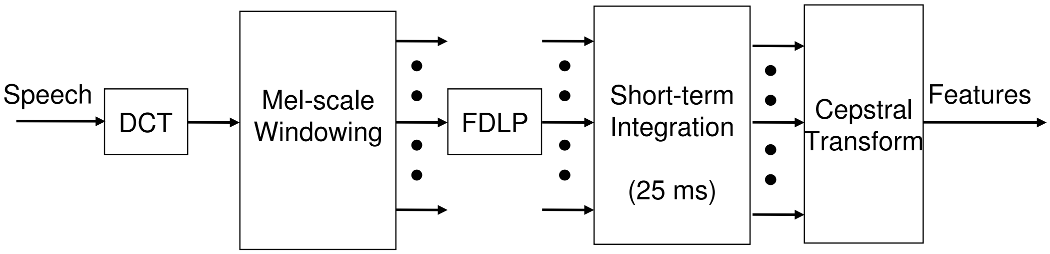
Speech
DCT
Mel-scale Windowing
FDLP
Short-term Integration
(25 ms)
Cepstral Transform
Features

FIG. 3.