APPLICATIONS OF SIGNAL ANALYSIS USING AUTOREGRESSIVE MODELS FOR AMPLITUDE MODULATION

Sriram Ganapathy[1], *Samuel Thomas*[1], *Petr Motlicek*[2] *and Hynek Hermansky*[1,3]

[1]Department of Electrical and Computer Engineering, Johns Hopkins University, USA
[2]Idiap Research Institute, Martigny, Switzerland
[3]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{ganapathy,samuel,hynek}@jhu.edu, motlicek@idiap.ch

ABSTRACT

Frequency Domain Linear Prediction (FDLP) represents an efficient technique for representing the long-term amplitude modulations (AM) of speech/audio signals using autoregressive models. For the proposed analysis technique, relatively long temporal segments (1000 ms) of the input signal are decomposed into a set of sub-bands. FDLP is applied on each sub-band to model the temporal envelopes. The residual of the linear prediction represents the frequency modulations (FM) in the sub-band signal. In this paper, we present several applications of the proposed AM-FM decomposition technique for a variety of tasks like wide-band audio coding, speech recognition in reverberant environments and robust feature extraction for phoneme recognition.

Index Terms— Frequency Domain Linear Prediction (FDLP), AM-FM decomposition, Wide-band audio coding, Robust features for speech recognition.

1. INTRODUCTION

Conventionally, signal analysis techniques for speech/audio signals start with estimating the spectral content of relatively short (about 10-40 ms) segments of the signal (short-term spectral or transform domain coefficients). Each estimated vector of spectral components represents a sample of the underlying dynamic process in production of these signals at a given time-frame. Stacking such estimates of the short-term spectra in time provides a two-dimensional (time-frequency) representation of these signals that forms the basis for many speech and audio processing systems (for example AAC [1], PLP [2]). However, the problems of time-frequency resolution and efficient sampling of the short-term representation are addressed in an ad-hoc manner.

Alternatively, 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 signals (for example LP-TRAP [3]). Spectral representation of sub-band energies, also called "Modulation Spectra", have been used in many engineering applications. Early work done in [4] for predicting speech intelligibility and characterizing room acoustics are now widely used in the industry [5]. Recently, there has been many applications of such concepts for robust speech recognition [6], audio coding [7] and noise suppression [8].

In this paper, we propose to connect the modulation based approaches for speech/audio applications using Frequency Domain Linear Prediction (FDLP). Our approach is based on the assumption that speech/audio signals in critical bands can be represented as a modulated signal [11], with the AM component obtained using Hilbert envelope estimate and the FM component obtained from the Hilbert carrier. The sub-band temporal envelopes are estimated using FDLP, which forms an efficient technique for autoregressive modelling of temporal envelopes of the signal [9, 10]. FDLP exploits the predictability of the slowly varying long-term AM envelopes of speech/audio signals in critical bands. The FDLP residual signal forms the sub-band FM component.

This paper presents various applications of the proposed AM-FM decomposition for signal analysis in speech and audio processing systems. Modulation spectral components, derived using FDLP envelopes, are used as features for phoneme recognition task in noisy speech [12]. For speech recognition in reverberant environments, AM envelopes extracted from narrow sub-bands of long segments of the signal are gain normalized to alleviate the effects of reverberation artifacts in speech signal. Short-term features, derived by integrating the gain normalized temporal envelopes, provide good robustness in reverberant environments [13]. For wide-band audio coding applications, the AM and FM components are quantized and transmitted [14]. Subjective and objective quality evaluations show that the FDLP based audio codec at $\sim$ 48 kbps provides similar results compared to the state-of-art codecs at this bit-rate.

An overview of the engineering applications of the FDLP analysis technique is shown in Fig. 1. Long term segments of the input speech/audio signal are decomposed into a set of sub-bands. FDLP is used for AM-FM decomposition in each sub-band (Sec. 2). The sub-band AM envelopes are gain normalized for the task of speech recognition in reverberant environments (Sec. 3). The AM envelopes are converted into modulation spectral components for phoneme recognition in noisy speech (Sec. 4). For wide-band audio coding applications, the AM and FM components are quantized and encoded (Sec. 5).

2. FREQUENCY DOMAIN LINEAR PREDICTION

Typically, autoregressive (AR) models have been used in speech/audio applications for representing the envelope of the power spectrum of the signal by performing the operation of Time Domain Linear Prediction (TDLP) [15]. This paper utilizes AR

This work was partially supported by grants from ICSI Berkeley, USA; European IST Programme DIRAC Project FP6-0027787; the Swiss National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM)2"

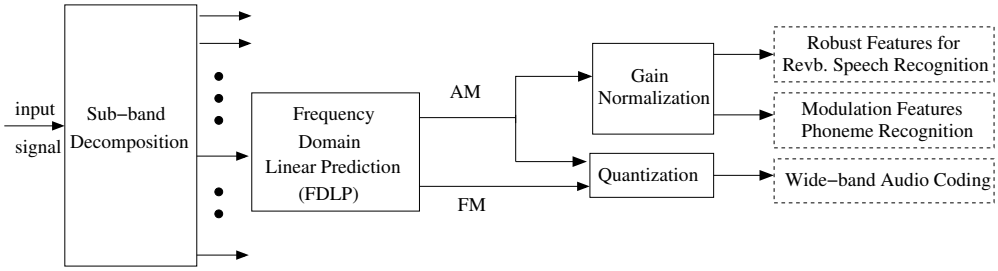

Figure 1: Application of the FDLP technique for speech and audio processing systems - Reverberant speech recognition (Sec. 3), Modulation features for phoneme recognition (Sec. 4) and Wide-band audio coding (Sec. 5).

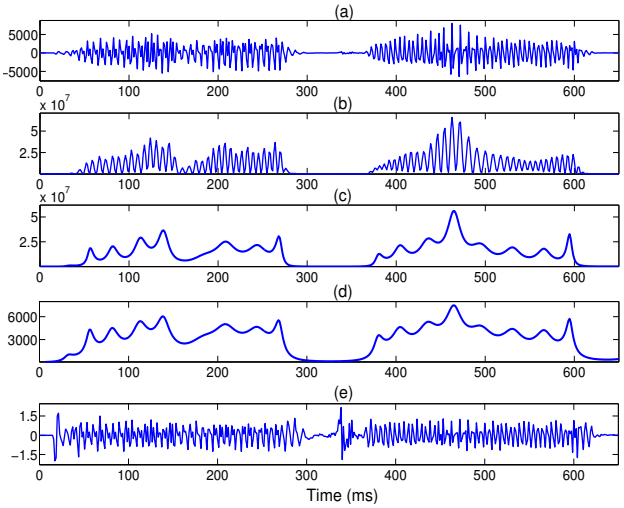

Figure 2: Illustration of the all-pole modeling property of FDLP. (a) a portion of the sub-band speech signal, (b) its Hilbert envelope (c) all pole model obtained using FDLP, (d) sub-band AM envelope estimate and (e) the sub-band FM component.

models for obtaining smoothed, minimum phase, parametric models of temporal rather than spectral envelopes. The duality between time and frequency domains means that AR modeling can be applied equally well to discrete spectral representations of the signal instead of time-domain signal samples. 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 by TDLP [15]).

In our experiments, we implement FDLP in two steps. Long term segments of the input speech/audio signal (1000 ms) are decomposed into a set of sub-bands by windowing the discrete cosine transform (DCT). Then, we apply linear prediction on the sub-band DCT components to derive the AR models of Hilbert envelopes [10].

For many modulated signals in the real world, the quadrature version of a real input signal and its Hilbert transform are identical [16]. This means that the Hilbert envelope approximates the squared AM envelope of the signal. Thus, FDLP estimates the AM envelope of the signal and the FDLP residual contains the FM component of the signal. Acoustic signals in sub-bands are modulated signals [11] and hence, FDLP is used for AM-FM decomposition of sub-band signals. Fig. 2 shows (a) a portion of sub-band speech signal, (b) its Hilbert envelope, (c) an all pole model of the Hilbert Envelope using FDLP, (d) the AM envelope estimate obtained as the square root of FDLP envelope and (e) the FDLP residual signal representing the sub-band FM component.

3. REVERBERANT SPEECH RECOGNITION

For signal analysis in long temporal windows with narrow sub-bands, the spectral autocorrelation function of the reverberant speech is the multiplication of spectral autocorrelation function of the clean speech with that of the room impulse response. For the room impulse response, the spectral autocorrelation function in narrow 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 the reverberant speech [13].

For feature extraction, segments of the input speech signal (of the order of 1000 ms) are decomposed into a number of narrow sub-bands, where FDLP is applied to obtain a parametric model of the temporal envelope. These temporal envelopes are gain normalized to suppress the reverberation artifacts in speech signal. The whole set of sub-band temporal envelopes forms a two dimensional (time-frequency) representation of the input signal energy. This two-dimensional representation is convolved with a rectangular window of duration 25 ms and re-sampled at a rate of 100 Hz (10 ms intervals, similar to the estimation of short term power spectrum in conventional feature extraction techniques). These sub-sampled short-term spectral energies are converted to short-term cepstral features similar to the PLP feature extraction technique.

We apply the proposed features for a connected word recognition task on a digits corpus [13] using the Aurora evaluation system [17]. The models are trained using TIDIGITS training dataset which contains 8400 clean speech utterances. These models are tested on the digits corpus recorded using far-field microphones which forms part of Aurora-5 speech database [18] (ICSI Meeting task). The test data consists of four sets with 2790 utterances each. Each of these sets correspond to speech recorded simultaneously using four different far-field microphones. The results for the proposed FDLP technique, along with those obtained for other robust feature extraction techniques namely Cepstral Mean Subtraction (CMS) [19], Long Term Log Spectral Subtraction (LTLSS) [20] and PLP [2], is shown in Table 1. In reverberant speech recognition experiments, FDLP features provide a relative improvement

	PLP	CMS	LTLSS	FDLP
Clean	99.7	99.7	99.6	99.1
Revb.	69.1	73.6	76.8	87.0

Table 1: Word accuracies (%) for clean and reverberated speech tested on models trained with clean utterances.

	PLP	ETSI	MRASTA	FDLP
Clean	64.9	63.1	63.9	65.4
Tel.	34.4	47.7	47.5	52.7

Table 2: Phoneme recognition accuracies (%) for clean and telephone speech tested on models trained with clean utterances.

of about 44% over the other feature extraction techniques.

4. MODULATION FEATURES FOR PHONEME RECOGNITION

The long-term sub-band FDLP envelopes form a compact representation of the temporal dynamics over long regions of the speech signal. For the task of phoneme recognition, the sub-band temporal envelopes are compressed using a static compression scheme which is a logarithmic function and a dynamic compression scheme [12]. The dynamic compression is realized by an adaptation circuit consisting of five consecutive nonlinear adaptation loops [21]. 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. Sudden transitions in the sub-band envelope that are very fast compared to the time constants of the adaptation loops are amplified linearly at the output due to the slow changes in the low pass filter output, whereas the slowly changing regions of the input signal are compressed. 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 $0 - 70$ Hz region with a resolution of 5 Hz. The static and adaptive modulation features for each sub-band are stacked together to obtain modulation features for each sub-band.

The proposed features are used for a phoneme recognition task on the HTIMIT database [22]. We use a phoneme recognition system based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [23]. The models are trained on clean speech using the TIMIT database downsampled to 8 kHz. The training data consists of 3000 utterances, cross-validation data set consists of 696 utterances and the test data set consists of 1344 utterances. For phoneme recognition experiments in telephone channel, speech data collected from 9 telephone sets in the HTIMIT database are used, which introduce a variety of channel distortions in the test signal. For each of these telephone channels, 842 test utterances, having clean recordings in the TIMIT test set, are used. The system is trained only on the 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 [12]. Table 2 shows results for the proposed technique compared with those obtained for other robust feature extraction techniques namely, Multi-resolution RASTA (MRASTA) [24], and the Advanced-ETSI (noise-robust) distributed speech recognition front-end [25]. For the task of phoneme recognition in telephone speech, the proposed features, on the average, provide a relative improvement of about 10% over the other feature extraction techniques considered.

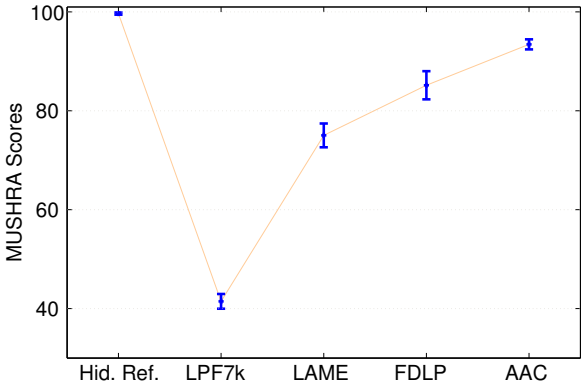

Figure 3: MUSHRA results for 6 speech/audio samples using three coded versions at 48 kbps FDLP codec (FDLP), MPEG-4 HE-AAC (AAC) and LAME-MP3 (LAME)), hidden reference (Hid. Ref.) and 7 kHz low-pass filtered anchor (LPF7k).

5. WIDE-BAND AUDIO CODING

In wide-band audio coding, long segments of the input speech/audio signals are analyzed using a non-uniform Quadrature Mirror Filter (QMF) bank to decompose the signal into frequency sub-bands [14]. For each sub-band signal, the Line Spectral Frequency (LSF) parameters of the AR model are quantized using Vector Quantization (VQ). The remaining sub-band FM components are split into relatively short frames (50 ms) and transformed using the Modified Discrete Cosine Transform (MDCT). We use the sine window with 50 % overlap for the MDCT analysis. The MDCT coefficients of the FM signal are quantized using the split VQ approach. Although the split VQ approach is suboptimal compared to a full search VQ, it reduces the computational complexity and memory requirements to manageable limits without severely degrading the VQ performance. The VQ quantization levels are Huffman encoded for further reduction of bit-rates. At the decoder, quantized MDCT coefficients of the FDLP residual signals are reconstructed and transformed back to the time-domain using inverse MDCT. The reconstructed FDLP envelopes (from LSF parameters) are used to modulate the corresponding sub-band residual signals. Finally, sub-band synthesis is applied to reconstruct the full-band signal. Without the use of any psycho-acoustic models, the FDLP codec provides efficient audio compression for speech/audio content at 48 kbps.

The subjective evaluation of the proposed audio codec is performed using single channel audio signals (sampled at 48 kHz) present in the framework for exploration of speech and audio coding [26]. This database comprises of speech, music and speech over music recordings. Recently, these audio samples were used for the development of a low bit-rate unified speech/audio codec [27]. The MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) methodology for subjective evaluation is em-

ployed. It is defined by ITU-R recommendation BS.1534 [29]. We perform the MUSHRA tests on 6 speech/audio samples from the database with 6 listeners using the FDLP codec along with LAME-MP3 (MPEG 1, layer 3) [28] at 48 kbps denoted as LAME, and MPEG-4 HE-AAC v1 [1] at 48 kbps denoted as AAC. The HE-AAC coder is the combination of spectral band replication (SBR) [30] and advanced audio coding (AAC). The mean MUSHRA scores (with 95% confidence interval) for the subjective listening tests, shown in Fig. 3, indicate that the subjective quality of the proposed FDLP codec is slightly poorer than the AAC codec but better than the LAME codec.

6. CONCLUSIONS

We have presented a novel signal analysis technique based on AR modelling of amplitude modulations. This technique provides an AM-FM decomposition of speech/audio signals in sub-bands. We have applied the proposed AM-FM decomposition technique for a variety of tasks like wide-band audio coding and speech recognition in noisy environments. The results show the usefulness of the proposed signal analysis technique for these speech and audio applications.

7. REFERENCES

[1] J. Herre and J.M. Dietz, "MPEG-4 high-efficiency AAC coding", *IEEE Signal Proc. Magazine*, Vol. 25(3), pp. 137-142, 2008.

[2] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", *J. Acoust. Soc. Am.*, Vol. 87(4), pp. 1738-1752, 1990.

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

[4] T. Houtgast, H.J.M. Steeneken and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function, I. General room acoustics", *Acoustica 46*, pp. 60-72, 1980.

[5] IEC 60268-16, "Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index", *<http://www.iec.ch>*.

[6] 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.

[7] M.S. Vinton and L.E. Atlas, "Scalable and progressive audio codec", *Proc. of ICASSP*, pp. 3277-3280, 2001.

[8] T.H. Falk, S. Stadler, W.B. Kleijn and W. Chan, "Noise Suppression Based on Extending a Speech-Dominated Modulation Band", *Interspeech 2007*, pp. 970-973, 2007.

[9] J. Herre, and J.H. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)", *Aud. Engg. Soc.*, 101st Convention, 1996.

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

[11] P. Maragos, J.F. Kaiser and T.F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. on Signal Proc.*, Vol. 41(10), pp. 3024-3051, 1993.

[12] S. Ganapathy, S. Thomas, and H. Hermansky, "Modulation spectrum based features for phoneme recognition in noisy speech", *JASA Express Letters* , Vol. 125 (1), pp. EL8-EL12, 2009.

[13] S. Thomas, S. Ganapathy and H. Hermansky, "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction", *IEEE Signal Proc. Letters* , Vol. 15, pp. 681-684, 2008.

[14] S. Ganapathy, P. Motlicek, H. Hermansky and H. Garudadri, "Autoregressive modelling of Hilbert Envelopes for Wide-band Audio Coding", *Aud. Engg. Soc.*, 124th Convention, 2008.

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

[16] A.H. Nuttal and E. Bedrosian, "On the Quadrature Approximation to the Hilbert Transform of modulated signals", *Proc. of IEEE*, Vol. 54 (10), pp. 1458-1459, 1966.

[17] H.G. Hirsch and D. Pearce,"The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", *Proc. of ISCA ITRW ASR 2000*, pp. 18-20, 2000.

[18] "Aurora-5", *<http://aurora.hsnr.de/aurora-5.html>*.

[19] A.E. Rosenberg, C. Lee and F.K. Soong,"Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", *Proc. of ICSLP*, pp. 1835-1838, 1994.

[20] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition", *Proc. of ICSLP*, pp. 2185-2188, 2002.

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

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

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

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

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

[26] ISO/IEC JTC1/SC29/WG11, "Framework for Exploration of Speech and Audio Coding", *MPEG2007/N92s54*, 2007.

[27] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre and B. Grill,"Unified speech and audio coding scheme for high quality at low bitrates," *Proc. of ICASSP*, Apr. 2009.

[28] LAME MP3 codec, *<http://lame.sourceforge.net>*

[29] ITU-R Recommendation BS.1534, "Method for the subjective assessment of intermediate audio quality", 2001.

[30] M. Dietz, L. Liljeryd, K. Kjorling and O. Kunz, "Spectral Band Replication, a novel approach in audio coding," *Audio Engg. Soc.*, 112th Convention, Munich, May 2002.