COMPARISON OF MODULATION FEATURES FOR PHONEME RECOGNITION

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 compare several approaches for the extraction of modulation frequency features from speech signal using a phoneme recognition system. The general framework in these approaches is to decompose the speech signal into a set of sub-bands. Amplitude modulations (AM) in the sub-band signal are used to derive features for automatic speech recognition (ASR). Then, we propose a feature extraction technique which uses autoregressive models (AR) of sub-band Hilbert envelopes in relatively long segments of speech signal. AR models of Hilbert envelopes are derived using frequency domain linear prediction (FDLP). Features are formed by converting the FDLP envelopes into static and dynamic modulation frequency components. In the phoneme recognition experiments using the TIMIT database, the FDLP based modulation frequency features provide significant improvements compared to other techniques (average relative improvement of 7.5 % over the base-line features). Furthermore, a detailed analysis is performed to determine the relative contribution of various processing stages in the proposed technique.

Index Terms**—** Frequency domain linear prediction (FDLP), Modulations, Feature Extraction, 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 is represented by a sequence of short-term feature vectors with each vector forming a sample of the underlying process. Most of the information contained in these acoustic features relate to formants which provide important cues for recognition of some of the basic speech units.

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 [3]. Spectral representation of amplitude modulation in sub-bands, 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 audio coding [6] and noise suppression [7].

It has been shown that important information for speech perception lies in the $1 - 16$ Hz range of the modulation frequencies [8]. Even when the spectral information is limited, the use of temporal amplitude modulations alone provides good human speech recognition [9]. These studies suggest that amplitude modulations could provide alternative feature representations for ASR.

Several techniques have been proposed for which use modulation spectrum for feature extraction [10, 11, 12, 13]. Here, the speech signal is divided into a set of sub-bands. In each sub-band, an AM demodulation procedure is carried out to derive the sub-band AM envelopes. These envelope are converted to modulation frequency components and are used for speech recognition. These techniques mainly differ in the AM demodulation procedure (for example, the half-wave rectification [10, 12], Hilbert envelope approaches [13], and long-term sub-band energy based approaches [11]). In this paper, we briefly review these techniques for the task of phoneme recognition.

We present a feature extraction technique that tries to capture fine temporal dynamics along with static modulations using sub-band temporal envelopes [14]. The input speech signal is decomposed into a set of critical bands (Bark scale decomposition) and long temporal envelopes of sub-band signals are extracted using the technique of frequency domain linear prediction (FDLP) [13]. The sub-band temporal envelopes of the speech signal are then processed by a static compression stage and a dynamic compression stage. The static compression stage is a logarithmic operation and the adaptive compression stage uses the adaptive compression loops proposed in [15]. The compressed sub-band envelopes are transformed into modulation frequency components and used as features for the phoneme recognition system.

A hybrid Hidden Markov Model - Artificial Neural Network (HMM-ANN) phoneme recognition system is used for all the experiments [16]. In the phoneme recognition experiments on the TIMIT database, the proposed features provide significant improvements over other techniques. The rest of the paper is organized as follows. In Sec. 2, we briefly review various modulation spectrum approaches proposed in the past for ASR. In Sec. 3, the proposed FDLP technique for deriving static and dynamic modulation features is explained. Experiments with the modulation features for phoneme recognition task are reported in Sec. 4. A detailed analysis of the various parameters used in the FDLP technique is described in Sec. 5. In Sec. 6, we conclude with a discussion of the proposed features.

2. PAST MODULATION APPROACHES FOR ASR

In this section, a few techniques proposed in the past for the use of modulation spectra in ASR are described.

- Modulation Spectrogram (MSG) - A speech representation is developed that emphasizes the low-frequency (below 16

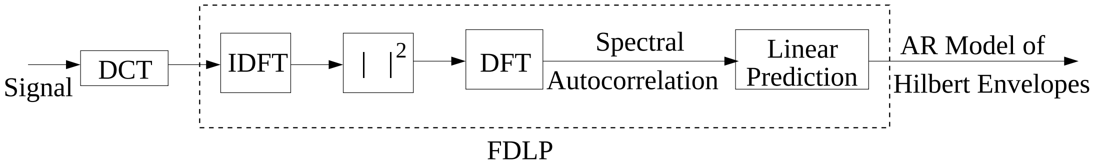

Fig. 1. Block schematic for the frequency domain linear prediction (FDLP)

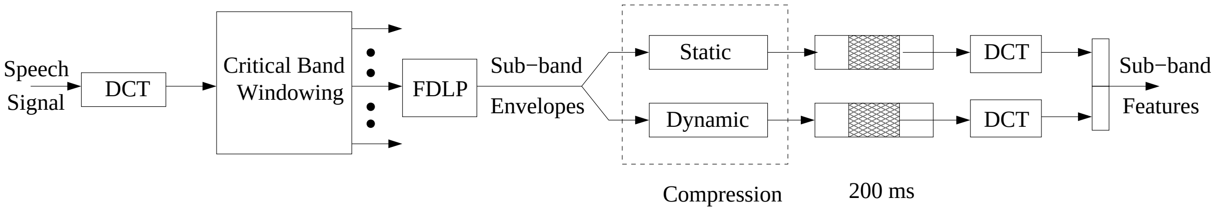

Fig. 2. Block schematic for the FDLP based modulation spectrum feature extraction technique.

Hz) amplitude modulations in sub-band channels [10]. Here, a spectral analysis into critical-band-wide channels is performed on an incoming speech signal. In each channel, an amplitude-envelope signal is computed by half-wave rectification and low-pass filtering with a cutoff frequency of 28 Hz. Each amplitude envelope signal is then downsampled by a factor of 100 and the slow modulations in each envelope signal are then analyzed by filtering the signal through a complex bandpass filter. These modulation components are used as features for ASR [10].

- MRASTA - A speech feature extraction based on multiple filtering of temporal trajectories of speech energies in frequency sub-bands is developed [11]. These filters emphasize different regions of the modulation spectrum in the range from 0-30 Hz. Further, the filters are designed to have zero-mean property which imply robustness to linear distortions of the signal and to changes in spectral tilt. The output of the filters, applied on long temporal trajectories (1000 ms) of sub-band energies, are used as features for speech recognition [11].
- Fepstum - Here, speech signal is analyzed in narrow sub-bands and an analytic signal is estimated in each sub-band [12]. The logarithm of the absolute magnitude of the sub-band analytic signal is used as an estimate of the AM signal. This is downsampled by a factor of 100 and lower discrete cosine transform (DCT) coefficients are used as form modulation spectral components. The modulation components from various sub-bands are collected and dimensionality reduced. A temporal context of 9 frames of Fepstrum features is used for phoneme recognition tasks [12].

3. AR MODELS OF HILBERT ENVELOPES

FDLP forms an efficient method for obtaining smoothed, minimum phase, parametric models of temporal rather than spectral envelopes. Being an auto-regressive (AR) modelling technique, FDLP captures the high signal-to-noise ratio (SNR) peaks in the temporal envelope. 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 [13]. 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 [17]).

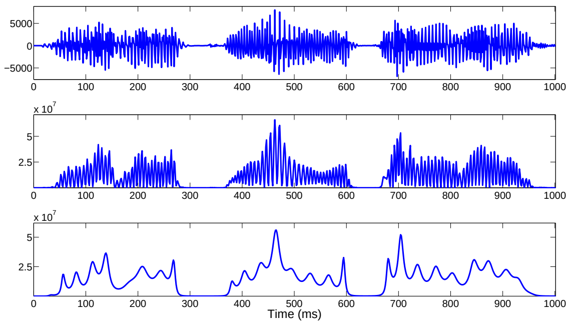

Fig. 3. Illustration of the all-pole modelling property of FDLP. (a) a portion of the sub-band speech signal, (b) its Hilbert envelope (c) all pole model obtained using FDLP

Fig. 3 shows the AR modelling property of FDLP. It shows (a) a portion of speech signal, (b) its Hilbert envelope and (c) an all pole approximation to the Hilbert Envelope using FDLP.

The block schematic for the proposed feature extraction technique is shown in Fig. 2. Long segments of speech signal are analyzed in critical bands using the technique of FDLP. The sub-band temporal envelopes are then compressed using a static compression scheme which is a logarithmic function and a dynamic compression scheme [14]. The use of the logarithm is to model the overall nonlinear compression in the auditory system which covers the huge dynamical range between the hearing threshold and the uncomfortable loudness level. The adaptive 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

Table 1. Phoneme Recognition Accuracies (%) for PLP features and various modulation features on TIMIT database.

PLP-9	Fepstrum	MSG	MRASTA	FDLP
66.8	61.1	62.4	64.5	69.3

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.

Conventional speech recognizers require speech features sampled at 100 Hz (i.e one feature vector every 10 ms). For using our speech representation in a conventional recognizer, the compressed temporal envelopes are divided into 200 ms segments with a shift of 10 ms. Discrete Cosine Transform (DCT) of both the static and the dynamic segments of temporal envelope yields the static and the dynamic 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.

4. EXPERIMENTS AND RESULTS

The proposed features are used for a phoneme recognition task on the TIMIT database. We use a phoneme recognition system based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [16] trained on the TIMIT database sampled at 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 [18].

The baseline system for these experiments uses the conventional Perceptual Linear Prediction (PLP) features [2] with a context of 9 frames [18] (351 dimensional features denoted as PLP-9). In the past, some of the modulation feature techniques have been used as additional sources of information by combining the modulation spectrum with conventional short-term PLP or MFCC features (for example Fepstrum [12], MSG [10]). However, in our experiments we report the recognition performance of the modulation features independently without any combination. This is done in order to illustrate the use of modulation spectrum as alternate representation compared to the conventional short-term spectral features.

In our implementation, Fepstrum features consist of 5 modulation frequency components in the $0-25$ Hz range from 40 mel bands yielding 200 dimensional vector for each frame. These features are dimensionality reduced to 60 dimensional features [12]. A context of 9 frames gives a 540 dimensional feature vector at the input of the phoneme recognition system. MSG features consist of 9 modulation components from 36 sub-bands resulting in 324 dimensional features for every speech frame [10]. MRASTA features use 19 critical bands with 14 modulation filters. These are appended with frequency derivatives yielding 504 dimensional features [11]. For the FDLP based modulation features, 21 critical bands are used with 14 static modulation spectral components and 14 dynamic modulation spectral components. This gives 588 dimensional features at the input vector.

Table 1 summarizes the results for the phoneme recognition experiments with various modulation features. Among the past modulation approaches, MRASTA features provide the best phoneme recognition performance. FDLP based features using static and dynamic modulation spectrum provides a relative improvement of 7.5 % compared to the baseline PLP features.

Table 2. Phoneme Recognition Accuracies (%) for various modifications of the proposed feature extraction technique.

AM Demodulation		
Half-Wave	Energy	FDLP
67.0	67.7	69.3

Temporal Context (ms)			
100	200	300	400
68.7	69.3	68.0	66.2

Modulation Extent (Hz)			
15	25	35	45
67.1	69.1	69.3	69.1

Type of Modulation		
Stat.	Dyn.	Stat. + Dyn.
67.9	64.6	69.3

5. RELATIVE CONTRIBUTION OF VARIOUS PROCESSING STEPS

The previous section showed that the proposed feature extraction provides promising phoneme recognition performance on TIMIT database. In-order to analyze the relative contribution of various stages of the proposed feature extraction, we perform a set of phoneme recognition experiments with different modifications to the proposed features. These modifications are:

Choice of AM demodulation

The proposed features use FDLP technique for AM demodulation of sub-band signals. As mentioned in Sec. 2, other methods of AM demodulation have been used in the past. We compare the phoneme recognition performance of FDLP approach with the half-wave rectification technique [10] and the sub-band energy trajectory approach [11]. All the other processing stages in the proposed features (like the sub-band decomposition, static and dynamic modulation spectrum etc) are retained. These results are shown in Table 2. In these experiments, FDLP based AM demodulation provides the best phoneme recognition.

Duration of Temporal Context

The temporal analysis window for the extraction of static and dynamic modulations is modified in these experiments from 100 to 400 ms. FDLP based sub-band processing is used and static and dynamic modulation features are derived. These results are shown in the second row of Table 2. It is interesting to note that the best phoneme recognition performance is obtained for a context of 200 ms, which also corresponds to the average syllabic rate of human speech.

Extent of Modulation Information

In these experiments, the extent of modulation spectrum used for feature extraction is varied from 15-45 Hz. The duration of modulation analysis on the FDLP envelopes is fixed at 200 ms and the number of DCT coefficients is varied. Static and dynamic modulations are used for phoneme recognition. These results, reported in the third row of Table 2, show that the phoneme recognition performance peaks for a modulation content in the range 0-35 Hz.

Type of Modulation Spectrum

As mentioned before, we derive modulation information from two types of envelope compression scheme. Static modulations are derived using a logarithmic compression and the dynamic modulations are derived using adaptive loops. FDLP envelope with a temporal context of 200 ms is used for deriving the modulations in the range of 0-35 Hz. These results are shown at the bottom of Table 2. The static modulation features provide good phoneme recognition for fricatives and nasals (which is due to modelling property of the signal peaks in static compression) whereas the dynamic modulation features provide good performance for plosives and affricates (where the fine temporal fluctuations like onsets and offsets carry the important phoneme classification information) [14]. Hence, the combination of these feature streams results in considerable improvement in performance for most of the phoneme classes.

From all these experiments, it is found that the feature extraction technique which uses static and dynamic modulation spectrum in 0-35 Hz range obtained from 200 ms of FDLP envelopes provides the best phoneme recognition performance.

6. SUMMARY

In this paper, we have compared some of the modulation approaches for phoneme recognition task. We have also proposed a feature extraction technique based on the modulation spectrum. Here, Hilbert envelopes of frequency sub-bands are modelled using FDLP. These temporal envelopes are compressed using an adaptive and static compression and are converted to modulation frequency components. These features provide significant improvements for phoneme recognition tasks. The results are promising and encourage us to experiment on other tasks with different test conditions.

7. ACKNOWLEDGEMENTS

The authors would like to thank the Medical Physics group at the Carl von Ossietzky-Universitat Oldenburg for code fragments implementing adaptive compression loops and K.V.S Narayana for his help in implementation of Fepstrum features.

8. 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. 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.

[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] M. S. Vinton and L. E. Atlas, “Scalable and progressive audio codec,” *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 5, pp. 3277-3280, Salt Lake City, USA, Apr. 2001.

[7] T. H. Falk, S. Stadler, W. B. Kleijn and Wai-Yip Chan, “Noise Suppression Based on Extending a Speech-Dominated Modulation Band,” *Interspeech 2007*, Antwerp, Belgium, Aug. 2007.

[8] 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.

[9] R.V Shannon, F.G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech Recognition with Primarily Temporal Cues”, *Science*, Vol. 270(5234), pp. 303-304, 1995.

[10] 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.

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

[12] V. Tyagi, “Tandem Processing of Fepstrum Features,” *Proc. of Interspeech*, Brisbane, Sept. 2008.

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

[14] 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.

[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), pp. 2040-2050, 1999.

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

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

[18] 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.