Static and Dynamic Modulation Spectrum for Speech Recognition

Sriram Ganapathy[1], *Samuel Thomas*[1] *and 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

We present a feature extraction technique based on static and dynamic modulation spectrum derived from long-term envelopes in sub-bands. Estimation of the sub-band temporal envelopes is done using Frequency Domain Linear Prediction (FDLP). These sub-band envelopes are compressed with a static (logarithmic) and dynamic (adaptive loops) compression. The compressed sub-band envelopes are transformed into modulation spectral components which are used as features for speech recognition. Experiments are performed on a phoneme recognition task using a hybrid HMM-ANN phoneme recognition system and an ASR task using the TANDEM speech recognition system. The proposed features provide a relative improvements of 3.8 % and 11.5 % in phoneme recognition accuracies for TIMIT and conversation telephone speech (CTS) respectively. Further, these improvements are found to be consistent for ASR tasks on OGI-Digits database (relative improvement of 13.5 %).
Index Terms: Frequency Domain Linear Prediction (FDLP), Modulation spectrum, Adaptive compression, Feature extraction for speech recognition.

1. Introduction

Conventionally, acoustic features for ASR are extracted by estimating the spectral content of relatively short (about 10-30 ms) segments of speech (for example [1]). Each estimated vector of spectral components represents a sample of the underlying dynamic speech production process. Stacking these estimates in time provides a two-dimensional (time-frequency) representation. Most of the information contained in these acoustic features relate to formant information in speech.

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 [2]. Even when the spectral information is limited, the use of temporal amplitude modulations alone provides good human speech recognition [3]. These studies suggest that amplitude modulations could provide alternative feature representations for ASR.

Spectral components of long-term amplitude modulations in individual frequency sub-bands are called modulation spectra. The modulation spectral representations have been used in the past for predicting speech intelligibility in reverberant environments [4]. They are now widely applied in many engineering applications (for example audio coding [5], noise suppression [6], etc). Feature extraction techniques based on modulation spectrum have also been proposed for ASR (for example [7, 8]).

For phoneme recognition task, the techniques that are based on deriving long-term modulation frequencies may not preserve fine temporal events like onsets and offsets. On the other hand, signal adaptive techniques which try to represent local temporal fluctuation, cause strong attenuation of higher modulation frequencies [10].

In our previous work [11], we have shown that a combination of static and dynamic modulation spectral features perform well in mismatched train and test conditions. The input speech signal is decomposed into a number of critical bands. In each sub-band, long term envelopes are extracted using Frequency Domain Linear Prediction (FDLP). FDLP envelopes are compressed using a static and a dynamic compression. The static compression stage is a logarithmic operation and dynamic compression stage uses adaptive compression loops [10]. The compressed envelopes are transformed into modulation spectral components which are used as features for a phoneme recognition system.

In this paper, we extend these modulation frequency features for phoneme recognition with matched conditions in clean and conversational telephone speech (CTS). We use a hybrid Hidden Markov Model - Artificial Neural Network (HMM-ANN) phoneme recognition system [12]. The proposed features provide considerable improvements in phoneme recognition accuracies for TIMIT and conversation telephone speech (CTS) databases. We also show the application of the proposed features for speech recognition tasks using the TANDEM system [9]. The improvements obtained in phoneme recognition are consistent for digit recognition tasks.

The rest of the paper is organized as follows. In Sec. 2, we describe the FDLP technique for the estimation of the temporal envelopes using linear prediction in spectral domain. The extraction of modulation frequency features from the temporal envelopes is given in Sec. 3. Experiments performed with the modulation frequency features for phoneme and word recognition tasks are reported in Sec. 4. In Sec. 5, we conclude with a discussion of the proposed features.

2. Frequency Domain Linear Prediction

The Hilbert envelope, which is the squared magnitude of the analytic signal, represents the instantaneous energy of a signal in the time domain. Hilbert envelopes are typically computed using the Hilbert transform operator in the time domain or by exploiting the causality of Discrete Fourier Transforms (DFT) [15]. For feature extraction in speech recognition, we use a parametric model of the Hilbert envelopes.

FDLP is an efficient technique for auto regressive (AR)

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

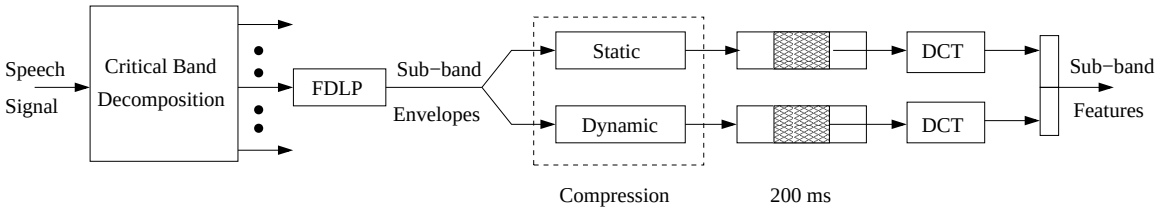

Figure 2: Block schematic for the modulation spectrum based feature extraction technique.

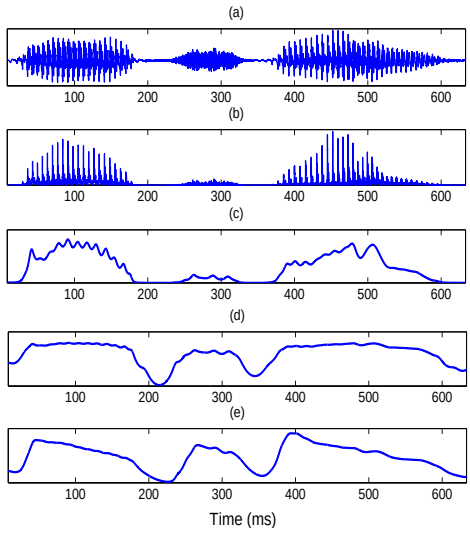

Figure 1: Static and dynamic compression of the temporal envelopes: (a) a portion of speech signal, (b) the temporal envelope extracted using the Hilbert transform [15], (c) the FDLP envelope, which is an all-pole approximation to (b) estimated using FDLP, (d) static compression of the FDLP envelope and (e) dynamic compression of the FDLP envelope.

modelling of temporal envelopes of a signal [14]. It represents a dual technique to the conventional Time Domain Linear Prediction (TDLP). In the case of TDLP, the AR model approximates the power spectrum of the input signal, whereas FDLP fits an all pole model to the Hilbert envelope. Fig. 1 shows the AR modelling property of FDLP. It shows (a) a portion of speech signal, (b) its Hilbert envelope computed using the Fourier transform technique [15] and (c) an all pole approximation for the Hilbert Envelope using FDLP.

3. Feature extraction

The block schematic for the modulation spectrum based feature extraction technique is shown in Fig. 2. Long segments of the speech signal (hundreds of milliseconds) are decomposed into frequency sub-bands by windowing the discrete cosine transform (DCT). In our experiments, we use a critical band decomposition. Using FDLP, an all-pole estimate of the temporal envelope in each sub-band is obtained. The whole set of sub-band temporal envelopes forms a two dimensional (time-frequency) representation of the input signal energy. The sub-band temporal envelopes are then compressed using a static compression which is a logarithmic function and a dynamic compression

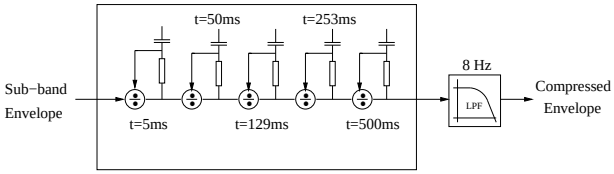

Figure 3: Dynamic compression of the sub-band FDLP envelopes using adaptive compression loops [10].

scheme [10]. The dynamic compression, shown in Fig. 3, is realized by an adaptation circuit consisting of five consecutive nonlinear adaptation loops [10]. Each of these loops consists of a divider and a low-pass filter with time constants ranging from 5 ms to 1000 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 fast compared to the time constants of the adaptation loops are amplified linearly at the output, whereas the slowly changing regions of the input signal are suppressed. In this way, changes in the input signal like onsets and offsets are emphasized in the dynamic compression stage. This is also illustrated in Fig. 1, where we show the static and dynamic compression of the FDLP envelopes. The dynamic compression stage is followed by a low pass filter [10].

Since speech recognition system require speech features sampled at 100 Hz (i.e one feature vector every 10 ms), the compressed temporal envelopes are divided into 200 ms segments with a shift of 10 ms. The temporal envelopes from the two compression streams are then converted into modulation spectral components using DCT, corresponding to the static and the dynamic modulation spectrum. We use 14 modulation frequency components from each of these streams, yielding modulation spectrum in the $0-35$ Hz range with a resolution of 2.5 Hz. This choice of modulation frequencies is obtained using phoneme recognition experiments on the cross validation data in TIMIT database.

4. Experiments and results

4.1. Phoneme recognition in clean speech

The phoneme recognition system is based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [12]. The MLP 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 taken with a window of certain frames. The relation between the posterior probability $P(q_t = i|x_t)$ and the likelihood

Table 1: Phoneme Recognition Accuracies (%) for different feature extraction techniques on TIMIT database.

PLP	MSG	MRASTA	FDLP
68.4	63.1	65.5	69.6

Table 2: Recognition Accuracies (%) of broad phonetic classes obtained from confusion matrix analysis on TIMIT database

Class	PLP	MSG	MRASTA	FDLP
Vowel	92.8	91.7	91.9	92.7
Plosive	84.7	81.3	84.4	85.5
Fricative	87.3	82.1	85.2	88.2
Semi Vowel	76.9	74.4	74.8	77.6
Nasal	85.9	81.6	83.0	86.2
Avg.	85.5	82.2	83.9	86.0

$P(x_t|q_t = i)$ is given by the Bayes rule,

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

It is shown in [12] that the neural network with sufficient capacity and trained on enough data estimates the true Bayesian a-posteriori probability. 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.

Experiments are performed on TIMIT database containing speech sampled at 16 kHz. The 'sa' dialect sentences are excluded in the experiments. 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].

As explained in Sec. 3, static and dynamic modulation frequency features are extracted for every frame. A three layered multi-layer perceptron (MLP) is used to estimate the phoneme posterior probabilities. The network is trained using the standard back propagation algorithm with cross entropy error criteria. 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 output neurons (with soft max nonlinearity) representing the phoneme classes. The performance of phoneme recognition is measured in terms of phoneme accuracy. In the decoding step, all phonemes are considered equally probable (no language model). The optimal phoneme insertion penalty that gives maximum phoneme accuracy on the cross-validation data is used for the test data.

Table 1 summarizes the results for the experiments with FDLP based modulation features. In these experiments, the proposed features are compared with other feature extraction techniques namely PLP features with a 9 frame context [16], and other modulation spectrum based features like MRASTA features [17] and modulation spectrogram (MSG) features [18]. The proposed feature extraction technique provides relative improvement of 3.8 % compared to the PLP features. We also report the results for recognition of broad phonetic classes using confusion matrix analysis (Table 2). The proposed technique of combining static and dynamic modulation spectrum provides good performances for most of the broad phonetic classes.

Table 3: Phoneme Recognition Accuracies (%) for different feature extraction techniques on CTS database.

PLP	RASTA	MRASTA	ETSI	FDLP
52.3	52.8	52.2	54.0	59.3

Table 4: Recognition Accuracies (%) of broad phonetic classes obtained from confusion matrix analysis on CTS database

Class	PLP	RASTA	MRASTA	ETSI	FDLP
Vowel	70.5	71.0	69.4	72.2	75.3
Plosive	82.7	84.3	83.6	83.9	85.4
Fricative	71.5	72.4	72.1	72.6	75.4
Semi Vowel	71.3	73.8	73.0	74.8	77.9
Nasal	65.5	66.0	65.8	66.2	69.4
Avg.	72.3	73.5	72.8	73.9	76.7

4.2. Phoneme recognition in CTS

The CTS database consists of 300 hours of conversational speech recorded over a telephone channel at 8 kHz [19]. The training data consists of 250 hours of speech from 4538 speakers, cross-validation data set consists of 40 hours of speech from 726 speakers and the test data set consists of 10 hours from 182 speakers. It is labeled using 45 phonemes. The phoneme labels are obtained by force aligning the word transcriptions to the previously trained HMM/GMM models [19].

We use the phoneme recognition system based on HMM-ANN system (described in Sec. 4.1). Here, the MLP consists of 8270 hidden neurons, and 45 output neurons (with soft max nonlinearity) representing the phoneme classes. Table 3 reports the results for the phoneme recognition experiments on CTS database. We compare the proposed FDLP features with other features like PLP, RASTA [20], MRASTA and Advanced-ETSI (noise-robust) distributed speech recognition front-end [21]. The proposed modulation features result in improved phoneme recognition rate for all the broad phonetic classes (Table 4) and hence, provide significant improvements in individual phoneme recognition rate. We obtain a relative improvement of 11.5 % compared to the ETSI feature extraction technique.

4.3. Word Recognition on OGI-Digits

Experiments are performed with small vocabulary continuous digit recognition task (OGI-Digits database). The vocabulary consists of eleven ($0 - 9$ digits and "Oh") digits in 28 different pronunciations. Features extracted from speech for every 10 ms are used to train an ANN with 1800 hidden nodes. The ANN estimates posterior probabilities of 29 English phonemes [17]. The training data consists of the whole Stories database plus the training part of the Numbers95 database. Around 10 % of the data is used for cross-validation. Log and Karhunen Loeve (KL) transforms are applied on these features. This is done in order to convert the phoneme posterior probabilities into features appropriate for a conventional HMM recognition system [9]. The HMM based recognizer, trained on the training part of the OGI-Digits database, is used for classification.

The performance of the proposed features is compared with

Table 5: Word Recognition Accuracies (%) for different feature extraction techniques on OGI-Digits database.

PLP-D-A	PLP	MSG	MRASTA	FDLP
95.9	96.2	96.0	96.3	96.8

other features like PLP, MRASTA and MSG features (Table 5). We also report the base-line performance with 39 dimensional PLP features (PLP-D-A) on the HMM-GMM system (without the use of TANDEM setup) The proposed features provide a relative improvement of about 13.5 % compared to the MRASTA features.

5. Conclusions

We have presented a feature extraction technique based on deriving static and dynamic modulation spectrum from speech signal. The input speech signal is analyzed in critical bands and sub-band temporal envelopes are estimated using FDLP. These envelopes are compressed using a static and dynamic compression scheme. The compressed envelopes are transformed using DCT to obtain static and dynamic modulation frequency features. In the presence of telephone noise, these features provide significant robustness for all the broad phonetic classes. These features are currently investigated for speech recognition in additive noise.

6. Acknowledgments

The authors would like to thank the Medical Physics group at the Carl von Ossietzky-Universitat Oldenburg for code fragments implementing adaptive compression loops, Joel Pinto for his help in phoneme recognition task on CTS database and Sivaram Garimella for the code fragments implementing the TANDEM based ASR system.

7. References

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

[2] 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.

[3] 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.

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

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

[7] H. Hermansky and S. Sharma, "TRAPS - Classifiers of Temporal Patterns", *Proc. of ICSLP*, Sydney, Australia, Vol. 3, pp. 1003-1006, 1998.

[8] 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.

[9] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems", *Proc. of ICASSP*, Vol. 3, pp. 1635-1638, 2000.

[10] 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.

[11] 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.

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

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

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

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

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

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

[18] S. Greenberg and B.E.D. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech", *Proc. ICASSP*, Vol. 3, pp. 1647-1650, 1997.

[19] T. Hain *et al.*, "The Development of AMI System for Transcription of Speech in Meetings", *Proc. of MLMI*, pp. 344356, 2005.

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

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