PHONEME RECOGNITION USING SPECTRAL ENVELOPE AND MODULATION FREQUENCY FEATURES

Samuel Thomas, Sriram Ganapathy and Hynek Hermansky

Idiap Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{tsamuel,ganapathy}@idiap.ch, hermansky@ieee.org

ABSTRACT

We present a new feature extraction technique for phoneme recognition that uses short-term spectral envelope and modulation frequency features. These features are derived from sub-band temporal envelopes of speech estimated using Frequency Domain Linear Prediction (FDLP). While spectral envelope features are obtained by the short-term integration of the sub-band envelopes, the modulation frequency components are derived from the long-term evolution of the sub-band envelopes. These features are combined at the phoneme posterior level and used as features for a hybrid HMM-ANN phoneme recognizer. For the phoneme recognition task on the TIMIT database, the proposed features show an improvement of 4.7% over the other feature extraction techniques.

Index Terms— Spectral envelope and Modulation frequency features, Phoneme Recognition, Frequency Domain Linear Prediction

1. INTRODUCTION

Time-varying spectrum of speech is usually derived as a sequence of short-term spectral vectors, each vector representing instantaneous values of spectral magnitudes at the individual carrier frequencies. An alternative functionally equivalent representation is a collection of temporal envelopes of spectral energies at the individual carrier frequencies. The Fourier transform of these time-varying temporal envelopes yields a set of modulation spectra of speech, where each modulation spectral value represents the dynamics of the signal at the given carrier frequency.

Conventional acoustic features for Automatic Speech Recognition (ASR) systems are typically based on the first of the two representations, i.e. on the short-term spectrum. They are extracted by applying Bark or Mel scale integrators on power spectral estimates in short analysis windows ($10 - 30$ ms) of the speech signal. The signal dynamics are represented by a sequence of short-term feature vectors with each vector forming a sample of the underlying process. These features are appended with derivatives of the spectral trajectory at each instant to enhance the local speech variations. Typical examples of such features are the Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) [2].

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 [3]. Furthermore, in the presence of limited spectral information, it has been shown that the use of temporal amplitude modulations alone provides nearly perfect human speech recognition [4]. This further emphasizes the importance of exploiting temporal amplitude modulations as alternative feature representations for ASR. In order to exploit the information in these modulation frequencies, relatively long segments of speech signal need to be analyzed [5]. For example, the more recent feature extraction techniques like [6, 7] use the long-term dynamics of the sub-band energies for phoneme recognition. Combining the short-term spectral information with modulation frequency components has also shown to improve phoneme recognition performance [8].

In this paper, we develop a feature extraction technique that combines the short-term spectral envelope features and the long term modulations features. The spectral envelope features and modulation frequency features are both derived from the same initial two-dimensional (time-frequency) representation of speech that is formed by sub-band temporal envelopes. Specifically, speech signals in frequency sub-bands are analyzed over long temporal segments using the Frequency Domain Linear Prediction (FDLP). The FDLP technique fits an all pole model to the squared Hilbert envelope of the signal [9]. These representations of the speech signal are able to capture fine temporal events associated with transient events like stop bursts while at the same time summarize the signal's gross temporal evolution in timescales of several hundred milliseconds [10].

In our case, the auditory spectrogram, which is a two-dimensional representation of the input signal, is obtained by stacking the sub-band temporal envelopes in frequency (similar to the stacking of short-term spectral estimates in time for the conventional features). The short-term spectral envelopes are derived by integrating the auditory spectrogram in short analysis windows and the modulation frequency components are obtained by the application of cosine transform on the compressed (static and adaptive compression) long-term sub-band temporal envelopes. The spectral envelope features and the modulation features are combined at the phoneme posterior level and used as features for the hybrid Hidden Markov Model - Artificial Neural Network (HMM-ANN) phoneme recognition system [11]. Experiments on a phoneme recognition task using the TIMIT database compare the proposed features with other feature extraction techniques.

The rest of the paper is organized as follows. In Sec. 2, the FDLP technique for deriving sub-band envelopes is described. The conversion of these sub-band envelopes into spectral envelope and modulation frequency features is explained in Sec. 3. Experiments with the proposed features for a phoneme recognition task are reported in Sec. 4 along with a comparison of the results for the other feature extraction techniques in the literature. In Sec. 5, we conclude with a discussion of the proposed features.

This work was supported by the European Union 6th FWP IST Integrated Project AMIDA and the Swiss National Science Foundation through the Swiss NCCR on IM2.

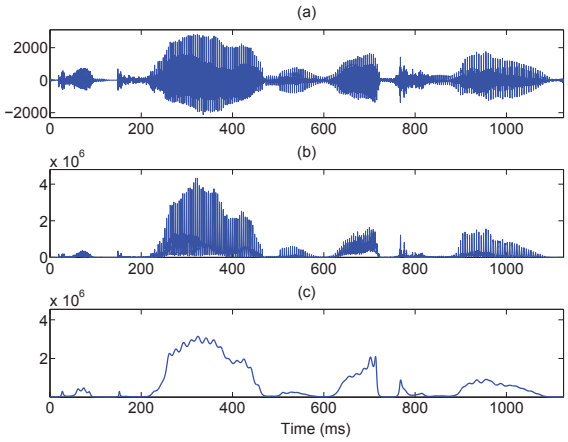

Fig. 1. Illustration of the all-pole modeling property of FDLP. (a) a portion of the speech signal, (b) its Hilbert envelope (c) all pole model obtained using FDLP

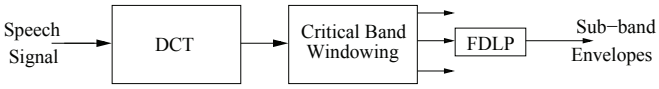

Fig. 2. Deriving sub-band temporal envelopes from speech signal using FDLP

2. FREQUENCY DOMAIN LINEAR PREDICTION

FDLP is an efficient technique for auto regressive (AR) modeling of temporal envelopes of a signal [10]. 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 (squared magnitude of the analytic signal).

The FDLP technique is implemented in two parts - first, the discrete cosine transform (DCT) is applied on long segments of speech to obtain a real valued spectral representation of the signal. Then, linear prediction is performed on the DCT representation to obtain a parametric model of the temporal envelope. Fig. 1 illustrates the AR modeling of FDLP. It shows (a) a portion of speech signal, (b) its Hilbert envelope computed using the Fourier transform technique [12] and (c) an all pole approximation to the Hilbert Envelope using FDLP. The block schematic for extraction of sub-band temporal envelopes from speech signal is shown in Fig. 2. Long segments of the input speech signal are transformed using DCT. The sub-band DCT components are obtained by windowing the input signal DCT on a bark scale. FDLP is applied on these sub-band DCT components to estimate the sub-band temporal envelopes.

3. DERIVING FEATURES FROM SUB-BAND TEMPORAL ENVELOPES

The sub-band temporal envelopes, estimated using FDLP, are converted into spectral envelope and modulation frequency features.

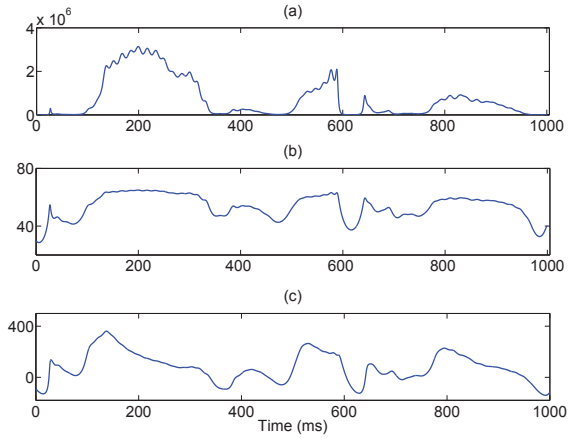

Fig. 3. Static and dynamic compression of the temporal envelopes. (a) a portion of the temporal envelope of a speech signal, (b) static compression (logarithm) and (c) adaptive compression using adaptive loops.

3.1. Spectral envelope features

The Hilbert envelope, which is the squared magnitude of the analytic signal, represents the instantaneous energy of a signal in the time domain. Since integration of signal energy is identical in time and frequency domain, the sub-band Hilbert envelopes can equivalently be used for obtaining the sub-band energy based short-term spectral envelope features. This is achieved by integrating the sub-band temporal envelopes in short term frames (of the order of 25 ms with a shift of 10 ms). These short term sub-band energies are then converted into 13 cepstral features along with their first and second derivatives (similar to 39 dimensional PLP features [2]). Each frame of these spectral envelope features is used with a context of 9 frames for training a phoneme posterior probability estimator [13].

3.2. Modulation features

The long-term sub-band envelopes from the FDLP form a compact representation of the temporal dynamics over long regions of the speech signal. The sub-band temporal envelopes are compressed using a static compression scheme which is a logarithmic function and a dynamic compression scheme [14]. The dynamic compression is realized by an adaptation circuit consisting of five consecutive non-linear adaptation loops [14]. 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. This is illustrated in Fig. 3 which shows (a) a portion of temporal envelope of a speech signal, (b) logarithmically compressed temporal envelope and (c) the temporal envelope compressed with the adaptive compression scheme.

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

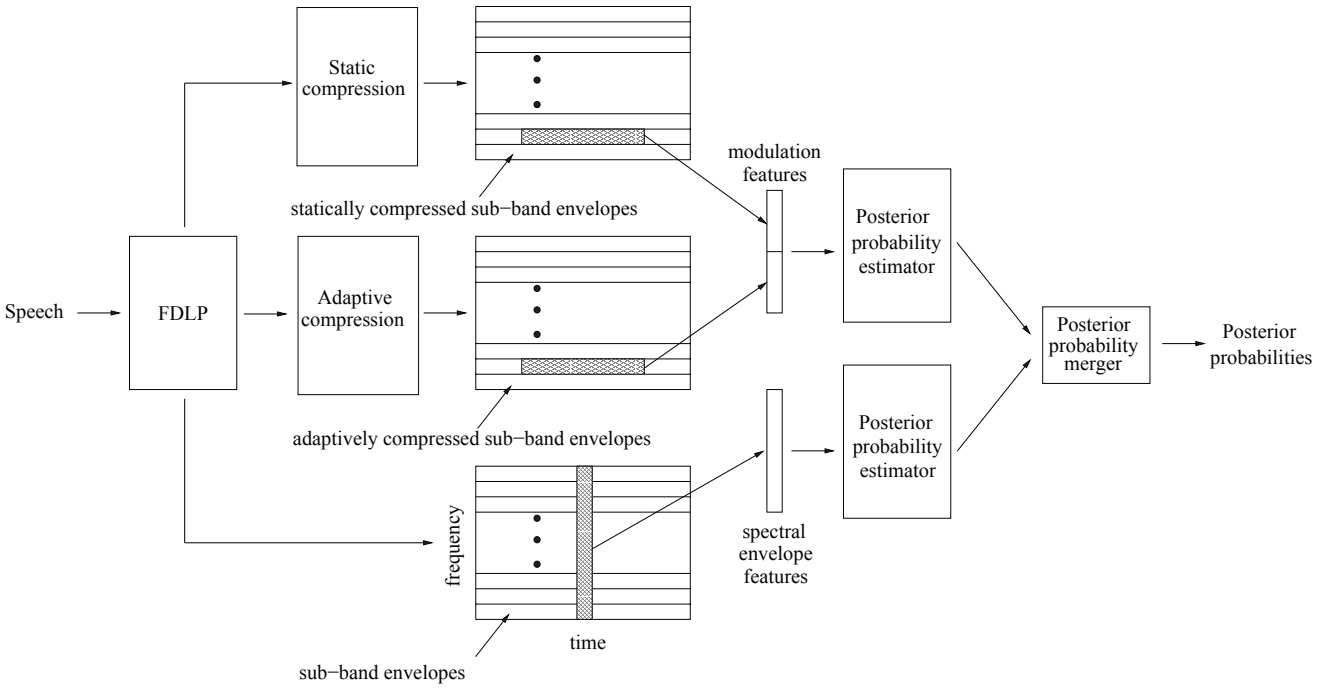

Fig. 4. Schematic of the joint spectral envelope-modulation features for posterior based ASR

Table 1. Recognition Accuracies (%) of broad phonetic classes obtained from confusion matrix analysis

Class	PLP	FDLP-S	M-RASTA	FDLP-M	PLP+M-RASTA	FDLP-S+FDLP-M
Vowel	85.3	84.9	82.4	85.7	86.1	87.3
Diphthong	78.2	79.1	74.2	76.8	78.4	79.8
Plosive	83.8	82.8	81.6	84.1	84.6	85.4
Affricative	73.5	74.4	68.6	75.6	72.9	78.0
Fricative	85.8	85.9	83.5	86.8	86.4	88.0
Semi Vowel	76.2	74.9	72.9	77.1	77.8	79.0
Nasal	84.2	82.8	80.4	84.9	85.8	86.6
Avg.	81.0	80.7	77.7	81.6	81.7	83.4

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. This choice is a result of series of optimization experiments (which are not reported here). The static and adaptive modulation features for each sub-band are stacked together to obtain modulation features for each sub-band and fed to the posterior probability estimator.

We combine the spectral envelope and modulation frequency features at the phoneme posterior level using the Dempster Shafer (DS) theory of evidence [15]. Fig. 4 shows the schematic of the proposed feature extraction technique.

4. EXPERIMENTS AND RESULTS

The phoneme recognition system is based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [11]. 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. The relation between the posterior probability $P(q_t = i|x_t)$ and the likelihood $P(x_t|q_t = i)$ is given by the Bayes rule,

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

A neural network, with sufficient capacity, trained on enough data estimates the true Bayesian a-posteriori probability [11]. 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 the TIMIT database, excluding 'sa' dialect sentences. All speech files are 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 [13]. A three layered MLP is used to estimate the phoneme posterior probabilities. The network

Table 2. Phoneme Recognition Accuracies (%) for different feature extraction techniques

PLP	68.3
FDLP-S	68.1
M-RASTA	64.9
FDLP-M	69.3
PLP+M-RASTA	70.0
FDLP-S+FDLP-M	71.4

is trained using the standard back propagation algorithm with cross entropy error criteria. The learning rate and stopping criterion are controlled by the error in the frame-based phoneme classification 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 as well as the recognition accuracy of broad phonetic classes. 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 spectral envelope features and modulation features for the recognition of broad phonetic classes. In the base-line experiments, the proposed features are compared with two other feature extraction techniques on the same task - the PLP features with a 9 frame context [13] which are similar to spectral envelope features derived using FDLP (FDLP-S) and M-RASTA features [6] which are similar to features derived using FDLP from the modulation spectra (FDLP-M). We combine the spectral envelope and modulation frequency features using the DS theory of evidence to obtain two more feature sets - PLP features with M-RASTA features (PLP+M-RASTA) and FDLP-S features with FDLP-M features (FDLP-S+FDLP-M). Table 2 shows the results for phoneme recognition accuracies across all individual phoneme classes for these techniques. The FDLP-S features provide comparable results as the PLP-9 features. The modulation features (FDLP-M) result in improved phoneme recognition rate for all the broad phonetic classes compared to the M-RASTA features and hence, provide significant improvements in individual phoneme recognition rate (Table 2). Further, the joint spectral envelope and modulation features yield improved phoneme class recognition for all the broad phonetic classes compared to the baseline system. We obtain a relative improvement of 9.2 % over the baseline system for recognition of broad phonetic classes and an improvement of 4.7 % (which is statistically significant) in the individual phoneme recognition rate.

5. CONCLUSIONS

We have proposed a novel method of extracting spectral envelope and modulation features for ASR. The spectral envelope features derived from sub-band temporal envelopes are comparable to conventional features that are derived from short-term power spectral estimates. The FDLP based modulation features are significantly better than other features based on the modulation spectrum. Combining the spectral envelope and modulation features provides significant improvements over the base-line system for phoneme recognition tasks. The results on clean conditions are promising and encourage us to experiment on other tasks in noisy conditions.

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.

7. REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[3] R. Drullman, J.M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, pp. 2670–2680, 1994.

[4] R.V. Shannon, F.G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech Recognition with Primarily Temporal Cues," *Science*, vol. 270, no. 5234, pp. 303, 1995.

[5] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," *IEEE ASRU*, pp. 140–147, 1997.

[6] H. Hermansky and P. Fousek, "Multi-Resolution RASTA Filtering for TANDEM-Based ASR," in *ISCA INTERSPEECH*, 2005, pp. 361–364.

[7] H. Hermansky and S. Sharma, "TRAPS - classifiers of temporal patterns," in *ISCA ICSLP*, 1998, pp. 1003–1006.

[8] S.R.M. Prasanna, B. Yegnanarayana, J.P. Pinto, and H. Hermansky, "Analysis of Confusion Matrix to Combine Evidence for Phoneme Recognition," *IDIAP Research Report (IDIAP-RR-07-27)*, 2007.

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

[10] M. Athineos and D.P.W. Ellis, "Autoregressive Modeling of Temporal Envelopes," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5237–5245, 2007.

[11] H. Bourlard and N. Morgan, *Connectionist speech recognition: A hybrid approach*, Kluwer Academic Publishers, 1994.

[12] L.S. Marple, "Computing the discrete-time "analytic" via FFT," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 47, no. 9, pp. 2600–2603, 1999.

[13] J. Pinto, B. Yegnanarayana, H. Hermansky, and M.M. Doss, "Exploiting contextual information for improved phoneme recognition," in *ISCA INTERSPEECH*, 2007, pp. 1817–1820.

[14] T. Dau, D. Pueschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *The Journal of the Acoustical Society of America*, vol. 99, pp. 3615–3622, 1996.

[15] F. Valente and H. Hermansky, "Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence," in *IEEE ICASSP*, 2007, pp. 1129–1132.