Modulation frequency features for phoneme recognition in noisy speech

Sriram Ganapathy, Samuel Thomas, and Hynek Hermansky

Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland

Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

Email: ganapathy@idiap.ch, tsamuel@idiap.ch, hermansky@ieee.org

Abstract

In this letter, a new feature extraction technique based on modulation spectrum derived from syllable-length segments of sub-band temporal envelopes is proposed. These sub-band envelopes are derived from auto-regressive modelling of Hilbert envelopes of the signal in critical bands, processed by both a static (logarithmic) and a dynamic (adaptive loops) compression. These features are then used for machine recognition of phonemes in telephone speech. Without degrading the performance in clean conditions, the proposed features show significant improvements compared to other state-of-the-art speech analysis techniques. In addition to the overall phoneme recognition rates, the performance with broad phonetic classes is reported.

PACS numbers: 43.72.Ne, 43.72.Ar

1. Introduction

Conventional speech analysis techniques start with estimating the spectral content of relatively short (about 10-20 ms) segments of the signal (short-term spectrum). Each estimated vector of spectral energies represents a sample of the underlying dynamic process in production of speech at a given time-frame. Stacking such estimates of the short-term spectra in time provides a two-dimensional (time-frequency) representation of speech that represents the basis of most speech features (for example [Hermansky, 1990]). 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 forms a two-dimensional representation of speech (for example [Athineos et al., 2004]).

For machine recognition of phonemes in noisy speech, the techniques that are based on deriving long-term modulation frequencies do not preserve fine temporal events like onsets and offsets which are important in separating some phoneme classes. On the other hand, signal adaptive techniques which try to represent local temporal fluctuation, cause strong attenuation of higher modulation frequencies which makes them less effective even in clean speech [Tchorz and Kollmeier, 2004].

In this letter, we propose a feature extraction technique for phoneme recognition that tries to capture fine temporal dynamics along with static modulations using sub-band temporal envelopes. The input speech signal is decomposed into 17 critical bands (Bark scale decomposition) and long temporal envelopes of sub-band signals are extracted using the technique of Frequency Domain Linear Prediction (FDLP) [Athineos and Ellis, 2007]. 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 [Dau et al., 1996]. The compressed sub-band envelopes are transformed into modulation frequency components and used as features for hybrid Hidden Markov Model - Artificial Neural Network (HMM-ANN) phoneme recognition system [Bourlard and Morgan, 1994]. The proposed technique yields more accurate estimates of phonetic values of the speech sounds than several other state-of-the-art speech analysis techniques. Moreover, these estimates are

much less influenced by distortions induced by the varying communication channels.

2. Feature extraction

The block schematic for the proposed feature extraction technique is shown in Fig. 1. Long segments of speech signal are analyzed in critical bands using the technique of FDLP [Athineos and Ellis, 2007]. 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. The whole set of sub-band temporal envelopes, which are obtained by the application of FDLP on individual sub-band signals, forms a two dimensional (time-frequency) representation of the input signal energy.

The sub-band temporal envelopes are then compressed using a static compression scheme which is a logarithmic function and a dynamic compression scheme [Dau et al., 1996]. 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 [Dau et al., 1996]. 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. 2, which shows (a) a portion of 1000 ms of full-band speech signal, (b) the temporal envelope extracted using the Hilbert transform, (c) the FDLP envelope, which is an all-pole approximation to (b) estimated using FDLP, (d) logarithmic compression of the FDLP envelope and (e) adaptive compression of the FDLP envelope.

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 $0 - 70$ Hz region with a resolution of 5 Hz. This choice is a result of series of optimization experiments (which are not reported here).

3. Experiments and results

The proposed features are used for a phoneme recognition task on the HTIMIT database [Reynolds, 1997]. 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 downsampled to 8 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 [Pinto et al., 2007]. 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, also having clean recordings in the TIMIT test set, are used. The system is trained only on the original 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.

The results for the proposed technique are compared with those obtained for several other robust feature extraction techniques namely RASTA [Hermansky and Morgan, 1994], auditory model based front-end (Old.) [Tchorz and Kollmeier, 2004], Multi-resolution RASTA (MRASTA) [Hermansky and Fousek, 2005], and the Advanced-ETSI (noise-robust) distributed speech recognition front-end [ETSI, 2002]. The results of these experiments on the clean test conditions are shown in the top panel of Table 1. The conventional Perceptual Linear Prediction (PLP) feature extraction used with a context of 9 frames [Pinto et al., 2007] is denoted as PLP-9. RASTA-PLP-9 features use 9 frame context of the PLP features extracted after applying the RASTA filtering [Hermansky and Morgan, 1994]. Old.-9 refers to the 9 frame context of the auditory model based front-end reported in [Tchorz and

Kollmeier, 2004]. The ETSI-9 corresponds to 9 frame context of the features generated by the ETSI front-end. The FDLP features derived using static, dynamic and combined (static and dynamic) compression are denoted as FDLP-Stat., FDLP-Dyn. and FDLP-Comb. respectively (Sec.2). The performance on clean conditions for the FDLP-Dyn. and Old.-9 features validates the claim in [Tchorz and Kollmeier, 2004] regarding the effects of the distortions introduced by adaptive compression model on the higher signal modulations. The experiments on clean conditions also illustrate the gain obtained by the combination of the static and dynamic modulation spectrum for phoneme recognition. The bottom panel of Table 1 shows the average phoneme recognition accuracy (100 - PER, where PER is the phoneme error rate [Pinto et al., 2007]) for all the 9 telephone channels. The proposed features, on the average, provide a relative error improvement of about 10% over the other feature extraction techniques considered.

4. Discussion

Table 2 shows the recognition accuracies of broad phoneme classes for the proposed feature extraction technique along with a few other speech analysis techniques. For clean conditions, the proposed features (FDLP-Comb.) provide phoneme recognition accuracies that are competent with other feature extraction techniques for all the phoneme classes. In the presence of telephone noise, the FDLP-Stat. features provide significant robustness for fricatives and nasals (which is due to modelling property of the signal peaks in static compression) whereas the FDLP-Dyn. features provide good robustness for plosives and affricates (where the fine temporal fluctuations like onsets and offsets carry the important phoneme classification information). Hence, the combination of these feature streams results in considerable improvement in performance for most of the broad phonetic classes.

5. Summary

We have proposed a feature extraction technique based on the modulation spectrum. Sub-band temporal envelopes, estimated using FDLP, are processed by both a static and a dynamic compression and are converted to modulation frequency features. These features provide good robustness properties for phoneme recognition tasks in telephone speech.

Acknowledgments

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. The authors would like to thank the Medical Physics group at the Carl von Ossietzky-Universitat Oldenburg for code fragments implementing adaptive compression loops.

References and links

Athineos, M., Hermansky, H. and Ellis, D.P.W. (**2004**). "LP-TRAPS: Linear predictive temporal patterns," Proc. of INTERSPEECH, pp. 1154-1157.

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

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

Dau, T., Püschel, D. and Kohlrausch, A. (**1996**). "A quantitative model of the "effective" signal processing in the auditory system: I. Model structure," J. Acoust. Soc. Am., Vol. 99(6), pp. 3615-3622.

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

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752.

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

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

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

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

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

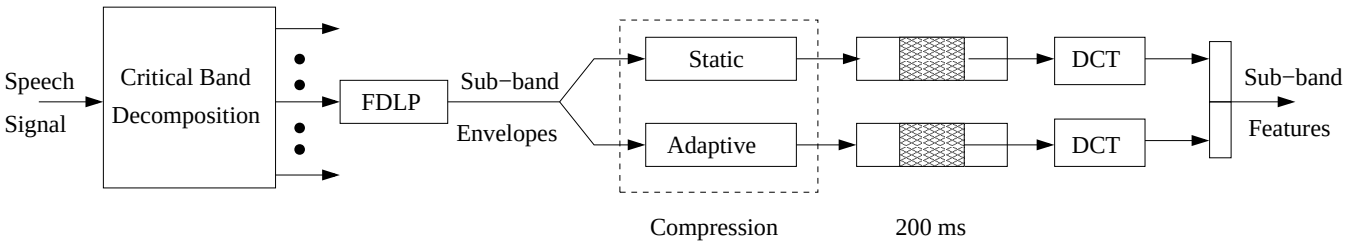
Table 1. Recognition Accuracies (%) of individual phonemes for different feature extraction techniques on clean and telephone speech

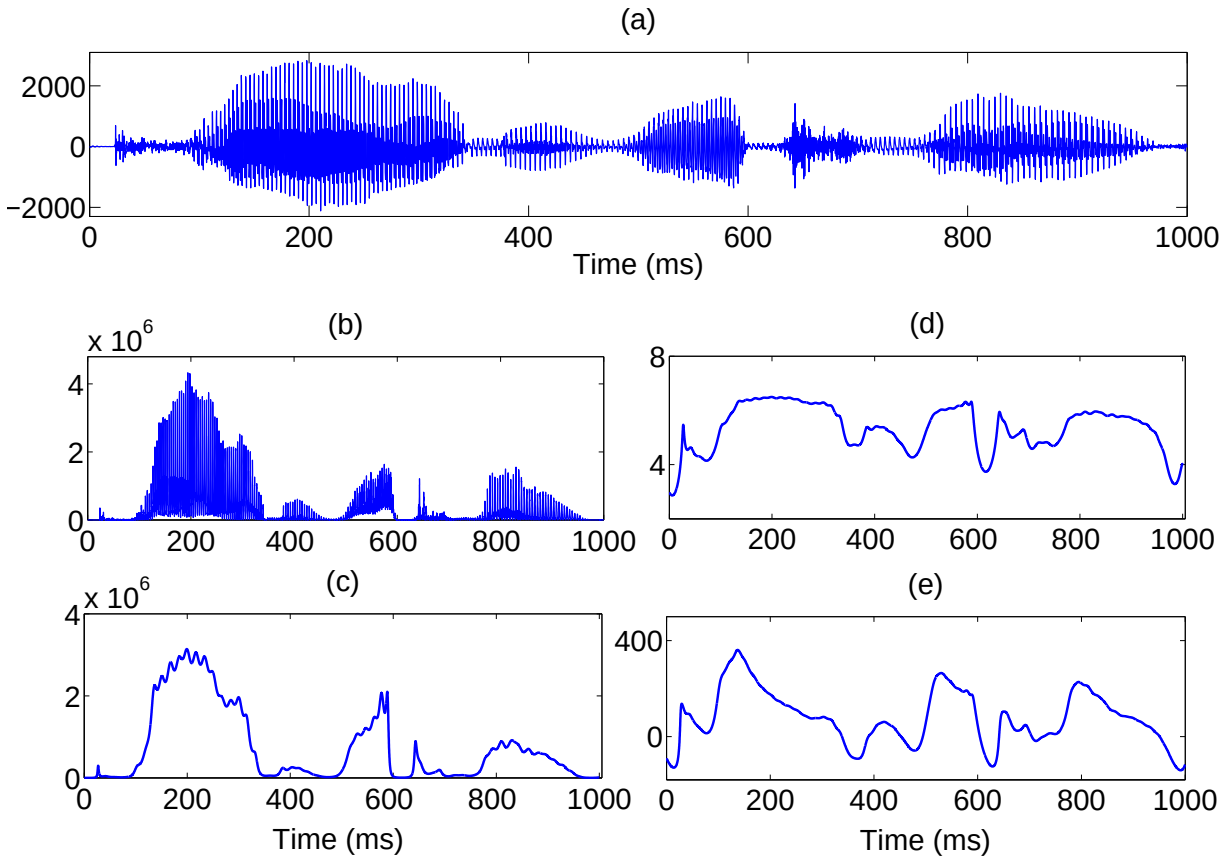
Clean Speech							
PLP-9	R-PLP-9	Old.-9	MRASTA	ETSI-9	FDLP-Stat.	FDLP-Dyn.	FDLP-Comb.
64.9	61.2	60.3	63.9	63.1	63.1	59.7	**65.4**
Telephone Speech							
PLP-9	R-PLP-9	Old.-9	MRASTA	ETSI-9	FDLP-Stat.	FDLP-Dyn.	FDLP-Comb.
34.4	46.2	45.3	47.5	47.7	50.8	48.7	**52.7**

Table 2. Recognition Accuracies (%) of broad phonetic classes obtained from confusion matrix analysis on clean and telephone speech

Clean Speech					
Class	PLP-9	MRASTA	FDLP-Stat.	FDLP-Dyn.	FDLP-Comb.
Vowel	83.3	81.9	82.7	81.3	**83.8**
Diphthong	**75.1**	73.0	70.7	67.9	74.2
Plosive	**81.6**	80.5	79.5	78.2	**81.6**
Affricative	69.1	68.8	64.6	62.5	**69.9**
Fricative	81.8	80.1	80.0	77.8	**81.9**
Semi Vowel	72.2	71.6	70.7	69.5	**73.5**
Nasal	80.4	79.2	80.8	77.7	**82.4**
Telephone Speech					
Class	PLP-9	MRASTA	FDLP-Stat.	FDLP-Dyn.	FDLP-Comb.
Vowel	61.1	74.2	77.5	77.6	**79.8**
Diphthong	51.1	**68.2**	63.4	61.7	67.2
Plosive	46.9	52.5	56.1	**59.0**	**59.0**
Affricative	28.0	38.5	35.7	36.9	**39.8**
Fricative	63.3	70.7	78.5	74.0	**79.4**
Semi Vowel	55.8	61.3	60.5	60.7	**63.8**
Nasal	35.4	57.7	66.6	64.9	**68.7**

List of figures


Speech
Signal
Critical Band
Decomposition
FDLP
Sub−band
Envelopes
Static
Adaptive
DCT
DCT
Sub−band
Features
Compression
200 ms


(a)
2000
0
−2000
0
200
400
600
800
1000
Time (ms)
(b)
x 10^6
4
2
0
0
200
400
600
800
1000
(d)
8
4
0
200
400
600
800
1000
(c)
x 10^6
4
2
0
0
200
400
600
800
1000
Time (ms)
(e)
400
0
0
200
400
600
800
1000
Time (ms)