Introducing Temporal Asymmetries in Feature Extraction for Automatic Speech Recognition

G.S.V.S. Sivaram and Hynek Hermansky

IDIAP Research Institute, Martigny
Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland
{sgarimel, hynek}@idiap.ch

Abstract

We propose a new auditory inspired feature extraction technique for automatic speech recognition (ASR). Features are extracted by filtering the temporal trajectory of spectral energies in each critical band of speech by a bank of finite impulse response (FIR) filters. Impulse responses of these filters are derived from a modified Gabor envelope in order to emulate asymmetries of the temporal receptive field (TRF) profiles observed in higher level auditory neurons. We obtain 11.4% relative improvement in word error rate on OGI-Digits database and, 3.2% relative improvement in phoneme error rate on TIMIT database over the MRASTA technique.

Index Terms: feature extraction, auditory neurons, speech recognition

1. Introduction

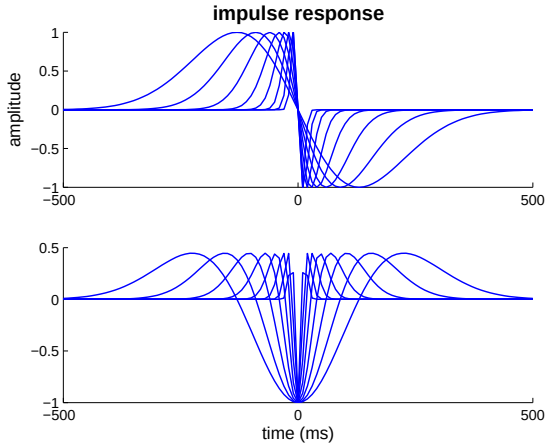

Figure 1: Normalized impulse responses of the MRASTA filters, $\sigma = 8 - 130$ ms.

MRASTA ([2]) technique extracts features by filtering the temporal trajectory of each critical band energy of speech by a bank of finite impulse response (FIR) filters. Thus each feature represents the convolution of the corresponding input critical band trajectory with the impulse response of a filter. The impulse response of each FIR filter is symmetric (even or odd) around the center as shown in figure 1. The impulse responses in MRASTA feature extraction attempt to emulate variable lengths of the temporal envelopes of spectro-temporal receptive fields (STRFs) of auditory neurons at various frequencies [4, 7, 8].

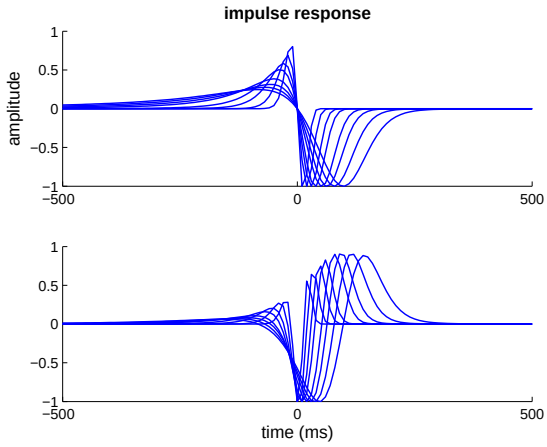

Figure 2: Normalized impulse responses of the asymmetric filters, $m = -140$.

In this paper, we propose modifications to these impulse responses, motivated by the asymmetries of the temporal envelopes of STRFs of the higher level auditory neurons, as shown in the figure 2. The rest of the paper is organized as follows. The motivation for this work is presented in the section 2. In section 3, we give an overview of the MRASTA feature extraction technique and describe our proposed technique to emulate asymmetries of the TRF profiles. Then we discuss experimental results in section 4. Finally we conclude in section 5.

2. Motivation

The peripheral auditory system encodes the acoustic waveform into a neural code in the auditory nerve. This neural code is then interpreted by the central auditory pathways to identify various sounds. Neurons in central auditory stations are sensitive to dynamic variations in the temporal, spectral and intensity composition of the sensory stimulus.

MRASTA approach is motivated to some extent by the recent findings ([4] and [5]) in brain physiology of some mammal species, where spectro-temporal receptive fields (STRFs) are used to characterize some of the higher level auditory neurons. STRF, a linear model, describes the spectro-temporal features of the stimulus (speech) that most likely activate the neuron. Efforts were made in the past to emulate these STRFs using multiple 2-D Gabor filters [8]. However, as in MRASTA, their method did not emulate asymmetry in time which is of interest to this paper.

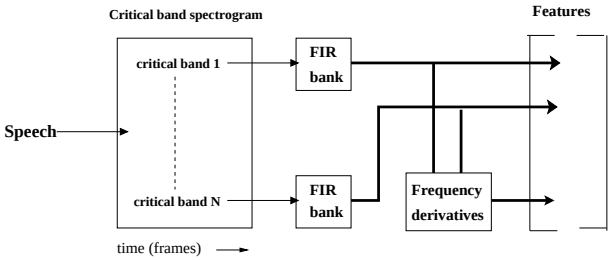

Figure 3: Schematic of the feature extraction.

We hypothesize that the higher level auditory neurons might encode information about acoustic objects such as sounds of speech in the form of neural firing rate. Furthermore, it is possible to predict the neural firing rate of a neuron due to an arbitrary stimulus (speech) by convolving (2-D) the corresponding STRF with the input spectrogram of speech as given by equation 1 ([7]).

$$r_{pre}\left(t\right) \;=\; \sum_{i=1}^{nf} \int h_i\left(\tau\right)\; S_i\left(t \;-\; \tau\right)\; d\tau \qquad (1)$$

where $r_{pre}\left(t\right)$ – predicted firing rate,
nf – number of critical bands,
$h_{\{i\}}\left(t\right)$ – STRF,
$h_i\left(t\right)$ – temporal receptive field of i^{th} frequency channel (critical band),
$S_i\left(t\right)$ – i^{th} critical band trajectory of speech.

One can think of this 2-D convolution as several 1-D convolutions at various critical band trajectories of speech and temporal receptive field (TRF) profiles of the STRF, and subsequent summation of all such convolutions. The TRF profile is obtained by slicing through the STRF at a particular frequency. Additionally, we note that these profiles ($h_i\left(t\right)$) are not symmetric ([6]) for higher level auditory neurons. In fact, [6] uses a modified Gabor envelope to model these asymmetries in time. However, MRASTA feature extraction technique fails to emulate these asymmetries as each of its filter has a symmetric impulse response. This observation motivates us to study the effect of using asymmetric filters in MRASTA feature extraction technique.

3. Feature Extraction

3.1. MRASTA overview

Detailed description of this technique can be found in [2]. In this section, we describe only the FIR filter bank.

Energy in each critical band is extracted from 25 ms windowed speech for every 10 ms as described in [1]. Features are extracted for each frame (10ms) by filtering each of the 15 temporal trajectories of critical band spectral energies (OGI-Digits database) by a bank of 16 FIR filters (shown in the figure 1). Thus the total number of features per frame are $15 \times 16 = 240$. Typically, three tap FIR filter with impulse response $\{-1, 0, 1\}$ is used for computing the first frequency derivatives ($16 \times 13 = 208$ features). Dimensionality is further increased by appending these frequency derivatives to the above described features ($240 + 208 = 448$ features). The schematic of this feature extraction technique is shown in figure 3.

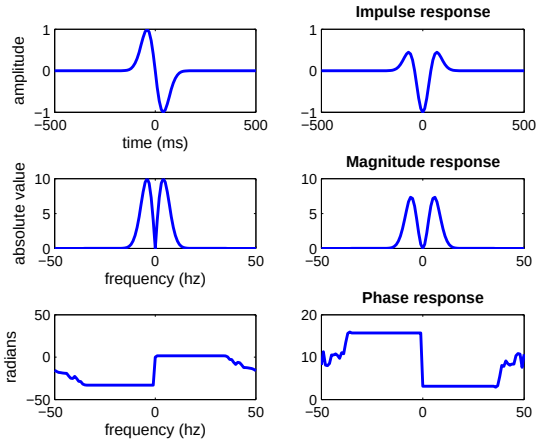

Figure 4: Impulse, magnitude and phase responses of MRASTA filters ($\sigma = 40$ ms), left column: first Gaussian derivative, right column: second Gaussian derivative.

In MRASTA, impulse response of each filter in the FIR filter bank is a discrete version of either first or second analytic derivative of a Gaussian function and is given by equation 2 or 3.

$$g1\left[x\right] \propto -\frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) \qquad (2)$$

$$g2\left[x\right] \propto \left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2}\right) \qquad (3)$$

where x is time, $x \;\in (-500,\; 500)$ ms with the step of 10 ms; standard deviation σ determines the effective width of the Gaussian. Filters with low σ values have finer temporal resolution whereas high σ filters cover wider temporal context and yield smoother trajectories. The impulse response of each filter is shown in figure 1 (total eight different σ values are used). Length of all filters is fixed at 101 frames, corresponding to 1010 ms.

Figure 4 shows the impulse, magnitude and phase responses of few MRASTA filters for $\sigma \;= 40$ ms. Note that each filter has a zero-phase phase response in the passband as the corresponding impulse response is symmetric (even or odd) around the center. Since interval between the frames is 10 ms, the highest frequency (modulation) component is 50 Hz as shown in the figure 4. Therefore one can view this MRASTA technique as performing multiple filtering in modulation spectral domain of speech. Modulation spectral domain is the Fourier domain of the temporal trajectory of a critical band energy.

3.2. Asymmetric filters (proposed technique)

To fit the observed temporal asymmetry of the TRF profile, [6] uses a modified Gabor function. Their idea is to first skew the time axis and then to fit a symmetric Gabor function. Generalized version of their Gabor envelope is given by the equation 4.

$$g\left(x\right) = \exp\left(-\frac{\left(a\;\tan^{-1}(bx) - m\right)^2}{2c^2}\right) \qquad (4)$$

$$x_{peak} = \frac{\tan\left(\frac{m}{a}\right)}{b},\; when\;\; |m| < \left|\frac{a\pi}{2}\right| \qquad (5)$$

The envelope (equation 4) shows asymmetry about its peak point for non zero values of m and the degree of asymmetry

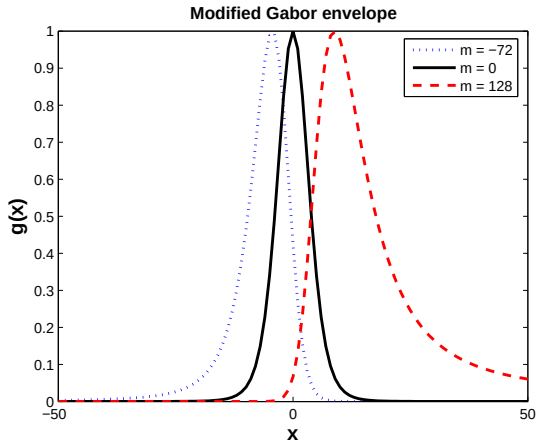

Figure 5: Modified Gabor envelope for $m = -72,\ 0\ and\ 128$, $a = 600/\pi$, $b = 0.09$ and $c = 55$.

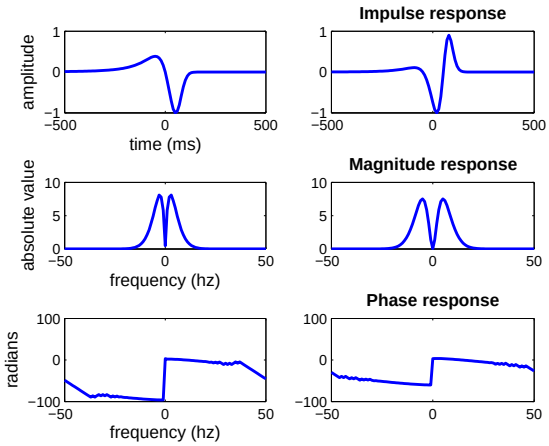

Figure 6: Impulse, magnitude and phase responses of asymmetric filters ($a = 600/\pi$, $(b, c) = (0.09,\ 55)$ and $m = -140$), left column: first derivative (equation 8), right column: second derivative (equation 9).

increases with absolute value of m as shown in the figure 5. The value of x for which envelope reaches its peak is given by equation 5. Note from the equation 4 that $g\,(x)$ is an even function of x when $m\ = 0$. The first and second derivatives of the above envelope are given by the equations 6 and 7 respectively.

$$g^{'}\,(x) = -\frac{a\ b\ \exp\left(-\frac{\left(a\ \tan^{-1}(bx)-m\right)^2}{2c^2}\right)\left(a\ \tan^{-1}(bx)-m\right)}{c^2\ \left(1+b^2x^2\right)} \tag{6}$$

$$g^{''}\,(x) = \frac{2\ a\ b^3\ \exp\left(-\frac{\left(a\ \tan^{-1}(bx)-m\right)^2}{2c^2}\right)\left(a\ \tan^{-1}(bx)-m\right)x}{c^2\ \left(1+b^2x^2\right)^2} + \frac{a^2\ b^2\ \exp\left(-\frac{\left(a\ \tan^{-1}(bx)-m\right)^2}{2c^2}\right)\left(a\ \tan^{-1}(bx)-m\right)^2}{c^4\ \left(1+b^2x^2\right)^2} - \frac{a^2\ b^2\ \exp\left(-\frac{\left(a\ \tan^{-1}(bx)-m\right)^2}{2c^2}\right)}{c^2\ \left(1+b^2x^2\right)^2} \tag{7}$$

The impulse responses of the asymmetric filters are derived from these derivatives as per the equations 8 and 9.

$$g1'\,[x] = g^{'}\left(\frac{x}{10}+x_{peak}\right) \tag{8}$$

$$g2'\,[x] = g^{''}\left(\frac{x}{10}+x_{peak}\right) \tag{9}$$

where x is time, $x\ \in (-500,\ 500)$ ms with the step of 10 ms; x_{peak} is given by the equation 5. Furthermore, these impulse responses are symmetric for $m\ = 0$. We choose a set of parameters (not unique) $a = 600/\pi$ and (b, c)={(0.09, 13), (0.09, 20), (0.09, 29), (0.09, 38), (0.09, 55), (0.09, 70), (0.08, 80), (0.07, 90)} such that for each combination, the variance of the envelope $g\,(x)$ (equation 4) with $m\ = 0$ matches that of the underlying Gaussian function (i.e., σ^2) of MRASTA. Parameter m can be used to control the degree of asymmetry after fixing these remaining parameters. Figure 2 shows asymmetric impulse responses for the above choice of parameters and $m = -140$. Magnitude and phase responses of some of these asymmetric filters are shown in figure 6. Note that zero mean property of the impulse response is preserved[1] but we no longer have the zero-phase response as the impulse response is asymmetric around center.

Features are extracted from speech by using these asymmetric filters for different values of m. These features are compared against baseline MRASTA features in terms of ASR performance in the following section. Note that the configuration of the ASR system remains same for both proposed and MRASTA features.

4. Experiments

Initial set of experiments consists of small vocabulary continuous digit recognition (OGI Digits database). Recognized words are eleven ($0-9$ and $zero$) digits in 28 pronunciation variants. Features are extracted from speech every 10 ms as described in section 3. Multi-layer perceptron feed forward neural net (MLP) with 1800 hidden nodes is trained on the whole Stories database plus training part of the Numbers95 database to estimate posterior probabilities of 29 English phonemes. Around 10% of the data is used for cross-validation. Log and Karhunen Loeve (KL) transforms are applied on these features in order to convert them into features appropriate for a conventional HMM recognizer ([3]). The HMM based recognizer, trained on training part of the Numbers95 database, is used for classification. The performance of the proposed features is compared against the baseline MRASTA features in terms of word error rate (WER) below.

The WER of baseline MRASTA features on OGI-Digits database is **3.5%**. Figure 7 shows WER of proposed features for different values of the parameter m (values of other parameters are as in section 3.2). Though variances of the envelopes are matching, however, the baseline performance is slightly better than that of the proposed features when $m = 0$. This can be attributed to the fact that time axis is skewed by *arctan* asymmetry resulting in different rate of change. Observe from the

[1] *Except the largest two first derivatives as they are not close to zero at the point of truncation.*

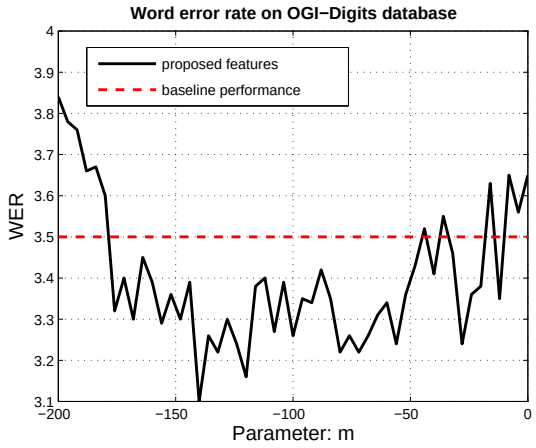

Figure 7: Word error rate as a function of parameter m on OGI-Digits database, optimal $m = -140$.

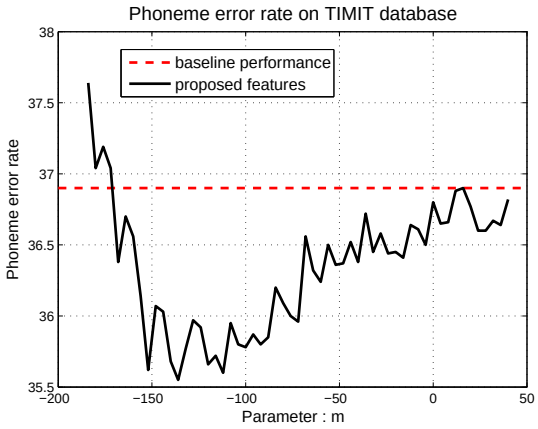

Figure 8: Word error rate as a function of parameter m on TIMIT database, optimal $m = -136$.

figure that the best WER of about **3.1%** corresponds to the parameter value $m = -140$. *–a relative improvement in WER of over* 11.4% *on OGI-Digits database.* The impulse responses of the asymmetric filters corresponding to these parameters are shown in figure 2.

Table 1: Comparison of performances (in $\%$) of proposed features and baseline MRASTA features.

	Asymmetric filters	MRASTA (baseline)
OGI-Digits (WER)	3.1	3.5
TIMIT (PER)	35.7	36.9

In order to test the effectiveness of the proposed features on a different database, phoneme classification experiments are conducted on TIMIT. The training data set 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. MLP with 1000 hidden nodes is trained to convert input speech features into posterior probabilities of phoneme classes (standard set of 39) and decisions are made based on these probabilities (Viterbi decoding). Phoneme error rate (PER) is used as a measure to evaluate performance of the features. The PER of the baseline MRASTA features is **36.9%** while that of the proposed features ($m = -140$) is **35.7%**. Thus the proposed features yield *a relative improvement of about* 3.2% *over the baseline features on TIMIT database.* We summarized the results in table 1. Figure 8 shows PER as a function of the parameter m on TIMIT database and the optimal value of m is -136. Thus the optimal parameter values on two different databases ($m = -140,\ -136$ on OGI-Digits and TIMIT respectively) are matching when optimized the performance with respect to the parameter m. This shows that found asymmetry applies equally well to different databases.

5. Conclusions

A new auditory inspired feature extraction technique, motivated by the asymmetries of the TRF profiles of higher level auditory neurons, has been proposed and tested for an ASR task. Results from the experiments on different databases seem to be promising, suggesting that careful emulation of STRFs of higher level auditory neurons would lead to better performance. With the proposed approach, we obtained about 11.4% relative improvement in WER on OGI-Digits database and 3.2% relative improvement in PER on TIMIT database. Experimental results indicate that the proposed asymmetric filters generalize well over different databases.

6. Acknowledgments

This work was supported by the European Union (EU) under the integrated project DIRAC, Detection and Identification of Rare Audio-visual Cues, contract number FP6-IST-027787, and by DARPA GALE program.

7. References

[1] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", *J. Acoust. Soc. Am.*, Vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

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

[3] H. Hermansky , D.P.W. Ellis , S. Sharma, "Tandem connectionist feature extraction for conventional HMMsystems," *Proc. of ICASSP*, Istanbul, Turkey, 2000.

[4] D.A. Depireux , J.Z. Simon , D.J. Klein , S.A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of Neurophysiology*, Vol. 85, pp. 1220-1234, 2001.

[5] C.E. Schreiner , H.L. Read , M.L. Sutter, "Modular Organization of Frequency Integration in Primary Auditory Cortex," *Annual Review of Neuroscience*, Vol. 23, pp. 501529, Mar. 2000.

[6] A. Qiu , C.E. Schreiner , M.A. Escabi, "Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition," *Journal of Neurophysiology*, Vol. 90, 2003.

[7] F.E. Theunissen , K. Sen , A.J. Doupe, "Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds," *Journal of Neurophysiology*, pp. 20: 2315-2331, Mar. 2000.

[8] M. Kleinschmidt , D. Gelbart, "Improving Word Accuracy with Gabor Feature Extraction," *Proc. of ICSLP*, Colorado, USA, 2002.