EMULATING TEMPORAL RECEPTIVE FIELDS OF AUDITORY MID-BRAIN NEURONS 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

This paper proposes modifications to the Multi-resolution RASTA (MRASTA) feature extraction technique for the automatic speech recognition (ASR). By emulating asymmetries of the temporal receptive field (TRF) profiles of auditory mid-brain neurons, we obtain more than 13% relative improvement in word error rate on OGI-Digits database. Experiments on TIMIT database confirm that proposed modifications are indeed useful.

1. INTRODUCTION

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. Note that impulse response of each FIR filter is symmetric (even or odd) around the center as shown in the figure 1.

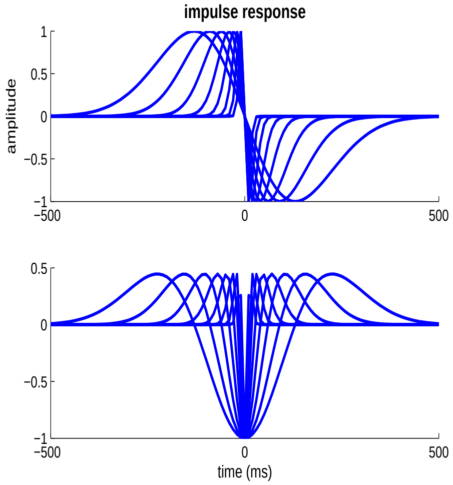

Figure 1: Normalized impulse responses of the MRASTA filters, $\sigma = 8 - 130$ ms.

In this paper, we propose modifications to these impulse responses, motivated by the asymmetries of the auditory mid-brain neurons, as shown in the figure 2. These filters give more importance to the past than to the future. For content based audio classification task, use of spectro-temporal features has been recently demonstrated in [9].

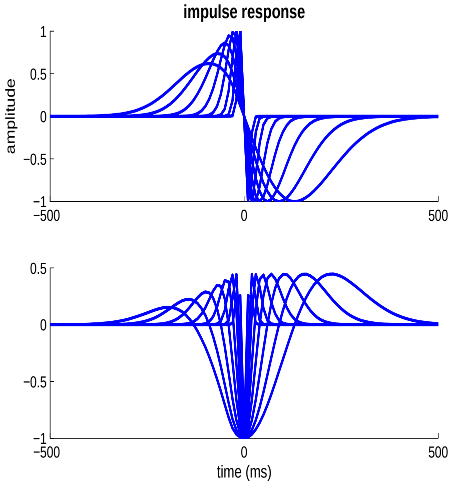

Figure 2: Normalized impulse responses of the asymmetric filters, $\sigma = 8 - 130$ ms, $a = -15$ and $c = -36$

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.

It is believed that these higher level auditory neurons encode information pertained to the speech recognition 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 the equation 1 ([7]).

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

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

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 ($h_i(t)$) are not symmetric ([6]). 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 features described above ($240 + 208 = 448$ features). The schematic of this feature extraction technique is shown in the figure 3.

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

$$g1[x] \propto -\frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$
$$g2[x] \propto \left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad (2)$$

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 the figure 1 (total eight different σ values are used). Length of all filters is fixed at 101 frames, corresponding to roughly 1000 ms in time.

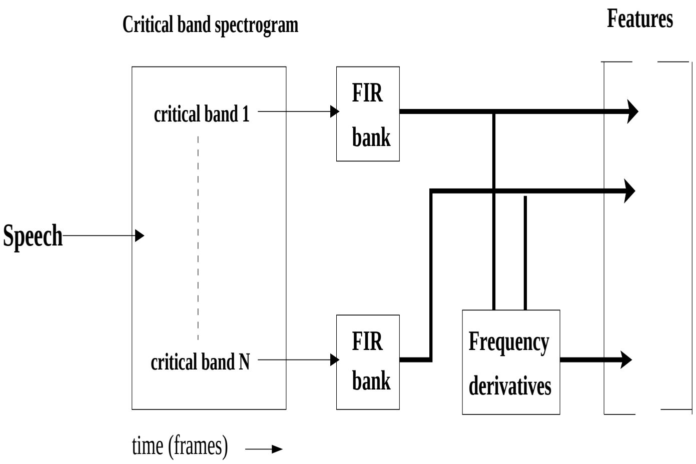

Figure 3: Schematic of the feature extraction

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)

Impulse response of each MRASTA filter is made asymmetric (shown in the figure 2) by multiplying one half of it with warped sigmoid decay function. This makes asymmetric filter impulse response to be smooth around the center. The weights ($W[i]$, $-50 \le i \le 50$) used for multiplication are computed as below.

$$W[i] = 1,\ i \geq 0$$
$$W[i] = \frac{1}{1 + \exp(Q[i])},\ otherwise\ , \qquad (3)$$

where $Q[i]$ represents the time warping function and is given by equation 4 (it has two parameters a and c such that $-50 < c \le a < 0$).

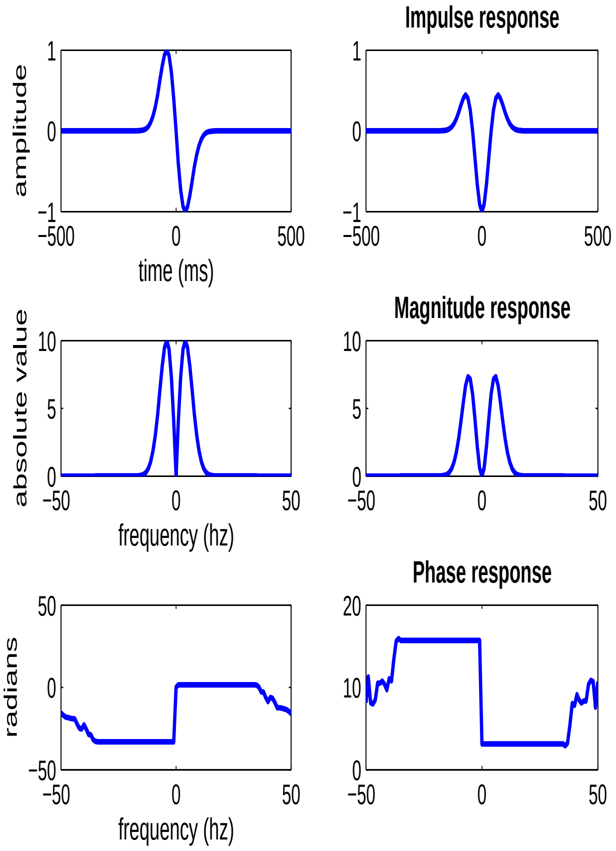

Figure 4: Impulse, magnitude and phase responses of MRASTA filters ($\sigma = 40$ ms), left column: first Gaussian derivative, right column: second Gaussian derivative

$$Q[i] = \tan\left(\frac{\pi(i-a)}{2(a+1)}\right),\ i\ \geq\ a$$

$$Q[i] = \frac{\pi(i-a)}{2(a+1)},\ a\ >\ i\ >\ c$$

$$Q[i] = \frac{\pi(c-a)}{2(a+1)} + \tan\left(\frac{\pi(i-c)}{2(-50-c)}\right),\ otherwise \qquad (4)$$

The impulse responses of asymmetric filters are obtained (from equations 2 and 3) as per the equation 5.

$$g1'[x] = g1[x] \times W\left[\frac{x}{10}\right]$$

$$g2'[x] = g2[x] \times W\left[\frac{x}{10}\right], \qquad (5)$$

where x is time, $x\ \in (-500,\ 500)$ ms with the step of 10 ms; Figure 2 shows these asymmetric impulse responses for a particular case ($a = -15$ and $c = -36$). Magnitude and phase responses of some of these asymmetric filters are shown in the figure 5. Note that we no longer have the zero-phase response as the impulse response is asymmetric around the center.

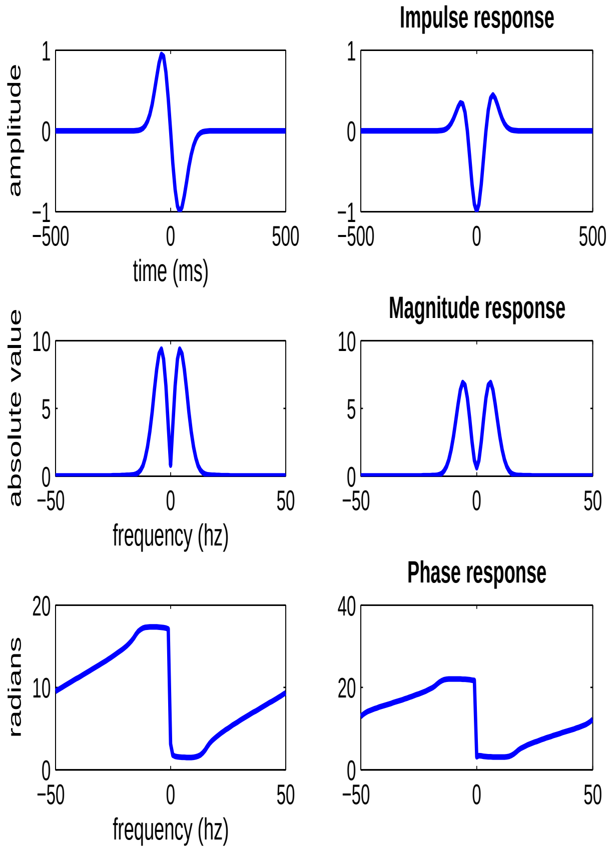

Figure 5: Impulse, magnitude and phase responses of asymmetric filters ($\sigma = 40$ ms, $a = -15$ and $c = -36$), left column: first Gaussian derivative, right column: second Gaussian derivative

Features are extracted from speech by using these asymmetric filters. The section below describes the ASR experiments conducted on different databases and lists the performances of the proposed approach and the baseline MRASTA technique.

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 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 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%**. Table 1 shows the WER of proposed features for different warping parameter values. Note from the table that the proposed features perform better than the baseline features in many occasions. Additionally, the best WER of about 3.0% corresponds to the parameter values $a = -15$ and $c = -36$ *–a relative improvement in WER of over* 13% *on OGI-Digits database.* A bootstrap method for significance analysis ([10]) confirms that difference in performances is statistically significant with 99.98% confidence. The impulse responses of the asymmetric filters corresponding to the optimal parameters are shown in the figure 2.

Table 1: WER (%) for different warping parameters, OGI-Digits database

a/c	-30	-33	-36	-39	-42	-45
-7	3.48	3.51	3.39	3.35	3.37	3.29
-10	3.34	3.25	3.57	3.51	3.26	3.45
-12	3.32	3.19	3.36	3.3	3.2	3.29
-15	3.49	3.45	**3.04**	3.57	3.51	3.26
-17	3.43	3.23	3.42	3.43	3.35	3.14

Table 2: Comparison of performances (in %) of proposed features and baseline MRASTA features.

	Asymmetric filters	MRASTA
OGI-Digits (WER)	3.0	3.5
TIMIT (PER)	35.5	36.9

In order to test the effectiveness of the proposed features on a different database, phoneme classification experiments are conducted on TIMIT. MLP with 1000 hidden nodes is trained to convert input speech features into posterior probabilities of phoneme classes 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 ($a = -15$ and $c = -36$) is 35.5%. Thus the proposed features yield a relative improvement of about 3.8% over the baseline features on TIMIT database. We summarized the results in table 2. The above results indicate that asymmetry in filter shapes is indeed desired for speech recognition task.

5. CONCLUSIONS

Modifications, motivated by the asymmetries of the TRF profiles of auditory mid-brain neurons, to the MRASTA feature extraction technique 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 more than 13% relative improvement in performance on OGI-Digits database. The proposed features also performed better when tested on different (TIMIT) database.

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.

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 and P. Fousek, “Multi-resolution RASTA filtering for TANDEM-based ASR,” *INTERSPEECH*, pp. 361-364, Sep. 2005.

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

[4] D.A. Depireux and J.Z. Simon and D.J. Klein and 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 and H.L. Read and M.L. Sutter, “Modular Organization of Frequency Integration in Primary Auditory Cortex,” *Annual Review of Neuroscience*, Vol. 23, pp. 501-529, Mar. 2000.

[6] A. Qiu and C.E. Schreiner and 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 and K. Sen and 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 and D. Gelbart, “Improving Word Accuracy with Gabor Feature Extraction,” *Proc. of ICSLP*, Colorado, USA, 2002.

[9] N. Mesgarani and M. Slaney and S.A. Shamma, “Content-based audio classification based on multi-scale spectro-temporal features,” *IEEE Transactions on Speech and Audio processing*, Vol. 14, Issue 3, pp. 920-930, May. 2006.

[10] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in ASR performance evaluation,” *Proc. of ICASSP*, Quebec, Canada, 2004.