IDIAP RESEARCH REPORT

Reverse Correlation for analyzing MLP Posterior Features in ASR

Joel Pinto [a] [b] G.S.V.S Sivaram [a] [b]
Hynek Hermansky [a] [b]

IDIAP–RR 08-13

April 2008

soumis à publication

[a] IDIAP Research Institute, Martigny, Switzerland
[b] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Rapport de recherche de l'IDIAP 08-13

Reverse Correlation for analyzing MLP Posterior Features in ASR

Joel Pinto G.S.V.S Sivaram Hynek Hermansky

April 2008

soumis à publication

Résumé. In this work, we investigate the reverse correlation technique for analyzing posterior feature extraction using an multilayered perceptron trained on multi-resolution RASTA (MRASTA) features. The filter bank in MRASTA feature extraction is motivated by human auditory modeling. The MLP is trained based on an error criterion and is purely data driven. In this work, we analyze the functionality of the combined system using reverse correlation analysis.

1 Introduction

Posterior based features figure prominently in the current state-of-the-art large vocabulary continuous speech recognition systems [1] [2]. Here, a multilayered perceptron is discriminatively trained on conventional features (MFCC, PLP, etc) to estimate the posterior probability of phonemes for every frame (typically 10 ms). The posterior probabilities are used as features in subsequent modeling and hence the name posterior features. The posterior features can be used either stand alone [3] or in conjunction with other traditional features [4].

While posterior based features have shown to improve the ASR performance, understanding of its working is limited as neural networks are considered black-boxes and the trained weights do not reflect any properties of speech/features. After the MLP is trained, its properties are typically not further analyzed. It would be useful to develop techniques that would allow to evaluate the trained MLP other than applying it in the target ASR system. This paper aims to contribute to the development of such objective evaluation techniques.

The trained MLP is treated as a nonlinear "black box" in a manner similar to the treatment of the nonlinear perceptual systems in biology. Namely, the reverse correlation technique [10], often applied for obtaining the linear time-invariant (LTI) approximation of the unknown system under consideration [10]. In this work, the MLP is trained using MRASTA [5] features. As shown in Fig. 1, we treat the MRASTA filters followed by MLP as the unknown system taking critical band energies as input and estimating posterior probabilities at the output. We consider MRASTA features because (a) average stimuli derived from reverse correlation analysis can be compared to the expected time-frequency pattern and interpreted in terms of formant energies, and (b) have successfully been applied in various state-of-the-art ASR systems [4] and hence the usefulness of the analysis.

To draw analogy to the reverse correlation studies in physiology [10], we can loosely compare the MRASTA-MLP system to the human auditory system. The variable frequency response in MRASTA feature extraction attempts to emulate the property that each particular higher level neuron in the auditory cortex is the most sensitive to a particular modulation frequency of the signal [7] [8] [9]. Since we do not know exactly how the human brain is integrating this information to perceive speech sounds, we conveniently assume that the MLP learns the transformation. However, human auditory system is far superior compared to the simple MRASTA-MLP system. For example, humans do not perceive random time frequency pattern (away from the speech classes) as speech sounds whereas, MLP could assign a high posterior probability depending on its distance from decision boundary. This model deficiency clearly shows up in the reverse correlation experiments using white noise stimulus (section 3.3). One way to overcome this deficiency is to use generative models for speech (or phonemes) such as GMM, as it restricts the boundary of a speech classes.

The rest of the paper is organized as follows. In section 2, we briefly describe the MRASTA-MLP system that we analyze in this paper. In section 3, we review the reverse correlation technique and use the same to analyze the basic system for various stimuli, namely speech and white noise. Section 4 describes the deficiency of the MRASTA-MLP system in white noise analysis and discusses the generative GMM model.

2 MRASTA-MLP System

The block diagram of a posterior feature extraction using MRASTA features is shown in Fig. 1.

2.1 Critical Band Analysis

Speech is first frame blocked into 25 ms windows with a frame shift of 10ms. Spectral analysis is performed on the windowed speech signal and energies in the critical bands are computed. The center frequency and bandwidth of the critical bands are based on the perceptual modeling of speech. The

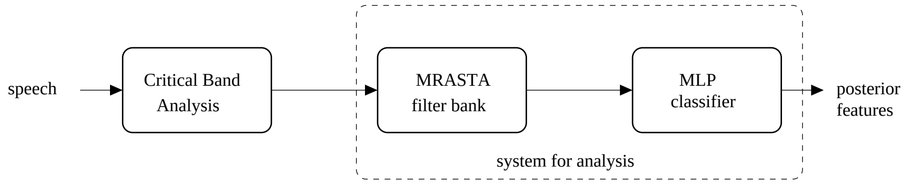

Fig. 1 – *Block diagram of computing posterior features using MRASTA feature extraction.*

trajectory of the log-energy in each of the 19 critical bands is then filtered independently using a bank of MRASTA filters.

2.2 MRASTA Filters

MRASTA filters [5] are zero-mean, 101-tap finite impulse response filters whose shape is that of either the first or second derivative of a Gaussian function. The variance of the Gaussian function controls the resolution of each filter. Our implementation of an MRASTA filter-bank includes 8 first derivatives and 8 second derivatives of Gaussian functions with standard deviations between 8ms and 130 ms. Furthermore, the frequency derivatives are appended to the base features.

2.3 MLP Classifiers

We consider a three layered MLP classifier, where the features presented at the input layer are projected to a higher dimensional hidden layer. The nodes in the output layer represent the phoneme classes. The hidden nodes have a static non-linearity function such as sigmoid, tanh etc. The output layer has a softmax nonlinearity, which enforces the constraint that the outputs sum to unity. Cross entropy error criterion is used to train the MLP. It has been shown that MLPs with sufficient capacity estimate the Bayesian *a posteriori* probability provided that, the network is trained on sufficient training data and classes are taken with the correct *a priori* probabilities [6].

3 Reverse Correlation

Reverse correlation can be used to identify linear time-invariant (LTI) systems. If an LTI system is presented with white noise as input and yields spikes at the output, its impulse response function can be recovered by a simple spike-triggered average of the noise stimulus preceding the spikes. Section 3.1 describes the theory of reverse correlation for a linear system. In 3.2, we investigate its possible extension to analyzing a MLP using speech signal as input. In section 3.3, we apply reverse correlation by presenting white noise as input to the system.

3.1 Reverse correlation on LTI system

Suppose that an unknown linear system with impulse response $h(t)$ and frequency response $H(\omega)$ is to be identified. Suppose that when the system is presented with white noise, spikes are produced at times times $t_1, t_2 \cdots t_N$. Denoting $x(t)$ and $y(t)$ as the input and output to the system, the power spectrum of the system can be written as

$$H(\omega) = \frac{S_{xy}(\omega)}{S_{xx}(\omega)}, \tag{1}$$

where, $S_{xy}(\omega)$ is the cross power spectral density and $S_{xx}(\omega) = \sigma^2$ is the power spectral density of the white noise input. Hence, the impulse response of the unknown system can be written as

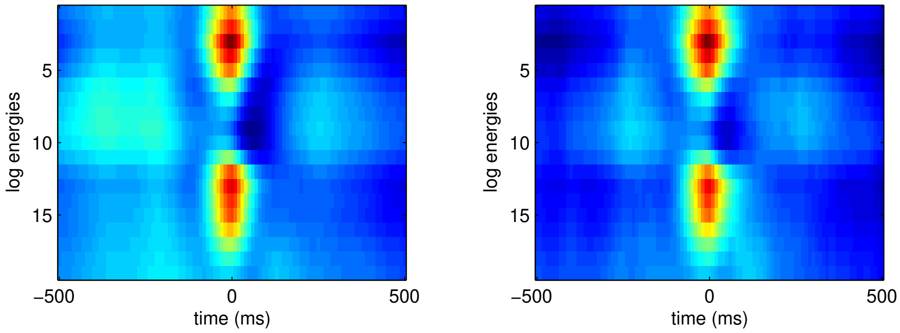

Fig. 2 – *The true average time-frequency pattern (left) and the average pattern estimated by reverse correlation analysis for the phoneme /iy/.*

$$
\begin{aligned}
h(t) &= \frac{1}{\sigma^2} r_{xy}(t) = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} x(\tau - t) y(\tau) d\tau \\
&= \frac{1}{\sigma^2} \int_{-\infty}^{\infty} x(\tau - t) \sum_{k=1}^{N} \delta(\tau - t_k) d\tau \\
&= \frac{1}{\sigma^2} \sum_{k=1}^{N} x(t_k - t)
\end{aligned}
$$

This is the reverse-correlation formula which states that the impulse response $h(t)$ of an LTI system can be obtained as the average of the stimulus preceding the spikes.

Reverse correlation analysis is valid only for a linear system that produces spikes when presented with white noise input. Since the MRASTA-MLP system is a nonlinear system with memory, its impulse response is not defined. Nevertheless, this method can be used to estimate an average pattern in the time-frequency (critical band energy) plane that represents patterns likely to trigger the output neuron for a phoneme. In this direction, we perform reverse correlation studies using actual speech signal and white noise as input. This is explained in the following sections.

3.2 Reverse correlation on MLP (Speech input)

We present speech signal from the test set and average all time-frequency patterns that give a posterior probability greater than certain threshold (e.g. 0.9) for a particular phoneme. Reverse correlation analysis on the TIMIT database shows that the average time-frequency pattern thus obtained is consistent with the expected time-frequency pattern derived using the ground truth label information as shown in Fig. 2. While the average pattern obtained by reverse correlation analysis is consistent with the expected pattern, this is in the average sense (first order approximation) and this does not indicate that the trained system is perfect. Moreover, such a result is not surprising as the neural network is trained to do so.

Reverse correlation analysis using speech as input will reveal the behavior of the system for time-frequency patterns that closely match those that are seen during training. This analysis will not reveal the true functionality of the system as the stimulus space is restricted to be speech like. Reverse correlation analysis with white noise as critical band energies would reveal the behavior of the system in the average sense. White noise analysis is also motivated by the following two factors. Firstly, in the reverse correlation analysis explained in Section 3.1, impulse response of a linear system can be estimated as the average of the noise stimulus preceding the spikes. Secondly, in physiology experiments, spectro-temporal receptive field (STRF) of a neuron can be estimated for white noise stimulus

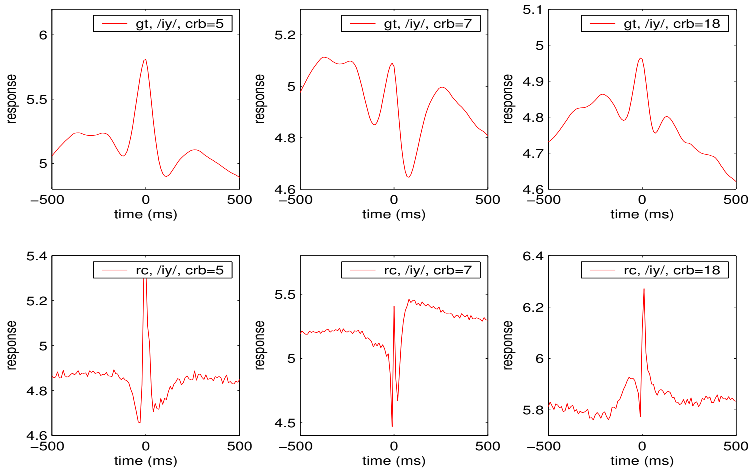

FIG. 3 – *Critical band trajectories for phoneme /iy/, estimated based on ground truth (gt) (top) and reverse correlation (rc) (bottom) for critical bands 5, 7, and 18*

by using reverse correlation technique [10].

3.3 Reverse correlation on MLP (White noise input)

We present uniform noise as critical band energies to the MRASTA-MLP system and perform reverse correlation analysis. The minimum and maximum value of the uniform noise for each critical band is estimated from the training data. In this way, we bound the stimulus space. Noise is presented as critical band energies and not as the actual speech signal. This is because we are interested in identifying response of the system that estimates posterior probabilities from time frequency plane as this can be compared to the formant structure observed in a spectrogram.

Experiments were conducted on the TIMIT database. The average stimuli pattern obtained by reverse correlation is noisy and a plot similar to Fig. 2 will not be informative. Hence, we plot the trajectories of the individual critical bands obtained from reverse correlation as shown in Fig 3. It can be observed from the figure that the trajectories obtained from reverse correlation have similar shape to the expected trajectory for all phonemes. This enables us to devise strategies to compare different systems (e.g. trained on different amounts of data, different capacity, various languages, etc) without having to actually run ASR experiments.

The average pattern is still very noisy when compared to the one derived using speech as input. This can be attributed to the inherent nature of modeling in the MLP as explained in the following section. On the other hand, human auditory system is robust to white noise and will not associate noise patterns to any phoneme.

4 Generative Vs Discriminative Modeling

An MLP is trained using an error criterion which minimizes the classification error on the training set. This is achieved by adjusting the decision boundaries to maximally separate the data points corresponding to the classes. This leaves huge voids within the stimulus space, where a posterior probability of close to unity is assigned to data points even falling away from its distribution. Fig. 4 is the block schematic diagram illustrating discriminative and generative modeling in the critical band space. Here, the data point X falls outside the data points of phonemes $P1$ and $P2$. However, the

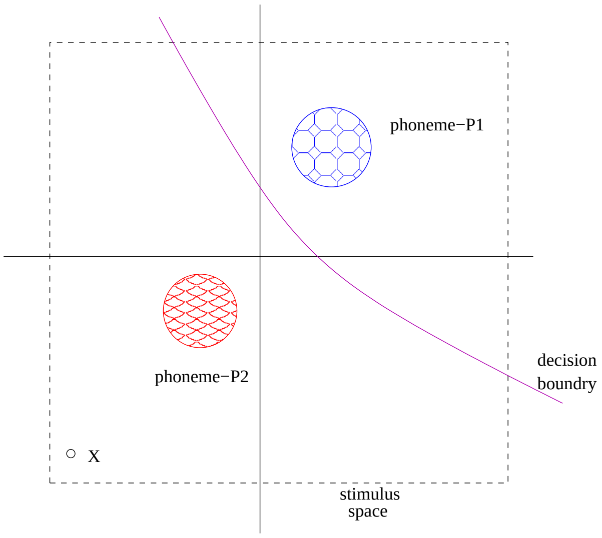

Fig. 4 – *Block schematic illustrating discriminative and generative modeling in the critical band space.*

MLP will assign it to class $P2$ with probability close to unity. This is reason why reverse correlation analysis with white noise fails to give a time-frequency pattern close the one computed using ground truth in Fig. 2. On the contrary, human auditory system is robust to white noise and will not associate noise patterns to any phoneme.

Generative models like Gaussian mixture model (GMM) may be more robust when presented with white noise. If reverse correlation analysis is performed by thresholding the likelihoods, the data point X in Fig. 4 will not be assigned to any phoneme class. Let S be the stimulus space in the critical band energy space. Let $S_M(q, \tau)$ denote the subset of the stimulus space such that every point in S_M will give a MLP posterior probability estimate for phoneme q exceeding threshold τ. Similarly, let $S_G(q, \tau)$ denote the subset of the stimulus space such that every point in S_G will give a GMM likelihood for phoneme q exceeding threshold τ.

$$S_M(q, \tau) = \{x \subset S \mid P(q|x) > \tau\} \quad (2)$$

$$S_G(q, \tau) = \{x \subset S \mid p(x|q) > \tau\} \quad (3)$$

In the case of generative GMM model, by selecting sufficiently high threshold τ, the volume of S_G can be shrunk so that reverse correlation analysis will give an average pattern close to the one obtained with speech input. On the other hand, in the case of discriminative MLP, even though a high τ (close to unity) is fixed, the volume of S_M will be still large as points far of from decision boundary will give an high posterior probability. Reverse correlation studies on GMM model is practically impossible as the volume of S_G will be significantly smaller than stimulus space S especially as the dimension of the feature vector increases. If infinite noise samples are generated, then we can expect an average pattern close to that obtained with speech input.

5 Conclusions

In this work, we present preliminary experiments on the use of reverse correlation for analyzing the system consisting of MRASTA filter banks followed by an MLP. Reverse correlation was performed using two stimuli sources namely, speech and white noise. In the case of speech stimuli, as expected the average time frequency pattern obtained by reverse correlation is close to the expected pattern derived from ground truth. Even in the case of white noise stimuli, the reverse correlation gives time-frequency patterns which are similar to the expected patterns. Reverse correlation with white noise input assumes significance as this could lead to various strategies to analyzing different MLPs (trained on different data sizes, different capacities, different languages, etc.) without actually having to run ASR experiments. In this work, we chose MRASTA feature extraction. In general, reverse correlation analysis can be applied to any feature extraction technique.

6 Acknowledgements

This work was supported in parts by the Swiss National Science Foundation under the Indo-Swiss joint research program KEYSPOT, the European Union under the DIRAC integrated project, contract No. FP6-IST-027787 as well as DARPA under the GALE program, contract No. HR0011-06-C-0023. Any findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of funding agencies.

Références

[1] Q. Zhu, A. Stolcke, B. Chen, N. Morgan "Using MLP Features in SRI's Conversational Speech Recognition System", *Proc. of Interspeech*, pp 2141-2144, 2005.

[2] Q. Zhu, B. Chen, N. Morgan, A. Stolcke "On Using MLP Features in LVCSR", *Proc. of Interspeech*, pp. 921-924, 2004.

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

[4] F. Valente, et al."Hierarchical Neural Networks Feature Extraction for LVCSR system", *Proc. of Interspeech*, 2007.

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

[6] M.D. Richard, R.P. Lippmann, "Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities", *Neural Computation*, pp. 461-483, vol. 3, 1991.

[7] 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.

[8] 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.

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

[10] D.J. Klein, D.A. Depireux, J.Z. Simon, S.A. Shamma, "Robust Spectrotemporal Reverse Correlation for the Auditory System : Optimizing Stimulus Design," *Journal of Computational Neuroscience*, Vol. 9, pp. 85-111, July. 2000.