Towards Fusion of Feature Extraction and Acoustic Model Training: A Top Down Process for Robust Speech Recognition

Yu-Hsiang Bosco Chiu, Bhiksha Raj and Richard M. Stern

Department of Electrical and Computer Engineering and Language Technologies Institute
Carnegie Mellon University, Pittsburgh PA 15213 USA
{ychiu,bhiksha,rms}@cs.cmu.edu

Abstract

This paper presents a strategy to *learn* physiologically-motivated components in a feature computation module discriminatively, directly from data, in a manner that is inspired by the presence of efferent processes in the human auditory system. In our model a set of logistic functions which represent the rate-level nonlinearities found in most mammal hearing system are put in as part of the feature extraction process. The parameters of these rate-level functions are estimated to maximize the *a posteriori* probability of the correct class in the training data. The estimated feature computation is observed to be robust against environmental noise. Experiments conducted with the CMU Sphinx-III on the DARPA Resource Management task show that the discriminatively estimated rate-nonlinearity results in better performance in the presence of background noise than traditional procedures which separate the feature extraction and model training into two distinct parts without feed back from the latter to the former.

Index Terms: automatic speech recognition, discriminative training, auditory model, data analysis

1. Introduction

Automatic Speech Recognition (ASR) systems strive to achieve high word recognition accuracy, while being robust to variations in the environmental or acoustic conditions that the speech is recorded in. The first step in a speech recognition system is to compute a set of feature vectors from the incoming speech. Traditional features such as mel-frequency cepstral coefficients (MFCC) [1] or perceptual linear prediction (PLP) [2] are known to provide good speech recognition accuracy when the recording conditions of the "test" data used to train the recognizer and the speech being recognized are similar. However when the training and test data are not matched in this manner, recognition accuracy degrades.

In contrast, humans are very good at recognizing speech even under highly noisy conditions that they have not been exposed to earlier. Some of this robustness to noise is attributable to a top-down process that is part of human auditory perception. Specifically, in the human auditory system, there are not only nerve fibers which send information from the ear to the brain, but also *efferent* fibers travel down the nerve and synapses in the cochlea that control the motion of cochlear hair cells [3] so as to refine the "features" extracted from the basilar membrane.

Using a similar principle, several feature extraction strategies have been proposed that integrate information from the output of a recognizer to refine the feature generation process. For example, Biem et al. proposed a discriminative feature extraction procedure which refines the filter bank that is a major component of most feature computation schemes, by using a smoothed binary loss [4]. Kinnunen used the F-ratio to design a filter bank for improving the speaker recognition performance [5]. These methods have primarily addressed data-driven optimization of the frequency analysis of the speech signal. Other methods have employed the same principle to compute *transformations* of features to optimize recognition, *e.g.* [6, 7].

In this paper, we investigate a technique for the design of a physiologically-motivated processing stage in feature computation, that is optimized for recognition accuracy. In previous work [8] we have determined that the *rate-level* nonlinearity that models the non-linear relationship between input signal level and the auditory neural spike rate is a major contributor to robustness in speech recognition. In other physiological studies in cats it has been observed that the distribution of different types of auditory neurons (neurons with high, medium or low spontaneous rate of spike generation) that affect spike rate depends on the noise in the environment that the animal was raised in [9], indicating that the rate levels are at least partially a function of the "training" data the animal was exposed to. Motivated by both facts, we investigate a technique for automatically learning the parameters of a non-linear compressive function that mimics the rate-level nonlinearity to optimize recognition performance in noise.

The rest of this paper is organized as follows. In Section 2, we describe the feature computation scheme we employ. In Section 3 we describe the learning algorithm that learns the relevant parameters of the feature computation. In Section 4 we summarize the learning procedure. In Section 5 we describe experiments conducted on the DARPA Resource Management database. Finally, Section 6 concludes the paper.

2. Feature Computation with Equal-loudness and Rate-level Nonlinearity

We parameterize speech signals using a feature computation scheme proposed by Chiu and Stern [8, 10]. The overall scheme is shown in Fig.1. Each analysis frame of the incoming speech signal is analyzed by a fast Fourier transform. The resulting spectrum is integrated into a smaller number of Mel-spectral values using a Mel-frequency filter bank. Each Mel-spectral value is compressed using a logarithmic compression. The log-compressed Mel-spectral values are passed through a sigmoidal nonlinearity that represents the rate-level nonlinearity. The sigmoidal nonlinearity is given by

$$x_i[t] = \frac{\alpha[i]}{1 + exp(w_1[i] \cdot y_i[t] + w_0[i])} \qquad (1)$$

where $y_i[t]$ is the i^{th} log Mel-spectral value and $x_i[t]$ is the corresponding sigmoid-compressed value of frame t. In [10] the parameters of the non-linearity, $\alpha[i] = 0.05$; $w_0[i] = 0.613$; $w_1[i] = -0.521\ \forall i$ were obtained by fitting it to physiological measurements followed by further hand refinement, in order to mimic the effect of measured non-linear neural response. Note that these values are the same for all Mel-frequency components, *i.e.* they are frequency independent.

The compressed values are then projected down to a 13-dimensional cepstrum by a Discrete Cosine Transform (DCT) and used for further recognition.

An additional aspect of the feature computation that is not illustrated in Figure 1 is an *equal-loudness weighting* that is applied to every spectral component prior to the logarithmic compression. This is a frequency-dependent gain term, shown in Figure 2, that is derived from the equal-loudness curve [11] which characterizes loudness response of the auditory system. In reality, loudness response is a function of both frequency and the perceived loudness of the signal; however here we only use the mean response and assume it is only dependent on frequency.

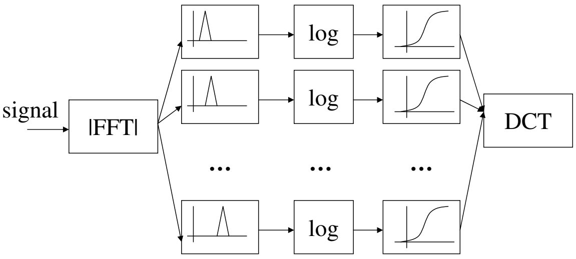

Figure 1: *Feature computation scheme.*

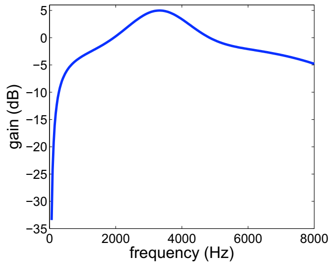

Figure 2: *Equal loudness weighting.*

The sigmoidal non-linearity serves two purposes. The primary purpose is to mimic rate-level nonlinearity in human auditory response. The secondary purpose relates to the equal-loudness weighting. In the absence of the sigmoidal non-linearity, equal loudness weighting emerges as an additive constant after the logarithmic compression and would get eliminated by the cepstral mean subtraction (CMS) that is routinely used in speech recognition. The sigmoidal non-linearity serves to combine the gain into the features in a non-linear manner such that it cannot be eliminated by CMS.

3. Learning the Non-linearity

We would like to optimize the parameters of the non-linearity to optimize recognition accuracy. However, the statistical models used for automatic speech recognition are highly complex, including hidden Markov models for the various phonemes and a language model, and it is difficult to obtain a simple update mechanism that can relate recognition accuracy to the parameters of the sigmoidal non-linearity. Instead, we use a simple Bayesian classifier for sound classes in the language as a simple substitute for the recognizer itself. Each sound class is modeled by a Gaussian distribution, computed from the training data for that sound class. We use a maximum-mutual information (MMI) criterion to estimate the parameters of the nonlinearity such that the posterior probabilities of the sound classes on the training data are maximized. The actual optimization is performed using a gradient descent algorithm. This is illustrated by Figure 3.

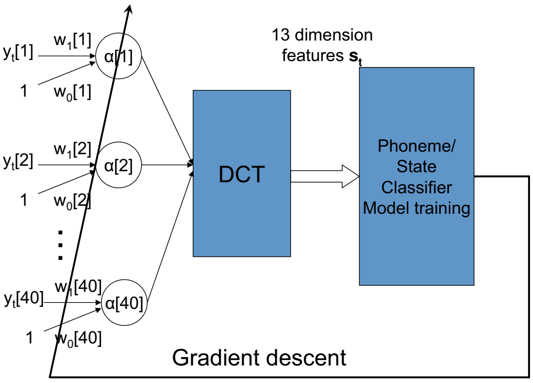

Figure 3: *The integrated system to refine the features extracted.*

The procedure for optimizing the non-linearity is as follows. Let $\boldsymbol{\mu}_C$ be mean vector and $\boldsymbol{\sigma}_C$ be the covariance of the feature vectors for any sound class C. The likelihood of any vector **s**, as computed by the distribution for that sound class is assumed to be given by a Gaussian density $N(\mathbf{s}|\boldsymbol{\mu}_C, \boldsymbol{\sigma}_C)$.

The posterior probability of any sound class C, given a specific observation **s** is given by

$$P(C|\mathbf{s}) = \frac{P(\mathbf{s}|C)P(C)}{\sum_{C'} P(\mathbf{s}|C')P(C')} = \frac{P(\mathbf{s}|C)}{\sum_{C'} P(\mathbf{s}|C')} = \frac{N(\mathbf{s}|\boldsymbol{\mu}_C, \boldsymbol{\sigma}_C)}{\sum_{C'} N(\mathbf{s}|\boldsymbol{\mu}_{C'}, \boldsymbol{\sigma}_{C'})} \quad (2)$$

with the assumption that the prior probabilities of each class are equal.

We assume that we have a collection of training data, and that for each analysis frame of this data we know the identity of the correct sound class. We initialize the parameters of the feature computation with the values from [10]. Each recording from the training data is parameterized using the initial values. CMS is performed on every training recording, in order to stay consistent with the processing that is performed in a complete speech recognition system.

Let $\mathbf{s}_{u,t}$ be the feature vector obtained for the t^{th} analysis frame of the utterance u. Let $C_{u,t}$ be the sound class that the corresponding segment of speech belongs to.

The overall accumulated posterior probability of the entire training data is given by

$$P = \prod_{u,t} \frac{N(\mathbf{s}_{u,t}|\boldsymbol{\mu}_{C_{u,t}}, \boldsymbol{\sigma}_{C_{u,t}})}{\sum_C N(\mathbf{s}_{u,t}|\boldsymbol{\mu}_C, \boldsymbol{\sigma}_C)} \quad (3)$$

The parameters of the distributions of each sound class, and those of the sigmoidal non-linearity in the feature computation are now iteratively optimized to maximize $\log(P)$.

3.1. Estimating Sound Class Distribution Parameters

The model parameters $\boldsymbol{\mu}_C$ and $\boldsymbol{\sigma}_C$ for each sound class is obtained using the same objective criterion employed by the speech recognizer. For maximum-likelihood training, this is given by:

$$\boldsymbol{\mu}_C = \frac{1}{\sum_u \sum_t I(\boldsymbol{s}_{u,t} \in C)} \sum_u \sum_t I(\boldsymbol{s}_{u,t} \in C) \boldsymbol{s}_{u,t},$$

$$\boldsymbol{\sigma}_C = \frac{1}{\sum_u \sum_t I(\boldsymbol{s}_{u,t} \in C)} \sum_u \sum_t I(\boldsymbol{s}_{u,t} \in C) \cdot (\boldsymbol{s}_{u,t} - \boldsymbol{\mu}_C)(\boldsymbol{s}_{u,t} - \boldsymbol{\mu}_C)^T \quad (4)$$

where $I(\boldsymbol{s} \in C)$ is an indicator function that takes a value of 1 if $\boldsymbol{s}$ belongs to sound class C and 0 otherwise.

4. Estimating Sigmoidal Parameters

The parameters for the logistic function **F**= $\{\boldsymbol{\alpha}, \boldsymbol{w_0}, \boldsymbol{w_1}\}$ are estimated to maximize $log(P)$ using a gradient descent approach. Taking the derivative of the objective function with respect to **F** the nonlinear parameters are updated as (note that the step sizes are adjusted s.t. the converge rate for each individual set of parameters are roughly the same):

$$\boldsymbol{\alpha}^{new} = \boldsymbol{\alpha}^{old} + 0.00005 \frac{\partial \log P}{\partial \boldsymbol{\alpha}},$$

$$\boldsymbol{w}_0^{new} = \boldsymbol{w}_0^{old} + 0.05 \frac{\partial \log P}{\partial \boldsymbol{w}_0},$$

$$\boldsymbol{w}_1^{new} = \boldsymbol{w}_1^{old} + 0.01 \frac{\partial \log P}{\partial \boldsymbol{w}_1} \quad (5)$$

After each step of gradient descent as in previous equations on the noisy training set, the model parameters are updated by using Eq.(4) on the clean training set only. Finally, after training is done (the objective function has converged), only the nonlinear parameters **F**=$\{\boldsymbol{\alpha}, \boldsymbol{w_0}, \boldsymbol{w_1}\}$ are retained for the feature extraction process and the model parameters are retrained using the whole speech recognition system on the clean training set.

The entire learning algorithm can be shown as in algorithm 1. Here $\boldsymbol{y}_{u,t}$ represents the log Mel-spectral vector corresponding to the t^{th} analysis window of the u^{th} utterance, $\boldsymbol{s}_{u,t}$ the feature vector computed from it, and $C_{u,t}$ the corresponding sound class.

```
Input: $\mathbf{F}$,$\{(\boldsymbol{y}_{u,t}, C_{u,t}),\ u = 1..U, t = 1..T_U\}$
Output: $\mathbf{F}$
while not converged do
1    Compute feature vector $\{\mathbf{s}_{1,1}, ..., \mathbf{s}_{U,T_U}\}$ using
     Eq.(1) and DCT with CMS
2    Estimate $\{\boldsymbol{\mu}_C, \boldsymbol{\sigma}_C\}\ \forall C$ using Eq.(4) on clean
     training set
3    Compute $\log(P)$ using Eq.(3) on both clean and
     noisy training set
4    $\mathbf{F}_{new} \leftarrow \mathbf{F}_{old} + \frac{\partial \log P}{\partial \mathbf{F}}$ using Eq.(5) on both clean
     and noisy training set
end
```

Algorithm 1: *Algorithm for learning the parameters of the sigmoidal nonlinearity.*

Note that the learned parameters **F** are different for each Mel-spectral channel.

5. Experimental Results

Experiments were run on the DARPA Resource Management database to evaluate the proposed method. The Sphinx-III continuous-density HMM-based system was used in all experiments. HMMs with 1000 tied states, each modeled by a mixture of 8 Gaussians were trained for recognition experiments. The feature extraction employed a 40-filter Mel filter bank covering the frequency range 130Hz - 6800Hz.

In order to train the rate-level nonlinearity, the pink noise from NOISEX-92 was artificially added in to the original clean training set at 10dB SNR to create the noisy training set. The class labels were the 1000 tied states generated by force aligning the clean training set using previously trained models. The noisy testing sets were created by artificially adding babble noise from NOISEX92 and market, theater and restaurant noises from real environment recordings according to the corresponding SNR to the original clean testing set. The training terminates when the improvement of the current log posterior probability of the entire training set not exceeds 0.0001 of log posterior probability in previous iteration.

Figure 4 shows the rate-level nonlinearities learned. Fig. 4(a) is a 3-D plot showing the non-linearities for all 40 Mel-frequency channels. Figures 4(b) shows a few slides of this plot. Figures 4(c)-(e) show the individual parameters of the rate-level nonlinearities as a function of frequency. We note that the estimated optimal rate-level functions vary greatly across frequencies in all aspects, including gain, slope and attack. While we have not compared these to physiological measurements, we do note that these responses can be roughly clustered into low, mid and high-frequency responses.

Once the parameters of the feature computation module were learned, the feature computation module was employed to derive features from a *clean* version of the RM training set, from which HMM model parameters were retrained.

Recognition experiments were run on speech corrupted to various SNR levels by a variety of noises. The performance metric shown is recognition accuracy, which is computed as 100% minus the word error rate, where the latter includes insertion deletion and substitution errors. Figure 5 shows the recognition results obtained. Note that none of the noises used in these experiments were used to train the rate-level non-linearity. The plots of Figure 5 show three sets of recognition results. As a baseline, the recognition performance obtained using conventional Mel-frequency cepstral coefficients is shown. As a second comparator, the performance obtained with the implementation of [10], which also employed equal-loudness weighting and a rate-level nonlinearity is also shown. Here, however, both the equal-loudness weighting was set to be the *approximated* loudness weighting curve [11], while the rate-level nonlinearity was also set to model physiological data most closely. Finally the results obtained using the learned values for the rate-level nonlinearity are shown.

We note that even the equal-loudness weighting and rate-level non-linearity obtained from fit to physiological measurements greatly improve noise robustness. The automatically learned parameters, however, result in the best performance.

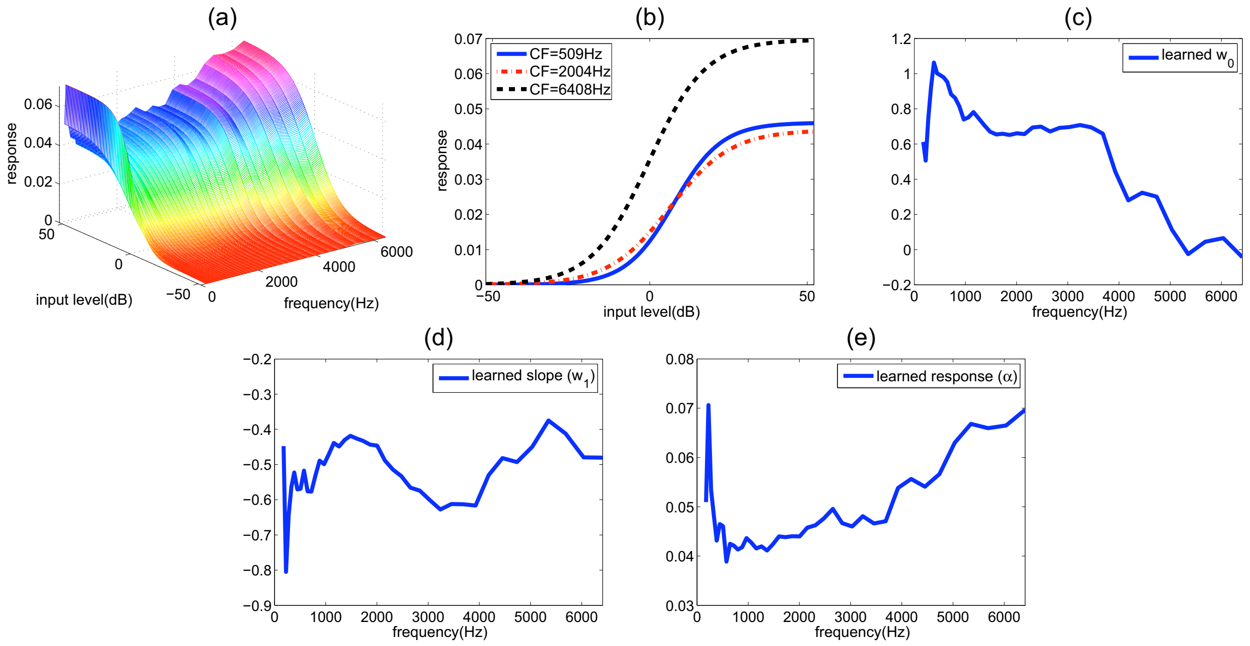

Figure 4: *(a)The trained RL nonlinear over channels. (b)Examples of trained RL nonlinear at low, mid and high frequency region: CF = 509Hz, CF = 2004Hz, CF = 6408Hz. (c)The trained w_0's over frequency channels. (d)The trained w_1's over frequency channels. (e)The trained α's over frequency channels.*

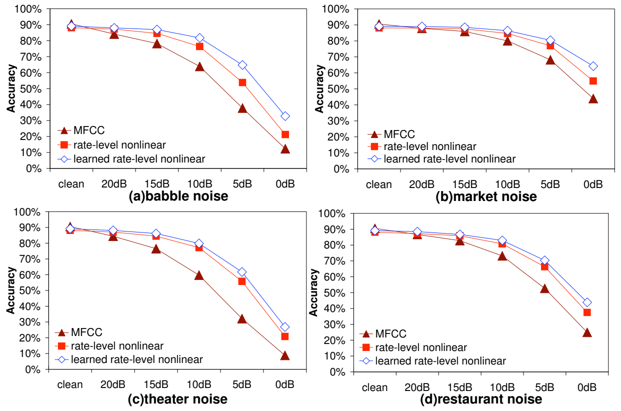

Figure 5: *Comparison of recognition accuracy for the same systems as in Fig.1 in the presence of four types of background noise using the RM corpus. WER under clean: MFCC: 9.45%, RL nonlinear: 11.88%, RL nonlinear from learning: 10.88%*

6. Conclusions

We have presented an algorithm for learning physiologically-motivated components of feature extraction for optimal speech recognition. The results obtained show that the learned feature extraction results in consistently improved speech recognition over conventional feature computation.

The current experiments are only to be considered a teaser, however. It remains unclear how the results will change with increase in data or learning model complexity. On the other hand, it also opens up the possibility that other aspects of human audition might be discovered similarly in a data-driven manner. These are topics of further investigation.

7. Acknowledgements

This work was partially supported by NSF Grant IIS-0420866.

8. References

[1] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, pp. 357-366, 1980.

[2] H. Hermansky,“Perceptual linear predictive (plp) analysis of speech”, J. Acoust. Soc. Am., vol. 87, pp. 1738-1752, 1990.

[3] J.J. Guinan Jr, ”Physiology of olivocochlear efferents”,The Cochlea, vol. 8, pp. 435-502, Springer, NewYork, 1996.

[4] A. Biem, S. Katagiri, E. McDermott and B.-H. Juang, ”An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition”, IEEE Trans. Acoust., Speech, Signal Processing, vol. 9, no. 2, pp. 96-110, 2001.

[5] T. Kinnunen, ”Design a Speaker-Discriminative Adaptive Filter Bank for Speaker Recognition”,Proc. ICSLP, Denver, Colorado, USA, September, 2002.

[6] X. Li and R. Stern, ”Pallel Feature Generation Based on Maximizing Normalized Acoustic Likelihood”, Proc. ICSLP, Jeju Island, Korea, October 2004.

[7] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau and G. Zweig, ”fMPE: Discriminatively Trained Features for Speech Recognition”, Proc. ICASSP, Philadelphia, USA, March 2005.

[8] Y.-H. Chiu and R. Stern, ”Analysis of Physiologically-Motivated Signal Processing for Robust Speech Recognition”, Proc. ICSLP, Brisbane, Australia, September 2008.

[9] M.C. Liberman, ”Auditory nerve response from cats raised in a low noise chamber”, J. Acoust. Soc. Am., vol. 63, pp. 442-455, 1978.

[10] Y.-H. Chiu and R. Stern, ”Minimum variance modulation filter for robust speech recognition”, Proc. ICASSP, Taipei, Taiwan, April 2009.

[11] E. Terhardt, ”Calculating virtual pitch”, Hearing Research, vol. 1, pp. 155-182, 1979.