DISCRIMINATIVE TRAINING OF AUDITORY FILTERS OF DIFFERENT SHAPES FOR ROBUST SPEECH RECOGNITION

Brian Mak, Yik-Cheung Tam†, and Roger Hsiao

Hong Kong University of Science and Technology
Department of Computer Science
Clear Water Bay, Hong Kong

ABSTRACT

The bank-of-filters spectrum analysis model is commonly used in the extraction of acoustic features for automatic speech recognition. The most critical component in the analysis model is a bank of bandpass filters. In this paper, we studied a data-driven approach to designing a bank of "optimal" filters of various shapes discriminatively so that the recognition error of a task is minimized. Three different shapes of varying degree of constraints were investigated: (1) parametric Gaussian filters; (2) non-parametric but constrained triangular-like filters; and (3) non-parametric and unconstrained free-formed filters. Filters were trained to derive the new robust auditory features recently proposed by the Bell Labs. In addition, both the filters (and thus the ensuing acoustic features) and the acoustic model parameters were discriminatively trained. The major result is that our proposed triangular-like filters perform at least as well as the free-formed filters and perform better than the Gaussian filters.

1. INTRODUCTION

One commonly used method of spectral analysis in the extraction of acoustic features for automatic speech recognition (ASR) is the bank-of-filters spectrum analysis model. It is motivated by the human auditory perception process that is believed to be doing spectral analysis through a bank of bandpass auditory filters. According to findings in psychoacoustics by Patterson and Moore *et al.* [1], the shape of auditory filters in the linear frequency scale is symmetric at moderate sound levels and may be approximated by Gaussian filters; and it becomes increasingly asymmetric at high sound levels with the low-frequency side getting shallower and the high-frequency side getting steeper. However, these findings were obtained with simple or mixed tones and the effect was usually measured on a single critical band. It is not clear that the Gaussian approximation to the shape of auditory filters is optimal in the perception of real speech. On the other hand, the computation of mel-frequency cepstral coefficients (MFCC) employs a bank of triangular filters. The triangular filters are a further approximation to the Gaussian approximation of humans' auditory filters, and are adopted for its computation efficiency.

Recently, we have been working on optimizing the parameters involved in the feature extraction of the new robust auditory features developed at the Bell Labs [2, 3]. (Hereafter, we will call it the Bell Labs features.) The Bell Labs feature is derived by mimicking closely the human peripheral auditory system. In particular, the filtering processes in the outer-middle ear and inner ear are explicitly modeled. The new auditory feature was found outperforming MFCC, LPCC, and PLP in noise environments [2, 3]. In our work, both feature extraction parameters and acoustic model parameters are discriminatively trained to minimize the recognition error of a specific task using the MCE/GPD framework [4, 5]. One of our major contributions is that better *discriminative auditory features* (DAF) are obtained through discriminative training of non-parametric auditory filters that are "*triangular-like*".

In this paper, we would like to study the effect of the shape of auditory filters in derving DAF. Auditory filters of three different shapes and varying degree of constraints were investigated:

1. parametric *Gaussian filters*. It is motivated by humans' Gaussian-like auditory filters and it serves as the basis for comparison.
2. non-parametric and weakly-constrained *triangular-like filters*. The triangular-like filters may be considered as a generalization of both triangular filters and Gaussian filters.
3. non-parametric and unconstrained *free-formed filters*. Except that all filter weights must be positive, they may take up any shapes even if they are not supported by any psychoacoustic evidence. This is simply mathematically motivated.

2. AUDITORY FILTERING

The extraction of the Bell Labs feature consists of the following major steps:

- frame blocking with a window of 25ms at every 10ms of speech;
- computing the FFT spectrum (in linear frequency domain);
- filtering by the outer-middle-ear transfer function;
- converting from the linear frequency scale to the Bark scale by linear interpolation to obtain a 128-point Bark spectrum;
- auditory filtering;
- de-correlation by DCT and computation of cepstrum; and
- computation of dynamic features.

There are 32 auditory filters in our system and they are equally spaced at an interval of 4 points apart in the Bark spectrum that covers 0–4kHz. After auditory filtering, the 128-point input Bark spectrum was converted to 32 channel energies from which ceptsra are computed using DCT.

†The co-author is now a graduate student at the Carnegie Mellon University, School of Computer Science, Pittsburgh, PA 15213, USA.

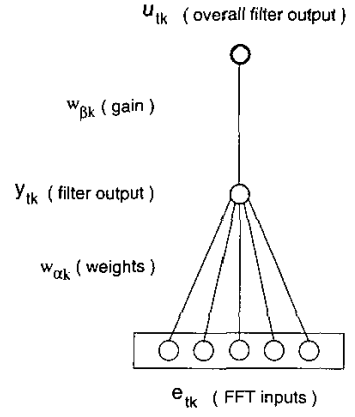

Fig. 1. The auditory filter of the k-th channel

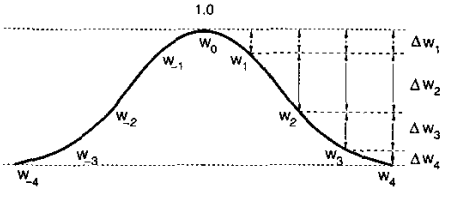

Fig. 2. Representation of a 9-point triangular-like filter

An auditory filter of our system has the general design of Fig. 1, one for each channel. It can be thought of as a two-layer perceptron without any nonlinearity. The weight $w_{\beta k}$ in the second layer perceptron is the gain of the auditory filter while the weights in the first layer are the normalized filter weights. Although the two-layer perceptron is equivalent to a single-layer perceptron, the design allows us to examine the resulting filter shapes and gains separately. All the filter weights are required to be positive.

2.1. Auditory Filters of Various Shapes

2.1.1. Triangular-like Filters

Motivated by findings from psychoacoustics, we propose to approximate humans' auditory filters by "triangular-like filters": all filter weights are positive with a maximum response of 1.0 in the middle, and their values taper off to both ends as depicted in Fig. 2. The triangular-like constraint are implemented by two successive parameter-space transformations. For a digital filter with $(2L+1)$ points, we associate the filter weights $\{ w_{-L}, \ldots, w_{-1}, w_0, w_1, \ldots, w_L \}$ with a set of *deltas*, $\{\delta_{-L}, \ldots, \delta_{-1}, \delta_1, \ldots, \delta_L\}$ so that after the parameter transformation and proper scaling, δ_i will be equivalent to Δw_i (see Fig. 2). Positively-indexed weights are related to the positively-indexed deltas mathematically as follows:

$$w_j = 1 - F(\sum_{i=1}^{j} H(\delta_i)) \quad , \quad j = 1, \ldots, L \tag{1}$$

where $F(.)$ and $H(.)$ are any monotonically increasing functions such that $0.0 \leq F(x) \leq 1.0$ and $0.0 \leq H(x)$. The negatively-indexed weights are similarly related to the negatively-indexed deltas.

In this paper, we used the exponential function as $H(x)$ and the sigmoid function as $F(x)$.

2.1.2. Gaussian Filters

A Gaussian filter with $(2L+1)$ points may be represented as

$$w_{\alpha ki} = \exp\left(-\frac{i^2}{\sigma_k^2}\right) \quad , \quad i = -L, \cdots, 0, \cdots, L\,. \tag{2}$$

Notice that unlike our general triangular-like filters, a Gaussian filter is always symmetric. Furthermore, there is only one parameter, the variance, to estimate.

2.1.3. Free-formed Filters

A free-formed filter is totally unconstrained except that all the filter weights are positive.

3. DISCRIMINATIVE TRAINING OF FILTERS

Although we are concerned only about the filter parameters, it is easier to describe their estimation in the larger context of discriminative training of *any* parameters ϕ that control the feature extraction process. Some of these parameters are illustrated in Fig. 3, and are denoted as follows:

$\boldsymbol{e_t}$	: Bark FFT inputs to auditory filters at time t
$\boldsymbol{u_t}$	: outputs from auditory filters at time t
$\boldsymbol{z_t}$	: channel outputs at time t
$\boldsymbol{x_t}$	: acoustic features at time t
$\boldsymbol{v_t}$	: static acoustic features at time t
$\boldsymbol{v'_t}$	: delta acoustic features at time t
$w_{\beta k}$	: gain of the filter in the k-th channel
$\boldsymbol{w_{\alpha k}}$	: weights of the k-th filter
$\boldsymbol{\delta_k}$	: supplementary deltas associated with $\boldsymbol{w_{\alpha k}}$
y_{tk}	: intermediate output of the k-th filter

As usual, vectors are bold-faced.

The empirical expected string-based misclassification error $\mathcal{L}$, is defined as

$$\mathcal{L}(\Theta) = \frac{1}{N_u}\sum_{u=1}^{N_u} \mathcal{L}_u(\Theta) = \frac{1}{N_u}\sum_{u=1}^{N_u} l(d(X_u)) \tag{3}$$

where Θ consists of any feature extraction parameters and acoustic model parameters; X_u is one of the N_u training utterances; $l(.)$ is the soft error-counting sigmoid function; and $d(X_i) = G_i(X_i) - g_i(X_i)$ measures the ratio between the log-likelihood of the correct string $g_i(X_i)$ and that of its competing hypotheses $G_i(X_i)$. To optimize any parameter ϕ, one finds the derivative of the loss function $\mathcal{L}$ w.r.t. ϕ for each training utterance X_i, which requires the partial derivative of g_i w.r.t. ϕ. If we assume independence between the feature extraction parameters and the model parameters, and the dynamic features $\boldsymbol{v'_t}$ are the linear regression

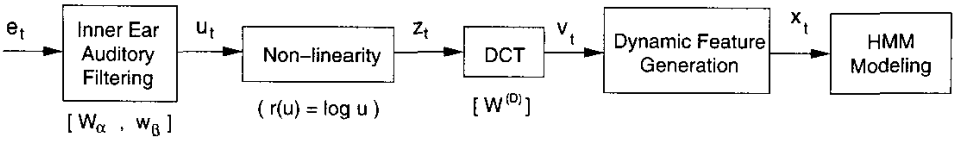

Fig. 3. Parameter notations in the extraction of our discriminative auditory feature

of the static features $\boldsymbol{v}_t$: $\boldsymbol{v}'_t = \sum_{m=-L_1}^{L_1} c'_m \boldsymbol{v}_{t+m}$, then

$$\frac{\partial g}{\partial \phi} = \sum_t \frac{1}{b_{q_t}(\boldsymbol{x}_t)} \sum_{j=1}^{2N} \frac{\partial b_{q_t}}{\partial x_{tj}} \cdot \frac{\partial x_{tj}}{\partial \phi} \quad (4)$$

$$= \sum_t \frac{1}{b_{q_t}(\boldsymbol{x}_t)} \sum_{j=1}^{N} \frac{\partial b_{q_t}}{\partial v_{tj}} \cdot \frac{\partial v_{tj}}{\partial \phi} + \sum_t \frac{1}{b_{q_t}(\boldsymbol{x}_t)} \sum_{j=1}^{N} \frac{\partial b_{q_t}}{\partial v'_{tj}} \left(\sum_{m=-L_1}^{L_1} c'_m \frac{\partial v_{(t+m)j}}{\partial \phi} \right) \quad (5)$$

Thus, all the computations boil down to finding $\frac{\partial v_{tj}}{\partial \phi}$.

3.1. Re-estimation of Filter Gains

By applying the chain rule on variables $\boldsymbol{z}_t$ and $\boldsymbol{u}_t$ (see Fig. 1 and Fig. 3), we have

$$\frac{\partial v_{tj}}{\partial w_{\beta k}} = \frac{\partial v_{tj}}{\partial z_{tk}} \cdot \frac{\partial z_{tk}}{\partial u_{tk}} \cdot \frac{\partial u_{tk}}{\partial w_{\beta k}} = W_{jk}^{(D)} \cdot \frac{1}{u_{tk}} \cdot y_{tk} \quad (6)$$

where $\boldsymbol{W}^{(D)}$ is the DCT matrix and $z_{tk} = \log(u_{tk})$.

3.2. Re-estimation of Filter Weights

3.2.1. Triangular-like Filters

The positively-indexed filter weights of the k-th channel $\boldsymbol{w}_{\alpha k}$ are re-estimated indirectly through the associated deltas $\delta_{kh}, h = 1, \ldots, L$. Using the chain rule, we obtain

$$\frac{\partial v_{tj}}{\partial \delta_{kh}} = \frac{\partial v_{tj}}{\partial z_{tk}} \cdot \frac{\partial z_{tk}}{\partial u_{tk}} \cdot \frac{\partial u_{tk}}{\partial y_{tk}} \cdot \frac{\partial y_{tk}}{\partial \delta_{kh}} \quad (7)$$

$$= W_{jk}^{(D)} \cdot \frac{1}{u_{tk}} \cdot w_{\beta k} \cdot H'(\delta_{kh}) \left[-\sum_{i=h}^{L} F' \cdot e_{tki} \right] \quad (8)$$

The actual filter weights $\boldsymbol{w}_{\alpha k}$ are obtained by the appropriate inverse transformations of δ_{kh}.

A similar formula may be derived for the negatively-indexed deltas.

3.2.2. Gaussian Filters

Let us simplify the notation by representing the Gaussian variance of the k-th (channel) filter σ_k^2 by ρ_k. The derivative of v_{tj} w.r.t. ρ_k may be derived in a similar way as Eqn.(7), and is given by

$$\frac{\partial v_{tj}}{\partial \rho_k} = \frac{\partial v_{tj}}{\partial z_{tk}} \cdot \frac{\partial z_{tk}}{\partial u_{tk}} \cdot \frac{\partial u_{tk}}{\partial y_{tk}} \cdot \frac{\partial y_{tk}}{\partial \rho_k} = W_{jk}^{(D)} \cdot \frac{1}{u_{tk}} \cdot w_{\beta k} \cdot \frac{\partial y_{tk}}{\partial \rho_k} \quad (9)$$

where, since the filter output $y_{tk} = \sum_{i=-L}^{L} e_{tki} \cdot \exp\left(-\frac{i^2}{\rho_k}\right)$,

$$\frac{\partial y_{tk}}{\partial \rho_k} = \frac{1}{\rho_k^2} \sum_{i=-L}^{i=L} e_{tki} \cdot i^2 \cdot \exp\left(-\frac{i^2}{\rho_k}\right). \quad (10)$$

3.2.3. Free-formed Filters

Each filter weight has to be trained in an unconstrained manner by finding the partial derivative of v_{tj} w.r.t. the weight $w_{\alpha ki}$.

$$\frac{\partial v_{tj}}{\partial w_{\alpha ki}} = \frac{\partial v_{tj}}{\partial z_{tk}} \cdot \frac{\partial z_{tk}}{\partial u_{tk}} \cdot \frac{\partial u_{tk}}{\partial y_{tk}} \cdot \frac{\partial y_{tk}}{\partial w_{\alpha ki}} \quad (11)$$

$$= W_{jk}^{(D)} \cdot \frac{1}{u_{tk}} \cdot w_{\beta k} \cdot e_{tki} \quad (12)$$

since $y_{tk} = \boldsymbol{w_{\alpha k}}^T \cdot \boldsymbol{e_{tk}}$.

4. EVALUATION

The effect of the filter shape on the discriminative auditory feature was investigated on Aurora2 [6]. The Aurora2 corpus consists of simulated telephone utterances of digit strings with additive noises at various signal-to-noise ratios. In this paper, only the multi-condition training mode was investigated and results were reported by combining the performance on all of its three test sets (A, B, and C) according to Aurora's evaluation standard.

4.1. Experimental Setup

The Bell Labs features were extracted from speech utterances every 10ms as described in [2] except that the auditory filters were replaced by our triangular-like filters, Gaussian filters, and free-formed filters. Each feature vector consisted of 13 MFCCs including c0, and their first- and second-order derivatives.

Regardless of their shapes, all our auditory filters had 11 weights, and each channel had its own filter. Triangular-like filters were generally asymmetric, whereas Gaussian filters are, by definition, symmetric. No restrictions were placed on free-formed filters, except that like all other filters, the filter weights must be positive.

Following the baseline setup in the ICSLP conference in 2002, each digit was represented by a context-independent whole-word hidden Markov models (HMM) and was trained using the EM algorithm to produce its initial ML estimates (MLE). Each model was a straightly left-to-right HMM with 16 states and 3 Gaussian components per state. The silence model had only 3 states, each with 6 mixture components. There was also a 1-state short-pause model tied to the middle state of the silence model. The HTK toolkit was used for both training the MLE models as well as for decoding. From the initial MLE models and auditory feature parameters, discriminative training was performed to obtain MCE estimates of the HMM parameters and/or MCE estimates of the filter parameters. Corrective training was employed using the 1-nearest competing hypotheses [7].

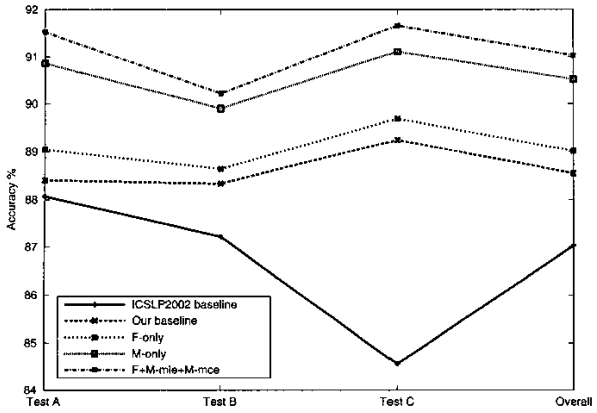

Fig. 4. Aurora2 performance of various training schemes using triangular-like filters (Reference: ICSLP2002 baseline with traditional MFCCs is 87.03%; our baseline using Bell Labs auditory features without discriminative training is 88.54%.)

4.2. Results and Discussion

Various ways to combine the discriminative training (DT) of the filter parameters with that of the HMM parameters were tried and we found that they should be done sequentially. Joint optimization did not work and its result was just equal to that due to discriminative training of the model parameters alone. This agrees with the experience from other researchers [8]. Thus, the following training schemes were finally investigated:

- *M-only*: discriminative training of HMM parameters *only*;
- *F-only*: discriminative training of filter parameters *only*;
- *F + M-mce*: discriminative training of filter parameters followed by an MCE re-estimation of the models using the new features (obtained with the new filter parameters);
- *F + M-mle*: discriminative training of filter parameters followed by an ML re-estimation of the models using the new features;
- *F + M-mle + M-mce*: same as the last one but followed by a subsequent discriminative training of HMM parameters.

4.2.1. *Effect of Various Training Schemes*

Fig. 4 shows the results of various training schemes of the two kinds of parameters using triangular-like filters. Similar trends are observed for Gaussian filters or free-formed filters. It is clear that the Bell Labs feature performs better than MFCC on the noise digits and reduces the word error rate (WER) by 11.6%. DT of model parameters alone (M-only scheme) is very effective and reduces the WER of our baseline by another 17.3%. DT of filter parameters alone (F-only scheme) is less effective and reduces the WER by only 4.1%. However, if it is followed by DT of the model parameters (F+M-mce), then the result is slightly better than that of M-only training. The biggest gain was obtained by first re-training the HMMs using the new features derived from the new filters followed by another round of DT of the model parameters. The final WER reduction is 21.7% over our baseline.

4.2.2. *Effect of Filter Shapes*

Table 1. Overall performance (in word accuracy in %) of DAF on Aurora2 using filters of different shapes.

Filters	M-only	F-only	F + M-mle	F + M-mce	F + M-mle + M-mce
Gaussian	90.19	88.72	88.75	90.16	90.33
free-formed	90.52	89.01	89.22	90.72	91.01
triangular-like	90.52	89.01	89.09	90.83	91.03

The results of discriminative auditory features (DAF) derived using filters of various shapes are summarized in Table 1. It can be seen that the Gaussian constraint limits the performance of DAF in all training schemes, whereas the free-formed filters have almost the same performance as our proposed triangular filters.

5. CONCLUSIONS

It is believed that humans' auditory filters may be approximated well by Gaussian filters, and the computation of MFCCs uses triangular filters. However, our study shows that the proposed triangular-like filters are more general than either type of filters, and in the Aurora2 task, they give better word recognition accuracy. There are no psychoacoustic grounds for the use of free-formed filters; but that they perform as well as our triangular-like filters shows that the triangular-like filters may be "optimal". We also believe that since our triangular-like filter is closer to humans' filter shape, it may be more robust to varying environments.

6. ACKNOWLEDGEMENTS

This work is supported by the Hong Kong Research Grants Council under the grant number HKUST6201/02E.

7. REFERENCES

[1] Brian C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 4th edition, 1997.

[2] Qi Li, Frank Soong, and Olivier Siohan, "A High-Performance Auditory Feature for Robust Speech Recognition," in *ICSLP*, 2000.

[3] Qi Li, Frank Soong, and Olivier Siohan, "An Auditory System-based Feature for Robust Speech Recognition," in *Eurospeech*, 2001.

[4] B. Mak, Y. C. Tam, and Q. Li, "Discriminative Auditory Features for Robust Speech Recognition," in *ICASSP*, Orlando, Florida, USA, 2002, vol. 1, pp. 381–384.

[5] B. Mak and Y. C. Tam, "Performance of Discriminatively Trained Auditory Features on Aurora2 and Aurora3," in *ICSLP*, Denver, Colorado, USA, Sept. 2002, vol. 1, pp. 33–36.

[6] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, September 2000.

[7] Y. C. Tam and B. Mak, "An Alternative Approach of Finding Competing Hypotheses for Better Minimum Classification Error Training," in *ICASSP*, Orlando, Florida, USA, 2002, vol. 1, pp. 101–104.

[8] A. Torre, A. M. Peinado, A. J. Rubio, J. C. Segura, and C. Benitez, "Discriminative Feature Weighting for HMM-based continuous Speech Recognizers," *Speech Communication*, vol. 38, no. 3–4, pp. 267–286, Nov. 2002.