DISCRIMINATIVE AUDITORY FEATURES FOR ROBUST SPEECH RECOGNITION

Brian Mak, Yik-Cheung Tam

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

Qi Li

Bell Labs, Lucent Technologies
600 Mountain Avenue, Murray Hill
NJ 07974, USA

ABSTRACT

Recently, Li *et al.* proposed a new auditory feature for robust speech recognition in noise environments. The new feature was derived by mimicking closely the function of human auditory process. Several filters were used to model the outer ear, middle ear, and cochlea, and the initial filter parameters and shapes were obtained from crude psychoacoustics results, experience, or experiments. Although one may adjust the feature parameters by hand to get better performance, the resulting feature parameters still may not be optimal in the sense of minimal recognition errors, especially for different tasks. To further improve the auditory feature, in this paper we apply discriminative training to optimize the auditory feature parameters with some guidance from psychoacoustic evidence but otherwise in a data-driven approach so as to minimize the recognition errors. One significant contribution over similar efforts in the past, such as discriminative feature extraction, is that we make no assumption on the parametric form of the auditory filters. Instead, we only require the filters to be smooth and triangular-like as suggested by psychoacoustics research. Our approach is evaluated on the Aurora database and achieves a word error reduction of 19.2%.

1. INTRODUCTION

In automatic speech recognition (ASR), the design of acoustic models involves two main tasks: feature extraction and data modeling. Acoustic features such as LPCC, MFCC, PLP are commonly used; and the most popular data modeling techniques in current ASR are based on hidden Markov modeling (HMM). Recently, Li *et al.* proposed a new auditory feature for robust speech recognition based on an analysis of the human peripheral auditory system [1]. In the approach, the auditory system is first divided into several modules, then each module is modeled from a signal processing point of view with a constraint on computational complexity. The feature computation is comprised of an outer-middle-ear transfer function, FFT, conversion from linear frequency scale to the Bark scale, auditory filtering, non-linearity, and discrete cosine transform (DCT). As reported in [1], the new auditory feature outperformed MFCC, LPCC, and PLP, in noise environments, and the major improvement was attributed to the new auditory filters. Although in the new auditory feature platform, the filter shapes and other parameters can be adjusted easily through experiments, the filters still may not be optimal in the sense of minimal recognition errors, especially under the context of different tasks.

Traditionally, in ASR, feature extraction and acoustic modeling are addressed separately, which may not result in an optimal recognition performance. Several approaches have been proposed to optimize feature parameters using *discriminative feature extraction* (DFE) along with the optimization of model parameters under the unified framework of MCE/GPD (minimum classification error and generalized probabilistic descent) [2, 3]. The past efforts on DFE may be divided into two major categories:

(1) Most DFE-related works were based on common features such as log power spectra [4], mel-filterbank log power spectra [5], and LPCC [6] and discriminatively trained a transformation network to obtain new discriminative features for the following data modeling process. Notice that these work did not touch the front-end signal processing module that derives inputs to their transformation networks.

(2) In contrast, Alain Biem *et al.* [7] applied joint discriminative training on both HMM parameters and filters in the front-end. Two kinds of filter parameterization were tried: Gaussian filters or free-formed filters. The tasks were relatively simple to today's standard, and the improvement was small. Furthermore, the free-formed filters performed worse than Gaussian filters.

In this paper, we attempt to design the auditory filters involved in the extraction of our new auditory features without making an assumption on the parametric form of the auditory filters. Instead, guided by psychoacoustic evidences, we only require the filters to be smooth and triangular-like. One of the challenges is to derive a mathematical expression for a filter satisfying the two constraints. We achieve this through two parameter space transformations.

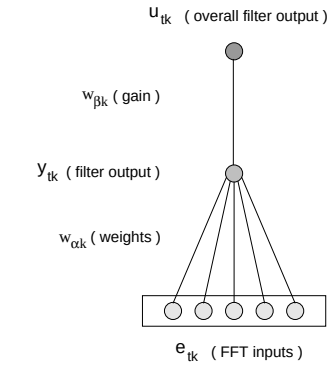
2. AUDITORY FILTER DESIGN

We postulate that the use of Gaussian auditory filters in [7] may be too restrictive; however, the suggestion of *absolutely* free-formed filters in [7] is not supported by psychoacoustic findings either. We believe that the shape of human auditory filters is not arbitrary and their properties should be observed in our discriminative auditory filter design. Based on the findings from psychoacoustics, we require our auditory filter response to satisfy the following two constraints:

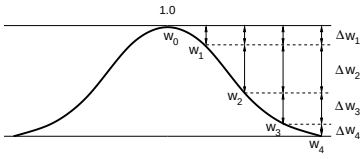
Constraint #1 : it is triangular-like. That is, all its weights must be positive with a maximum response of 1.0 somewhere in the middle, and then its values taper off to both ends; and,

Constraint #2 : it is differentiable.

In our feature extraction, a 128-point Bark spectrum from FFT and the outer-middle-ear transfer function was fed to 32 auditory filters as in the cochlea. The filters were equally spaced at an interval of 4 points apart in the spectrum. Thus, after auditory filtering,



(a) An auditory filter



(b) Mathematical representation

Fig. 1. A constrained auditory filter of the k-th channel

the 128-point input spectrum was converted to 32 channel energies from which ceptsra were computed using DCT. An auditory filter of our system has the design as depicted in Fig. 1(a), 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.

A filter satisfying the two aforementioned constraints can be implemented through 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 parameter transformation and proper scaling, δ_i will be equivalent to Δw_i (see Fig. 1(b)). 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

$$0.0 \le F(x) \le 1.0 \quad \text{and} \quad 0.0 \le H(x)\,. \tag{2}$$

Similarly, negatively-indexed weights are related to the negatively-indexed deltas. The motivation is that we want to subtract more positive quantities from the maximum weight $w_0 = 1$ as we move towards the two ends of the filter. Eqn.(1) involves two transformations: $H(.)$ is any monotonically increasing function which turns arbitrarily-valued deltas to positive quantities; and, $F(.)$ is any monotonically increasing function that restricts the sum of transformed deltas to less than unity. In this paper, we use the exponential function as $H(x)$ and the sigmoid function as $F(x)$.

3. DISCRIMINATIVE AUDITORY FEATURE (DAF)

In our acoustic modeling, there are two types of free parameters $\Theta = (\Lambda, \Phi)$: the HMM parameters Λ and the parameters Φ that control feature extraction (FE). The former include state transition probabilities and observation probability distribution functions; and, the latter consist of inner-ear auditory filters in our filter-bank-based feature extraction. All these parameters were trained in the discriminative framework of MCE/GPD.

3.1. Re-estimation formulas

Various feature extraction parameters are denoted as follows:

$\boldsymbol{e_t}$	: 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

These parameter notations are also illustrated in Fig. 2. 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 X_u is one of the N_u training utterances; $d(.)$ is a distance measure for misclassifications; and, $l(.)$ is a soft error-counting function. We followed the common practice of using the sigmoid function for counting soft errors and using log-likelihood ratio between the correct string and its competing hypotheses as the distance function. i.e. $d(X_i) = G_i(X_i) - g_i(X_i)$ in which the discriminant function $g(.)$ is the log-likelihood of a decoding hypothesis of an utterance and $G_i(X_i)$ is the log of the mean probabilities of its N_c competing strings that is defined as:

$$G_i(X_i) = \log\left[\frac{1}{N_c}\sum_{j=1;j\neq i}^{N_c} \exp(\eta g_j(X_i))\right]^{1/\eta} . \tag{4}$$

To optimize any parameter $\theta \in \Theta$, one finds the derivative of the loss function $\mathcal{L}$ w.r.t. θ for each training utterance X_i:

$$\frac{\partial \mathcal{L}(X_i)}{\partial \theta} = \frac{\partial l}{\partial d}\left[\frac{\sum_{j\neq i}^{N_c} \exp(\eta g_j(X_i))\left(\frac{\partial g_j}{\partial \theta} - \frac{\partial g_i}{\partial \theta}\right)}{\sum_{j\neq i}^{N_c} \exp(\eta g_j(X_i))}\right] . \tag{5}$$

To evaluate Eqn.(5), one has to find the partial derivative of g_i w.r.t. any trainable parameters. We will drop the utterance index i for clarity from now on. Also, since many works have been done on discriminative training of HMM parameters with MCE/GPD, one may refer the tutorial paper [3] for the re-estimation formulas of HMM parameters and we will only present those of feature

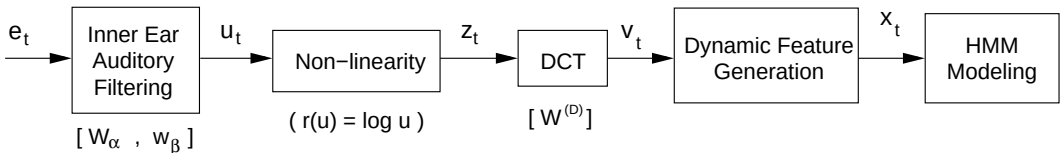


Fig. 2. Parameter notations in the extraction of our discriminative auditory feature

extraction parameters. Firstly, we assumed that the trainable FE parameters Φ are independent of HMM parameters Λ. Second, it is helpful to see that the log-likelihood of an utterance is related to an FE parameter $\phi \in \Phi$ through the static features, and the dynamic features are related to ϕ also through the static features. Let's assume that the final feature vector $\boldsymbol{x_t}$ at time t consists of N static features $\boldsymbol{v_t}$ and N dynamic features $\boldsymbol{v'_t}$ which are computed from $\boldsymbol{v_t}$ by the following regression formula

$$\boldsymbol{v'_t} = \sum_{m=-L_1}^{L_1} c'_m \boldsymbol{v_{t+m}} \cdot \tag{6}$$

Hence, the derivative of an utterance hypothesis log-likelihood g w.r.t an FE parameter ϕ is given by

$$\frac{\partial g}{\partial \phi} = \sum_t \frac{1}{b_{q_t}(\boldsymbol{x_t})} \sum_{j=1}^{N} \frac{\partial b_{q_t}}{\partial v_{tj}} \cdot \frac{\partial v_{tj}}{\partial \phi} + \sum_t \frac{1}{b_{q_t}(\boldsymbol{x_t})} \sum_{j=1}^{N} \frac{\partial b_{q_t}}{\partial v'_{tj}} \left(\sum_{m=-L_1}^{L_1} c'_m \frac{\partial v_{(t+m)j}}{\partial \phi} \right) \tag{7}$$

The computation of $\frac{\partial v_{tj}}{\partial \phi}$ depends on the nature of each trainable parameter ϕ and will be described below separately.

3.2. Re-estimation of Filter Gains

Gain of the k-the channel filter is represented by the weight $w_{\beta k}$ in the second layer of the filter shown in Fig. 1(a). Positivity of the gains are ensured by the transformation: $w_{\beta k} = \exp(\tilde{w}_{\beta k})$. Since the static feature $\boldsymbol{v_t}$ is related to the non-linearity function output $\boldsymbol{z_t}$ which in turn is related to the filter output $\boldsymbol{u_t}$, by applying the chain rule (see Fig. 1(a) and Fig. 2), one may obtain the derivative of each static feature v_{tj} w.r.t. $\tilde{w}_{\beta k}$ as follows:

$$\begin{aligned} \frac{\partial v_{tj}}{\partial \tilde{w}_{\beta k}} &= \frac{\partial v_{tj}}{\partial z_{tk}} \cdot \frac{\partial z_{tk}}{\partial u_{tk}} \cdot \frac{\partial u_{tk}}{\partial w_{\beta k}} \cdot \frac{\partial w_{\beta k}}{\partial \tilde{w}_{\beta k}} \\ &= W^{(D)}_{jk} \cdot \frac{1}{u_{tk}} \cdot y_{tk} \cdot e^{\tilde{w}_{\beta k}} \end{aligned} \tag{8}$$

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

3.3. Re-estimation of Filter Weights

Filter weights of the k-th channel $\boldsymbol{w_{\alpha k}}$ are re-estimated indirectly through the associated deltas. Again using the chain rule, the derivative of the j-th static feature w.r.t. the h-th positively-indexed delta in the filter of the k-th channel is given by,

$$\begin{aligned} \frac{\partial v_{tj}}{\partial \delta_{kh}} &= \frac{\partial v_{tj}}{\partial z_{tk}} \cdot \frac{\partial z_{tk}}{\partial u_{tk}} \cdot \frac{\partial u_{tk}}{\partial y_{tk}} \cdot \frac{\partial y_{tk}}{\partial \delta_{kh}} \\ &= W^{(D)}_{jk} \cdot \frac{1}{u_{tk}} \cdot w_{\beta k} \cdot H'(\delta_{kh}) \left[-\sum_{i=h}^{L} F' \cdot e_{tki} \right] \end{aligned} \tag{9}$$

3.3.1. Updates

Finally, a (locally) optimal model or feature extraction parameter $\theta \in \Theta$ may be found by the iterative procedure of GPD using the following update rule:

$$\theta(t+1) = \theta(t) - \epsilon(t) \cdot \left. \frac{\partial \mathcal{L}}{\partial \theta} \right|_{\theta=\theta_{(t)}} . \tag{10}$$

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

4. EVALUATION

The proposed discriminative auditory feature was evaluated on the Aurora task. Only the multi-condition training mode was investigated and results were reported by combining its performance on all three test sets according to Aurora's testing standard.

4.1. The Aurora Corpus

The Aurora corpus [8] was created for research in distributed speech recognition under noisy environments. Connected digits from the clean TIDIGITS database [9] were pre-filtered according to the frequency characteristics of common telecommunication channels (G.712 or MIRS) and realistic noises were then artificially added at six different signal-to-noise (SNR) ratios ranging from 20dB to -5dB at 5dB steps. Two training modes: clean training and multi-condition training, and three test sets were also defined to evaluate recognition technologies under matched and unmatched noises, and matched and unmatched channel characteristics.

4.2. Experimental Setup

Auditory features were extracted from speech utterances every 10ms as described in [1] except that the auditory filters were replaced by those depicted in Section 2. Each feature vector consisted of 13 MFCCs including c0, and their first- and second-order derivatives computed by regression.

Each auditory filter had 11 weights and the middle (6-th) weight was assumed maximum with the value of 1.0. However, each channel had its own filter and the filters were not assumed symmetrical.

Context-dependent head-body-tail (HBT) digit models [10] were trained using maximum likelihood estimation to produce the initial "ML estimates" (MLE) of the models. Each model was a straightly left-to-right HMM with no skips. Each head and tail HMM had three states and each body HMM had four states, all with four mixtures per state. There was also a 1-state silence model with 8 mixtures. 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 and competing hypotheses were obtained from 4-best decoding. As the HMM

and FE parameters were assumed independent, different learning rates were used to account for their different dynamic ranges. The following learning rates were found empirically to give good results: 1.0 for FE parameters and 442 for HMM parameters. As required by the GPD algorithm, these learning rates R decreased with iteration t as: $R(t) = R(0) \cdot (1 - t/50)$; and, we limited our maximum number of iterations to 50.

4.3. Results and Discussion

Since there are two kinds of trainable parameters: HMM or FE parameters, we combined their training in various ways as follows:

- "Our Baseline": ML estimation of the HBT digit models using the original auditory feature [1].
- "M only": discriminative training of HMM parameters *only*;
- "F + M-mle": discriminative training of FE parameters followed by an ML re-estimation of the models under the new feature space.
- "F + M-mle + M-mce": same as the last one but followed by a subsequent discriminative training of HMM parameters.

Discriminative training of FE parameters alone was not found helpful if without subsequent re-training of the models using new features generated by the new FE parameters. It seems to indicate that HMM parameters should "move" with the new feature space in order to make good use of the new features. The training mode "F + M-mle" tries to remedy the situation in two separate steps: first the FE parameters were discriminatively trained then new HMMs were re-estimated using the new features.

Table 1. Results on Aurora test sets A+B+C (Baseline results are in word accuracy %, and the rest are absolute % gains from "Our Baseline".)

dB	Aurora Baseline	Our Baseline	F + M-mle	M only	F + M-mle + M-mce	$(\text{F + M-mle})^2$ + M-mce
clean	98.52	99.13	-0.13	0.03	-0.47	0.17
20	97.35	98.66	-0.01	-0.08	0.16	0.39
15	96.29	97.81	0.01	0.26	0.12	0.55
10	93.78	95.55	0.17	0.82	0.81	1.16
5	85.51	88.86	0.76	1.81	1.93	2.38
0	58.99	68.65	2.38	3.53	4.64	5.24
-5	24.49	33.21	2.99	4.21	7.21	6.75
Ave	79.28	83.12	0.88	1.51	2.01	2.38
A-Ave	86.38 (86.38)	89.91 (89.91)	0.66 (90.57)	1.27 (91.18)	1.47 (91.38)	1.94 (91.85)

4.3.1. Discriminative Training of HMM and/or FE Parameters

Recognition performance of the various training modes is shown in columns 4–6 in Tables 1 together with the official Aurora baseline results given in [8]. Two different averages are reported: "Ave" represents the mean performance over all 7 SNRs, while "A.Ave" ignores clean speech and speech at -5dB in conformity to Aurora's evaluation metric. Furthermore, recognition results from discriminative auditory feature (DAF) estimation are reported in terms of their accuracy gains from *our baseline* results. The results show that our discriminative training all gave significant improvements in word accuracies, and the improvement was greater for noisier data. One obvious reason is that many training samples came from the noisier data as the recognizer made more errors with them and the model parameters were adjusted to fix those errors. It also shows that discriminative training of HMM parameters alone is more effective than DAF estimation alone. Nevertheless, DAF estimation followed by MCE training of the model parameters gave the best performance. Compared with our baseline word error rate (WER) using the Aurora averages, DAF estimation alone reduced WER by 6.54% (relative); and a further reduction of 8.59% (relative) was obtained if the resulting MLE models were subsequently re-trained using the MCE/GPD algorithm. That is, altogether for a WER reduction of 14.6%. On the other hand, if only HMM parameters were discriminatively trained, WER was reduced by 12.6%.

4.3.2. Number of DAF Iterations

We also explored the effect of more iterations for DAF estimation. Empirically we found that two iterations were enough and the results are shown in the rightmost column in Table 1. Compared with the Aurora baseline, our baseline improved WER by 25.9%; one iteration of DAF estimation followed by MCE model training improved WER by 36.7%; and, two iterations of DAF estimation followed by MCE model training improved WER by 40.2%.

5. FUTURE WORKS

In the future, we would like to apply discriminative training on other FE parameters such as the non-linearity function parameters. In addition, we would also like to remove the independence assumption between HMM and FE parameters in discriminative auditory feature estimation.

6. REFERENCES

[1] Q. Li, F. Soong, and O. Siohan, "An Auditory System-based Feature for Robust Speech Recognition," in *Proc. of Eurospeech*, 2001, vol. 1, pp. 619–622.

[2] B.H. Juang and S. Katagiri, "Discriminative Training for Minimum Error Classification," *IEEE Trans. on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, Dec 1992.

[3] W. Chou, "Discriminant-Function-Based Minimum Recognition Error Rate Pattern-Recognition Approach to Speech Recognition," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1201–1223, August 2000.

[4] J.S. Bridle and L. Doddi, "An Alphanet Approach to Optimising Input Transformations for Continuous Speech Recognition," in *Proc. of ICASSP*, 1991, vol. 1.

[5] R. Chengalvarayan and Li Deng, "HMM-Based Speech Recognition using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 243–256, May 1997.

[6] M. Rahim and C.-H. Lee, "Simultaneous ANN Feature and HMM Recognizer Design Using String-based Minimum Classification Error (MCE) Training," in *Proc. of ICSLP*, 1996.

[7] A. Biem, S. Katagiri, E. McDermott, and B.-H. Juang, "An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 96–110, Feb 2001.

[8] 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"*, Sept 2000.

[9] R.G. Leonard, "A Database for Speaker-Independent Digit Recognition," in *Proc. of ICASSP*, 1984.

[10] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum Error Rate Training of Inter-word Context Context-Dependent Acoustic Model Units in Speech Recognition," in *Proc. of ICSLP*, 1994, pp. 439–442.