PERFORMANCE OF DISCRIMINATIVELY TRAINED AUDITORY FEATURES ON AURORA2 AND AURORA3

Brian Mak and Yik-Cheung Tam

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

ABSTRACT

The design of acoustic models involves two main tasks: feature extraction and data modeling; and hidden Markov modeling (HMM) is commonly used in contemporary automatic speech recognition. In the past, discriminative training has been applied successfully to refine HMM parameters that are initially trained by EM algorithm. Recently, we applied discriminative training in the feature extraction process. We proposed a novel Discriminative Auditory Feature extraction method (DAF) in which filters are discriminatively trained from data. In DAF, we do not make any assumptions on the functional form of the auditory filters except that they have to be smooth and triangular-like. On the method of discriminative training, we also proposed an alternative approach to finding the competing hypotheses which we call N-nearest hypotheses (as opposed to the traditional N-best hypotheses). By applying the two new ideas and the new robust auditory features proposed by Li *et al.* of Bell Labs, we reduce the overall word error rate (WER) by 30.27% over ICSLP2002 Aurora2 baseline on multi-condition training. Similarly, we obtain a relative WER reduction of 38.42% over ICSLP2002 Aurora3 baseline.

1. INTRODUCTION

Discriminative training [1] has been widely used in automatic speech recognition (ASR) in the area of estimating HMM parameters [1], speaker verification [2], estimation of multi-band weightings [3], etc. Recently, we made two contributions to the area of discriminative training:

1. Discriminative Auditory Feature (DAF) [4]: discriminative training is applied to estimate the auditory filters during acoustic feature extraction. Most previous works of applying discriminative training on feature extraction involved feature transformation [5, 6, 7]. The most related work was done by Alain Biem *et al.* [8] who assumed the filters to be either Gaussians or totally free-formed, but the results were unsatisfactory. We postulate that the shape of human auditory filters is not arbitrary and their properties should be observed in auditory filter design. Guided by psychoacoustics evidence, we require the filters to be smooth and triangular-like. We derive a mathematical expression for such a filter through two parameter-space transformations.
2. N-nearest Hypotheses [9]: During MCE discriminative training, competing hypotheses against the correct one are commonly derived from the N-best algorithm. One problem with the N-best algorithm is that, in practice, some misclassified data can have very large misclassification distances from the N-best competitors and fall out of the steep trainable region of the sigmoid function, and thus cannot be utilized effectively. Recently, we proposed using the N nearest competing hypotheses instead to increase the amount of "effective" training data.

In this paper, we apply both techniques and derive discriminative auditory features (DAF) based on the robust auditory features recently proposed by Li *et al.* of Bell Labs [10]. HMM trained on the derived DAF are tested on both Aurora2 (multi-condition training) and Aurora3 corpora. In both corpora, we find that DAF gives additional improvement over Li's original auditory features.

2. DISCRIMINATIVE AUDITORY FEATURE

Our discriminative auditory feature extraction is a data-driven method to estimate auditory filters. It can be applied to any filters involved in any feature extraction process. For instance, one may replace the conventional triangular filters in the extraction of MFCCs using DAF. However, we will describe our DAF in the context of extracting Li's auditory features as the ensuing DAF were used for Aurora evaluation in this paper.

2.1. Auditory Filter Design

Li's new robust auditory features are designed by closely mimicking the feature extraction process of the human peripheral auditory system [10]. The new auditory feature was found to outperform MFCC, LPCC, and PLP in noisy environments, and the major improvement was attributed to the new auditory filters. Currently, Li obtains these filters from psychoacoustics literature and all channels of the cochlea share one filter.

In our DAF extraction, a 128-point Bark spectrum output from the outer and middle ear is fed to 32 auditory filters in the cochlea. The filters are equally spaced at an interval of 4 points apart in the spectrum. An auditory filter of our system has the design as depicted in Fig. 1(a). The weight $w_{\beta k}$ represents the gain of the auditory filter while the weights $\boldsymbol{w_{\alpha k}}$ are the normalized filter weights. To ensure the filter being triangular-like and differentiable, two successive parameter-space transformations are employed. For a digital filter with $(2L+1)$ points, we associate the weights $\{ w_{-L}, \ldots, w_{-1}, 1.0, w_1, \ldots, w_L \}$ 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 as follows:

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

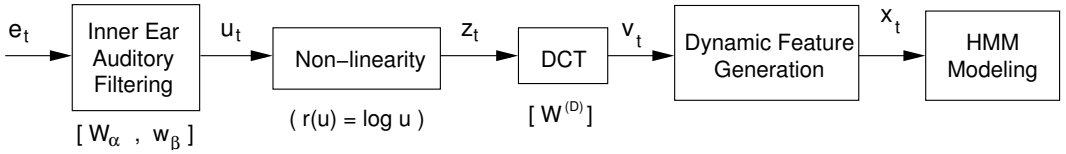
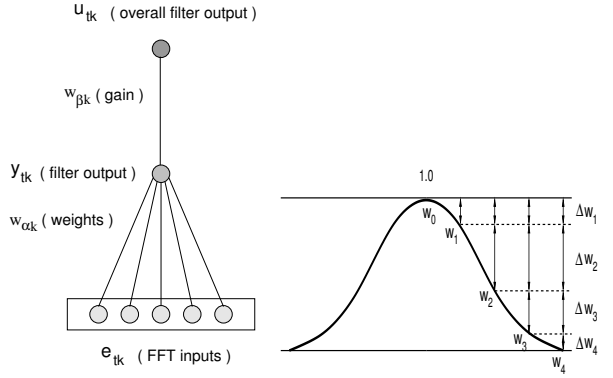


Fig. 2. Parameter notations in the extraction of our discriminative auditory feature



(a) An auditory filter (b) Math representation

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

where $F(\cdot)$ and $H(\cdot)$ are any monotonically increasing functions such that $0.0 \leq F(x) \leq 1.0$ and $0.0 \leq H(x)$. Negatively-indexed weights are related to the negatively-indexed deltas similarly. In Eqn.(1), $H(\cdot)$ turns arbitrarily-valued deltas to positive quantities; and, $F(\cdot)$ 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 $F(x)$.

2.2. Discriminative Training of Auditory Filters

In 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). All these parameters can be trained in the discriminative framework of MCE/GPD. Here we will discuss only the re-estimation of feature extraction parameters and assume that they are independent of HMM parameters Λ.

The various feature extraction parameters are denoted as shown in Fig. 1 and Fig. 2. The empirical expected string-based misclassification error $\mathcal{L}$ is defined as

$$\mathcal{L}(\Theta) = \frac{1}{N_u}\sum_{u=1}^{N_u}\mathcal{L}_u(\Theta) = \frac{1}{N_u}\sum_{u=1}^{N_u} l(d(X_u)) \qquad (2)$$

where X_u is one of the N_u training utterances, and $l(\cdot)$ is the soft-error-counting sigmoid function. $d(\cdot)$ is defined as $d(X_i) = G_i(X_i) - g_i(X_i)$, measuring the log-likelihood ratio between the correct string g_i and its competing hypotheses G_i. To optimize any parameter $\phi \in \Phi$, one finds the derivative of the loss function $\mathcal{L}$ w.r.t. ϕ for each training utterance X_i, which, in turn, requires the partial derivative of g_i w.r.t. ϕ as given by:

$$\frac{\partial g}{\partial \phi} = \sum_t \frac{1}{b_{q_t}(\boldsymbol{x_t})}\sum_{j=1}^{N}\frac{\partial b_{q_t}}{\partial x_{tj}}\cdot\frac{\partial x_{tj}}{\partial \phi} \qquad (3)$$

2.3. Re-estimation of Filter Gains

Positivity of the gains of the k-th channel filter is ensured by the transformation: $w_{\beta k} = \exp(\tilde{w}_{\beta k})$. By applying the chain rule on variables $\boldsymbol{z_t}$ and $\boldsymbol{u_t}$(see Fig. 1(a) and Fig. 2), we have

$$\begin{aligned}\frac{\partial x_{tj}}{\partial \tilde{w}_{\beta k}} &= \frac{\partial x_{tj}}{\partial z_{tk}}\cdot\frac{\partial z_{tk}}{\partial u_{tk}}\cdot\frac{\partial u_{tk}}{\partial w_{\beta k}}\cdot\frac{\partial w_{\beta k}}{\partial \tilde{w}_{\beta k}} \\ &= W^{(D)}_{jk}\cdot\frac{1}{u_{tk}}\cdot y_{tk}\cdot e^{\tilde{w}_{\beta k}} \end{aligned} \qquad (4)$$

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

2.4. Re-estimation of Filter Weights

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

$$\begin{aligned}\frac{\partial x_{tj}}{\partial \delta_{kh}} &= \frac{\partial x_{tj}}{\partial z_{tk}}\cdot\frac{\partial z_{tk}}{\partial u_{tk}}\cdot\frac{\partial u_{tk}}{\partial y_{tk}}\cdot\frac{\partial y_{tk}}{\partial \delta_{kh}} \\ &= W^{(D)}_{jk}\cdot\frac{1}{u_{tk}}\cdot w_{\beta k}\cdot H'(\delta_{kh})\left[-\sum_{i=h}^{L}F'\cdot e_{tki}\right] \end{aligned} \qquad (5)$$

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

3. N-NEAREST HYPOTHESES

In current MCE training, when the correct hypothesis is too far from the N-best hypotheses, it will fall into the un-trainable region of the sigmoid function. Although one may alleviate the problem by using a flatter sigmoid, training becomes so slow that one may have to increase the learning rate. However, if the learning rate is too large, overshoot may occur. In [9], we suggested using the N nearest competing hypotheses instead so as to keep the training data as close to the trainable region as possible. Consequently, the amount of "effective" training data is increased, and since there is no need to use a flatter sigmoid and a large learning rate, the training seems to be more stable. An approximation algorithm [9] based on beam search was also designed to locate the nearest competing hypothesis efficiently.

4. AURORA EVALUATION

Following the baseline setup as suggested by ICSLP2002's special session on "Speech Recognition in Noise", each of the 11 digits were modeled as strictly left-right whole-word HMMs, each with 16 states and each state consisted of 3 Gaussian mixtures. The silence model had only 3 states, each with 6 mixture components. The 1-state short pause model was tied to the second state of the

Table 1. Aurora2 multi-condition: WER on Test A using MFCC with different sigmoid slopes and different competing hypotheses

System	γ=0.5	γ=0.1	γ=0.02	γ=0.004
baseline(ML)	12.71%	12.71%	12.71%	12.71%
1-best	11.55%	**11.01%**	11.08%	12.07%
approximate 1-nearest	10.85%	**10.71%**	11.27%	12.27%
exact 1-nearest	10.46%	**10.45%**	10.92%	12.16%

Table 2. Aurora3: Word accuracies using 1-best and approximate 1-nearest competitors on the German corpus

Condition	MCE		DAF	
	1-best	1-nearest	1-best	1-nearest
WM	93.89%	94.43%	93.77%	93.97%
MM	85.29%	85.72%	84.99%	85.36%
HM	87.84%	87.97%	87.65%	87.79%
Overall	89.37%	89.77%	89.17%	89.41%

silence model. The HTK toolkit was used for both training the initial ML models and subsequent decoding. However, Li's auditory features were used.

Since it was found that for Aurora3 database, endpointing will greatly improve recognition performance, we first endpointed all training and testing files by the following procedure: The un-endpointed training data were first used to train a set of HMMs which was employed to segment all speech data by forced alignment; and, the final endpoints were obtained by adding 100ms of speech at each end. The endpointed training data were then used to re-train the final set of HMMs, which were then tested on the endpointed test data. The following systems were compared:

- **Auditory**: MLE-trained HMMs using Li's auditory features
- **MCE**: the Auditory system with additional MCE training
- **DAF**: the Auditory system with additional DAF extraction
- **DAF+MLE**: the DAF system followed by HMM re-training
- **DAF+MLE+MCE**: the DAF+MLE system with additional MCE training
- **DAF** [**+MLE**]**+MCE**: similar to DAF+MLE+MCE but MLE re-training is optional

4.1. Effectiveness of the N-nearest Hypotheses

Table 1 and Table 2 demonstrates the effectiveness of discriminative training using our new 1-nearest hypotheses.

4.2. Evaluation of DAF on Aurora2

For Aurora2 evaluation, speech files were not endpointed and only the multi-condition training mode was attempted. Summary of the results is shown in Table 3 and the details are shown in Table 7.

Table 3. Aurora2 multi-condition: Evaluation results

Condition	Test A	Test B	Test C	Overall
ICSLP baseline	11.93%	12.78%	15.44%	12.97%
DAF[+MLE]+MCE	8.47%	9.75%	8.34%	8.96%
WER reduction	28.98%	23.71%	45.98%	30.27%

4.3. Evaluation of DAF on Aurora3

Summary of results for Aurora3 evaluation are shown in Table 4–6 and the results for each language are given in Table 8–12.

Table 4. Aurora3: Word error rates of ICSLP2002 baseline

Condition	Finnish	Spanish	German	Danish	Average
WM	7.26%	7.06%	8.80%	12.72%	8.96%
MM	19.49%	16.69%	18.96%	32.68%	21.95%
HM	59.47%	48.45%	26.83%	60.63%	48.84%
Overall	24.59%	20.78%	16.86%	31.68%	23.48%

Table 5. Aurora3: Word error rate with DAF[+MLE]+MCE

Condition	Finnish	Spanish	German	Danish	Average
WM	3.39%	4.48%	5.47%	9.25%	5.65%
MM	11.49%	7.97%	13.91%	27.54%	15.23%
HM	26.82%	24.78%	12.26%	49.22%	28.27%
Overall	12.08%	10.78%	10.12%	25.64%	14.66%

Table 6. Aurora3: WER reduction with DAF[+MLE]+MCE

Condition	Finnish	Spanish	German	Danish	Average
WM	53.31%	36.54%	37.84%	27.28%	38.74%
MM	41.05%	52.25%	26.64%	15.73%	33.92%
HM	54.90%	48.85%	54.30%	18.82%	44.22%
Overall	49.41%	45.12%	38.03%	21.12%	38.42%

5. REFERENCES

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

[2] A.E. Rosenberg, O. Siohan, and S. Parathasarathy, "Speaker Verification Using Minimum Verification Error Training," in *Proceedings of ICASSP*, 1998, vol. 1, pp. 105–108.

[3] Y. C. Tam and B. Mak, "Development of an Asynchronous Multiband System for Continuous Speech Recognition," in *Proceedings of Eurospeech*, 2001, vol. 1, pp. 575–578.

[4] B. Mak, Y. C. Tam, and Q. Li, "Discriminative Auditory Features for Robust Speech Recognition," in *Proceedings of ICASSP*, 2002.

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

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

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

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

[9] Y. C. Tam and B. Mak, "An Alternative Approach of Finding Competing Hypotheses for Better Minimum Classification Error Training," in *Proceedings of ICASSP*, 2002.

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

Table 7. Aurora2 multi-condition: Average word accuracies of Test A, B, and C

Condition	ICSLP2002	Auditory	MCE	DAF	DAF+MLE	DAF[+MLE]+MCE
TestA	88.07%	88.40%	90.86%	89.04%	89.02%	91.53%
TestB	87.22%	88.33%	89.90%	88.63%	88.83%	90.22%
TestC	84.56%	89.24%	91.11%	89.69%	89.76%	91.66%
Overall	87.03%	88.54%	90.53%	89.01%	89.09%	91.03%
WER reduction		10.63%	26.22%	14.30%	14.95%	30.27%

Table 8. Aurora3: Overall word accuracies and WER reduction

	ICSLP2002	Auditory	MCE	DAF	DAF+MLE	DAF[+MLE]+MCE
Overall Acc	76.52%	84.17%	85.15%	84.25%	84.13%	85.34%
WER reduction		29.84%	36.76%	30.79%	30.07%	38.42%

Table 9. Aurora3: Word accuracy on the Finnish corpus

Condition	ICSLP2002	Auditory	MCE	DAF	DAF+MLE	DAF[+MLE]+MCE
WM	92.74%	94.44%	96.04%	94.86%	94.41%	96.61%
MM	80.51%	87.14%	88.58%	87.07%	87.35%	88.51%
HM	40.53%	73.53%	74.03%	72.79%	70.92%	73.18%
Overall	75.41%	86.66%	87.93%	86.62%	86.07%	87.92%
WER reduction		35.15%	46.76%	37.02%	34.26%	49.41%

Table 10. Aurora3: Word accuracy on the Spanish corpus

Condition	ICSLP2002	Auditory	MCE	DAF	DAF+MLE	DAF[+MLE]+MCE
WM	92.94%	94.30%	95.27%	94.20%	94.25%	95.52%
MM	83.31%	91.72%	91.94%	91.88%	91.72%	92.03%
HM	51.55%	73.11%	75.10%	73.23%	72.81%	75.22%
Overall	79.22%	88.10%	89.06%	88.15%	88.00%	89.22%
WER reduction		36.47%	43.45%	36.30%	36.03%	45.12%

Table 11. Aurora3: Word accuracy on the German corpus

Condition	ICSLP2002	Auditory	MCE	DAF	DAF+MLE	DAF[+MLE]+MCE
WM	91.20%	93.67%	94.43%	93.97%	93.81%	94.53%
MM	81.04%	84.99%	85.72%	85.36%	85.43%	86.09%
HM	73.17%	87.65%	87.97%	87.79%	87.56%	87.74%
Overall	83.14%	89.13%	89.77%	89.41%	89.31%	89.88%
WER reduction		32.01%	37.11%	34.19%	33.38%	38.03%

Table 12. Aurora3: Word accuracy on the Danish corpus

Condition	ICSLP2002	Auditory	MCE	DAF	DAF+MLE	DAF[+MLE]+MCE
WM	87.28%	90.01%	90.68%	89.96%	89.96%	90.75%
MM	67.32%	70.48%	71.61%	70.48%	71.33%	72.46%
HM	39.37%	48.51%	50.08%	48.67%	48.79%	50.78%
Overall	68.32%	72.80%	73.86%	72.82%	73.15%	74.36%
WER reduction		15.74%	19.70%	15.65%	16.61%	21.12%