LETTER

Evaluation of Combinational Use of Discriminant Analysis-Based Acoustic Feature Transformation and Discriminative Training

Makoto SAKAI[†,††a)], Norihide KITAOKA[††], *Members*, Yuya HATTORI[†], *Nonmember*, Seiichi NAKAGAWA[†††], *Fellow*, *and* Kazuya TAKEDA[††], *Member*

SUMMARY To improve speech recognition performance, acoustic feature transformation based on discriminant analysis has been widely used. For the same purpose, discriminative training of HMMs has also been used. In this letter we investigate the effectiveness of these two techniques and their combination. We also investigate the robustness of matched and mismatched noise conditions between training and evaluation environments.

key words: *speech recognition, feature extraction, discriminative training*

1. Introduction

To improve speech recognition performance, feature transformation such as linear discriminant analysis (LDA) [1] and heteroscedastic discriminant analysis (HDA) [2] are widely used to transform concatenated acoustic features. In a previous paper we proposed power linear discriminant analysis (PLDA) [3], which can describe various criteria including LDA and HDA as special cases. All these methods have improved speech recognition performance. Recently, in machine learning/vision communities, other discriminant analyses have been proposed. Several researchers proposed objective functions such as oriented discriminant analysis (ODA) and a heteroscedastic extension of LDA using Chernoff criterion [4]. All of these discriminant analyses transform features discriminatively in a feature space. On the other hand, various criteria for discriminative training of acoustic models have been studied. Maximum mutual information (MMI) and minimum phone error (MPE) criteria have been successfully applied to many speech recognition systems [5]–[7].

The feature transformation technique and the discriminative training technique aim to improve speech recognition performance at different levels. The combination of these two techniques can further improve speech recognition performance [8]–[11]. In this letter, we investigate combinations of discriminant analysis-based feature transformation and discriminative training through experiments using in-car speech [12]. We also investigate the robustness against mismatched noise conditions between training and evaluation environments.

2. Feature Transformation Based on Discriminant Analysis

This section briefly reviews five feature transformation techniques: LDA, HDA, PLDA, ODA and heteroscedastic extension of LDA using Chernoff distance.

2.1 Linear Discriminant Analysis (LDA)

Given n-dimensional features $\mathbf{x}_j \in \mathbb{R}^n (j = 1, 2, \ldots, N)$, for example, several successive speech frames, let us find a transformation matrix $\mathbf{B} \in \mathbb{R}^{n \times p}$ that transforms these features to p-dimensional features $\mathbf{z}_j \in \mathbb{R}^p$ $(p < n)$, where $\mathbf{z}_j = \mathbf{B}^T \mathbf{x}_j$, and N denotes the number of features.

To obtain an optimal transformation matrix $\mathbf{B}$, the objective function of LDA is defined as follows [1]:

$$J_{LDA}(\mathbf{B}) = \frac{\left|\mathbf{B}^T \mathbf{C}_b \mathbf{B}\right|}{\left|\mathbf{B}^T \mathbf{C}_w \mathbf{B}\right|}, \tag{1}$$

where $\mathbf{C}_b$ and $\mathbf{C}_w$ denote between-class and within-class covariance matrices, respectively. $\mathbf{C}_w = \sum_{k=1}^{c} P_k \mathbf{C}_k$, where $\mathbf{C}_k$ is the covariance matrix of class k, P_k is the class weight, and c is the number of classes. LDA finds a transformation matrix $\mathbf{B}$ that maximizes Eq. (1).

2.2 Heteroscedastic Discriminant Analysis (HDA)

The objective function of HDA is defined as follows [2]:

$$J_{HDA}(\mathbf{B}) = \prod_{k=1}^{c} \left(\frac{\left|\mathbf{B}^T \mathbf{C}_b \mathbf{B}\right|}{\left|\mathbf{B}^T \mathbf{C}_k \mathbf{B}\right|} \right)^{N_k}, \tag{2}$$

where N_k denotes the number of features labeled as class k. Maximization of Eq. (2) is performed using a numerical optimization technique.

2.3 Power Linear Discriminant Analysis (PLDA)

We have proposed PLDA with the following objective function which includes LDA and HDA as special cases [3]:

$$J_{PLDA}(\mathbf{B}, m) = \frac{\left|\mathbf{B}^T \mathbf{C}_b \mathbf{B}\right|}{\left|\left(\sum_{k=1}^{c} P_k (\mathbf{B}^T \mathbf{C}_k \mathbf{B})^m \right)^{1/m}\right|}, \tag{3}$$

Manuscript received September 2, 2009.
Manuscript revised October 30, 2009.

[†] The authors are with DENSO CORPORATION, Nisshin-shi, 470–0111 Japan.

[††] The authors are with Nagoya University, Nagoya-shi, 464–8603 Japan.

[†††] The author is with Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.

a) E-mail: msakai@rlab.denso.co.jp

DOI: 10.1587/transinf.E93.D.395

where m is a control parameter. By varying the control parameter m, the objective function can represent various objective functions. Especially, PLDA corresponds to LDA/HDA when m equals one/zero, respectively.

2.4 Oriented Discriminant Analysis (ODA)

ODA has adopted symmetric divergence as a measure of dissimilarity between two distributions [13]. The objective function is defined as follows:

$$J_{ODA}(\mathbf{B}) = -\sum_{i=1}^{c} tr((\mathbf{B}^T\mathbf{C}_i\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}_i\mathbf{B}),$$

where $\mathbf{A}_i = \sum_{j=1,j\neq i}^{c}(\mathbf{M}_{ij}+\mathbf{C}_j)$ and $\mathbf{M}_{ij} = (\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T$.

2.5 Heteroscedastic Linear Discriminant Analysis Using Chernoff Distance (HLDAC)

Loog et al. [4] proposed a heteroscedastic extension of LDA using the Chernoff criterion (HLDAC) †:

$$\begin{aligned} & J_{HLDAC}(\mathbf{B}) \\ & = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} P_iP_j tr\Big(\Big(\mathbf{B}^T\mathbf{C}_w\mathbf{B}\Big)^{-1}\mathbf{B}^T\mathbf{C}_w^{\frac{1}{2}} \\ & \times\Big(\frac{1}{\pi_i\pi_j}\Big(\log\Big(\mathbf{C}_{ij}^w\Big)-\pi_i\log\Big(\mathbf{C}_i^w\Big)-\pi_j\log\Big(\mathbf{C}_j^w\Big)\Big) \\ & + \Big(\mathbf{C}_{ij}^w\Big)^{-\frac{1}{2}}\mathbf{M}_{ij}^w\Big(\mathbf{C}_{ij}^w\Big)^{-\frac{1}{2}}\Big)\mathbf{C}_w^{\frac{1}{2}}\mathbf{B}\Big), \end{aligned}$$

where $\mathbf{C}_{ij}^w = \mathbf{C}_w^{-\frac{1}{2}}(\pi_i\mathbf{C}_i+\pi_j\mathbf{C}_j)\mathbf{C}_w^{-\frac{1}{2}}$, $\mathbf{C}_k^w = \mathbf{C}_w^{-\frac{1}{2}}\mathbf{C}_k\mathbf{C}_w^{-\frac{1}{2}}$, $\mathbf{M}_{ij}^w = \mathbf{C}_w^{-\frac{1}{2}}\mathbf{M}_{ij}\mathbf{C}_w^{-\frac{1}{2}}$, $\pi_i = P_i/(P_i+P_j)$, and $\pi_j = P_j/(P_i+P_j)$.

3. Discriminative Training

This section briefly reviews two discriminative training techniques: MMI [5], [6] and MPE [7].

3.1 Maximum Mutual Information (MMI)

The MMI criterion is defined as follows [5], [6]:

$$\mathcal{F}_{MMI}(\lambda) = \sum_{r=1}^{R}\log\frac{p_\lambda(O_r|s_r)^\kappa P(s_r)}{\sum_s p_\lambda(O_r|s)^\kappa P(s)},$$

where λ is the set of HMM parameters, O_r is the r'th training sentence, R denotes the number of training sentences, κ is an acoustic de-weighting factor which can be adjusted to improve the test set performance, $p_\lambda(O_r|s)$ is the likelihood given sentence s, and $P(s)$ is the language model probability for sentence s. The MMI criterion equals the multiplication of the posterior probabilities of the correct sentences s_r.

3.2 Minimum Phone Error (MPE)

MPE training aims to minimize the phone classification error (or maximize the phone accuracy) [7]. The objective function to be maximized by the MPE training is expressed as

$$\mathcal{F}_{MPE}(\lambda) = \sum_{r=1}^{R}\frac{\sum_s p_\lambda(O_r|s)^\kappa P(s)\mathrm{A}(s,s_r)}{\sum_s p_\lambda(O_r|s)^\kappa P(s)}, \tag{4}$$

where $\mathrm{A}(s,s_r)$ represents the raw phone transcription accuracy of the sentence s given the correct sentence s_r, which equals the number of correct phones minus the number of errors.

4. Combination of Feature Transformation and Discriminative Training

Feature transformation aims to transform high dimensional features to low dimensional features in a feature space while separating different classes such as monophones. Discriminative training estimates the acoustic models discriminatively in a model space. Because these two techniques are adopted at different levels, a combination of them is expected to have a complementary effect on speech recognition.

5. Experiments

We conducted experiments on CENSREC-3 database [14], which is designed as an evaluation framework for Japanese isolated word recognition in real in-car environments. Speech data were collected using two microphones: a close-talking (CT) microphone and a hands-free (HF) microphone attached to the driver's sun visor. The speech signals were sampled at 16 kHz. For training of HMMs, driver's speech of phonetically-balanced sentences was recorded under two conditions: while idling and driving on city streets under a normal in-car environment without air-conditioner noise. A total of 28,100 utterances spoken by 293 drivers (202 males and 91 females) were recorded with both microphones. We used all utterances recorded with CT and HF microphones for training. For evaluation, we used drivers speech of isolated words recorded with CT and HF microphones under three different conditions: an in-car environment without A/C noise (*normal*), with low fan-speed noise (*fan low*), and with high fan-speed noise (*fan high*). Tables 1 and 2 show the amount of data for evaluation in each condition (total six conditions) and the average SNR (Signal to Noise Ratio) in each recording condition for evaluation data [14], respectively.

[†] We let the function f of a symmetric positive definite matrix $\mathbf{A}$ equal $\mathbf{U}diag(f(\lambda_1),\ldots,f(\lambda_n))\mathbf{U}^T = \mathbf{U}(f(\mathbf{\Lambda}))\mathbf{U}^T$, where $\mathbf{A} = \mathbf{U\Lambda U}^T$, $\mathbf{U}$ denotes the matrix of n eigenvectors, and $\mathbf{\Lambda}$ denotes the diagonal matrix of eigenvalues, λ_i. Here, we define the function f as the logarithm of $\mathbf{A}$.

Table 1 Amount of evaluation data.

Microphone	In-car condition	# Utterances
CT	A/C off (*normal*)	2646
CT	A/C on, low (*fan low*)	2637
CT	A/C on, high (*fan high*)	2695
HF	A/C off (*normal*)	2646
HF	A/C on, low (*fan low*)	2637
HF	A/C on, high (*fan high*)	2695

Table 2 Average SNR of evaluation data in each environment (dB) [14].

Condition	Normal		Fan (low)		Fan (high)	
Microphone	CT	HF	CT	HF	CT	HF
Idling	41.19	16.75	32.86	11.01	25.76	5.47
Low speed	38.39	10.96	32.11	8.67	22.64	2.75
High speed	30.11	5.89	28.58	3.59	21.65	1.46

5.1 Experimental Setup

As for an evaluation procedure, we followed the CENSREC-3 baseline scripts except that fifty similar-sounding words (ex. *aim* for *game* and *tops* for *pops*) were added to the vocabulary. The total vocabulary size became 100. In CENSREC-3, the baseline scripts are designed to facilitate HMM training and evaluation by HTK. The acoustic models consist of triphone HMMs. Each HMM has five states and three of them have output distributions. Each distribution is represented with 32 mixture diagonal Gaussians. The total number of states with the distributions is 2,000. The feature vector consists of 12 MFCCs and log-energy with Δ and $\Delta\Delta$ (baseline). The frame length and the frame shift are 20 ms and 10 ms, respectively.

5.2 Feature Transformation Procedure

Feature transformation was performed using LDA, HDA, ODA, HLDAC, and PLDA for concatenated features. Eleven successive static frames (143 dimensions) were reduced to 39 dimensions, which are the same number of baseline feature dimensions. Although adding delta (and acceleration) coefficients to feature vectors to be processed may be regarded as finding a desirable projection, delta coefficients essentially have no additional information because they are a linear combination of static feature vectors around current time. Therefore, we did not add delta and acceleration to feature vectors. The number of classes was 43, corresponding to the number of monophones. MLLT [15] was applied after LDA, HDA, ODA and HLDAC. For PLDA, we assumed that projected class covariance matrices in Eq. (3) were diagonal. The optimal control parameter ($m = -1.5$) of PLDA was selected experimentally.

5.3 Discriminative Training Procedure

Discriminative training requires two lattices: one for the correct transcription of each training file and another derived from the recognition result of each training file. Having created these lattices using an initial set of models, the HMMs are re-estimated by 5 iterations of a parameter estimation procedure using the same set of lattices. Once these lattices were generated for each feature transformation technique, the same lattice was used to train HMMs with MMI and MPE criteria.

Table 3 Word error rates (%) on the evaluation set recorded under a *normal* condition.

	CT			HF		
	ML	MMI	MPE	ML	MMI	MPE
baseline	7.4	7.1	6.9	15.0	14.4	15.9
LDA	7.1	6.9	**3.9**	14.2	14.1	13.7
HDA	7.9	7.9	6.9	14.5	14.2	13.6
ODA	8.5	7.8	7.0	13.8	13.4	13.3
HLDAC	9.1	8.3	7.4	12.8	12.2	11.3
PLDA	6.2	6.0	5.0	10.7	10.3	**10.2**

Table 4 Word error rates (%) on the evaluation set recorded under a *fan low* condition.

	CT			HF		
	ML	MMI	MPE	ML	MMI	MPE
baseline	9.1	8.8	8.0	25.4	25.0	28.9
LDA	7.3	7.3	**4.4**	26.3	26.1	26.5
HDA	8.4	8.5	7.8	26.6	26.3	28.2
ODA	8.9	8.2	7.7	24.9	23.4	24.9
HLDAC	8.6	8.3	7.0	24.3	23.7	24.8
PLDA	6.4	6.1	4.9	19.7	19.7	**19.6**

Table 5 Word error rates (%) on the evaluation set recorded under a *fan high* condition.

	CT			HF		
	ML	MMI	MPE	ML	MMI	MPE
baseline	10.9	10.7	11.2	56.4	**55.9**	59.8
LDA	14.1	13.3	11.8	63.7	63.3	65.8
HDA	11.1	10.8	11.0	62.6	62.1	66.3
ODA	12.9	11.8	11.2	65.3	64.3	64.9
HLDAC	12.5	11.5	12.0	65.2	64.6	66.7
PLDA	11.3	11.0	**10.2**	61.4	63.2	62.4

5.4 Experimental Results

The experimental results are presented in Tables 3 to 5. The noise condition for the evaluation data used in Table 3 matches that for training data. The evaluation data used in Table 4 contain low air-conditioner noise. The data used in Table 5 contain high air-conditioner noise. These noises are not contained in training data. The best overall performance is shown in bold.

These results showed that both of feature transformations and two discriminative training techniques worked well under a matched noise condition between training and evaluation. In particular, combinations of feature transformations and MPE evidenced outstanding performance. On the other hand, under a mismatched noise condition, the results under a *fan low* noise condition and a *fan high* noise condition had a different tendency. Under a *fan low* condition, both of feature transformations and two discriminative trainings also worked well. This result comes from the fact that the difference between a *normal* condition and a *fan low* condition is slight because A/C noise with low fan-speed is small. Under a *fan high* noise condition, neither feature transformations nor MPE worked well for the data recorded with an HF microphone, as shown in Table 5. When noise in training differs considerably from that in evaluation, the degree of confusability of acoustic features among different classes would change. Therefore, no feature transformations estimated under a *normal* noise environment in training worked well under a *fan high* noise environment in evaluation. In terms of phone classification error among different classes, the data under a *normal* condition and the data under a *fan high* condition would have different optimal boundaries to minimize phone classification errors. Therefore, MPE had worse recognition performance than the other training criteria.

6. Conclusions

We have investigated the effectiveness of discriminant analysis-based feature transformation techniques and discriminative training techniques. Under a matched background noise condition between training and evaluation, both techniques achieved better results than the traditional one (MFCC+Δ+ΔΔ). In addition, a combination of these techniques obtained the best result. However, under a mismatched background noise condition, feature transformations, MPE and their combinations were not necessarily effective.

References

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, New York, 1990.

[2] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, “Maximum likelihood discriminant feature spaces,” Proc. ICASSP, pp.129–132, 2000.

[3] M. Sakai, N. Kitaoka, and S. Nakagawa, “Linear discriminant analysis using a generalized mean of class covariances and its application to speech recognition,” IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.478–487, March 2008.

[4] M. Loog and R. Duin, “Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion,” IEEE Trans. Pattern Anal. Mach. Intell., vol.26, no.6, pp.732–739, 2004.

[5] L. Bahl, P. Brown, P. de Sousa, and R. Mercer, “Maximum mutual information estimation of hidden Markov model parameters for speech recognition,” Proc. ICASSP, pp.49–52, 1986.

[6] P. Woodland and D. Povey, “Large scale MMIE training for conversational telephone speech recognition,” Proc. ICASSP, 2000.

[7] D. Povey and P. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” Proc. ICASSP, pp.105–108, 2002.

[8] E. Oja, Subspace Methods of Pattern Recognition, Letchworth: Research Studies Press, 1983.

[9] A. Biem, S. Katagiri, and B.H. Juang, “Pattern recognition using discriminative feature extraction,” IEEE Trans. Signal Process., vol.45, no.2, pp.500–504, 1997.

[10] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, “The IBM conversational telephony system for rich transcription,” Proc. ICASSP, pp.205–208, 2005.

[11] J.Z. Ma and S. Matsoukas, “Improvements to the BBN RT04 Mandarin conversational telephone speech recognition system,” Proc. Interspeech, pp.1625–1628, 2005.

[12] N. Kitaoka, M. Sakai, Y. Hattori, S. Nakagawa, and K. Takeda, “Evaluation of discriminant analysis-based feature transformation and discriminative traininig for speech recognition,” SPECOM, pp.47–50, 2009.

[13] F. de la Torre and T. Kanade, “Oriented discriminant analysis,” British Machine Vision Conference, pp.132–141, 2004.

[14] M. Fujimoto, K. Takeda, and S. Nakamura, “CENSREC-3: An evaluation framework for Japanese speech recognition in real driving-car environments,” IEICE Trans. Inf. & Syst., vol.E89-D, no.11, pp.2783–2793, Nov. 2006.

[15] R.A. Gopinath, “Maximum likelihood modeling with Gaussian distributions for classification,” Proc. ICASSP, 1998.