DISCRIMINATIVE LEARNING OF ADDITIVE NOISE AND CHANNEL DISTORTIONS FOR ROBUST SPEECH RECOGNITION

Jiqing Han[*,**], Munsung Han*[*], Gyu-Bong Park*[*], Jeongue Park*[*], Wen Gao*[**], Doosung Hwang*

*. Language Understanding Lab. , Systems Engineering Research Institute, ETRI, Korea
**. Department of Computer Science and Engineering, Harbin Institute of Technology, P.R. China
email { jqhan, mshan, gbpark, jgpark}@seri.re.kr, wgao@jdl.mcel.mot.com

ABSTRACT

Learning the influence of additive noise and channel distortions from training data is an effective approach for robust speech recognition. Most of the previous methods are based on *maximum likelihood estimation* criterion. In this paper, we propose a new method of discriminative learning environmental parameters, which is based on *Minimum Classification Error* (MCE) criterion. By using a simple classifier defined by ourselves and the *Generalized Probabilistic Descent* (GPD) algorithm, we iteratively learn environmental parameters. After getting the parameters, we estimate the clean speech features from the observed speech features and then use the estimation of the clean speech features to train or test the back-end HMM classifier. The best error rate reduction of 32.1% is obtained, tested on a Korean 18 isolated confusion words task, relative to conventional HMM system.

1. INTRODUCTION

One of main difficulties in robust speech recognition is to overcome additive noise and channel distortions in environment. Learning distribution changes of clean speech influenced by additive noise and channel distortions from training data and then compensating the changes is an effective approach, and has been widely discussed [1-3]. Most of the previous methods are based on *maximum likelihood estimation* criterion and do not generally lead to a minimum error rate result. Recently, a discriminative learning method was proposed[4], which used *Generalized Probabilistic Descent* (GPD) algorithm to iteratively adjust the classifier parameters according to the criterion of the *Minimum Classification Error* (MCE) and, therefore, directly minimize the misclassifications. The discriminative learning method is first used for training speech recognizer, e.g. HMM-based recognizer [5], and then applied to feature extraction for speech recognition, which is called *Discriminative Feature Extraction* (DFE) [6] and used to design a filter band [7], and to find the optimal linear transformation of mel log spectrum [8]. In all the cases, DFE has been shown to be a powerful tool for error rate reduction. However, DFE needs simultaneously estimate both the feature extractor and the back-end HMM classifier. This is because that any modification of the feature extractor affects the parameter estimation of the back-end HMM classifier.

In this paper, we propose a new DFE-like method to learn environmental parameters (additive noise and channel distortions). Different from the former assumption of speech features distribution, which supposes that the observed noisy speech features follow the Gaussian distribution and has been proved to be not suitable to the practice by some literature (e.g. [2]), we assume that the clean speech features follow the Gaussian distribution. By using a simple classifier defined by ourselves and the MCE/GPD algorithm, we iteratively learn the environmental parameters. After getting the parameters, we estimate the clean speech features from the observed noisy speech features and then use the estimation of the clean speech features to train or test the back-end HMM classifier. Compared with DFE, the proposed method gets the environmental parameters by using a simple classifier and does not affect the back-end HMM classifier, thus it is simpler than the current DFE methods. In our experiments, the best error rate reduction of 32.1% is obtained, tested on a Korean 18 isolated confusion words task, relative to conventional HMM system.

2. DISCRIMINATIVE LEARNING OF ENVIRONMENTAL PARAMETERS

2.1 Environmental Model

We assume that the clean speech signal is first passed through a channel distortions filter whose output is then corrupted by uncorrelated additive noise. If we use $y_m[k], x_m[k], h_m[k]$ and $n_m[k]$ to represent the mel-frequency log power spectrums of the observed speech, the clean speech, the channel distortions filter and the additive noise respectively, and k is a particular mel-frequency band, then we can get

$$x_m[k] = y_m[k] - h_m[k] + \log(1 - \exp(n_m[k] - y_m[k])) . \quad (1)$$

Most of the current speech recognizer use the cepstral vectors as the features. When the environmental influence is observed in the cepstral domain, the relationship between cepstrums of the speech, the noise and the channel distortions is a rather complicated non-linear function as follow

$$\mathbf{X} = \mathbf{Y} - \mathbf{h} + \mathbf{C}\{\log(\mathbf{I} - \exp(\mathbf{C}^{-1}(\mathbf{n} - \mathbf{Y})))\} \quad (2)$$

where **X**, **Y**, **n**, and **h** are the cepstral vectors of the clean speech, the observed speech, the noise and the channel distortions filter, respectively, **I** is the unity vector, **C** and $\mathbf{C}^{-1}$ are the cosine transform matrix and inverse cosine transform matrix, respectively.

We assume that there are many kinds of additive noise in the environment, and we can use the weighted combination of multiple types of noise to represent **n**

$$\mathbf{n} = \sum_{v=1}^{V} w_v \cdot \mathbf{n}_v \quad (3)$$

where $\mathbf{n}_v$ is the v_th type noise, w_v the weight of $\mathbf{n}_v$, V the number of types.

In order to maintain the constraint of the weight, i.e. $\sum_{v=1}^{V} w_v = 1$, the following parameter transformation is adopted during parameters estimation

$$w_v = \frac{e^{\tilde{w}_v}}{\sum_{v=1}^{V} e^{\tilde{w}_v}} \quad . \tag{4}$$

We also assume the channel distortions $\mathbf{h}$ are consisted of the channel distortions of the whole training data and the channel distortions of the current utterance, and it can be implemented as follow

$$\mathbf{h} = \alpha \cdot \mathbf{h}_{wh} + (1-\alpha) \cdot \mathbf{h}_{cu} \tag{5}$$

where $\mathbf{h}_{wh}$ and $\mathbf{h}_{cu}$ are the whole channel distortions and the current one, respectively, and α is an empirical constant.
In this case, the equation (2) can be rewritten as,

$$\begin{aligned} \mathbf{X} = \mathbf{Y} - (\alpha \cdot \mathbf{h}_{wh} + (1-\alpha) \cdot \mathbf{h}_{cu}) \\ + \mathbf{C}\{\log(\mathbf{I} - \exp(\mathbf{C}^{-1}(\sum_{v=1}^{V} w_v \cdot \mathbf{n}_v - \mathbf{Y})))\} \end{aligned} \tag{6}$$

If we use Φ to represent the environmental parameters $\mathbf{n}_v$, w_v and $\mathbf{h}_{wh}$, then using the estimated Φ and $\mathbf{Y}$, $\mathbf{X}$ can be estimated.

2.2 Learning of Environmental Parameters

Let's suppose we have a set of the observed speech cepstral training sequences $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_M\}$ and a set of classes $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. We further assume that the all estimated clean speech signal $\mathbf{X}_m$ that belong to λ_i can be segmented into some equal parts and each same part follows a Gaussian distribution $N_{\mathbf{X}_{m,a}}(\mu_{i,a}, \Sigma_{i,a})$ (for a_th segmentation of $\mathbf{X}_m$, a=1,2, …, A), and the classifier we propose for learning environmental parameters models the estimated clean speech by the probability density function

$$\begin{aligned} p(\mathbf{X}_m|\lambda_i) = \prod_{a=1}^{A} p(\mathbf{X}_{m,a}|\lambda_i) = \\ \prod_{a=1}^{A} \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_{i,a}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{X}_{m,a} - \mu_{i,a})^t \Sigma_{i,a}^{-1} (\mathbf{X}_{m,a} - \mu_{i,a})) \end{aligned} \tag{7}$$

where d is the dimension of $\mathbf{X}_m$, $\mu_{i,a}$ and $\Sigma_{i,a}$ can be calculated

$$\mu_{i,a} = E_i(\mathbf{X}_{m,a}) \tag{8}$$

$$\Sigma_{i,a} = E_i((\mathbf{X}_{m,a} - \mu_{i,a})(\mathbf{X}_{m,a} - \mu_{i,a})^t) \, . \tag{9}$$

The goal of discriminative learning is to reduce the number of misclassifications through a minimization of the average loss function, the steps are as follow.

1. *Discriminative function*: According to the model, the discriminative function is constructed as

$$\begin{aligned} & g_i(\mathbf{X}_m, \Phi) = \log p(\mathbf{X}_m|\lambda_i) \\ & = \log(\prod_{a=1}^{A} p(\mathbf{X}_{m,a}|\lambda_i)) = \sum_{a=1}^{A} (\log p(\mathbf{X}_{m,a}|\lambda_i)) \\ & = \sum_{a=1}^{A} (-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{i,a}| - \frac{1}{2}(\mathbf{X}_{m,a} - \mu_{i,a})^t \Sigma_{i,a}^{-1}(\mathbf{X}_{m,a} - \mu_{i,a})) \, . \end{aligned} \tag{10}$$

The implied decision rule for classification is defined as

$$\mathbf{X}_m \in \lambda_i, \qquad \text{if} \quad g_i(\mathbf{X}_m, \Phi) = \max_j g_j(\mathbf{X}_m, \Phi) \, . \tag{11}$$

2. *Misclassification Measure*: Given a discriminative function, the misclassification measure is

$$d_i(\mathbf{X}_m, \Phi) = -g_i(\mathbf{X}_m, \Phi) + g_\psi(\mathbf{X}_m, \Phi) \tag{12}$$

where λ_ψ is the most confusable class. $d_i(\mathbf{X}_m, \Phi) > 0$ implies misclassification and $d_i(\mathbf{X}_m, \Phi) \le 0$ means correct classification.

3. *Loss Function*: The lost function is defined as a sigmoid function of $d_i(\mathbf{X}_m, \Phi)$

$$\zeta_i(\mathbf{X}_m, \Phi) = 1/(1 + \exp(-d_i(\mathbf{X}_m, \Phi))) \, . \tag{13}$$

4. *Average Loss Function*: For all the training data $\mathbf{X}_m$ (m=1,2, … , M), the average lost function is defined as

$$L(\Phi) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N} \zeta_i(\mathbf{X}_m, \Phi) \; 1(\mathbf{X}_m \in \lambda_i) \tag{14}$$

where $1(\ell) = \begin{cases} 1, & \text{if } \ell \text{ is ture} \\ 0, & \text{otherwise} \end{cases}$

5. *Minimization*: The parameter Φ can be computed iteratively by minimizing the lost function $L(\Phi)$.

$$\Phi^t = \Phi^{t-1} - \eta(t) \nabla L(\Phi^t) \tag{15}$$

where Φ^t is the environmental parameter at the t_th iteration, $\nabla L(\Phi^t)$ is the gradient of the average loss function. To control the convergence of the training procedure, we set the learning step size $\eta(t) = 1/(T_c + 2t)$, with T_c being a prescribed large value (=50 in our case).

2.3 Gradient Calculation

The environmental parameters are adaptively adjusted to reduce the average loss function along a gradient descent direction. The gradient is obtained by computing the partial derivatives of $L(\Phi)$

$$\begin{aligned} \frac{\partial L(\Phi)}{\partial \Phi} &= \frac{\partial}{\partial \Phi}(\frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N} \zeta_i(\mathbf{X}_m, \Phi) \; 1(\mathbf{X}_m \in \lambda_i)) \\ &= \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N} (\frac{\partial \zeta_i(\mathbf{X}_m, \Phi)}{\partial \Phi}) \; 1(\mathbf{X}_m \in \lambda_i) \end{aligned} \tag{16}$$

where

$$\frac{\partial \zeta_i(\mathbf{X}_m,\Phi)}{\partial \Phi} = \frac{\partial \zeta_i(\mathbf{X}_m,\Phi)}{\partial d_i(\mathbf{X}_m,\Phi)} \frac{\partial d_i(\mathbf{X}_m,\Phi)}{\partial g_j(\mathbf{X}_m,\Phi)} \frac{\partial g_j(\mathbf{X}_m,\Phi)}{\partial \mathbf{X}_m(\Phi)} \frac{\partial \mathbf{X}_m(\Phi)}{\partial \Phi}. \quad (17)$$

The first factor in the right-hand-side of equation (17) can be simplified to

$$\frac{\partial \zeta_i(\mathbf{X}_m,\Phi)}{\partial d_i(\mathbf{X}_m,\Phi)} = \frac{\partial}{\partial d_i(\mathbf{X}_m,\Phi)}\left(1/(1+\exp(-d_i(\mathbf{X}_m,\Phi)))\right) = \zeta_i(\mathbf{X}_m,\Phi)(1-\zeta_i(\mathbf{X}_m,\Phi)). \quad (18)$$

The second factor of the right-hand-side of equation (17) can be simplified as follow

$$\frac{\partial d_i(\mathbf{X}_m,\Phi)}{\partial g_j(\mathbf{X}_m,\Phi)} = \frac{\partial}{\partial g_j}(-g_i(\mathbf{X}_m,\Phi)+g_\psi(\mathbf{X}_m,\Phi)) = \begin{cases} -1, & \text{if } j = i \\ 1, & \text{if } j = \psi \end{cases}. \quad (19)$$

The third factor of the right-hand-side of equation (17) can be modified to

$$\frac{\partial g_j(\mathbf{X}_m,\Phi)}{\partial \mathbf{X}_m(\Phi)} = \sum_{a=1}^{A}\left(\frac{\partial}{\partial \mathbf{X}_{m,a}}\left(-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{i,a}| - \frac{1}{2}(\mathbf{X}_{m,a}-\mu_{i,a})^t\Sigma_{i,a}^{-1}(\mathbf{X}_{m,a}-\mu_{i,a})\right)\right) = -\sum_{a=1}^{A}\left(\Sigma_{i,a}^{-1}(\mathbf{X}_{m,a}-\mu_{i,a})\right). \quad (20)$$

For the whole channel distortions and noise, the fourth factors in equation (17) are

$$\frac{\partial \mathbf{X}_m(\Phi)}{\partial \mathbf{h}_{wh}} = -\alpha \quad (21)$$

$$\frac{\partial \mathbf{X}_m(\Phi)}{\partial \mathbf{n}_v} = \gamma(w_v) \quad (22)$$

$$\frac{\partial \mathbf{X}_m(\Phi)}{\partial w_v} = \gamma(\mathbf{n}_v) \quad (23)$$

where

$$\gamma(\Omega) = \mathbf{C}\left\{\left(\frac{-\exp(\mathbf{C}^{-1}(\sum_{v=1}^{V} w_v \cdot \mathbf{n}_v - \mathbf{Y}_m))}{\mathbf{I}-\exp(\mathbf{C}^{-1}(\sum_{v=1}^{V} w_v \cdot \mathbf{n}_v - \mathbf{Y}_m))}\right)\cdot \mathbf{C}^{-1}(\Omega)\right\}. \quad (24)$$

Since the gradient descant search could produce local optimality problems, a good initialization is recommended for the estimated parameters. CMS[9] has been proved to be an effective channel compensation method, which uses the cepstral mean of the utterance being the channel distortions. Thus it is reasonable to use the cepstral mean of all the utterances in the training database as an initial value of the whole channel distortions.

Generally, noise follows a kind of distribution, we simplify the issue by using a set of VQ codebook to represent the noise initial distribution. By using noise sample data and LBG [10] algorithm, we can get the initial value of the noise. And we use 1/V as the initial value of the weight w_v.

3. EXPERIMENTS

In our experiments, a continuous density phoneme-based HMM speaker independent recognizer was used. Each feature vector consists of 12 mel frequency cepstral coefficients (MFCCs). The vocabulary consisted of the 18 Korean isolated confusion words recorded by 80 speakers over different telephone channels at different times. Isolated words were manually segmented and labeled, and 40 speakers' 1850 utterances were used for training and another 40 speakers' 1371 utterances for testing. The SNRs of training database and testing database are 14.07dB and 13.95dB, respectively. A series of experiments are designed to evaluate the proposed method.

Table.1 *Word error rates using various segmentation*

Segment	Two segment	Three segment	Four segment	Phoneme segment
Error rate	8.39%	9.04%	8.1%	7.73%

We first select the parameter α in the equation (6) by using four segmentation for all X_m (A=4) and four types of noise (V=4) , the results of ten iterations are shown in Fig. 1. It shows that α =0.3 exhibits an optimum, which is adopted in the following experiments. We find when α =1.0 the cepstral mean of the whole training data is used as the channel distortions and not consider the influence of the current utterance, it is not very good. And when α =0.0 the cepstral mean of the current speech is used as the channel distortions, which is very similar to CMS, the result is also not very good.

Table.2 *Word error rates using various methods*

Method	Baseline	CMS	RMFCC	DFE	EDFE
Error rate	11.2%	10.9%	8.4%	7.5%	7.6%
Error reduction	–	2.7%	25.0%	33.0%	32.1%

We then compare the results of equal segmentation for all words and the different segmentation for different words based on the number of phonemes, the word error rates for V=4 are listed in Table.1. It shows that the segmentation based on phoneme is good, which is used in the following experiments.

We also evaluate the performances under the various combination types of noise, and Fig.2 shows the results. It is seen that the types of noise V=8 is enough for our database. Finally, we compare the proposed method EDFE (Environment Discriminative Feature Extraction) with CMS, our previous method RMFCC [11], and DFE, the results are listed in Table.2.

It is seen that EDFE is better than CMS and RMFCC. Compared with DFE, EDFE gets near the same performance but needs almost half computational complexity. With respect to both the performance and the computational complexity, EDFE is the best one.

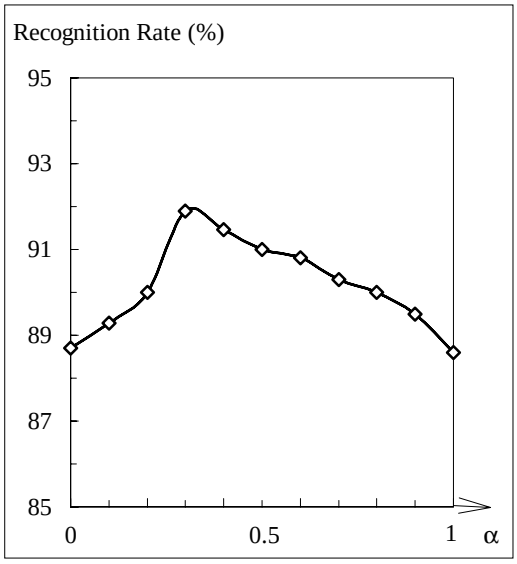

Fig.1 *Performances of using different channel combination parameter* α

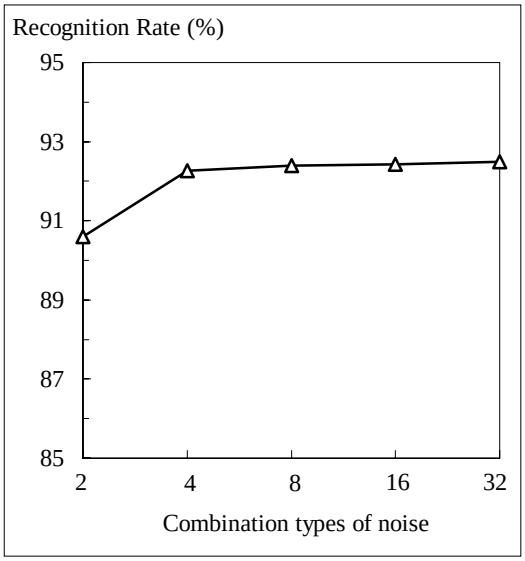

Fig.2 *Performances of various combination types of noise*

4. CONCLUSION

We have proposed a new DFE-like method, which learns the environmental parameters by using a simple classifier defined by ourselves and the MCE/GPD algorithm. After getting the parameters, we estimate the clean speech features from the observed speech features and then use the estimation of the clean speech features to train or test the back-end HMM classifier. Compared with DFE, the proposed method gets the environmental parameters by using a simple classifier and does not affect the back-end HMM classifier, thus it is simpler than the current DFE methods. In our experiments, the best error rate reduction of 32.1% is obtained, tested on a Korean 18 isolate confusion words task, relative to conventional HMM system.

5. REFERENCES

[1]. A.Acero, R.Stern, " Environmental Robustness in Automatic Speech Recognition ", ICASSP, 1990, PP.849-852.

[2]. J.Pedro, " Speech Recognition in Noisy Environments ", Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, April ,1996.

[3]. M.Gales, S.Young, " Robust Speech Recognition in Additive and Convolutional Noise using Parallel Model Combination ", Computer Speech and Language, 9, PP.289-307, 1995.

[4]. B.Juang, S.Katagiri, " Discriminative Learning for Minimum Error Classification ", IEEE Trans. On Signal Processing, Vol.40 (12), PP.3043-3054, 1992.

[5]. B. Juang, W. Chou, C. Lee, " Minimum Classification Error Rate Methods for Speech Recognition ", IEEE Trans. On Speech and Audio Processing, Vol.5 (3), PP.257-265, May 1997.

[6]. A.Biem, S.Katagiri, " Feature Extraction Based on Minimum Classification Error/Generalized Probabilistic Descent Method ", ICASSP, 1993, PP.II275-II278.

[7]. A.Biem, S.Katagiri, " Filter Bank Design Based on Discriminative Feature Extraction ", ICASSP, 1994, PP.I485-I488.

[8]. C.Rathinavelu, L.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 (3), PP.243-256, May 1997.

[9]. S.Furui, " Cepstral analysis technique for automatic speaker verification ", IEEE Trans. On ASSP, Vol. 29(4), PP.254-272, April 1981.

[10]. Y. Linde, A. Buzo, R. Gray, " An Algorithm for Vector Quantizer Design ", IEEE Trans. On Communication, Vol. 28, PP.84-95, 1980.

[11]. J. Han, M. Han, G. Park, J. Park, W. Gao, "Relative Mel-Frequency Cepstral Coefficients Compensation for Robust Telephone Speech Recognition", Eurospeech'97, PP. 1531-1534.