Discriminative Noise Adaptive Training Approach for an Environment Migration

Byung-Ok Kang, Ho-Young Jung and Yun-Keun Lee

Speech/Language Information Research Center,
Electronics and Telecommunications Research Institute, Daejeon, Korea
{bokang, hjung, yklee}@etri.re.kr

10.21437/Interspeech.2007-564

Abstract

A combined strategy of noise-adaptive training (NAT) and discriminative-based adaptation is proposed for effective migration of speech recognition systems to other noisy environments. NAT is an effective approach for real-field applications, but does not satisfy the minimum classification error (MCE) criterion for the recognition process and adapts poorly to new environments. The proposed method makes up for the weak points in discriminative adaptation strategies, and presents a new method for improving the MCE approach. Using this new method, experimental results show that the speech recognition system can successfully be migrated to other environments using specific-condition data of the target environment.

Index Terms: discriminative noise adaptive training, environment migration

1. Introduction

The mismatch between training and real recognition environments, mainly caused by additive noise and channel distortion in the real environments, degrades the performance of the automatic speech recognition system. It is commonly acknowledged that the main issue in developing practical speech recognition systems is to achieve robustness against environmental mismatches.

Many approaches have been proposed to handle this problem. Most of these can be classified into speech enhancement techniques and feature compensation or normalization schemes. While these methods are effective to obtain pseudo-clean speech through the estimation and suppression of noise, their primary concern is not the recovery of the masked phonetic information but the purgation of the corrupted signal. Speech enhancement and feature normalization methods distort the cues for speech recognition, and feature compensation methods may need to be jointly designed with the training process in order to guarantee good performance. A training strategy is therefore required which maximizes the consistency with noise reduction techniques. Hong proposed the robust environment-effects suppression training (REST) algorithm by which current acoustic models are updated with the enhanced speech [1]. Deng *et al* introduced a noise-adaptive training (NAT) which combines the multi-condition training concept with noise reduction techniques [2]. By adding various amounts and types of noisy training data and by applying a suitable noise reduction method to the noisy data, this approach can model the residual distortion effectively. As such, NAT is particularly attractive for use in real-field conditions. NAT however does not satisfy the minimum classification error (MCE) criterion for the recognition process and adapts insufficiently well to other environments. To address the first problem, Hong extends the REST algorithm by updating the current acoustic models using an MCE-based training scheme [3]. Wu and Huo applied the MCE criterion to the joint optimization of compensated training set and multi-condition training set to absorb the residual distortion. Maximum a posteriori (MAP) [4] or maximum likelihood linear regression (MLLR) [5] may be used to handle the second problem. Unfortunately, both these approaches require many speech data collected under various conditions and are thus over-fitted with respect to the specific environments used for the calibration data. To circumvent this problem, this paper proposes a new strategy for satisfying the MCE criterion as well as for making NAT more robust against environmental migrations.

The application of the MCE criterion to model adaptation has enjoyed a growing interest in the recent years. Martin *et al* showed that over a period of 6 months, MCE-based adaptation is better than MAP adaptation for speaker identification [6]. With data collected over several 1-month sessions, MCE-based adaptation was shown to be very effective for all sessions while MAP yielded an insignificant improvement for 2 to 4 sessions using adaptation data of session 1. While MAP suffers from the overspecialization problem mentioned earlier, MCE-based adaptation exhibits discriminative capability in various conditions using specific calibration data of the target environment. We therefore introduce a discriminative noise adaptive training (DNAT) approach that applies MCE-based adaptation to a NAT-based acoustic model in order to cope with the weaknesses of NAT. In addition, we propose a minimum phone classification error (MPCE) method to enhance the generalization capability of the MCE process for a large vocabulary. While MCE has a problem due to small calibration data for large vocabulary recognition, MPCE can prevent the over-specialization of any model unit by operating on the merged domain of final model units. We adopted the segmental generalized probabilistic descent (GPD) training algorithm using an N-best phone lattice [7][8]. The MPCE-adapted NAT-based acoustic model using calibration data of one particular condition (asphalt-paved road, 60km/h, closed window) not included in the training was evaluated for a car environment under 6 other driving conditions. Compared to conventional MCE and MAP, MPCE showed outstanding performance in all 6 driving conditions.

2. Noise Adaptive Training for Environment Migration

NAT is a combined approach of multi-condition training and variability-normalization. In this method, a noise reduction technique is applied to the various noisy training data, and the obtained pseudo-clean data are then used to construct an acoustic model [2]. NAT therefore depends on 3 assumptions:

absorption of various acoustic styles by the multi-condition training, compensation of the mismatch between clean and noisy data by noise reduction technique, and modeling of the residual distortion after compensation.

These assumptions are essentially correct if various amounts and types of noisy data are provided such that NAT can solve the non-acceptance of residual distortions of traditional multi-condition training and the non-absorptiveness of various acoustic environments of conventional clean training. However, it is impractical to collect all possible conditions of noisy data in a specific environment such as car or home. NAT needs to include the capability to adapt to new environments which were not considered in the training process. Furthermore, NAT does not satisfy the MCE criterion in the recognition process.

We present the DNAT method which transforms the NAT-based model into a robust model capable of recovering the nature of phonetic discriminative information masked by new noisy environments. Fig. 1 shows the general block diagram of the DNAT strategy. DNAT is based on discriminative adaptation using calibration data for the new environment, and a new discriminative adaptation method is proposed for an effective migration to new environments using a few conditions of data.

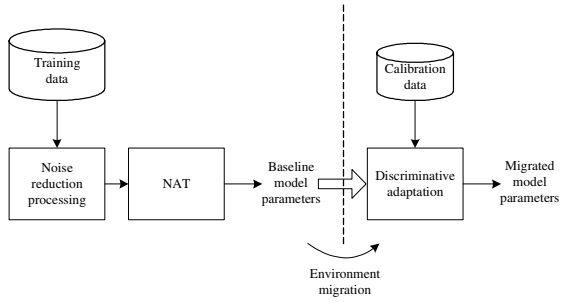

Fig. 1. Block diagram of the implementation of DNAT algorithm

3. Minimum Phone Classification Error Training

We describe the conventional MCE framework and introduce the MPCE approach which was designed to represent the new environment without over-specializing into the specific calibration conditions.

3.1. MCE training

The purpose of MCE training is to be able to correctly discriminate the observations of an HMM for best classification results rather than to find the best model for the distributions of the data. The difficulty of the MCE training approach lies in the derivation of an objective function which adequately reflects the performance measure and which is suitable for optimization. The error rate for a finite data set can be a good performance measure, but is a piecewise constant function of the classifier parameter Λ and a poor candidate for optimization by a simple numerical search method.

It is therefore necessary to embed the decision rule into a smooth-form loss function. A misclassification measure is defined by

$$d_i(\mathrm{X}) = -\log\left[P_i(\mathrm{X}\mid\Lambda)\right] + \log\left[\frac{1}{N}\sum_{j,\,j\neq i}^{j} e^{\log\left[P_j(\mathrm{X}\mid\Lambda)\right]\eta}\right]^{1/\eta}, \qquad (1)$$

where $P_i(X\mid\Lambda)$ is the class conditional likelihood function of the observation X, η is a positive number and N refers to the N-best incorrect classes [8]. This measure is a continuous function of the classifier parameter Λ and can be a good emulator of the decision rule. For an i_{th} class utterance X, a value of $d_i(X)>0$ implies a misclassification and $d_i(X)<0$ a correct decision. The general form of the loss function is then defined in terms of the misclassification measure using a zero-one loss function.

$$\ell_i(X;\Lambda) = \ell(d(\mathrm{X})), \qquad (2)$$

where the zero-one loss function can be the sigmoid function for example.

The optimal solution of the training process in the MCE approach is to minimize the expected loss. To this avail, the generalized probabilistic descent (GPD) algorithm [7] can be used. The update rule is described by

$$\Lambda_{n+1} = \Lambda_n - \varepsilon_n U_n \nabla\ell(\mathrm{X};\Lambda)|_{\Lambda=\Lambda_n}, \qquad (3)$$

where U_n is a positive definite matrix and ε_n indicates the learning rate. Using the GPD algorithm, the discriminative adjustment for the mean and variance of the Gaussian mixture can be shown to be as follows: [9]

$$\tilde{\mu}^{\,i}_{jkl}(n+1) = \tilde{\mu}^{\,i}_{jkl}(n) - \varepsilon \left.\frac{\partial \ell_i(\mathrm{X};\Lambda)}{\partial \tilde{\mu}^{\,i}_{jkl}}\right|_{\Lambda=\Lambda_n} \qquad (4)$$

$$\frac{\partial \ell_i(\mathrm{X};\Lambda)}{\partial \tilde{\mu}^{\,i}_{jkl}} = \gamma \ell_i(d_i)(1-\ell_i(d_i))\frac{\partial d_i}{\partial \tilde{\mu}^{\,i}_{jkl}} \qquad (5)$$

$$\frac{\partial d_i}{\partial \tilde{\mu}^{\,i}_{jkl}} = -\sum_{t=1}^{T}\delta(q_t - j)\frac{c^{\,i}_{jk}N(x_t;\mu^{\,i}_{jk};\sigma^{\,i}_{jk})}{b_j(t)}\left(\frac{x_{tl}}{\sigma^{\,i}_{jkl}} - \mu^{\,i}_{jkl}\right) \qquad (6)$$

$$\tilde{\sigma}^{\,i}_{jkl}(n+1) = \tilde{\sigma}^{\,i}_{jkl}(n) - \varepsilon \left.\frac{\partial \ell_i(\mathrm{X};\Lambda)}{\partial \tilde{\sigma}^{\,i}_{jkl}}\right|_{\Lambda=\Lambda_n} \qquad (7)$$

$$\frac{\partial \ell_i(\mathrm{X};\Lambda)}{\partial \tilde{\sigma}^{\,i}_{jkl}} = \gamma \ell_i(d_i)(1-\ell_i(d_i))\frac{\partial d_i}{\partial \tilde{\sigma}^{\,i}_{jkl}} \qquad (8)$$

$$\frac{\partial d_i}{\partial \tilde{\sigma}^{\,i}_{jkl}} = -\sum_{t=1}^{T}\delta(q_t - j)\frac{c^{\,i}_{jk}N(x_t;\mu^{\,i}_{jk};\sigma^{\,i}_{jk})}{b_j(t)}\left(\left(\frac{x_{tl}-\mu^{\,i}_{jkl}}{\sigma^{\,i}_{jkl}}\right)^2 - 1\right) \qquad (9)$$

Similar derivations for the mixture weights can easily be performed.

3.2. MPCE training for generalization capability of discriminative adaptation on noisy data

The aim of the MPCE approach is to provide the NAT training approach with the adaptation capability by solving the generalization problem of the conventional MCE method. The MCE method performs better in small-sized acoustic models than in large-sized ones, and works well on test data with characteristics similar to the training data. This is not suitable however for the environmental migration of a large-vocabulary speech recognition system using only few condition data of the target environment. By updating the model parameters of final context-dependent units using the loss function in the model of simplified phone-like units (PLU), MPCE allows for a simple migration to the new environment by means of discrimination enhancement using specific condition data for the target environment whilst preserving the generality for other environments.

Fig. 2 gives an overview of the procedure. First, the initial canonical model is trained using maximum likelihood (ML) estimation. This initial model consists of the final context-dependent (CD) model and the simplified PLU model which

has hierarchical correspondence with the final CD model. Since both models are trained with the same data, the simplified PLU model can serve as a good approximation of the final CD model. The discriminative adjustment, phone-alignment information, and loss function score obtained from simplified PLU models can therefore be used for discriminative adjustment of final CD models.

The MPCE training proceeds as follows. For the environment migration, an amount of calibration data is first collected from the target environment in which the speech recognition system will be used. After generating the *N*-best results for each utterance of the calibration set using the initial simplified PLU models and word-level lexicon, we can find the segmentation information of the correct simplified PLU sequence and the corresponding N-best sequences. Next, the update rules of the MCE algorithm can be applied for the discriminative adjustment of the simplified PLU models. Finally, for the discriminative adjustment of the final CD models, the following rules are applied, which use the loss function score obtained from the segments of simplified PLU models. The update rule for the mean of the Gaussian mixture is given by

$$\tilde{\mu}^{\,i}_{jkl}(n+1) = \tilde{\mu}^{\,i}_{jkl}(n) - \varepsilon \left. \frac{\partial \hat{\ell}_i(\mathrm{X};\Lambda)}{\partial \tilde{\mu}^{\,i}_{jkl}} \right|_{\Lambda=\Lambda_n} \tag{10}$$

and the variance update rule by

$$\tilde{\sigma}^{\,i}_{jkl}(n+1) = \tilde{\sigma}^{\,i}_{jkl}(n) - \varepsilon \left. \frac{\partial \hat{\ell}_i(\mathrm{X};\Lambda)}{\partial \tilde{\sigma}^{\,i}_{jkl}} \right|_{\Lambda=\Lambda_n} \tag{11}$$

where $\hat{\ell}_i(\mathrm{X};\Lambda)$ is the loss function obtained from the segments of simplified PLU models.

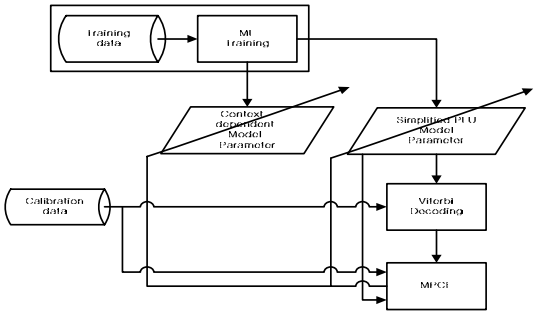

Fig. 2. MPCE training for model adaptation

From discriminative adaptation point of view, the proposed MPCE has an advantage over conventional MCE as follows. By assigning the discriminative adjustment obtained from the hierarchically upper nodes to all of the leaf nodes in the decision tree, MPCE can handle unseen units which are not included in the calibration data and overcome the problem of over-specialization. From the environment migration point of view, by expanding the discriminative power obtained from the specific calibration data of the target environment to leaf node models, the transformed model using MPCE training shows robustness to all conditions of the target environment. This is verified by the experimental results of Section 5.

The proposed MPCE training procedure is summarized as follows:

1) Initialize HMMs with ML-trained models. (simplified PLU HMMs and their hierarchical CD HMMs)
2) Generate the N-best phone sequence for each utterance of the calibration set using the simplified PLU HMMs.
3) Segment the correct reference sequences and all N-best competing sequences into states using the simplified PLU HMMs.
4) Adjust HMM parameters of the simplified PLU and final CD unit with the computed gradient of the loss function accumulated over the entire calibration data set.
   A. For the case of the simplified PLU HMMs, use (4)-(9).
   B. For the case of hierarchical CD HMMs, use (10)-(11).
5) Check convergence. Stop the iteration if the algorithm converges, otherwise go to step 1 with two types of HMMs obtained from step 4.

4. Experimental Setting

Our domain is 40k POI (point-of-interest) recognition. We are using a triphone based HMM which is a tied-state model of 1150 states, where each state is a mixture of 16 Gaussians. For the simplified PLU model in the proposed MPCE, we use a monophone based HMM which has 3 states and a mixture of 3 Gaussians per state. 39-dimensional feature vectors (13 MFCC including C0, and their first and second derivatives) are used.

The training DB consists of 2 speech corpora individually produced by SiTEC and ETRI. The speech corpus of SiTEC comprises 8,516 POI utterances recorded by 190 speakers. Both in the low speed (30~60 km/h) and high speed (70~90 km/h) driving environments, a AKG C400-BL microphone and Shure SM-10A headset were used. The speech corpus of ETRI consists of 94,566 POI utterances recorded by 433 speakers and in the various driving environment, microphone (AKG C400-BL) and headset (Altec Lansing AH302) are used. The training DB takes the various driving environments for multi-condition training in consideration and a noise-robust front-end composed of a Mel-warped Wiener filter and global CMS is used for noise reduction.

The test DB includes 2 sets labeled SET-TargetEnv and SET-GenEnv. SET-TargetEnv is the data set collected in the target environment where the speech recognition system will be used. It contains 1,652 POI utterances read by 6 male and 4 female speakers. It is recorded in the 2000cc New Sonata of Hyundai Motors using a mono channel microphone of type AKG C400-BL. In order to emulate real situations, it includes the 6 different driving conditions listed in Table 1. SET-GenEnv is the test set of the ETRI speech corpus used in the training DB. It consists of 2,074 POI utterances read by speakers who do not participate in the training corpus.

Table 1. 6 driving conditions of SET-TargetEnv

Name	Environment
Env1	Asphalt-paved road, 60 km/h, closed window
Env2	Asphalt-paved road, 60 km/h, open window
Env3	Asphalt-paved road, 100 km/h, closed window
Env4	Concrete-paved road, 60 km/h, closed window
Env5	Concrete-paved road, 60 km/h, open window
Env6	Concrete-paved road, 100 km/h, closed window

The calibration DB for the environmental migration consists of a total of 9,000 POI utterances read by 54 male and 36 female speakers. Speakers and uttered POI list distinct from those used in SET-TargetEnv. The actual calibration DB consisted of 2,000, 6,000 or all 9000 utterances. The recording environment is similar to the target environment, i.e. the same car (2000cc New Sonata of Hyundai Motors) and microphone (AKG C400-BL) are used. However, the only driving condition considered during recording is the standard condition of driving on an asphalt-paved road at the controlled speed of 60 km/h. This corresponds to the driving condition Env1 of SET-TargetEnv.

5. Experimental Results

Using the experimental setting described in section 4, the experiments are performed as follows. The proposed environmental migration using MPCE is evaluated and compared with the conventional MCE [7][9] and MAP [4][10] adaptation methods. These methods were implemented in the ETRI Speech Toolkit (ESTk) consisting of HMM training tools and a speech recognition engine. The parameters (η, λ, ε_t for MCE, MPCE and τ for MAP) applied to each method are optimized by means of experiments.

Table 2 shows the speech recognition results for the target environment for each of the adaptation methods. 9,000 POI utterances were used as calibration DB for the environmental adaptation of each method. In order to assess the performance in the various target environments, we used SET-TargetEnv described in section 4 as the test data and evaluated the speech recognition results in each of the 6 driving environments individually. As shown in Table 2, MPCE performed better than the NAT-based baseline model and conventional adaptation methods. MAP outperforms the other methods in the Env1 environment which is similar to the calibration environment, but shows serious degradation in Env2 and Env5 where road conditions and window states are not in accordance with the calibration conditions. These results show that MAP adaptation is a good method for specialization under specific calibration conditions, but also that it has the drawback of losing the property of the general environment. This is apparent from the results in Table 3. Overall, the environmental migration using MPCE shows good results, maintaining its efficiency in Env5 whose driving condition is diametrically different from the one used for calibration.

Table 2. Comparison of methods on SET-TargetEnv

	NAT	+MAP		+MCE		+MPCE	
	%Ra	%Ra	%Err	%Ra	%Err	%Ra	%Err
Env1	**90.88**	**94.39**	**38.49**	**92.98**	**23.03**	**94.04**	**34.65**
Env2	88.06	86.19	-15.66	88.43	3.10	90.67	21.86
Env3	92.34	93.49	15.01	92.72	4.96	94.25	24.93
Env4	89.73	90.41	6.62	90.75	9.93	92.47	26.68
Env5	**83.20**	**77.10**	**-36.31**	**84.73**	**9.11**	**83.59**	**2.32**
Env6	91.90	93.66	21.73	92.61	8.77	95.07	39.14
Total	89.41	89.35	-0.57	90.44	9.73	91.77	22.29

Table 3 shows the speech recognition results of SET-GenEnv for each of the methods discussed. It can be observed that MCE and our proposed method maintain the property of general environment.

Table 3. Comparison of methods on SET-GenEnv

	NAT	+MAP	+MCE	+MPCE
SET-GenEnv	89.25	84.91	89.34	89.15

Figure 3 shows the performance of MPCE and MCE in function of the POI utterance size of the calibration set. It can be observed that MPCE has a relatively good error rate reduction (ERR) performance of 17% when the calibration data set contains 2,000 utterances.

6. Conclusions

This paper proposed DNAT method to give NAT the effective migration of speech recognition system to other noisy environments. By including the MPCE adaptation module to NAT approach, we solved the lack of phone-discriminative power and adaptation ability. Compared with conventional MCE method, in addition, MPCE method could obtain the generality of a new environment using the data of specific condition of corresponding environment. When the system is migrated from training car environment to testing car one, experiment results showed that MPCE was very effective for six conditions using one condition data of testing car environment. As a future work, we will study the combination of MCE linear regression as a speaker adaptation and MPCE as an environmental migration.

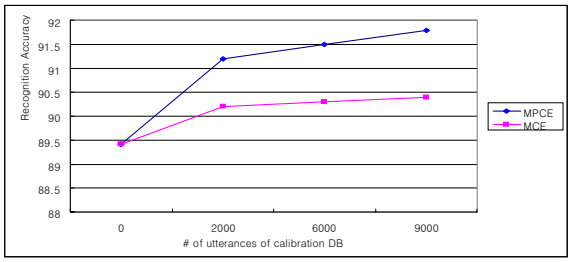

Fig. 3. Comparison of methods

7. References

[1] W.-Y. Hong and S.-H. Chen, "A Robust Training Algorithm for Adverse Speech Recognition," Speech Communication, vol. 30, no. 4, pp. 273-293, 2000.

[2] L. Deng, A. Acero, M. Plumpe, and X.-D. Huang, "Large-Vocabulary Speech Recognition Under Adverse Acoustic Environments," in Proc. ICSLP, 2000, pp. III-806-809.

[3] W.-T. Hong, "A Discriminative and Robust Training Algorithm for Noisy Speech Recognition," in Proc. ICASSP, 2003, pp. I-8-11.

[4] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. Speech Audio Processing, vol. 2, pp. 291-299, April 1994.

[5] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," Computer Speech and Language, 9, pp. 171-186, 1995.

[6] C. Martin del Alamo, J. Alvarez, C. de la Torre, F. J. Poyatos, and L. Hernandez, "Incremental Speaker Adaptation with Minimum Error Discriminative Training for Speaker Identification," in Proc. ICSLP, 1996, pp. III-312-315.

[7] W. Chou, B.-H. Juang, and C.-H. Lee, "Segmental GPD training of HMM based speech recognizer," in Proc. ICASSP'92, San Francisco, CA, 1992, pp. 473-476.

[8] J. Chen and F. K. Soong, "An N-Best Candidates-Based Discriminative Training for Speech Recognition Applications," IEEE Trans. Speech Audio Processing, vol. 2, pp. 206-216, Jan. 1994.

[9] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," IEEE Trans. Speech Audio Processing, vol. 5, pp. 257-265, May 1997.

[10] X. Wang and D. O'Shaughnessy, "Environmental Compensation Using ASR Model Adaptation by A Bayesian Parametric Representation Method," in Interspeech, Lisboa, Portugal, Sep. 2005, pp. 1801-1804.