Model Adaptation Using Discriminative Noise Adaptive Training Approach for New Environments

Ho-Young Jung, Byung-Ok Kang, and Yunkeun Lee

ABSTRACT—A conventional environment adaptation for robust speech recognition is usually conducted using transform-based techniques. Here, we present a discriminative adaptation strategy based on a multi-condition-trained model, and propose a new method to provide universal application to a new environment using the environment's specific conditions. Experimental results show that a speech recognition system adapted using the proposed method works successfully for other conditions as well as for those of the new environment.

Keywords—Discriminative adaptation, minimum phone classification error, speech recognition.

I. Introduction

The performance of automatic speech recognition systems is degraded by a mismatch between training and real recognition environments. The performance achieved by many approaches has not reached that of the matched condition [1]. Recently, two methods have shown effective strategies for obtaining a robust acoustic model. Based on noise reduction techniques to compensate a mismatch, the first strategy maximizes the consistency between training and real recognition environments. Hong proposed a robust environment-effects suppression training (REST) algorithm by which a robust reference acoustic model is directly trained from a database collected in adverse environments using bias and noise compensation [2]. Noise adaptive training (NAT), which combines the idea of multi-condition training and noise reduction techniques, was also introduced [3]. After each training utterance is enhanced using suitable noise reduction techniques, NAT can effectively absorb the residual distortion of various training conditions. However, REST and NAT cannot guarantee the objective of minimum classification error (MCE) for the recognition process and cannot adapt effectively to unknown environments. Hong also introduced the discriminative REST method, which re-trains discriminatively on phonetic variability using the acoustic model obtained from the REST algorithm [2]. Wu and Huo proposed an MCE criterion for the joint design of feature compensation and multi-condition training processes in NAT [3]. These approaches can enhance discrimination for the aim of an MCE but are not adaptable to new environments.

The second strategy is a discriminative adaptation scheme that uses discriminative objective functions in the adaptation process. This strategy combines the discriminative training concept with conventional adaptation techniques. Povey and others introduced the maximum mutual information (MMI)-maximum *a posteriori* (MAP) technique, which incorporates a prior distribution obtained from MAP with the statistics required by MMI estimation [4]. The MMI-MAP technique is reported to be effective for task adaptation as well as for generating gender-dependent models, but may adapt the models only to the specific condition data of the new domain.

Based on NAT and discriminative adaptation techniques, we present a new strategy to satisfy the generalization of MCE training on noisy adaptation data and to make NAT more robust against environmental adaptations. Although MCE-based adaptation shows some discriminative capability in other conditions using the specific condition data of a new environment, it has a problem due to the small adaptation data for a large vocabulary recognition system [5]. To cope with this, we propose a minimum phone classification error (MPCE) method that can prevent the over-fitting of any model unit by operating on the semi-tied level of the final model units. The proposed method is a discriminative NAT (DNAT) approach

Manuscript received Aug. 19, 2008; revised Oct. 21, 2008; accepted Oct. 28, 2008.

Ho-Young Jung (phone: + 82 42 860 1328, email: hjung@etri.re.kr), Byung-Ok Kang (email: bokang@etri.re.kr) and Yunkeun Lee (email: yklee@etri.re.kr) are with the Software & Content Research Laboratory, ETRI, Daejeon, Rep. of Korea.

applying MPCE-based adaptation to an NAT-based noise-compensated acoustic model using small noisy calibration data.

II. Discriminative Noise Adaptive Training

In NAT, a noise reduction technique is applied to compensate the noisy training data of various conditions, and the obtained pseudo-clean data is used to construct an acoustic model [3]. This indicates that NAT depends on three assumptions: absorption of various acoustic styles by multi-condition training, compensation of the mismatch between clean and noisy data by the noise reduction technique, and modeling of the residual distortion after compensation. These assumptions are substantially correct if various amounts and types of noisy data are provided. However, it is impractical to collect all conditions of noisy data in a new environment. NAT needs to include the ability to adapt to a new environment that is not treated in the training process. In addition, it requires a generalized discriminative adaptation ability to satisfy the objective of MCE and to avoid an over-fitting to specific calibration data.

Therefore, we propose the DNAT method to adapt the NAT-based model into a robust one with the recovery and nature of phonetic discriminative information masked in a different way by new environments. Figure 1 shows a block diagram of the DNAT strategy. The DNAT approach includes an MPCE training technique performing effective discriminative adaptation using small noisy calibration data of new environments based on noise-compensated acoustic models with various environmental characteristics.

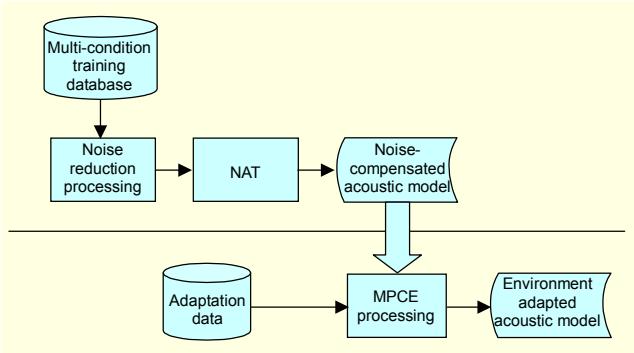

Fig. 1. Block diagram of environment adaptation using DNAT.

III. Minimum Phone Classification Error Training

Conventional MCE training performs better in small acoustic models than in large ones, and it works well on test data with characteristics similar to training data. This makes MCE unsuitable for the environmental adaptation of a large-vocabulary speech recognition system using little conditional data of the corresponding environment. Therefore, MPCE provides the adaptation ability for the NAT approach by solving the generalization problem of the conventional MCE method.

MPCE is based on a segmental generalized probabilistic descent (GPD) algorithm [6] and updates the model parameters of final context-dependent (CD) units using the misclassification measure obtained from semi-tied units (STUs). The STUs can be defined as nodes at high hierarchical positions of leaf nodes, which are final units from the decision tree designed for basic acoustic models. By this indirect model update with a hierarchical architecture, MPCE prevents a sparse adjustment of model parameters due to little adaptation data and leads to generalization of non-trained conditional data, which is a weakness of discriminative training approaches.

In the MPCE training process, STUs are first defined. The canonical model consists of a final CD model and an STU model. After extracting the N best results for each utterance of the calibration data using this canonical model and word-level lexicon, we can find a class misclassification measure based on the frame segmentation of the correct sequence and the corresponding N best sequences as

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

where $P_i(x|\Lambda)$ is the class conditional likelihood function of the observation x, η is a positive number, and N denotes the N best incorrect classes. After the sigmoid function is applied to (1), the loss functions of the STUs and final CD models are given by $\tilde{\ell}(\tilde{d}(X))$ and $\ell(d(X))$ using the definition of [6]. Next, the derivatives of $\tilde{\ell}(\tilde{d}(X))$ are computed based on the GPD algorithm, and the discriminative adjustment of the final CD model is performed using the computed derivative. For the acoustic model based on Gaussian mixture density, the new rules of the mean and variance updates are given by

$$\mu' = \mu - \varepsilon\xi \frac{\partial\tilde{\ell}(\tilde{d}(X;\Lambda))}{\partial\mu} \quad \text{and} \quad (2)$$

$$\sigma' = \sigma - \varepsilon\xi \frac{\partial\tilde{\ell}(\tilde{d}(X;\Lambda))}{\partial\sigma}, \quad (3)$$

where μ' and σ' are the parameters of the final CD models, ε is the step size of the adaptation, and ξ indicates the weighting factor given by $\tilde{\ell}/(\ell+\tilde{\ell})$.

IV. Large-Vocabulary Isolated Word Recognition

The DNAT approach was evaluated by the task of

recognizing 220,000 vocabulary items, namely, point-of-interest (POI) utterances for a car navigation system. As an acoustic model, we used a monophone-based hidden Markov model (HMM) with three states, where each state has a mixture of 32 Gaussian distributions. For the STU model, we used a monophone-based HMM with three states and three mixtures per state. In this case, as for the STU model, MPCE uses the small mixture model of the same units as the final model. The feature vectors consist of thirteen mel-frequency cepstral coefficients (MFCCs), including C0 and their first and second derivatives.

The training database (DB) consists of 8,516 POI utterances recorded from 190 speakers using an AKG C400-BL and Shure SM-10A in low- and high-speed driving conditions, and 94,566 POI utterances recorded from 433 speakers using an AKG C400-BL and Altec Lansing AH302 in various driving environments [5]. The training DB considers various noisy environments for a robust modeling. A noise-robust front-end composed of a Wiener filter and cepstral mean subtraction was used for noise reduction. The test DB consists of two sets (TSET1 and TSET2) collected in car environments that are not included in the training DB. The TSET1 comprises 731 utterances recorded from five speakers using an HP iPAQ PDA in various conditions. The TSET2 comprises 1,938 utterances obtained from 20 speakers using a low-cost embedded microphone.

The calibration DB for TSET1 was collected at 60 Km/h, and consists of 2,000 POI utterances recorded with an AKG C400-BL (CSET1-1) and 800 POI utterances from an HP iPAQ PDA (CSET1-2). The 3,000 POIs spoken by 30 speakers using a low-cost embedded microphone were used as a calibration DB for TSET2. The speakers and uttered POI list in each calibration DB are distinct from those used in the corresponding test set.

We implemented these methods on our large scale embedded speech recognizer, the ETRI Speech Toolkit (ESTk-laser). In this speech engine, a decoding process is divided into two stages (acoustic and lexical decoding), and the proposed method is used to adapt the models for the acoustic decoding stage. Because a comparison of MCE and MPCE is shown in [5], Table 1 shows the environmental adaptation results of MPCE for TSET1.

The calibration data obtained from a different speed and microphone from the test data improved the performance by 3.1%. The adaptation performance was more effective when the data recorded from the same iPAQ PDA was added. Table 2 shows the results for TSET2. Using the calibration data at a specific speed, MPCE reduced the error rate by 34.9% for test data at various speeds.

Table 1. Word correction rate (%) after adaptation for TSET1.

NAT	DNAT with MPCE CSET1-1	DNAT with MPCE CSET1-1 + CSET1-2
81.5	84.6	86.6

Table 2. Word correction rate (%) after adaptation for TSET2.

NAT	DNAT with MPCE
90.06	93.53

V. Conclusion

This paper proposed a DNAT method to give NAT an effective adaptation of speech recognition to other noisy environments. By applying the MPCE adaptation approach to the NAT framework, we solved the lack of phone-discriminative power and adaptation ability. The MPCE method played an important role in the environmental adaptation by providing a generality of other conditions using the data of the specific condition of the new environment. However, MPCE may have a problem when the acoustic channel of the new environment is very different from the training acoustic channel. In future work, we hope to use a priori information like the MMI-MAP approach to see if the abrupt changes between acoustic channels can be overcome.

References

[1] Y. Suh and H. Kim, "Class-Based Histogram Equalization for Robust Speech Recognition," *ETRI Journal*, vol. 28, 2006, pp. 502-505.

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

[3] J. Wu and Q. Huo, "An Environment-Compensated Minimum Classification Error Training Approach Based on Stochastic Vector Mapping," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, 2006, pp. 2147-2155.

[4] D. Povey et al., "MMI-MAP and MPE-MAP for Acoustic Model Adaptation," *Proc. Eurospeech*, 2003, pp. 1981-1984.

[5] B.O. Kang, H.Y. Jung, and Y.K. Lee, "Discriminative Noise Adaptive Training Approach for an Environment Migration," *Proc. Interspeech*, 2007, pp. 2085-2088.

[6] W. Chou, B.H. Juang, and C.H. Lee, "Segmental GPD Training of HMM Based Speech Recognizer," *Proc. ICASSP*, 1992, pp. 473-476.