PAPER

Combining Multiple Acoustic Models in GMM Spaces for Robust Speech Recognition*

Byung Ok KANG[†,††a)], *Nonmember and* **Oh-Wook KWON**[††b)], *Member*

SUMMARY We propose a new method to combine multiple acoustic models in Gaussian mixture model (GMM) spaces for robust speech recognition. Even though large vocabulary continuous speech recognition (LVCSR) systems are recently widespread, they often make egregious recognition errors resulting from unavoidable mismatch of speaking styles or environments between the training and real conditions. To handle this problem, a multi-style training approach has been used conventionally to train a large acoustic model by using a large speech database with various kinds of speaking styles and environment noise. But, in this work, we combine multiple sub-models trained for different speaking styles or environment noise into a large acoustic model by maximizing the log-likelihood of the sub-model states sharing the same phonetic context and position. Then the combined acoustic model is used in a new target system, which is robust to variation in speaking style and diverse environment noise. Experimental results show that the proposed method significantly outperforms the conventional methods in two tasks: Non-native English speech recognition for second-language learning systems and noise-robust point-of-interest (POI) recognition for car navigation systems.

key words: *noise-robust speech recognition, acoustic model, GMM combination, non-native speech recognition*

1. Introduction

With the significant progress of research on automatic speech recognition (ASR) over the past few decades, various applications of ASR have been successful beyond the boundary of research laboratories. In particular, large vocabulary continuous speech recognition (LVCSR) have been applied in many areas: Mobile voice search, broadcast news transcription, voice recording at call centers, lecture and meeting transcriptions, automatic speech translation, dialog-based information retrieval, and car navigation. Among them, a mobile voice search application is expanding its user base by improving the speech recognition accuracy under the gradationally matched conditions through utilizing enormous speech logs.

However, despite the widespread commercial adoption, LVCSR systems operating in real environments still often suffer from egregious recognition errors resulting from adverse background noise, channel distortion, diversity of speaking styles, disfluency in spoken dialogs, non-native speech, and out-of-vocabulary words. This is because the acoustic model of LVCSR systems is not robust to the mismatch between the training and real environments so that it cannot cope with various unexpected conditions in real environments. The mismatch problem could be solved if we had an infinite matched data covering all acoustic spaces of the target environments in terms of domain, speaking style, noise, and channel. In a specific application of mobile voice search services, the performance has been improved under the gradationally matched conditions because there has been an enormous accumulation of speech logs in which real conditions are reflected [1]. However, the performance improvement is rather limited when the accumulated speech logs are used for different tasks such as dialog-based car navigation systems.

Many approaches focusing on acoustic models (AMs) have been proposed to handle this problem. A multi-condition training method, currently the most common method, attempts to train the acoustic model by using a mixture of various amounts and types of noisy speech data [2]. For this purpose, a huge amount of speech data reflecting real environment noise is required. However, this requirement could not be achieved without deploying a real field service. On the other hand, Povey and others proposed a universal background model (UBM) framework composed of a large mixture of Gaussians covering the whole acoustic space, which is adapted to each context-dependent phone comprising the acoustic model of the target system [3]. A drawback of the UBM framework for speech recognition is the extremely large size of the Gaussian mixture models (GMMs). Recently, the subspace GMM (SGMM) [4] was proposed to replace the background model.

Many applications such as second-language learning or evaluation systems are increasingly exploiting speech technologies. Furthermore, speech user interfaces driven by an ASR engine are used globally across language borders. However, due to the different acoustic and linguistic characteristics from native speech, non-native speech generally degrades the performance of an ASR system. Many approaches have been proposed to handle this problem; acoustic modeling strategies can be classified in accordance with the amount of non-native speech available in the applied task. Nallasamy and others [5] proposed a polyphone decision-tree extension/adaptation method to accept new contextual variations identified in a small amount of

Manuscript received June 30, 2015.
Manuscript revised October 2, 2015.
Manuscript publicized November 24, 2015.

[†] The author is with SW Content Research Laboratory, ETRI, Daejeon, Rep. of Korea.
[††] The authors are with School of Electronics Engineering, Chungbuk National University, Cheongju, Rep. of Korea.

[*] This work was conducted during the research year of Chungbuk National University in 2015, Korea.

a) E-mail: bokang@etri.re.kr
b) E-mail: owkwon@cbnu.ac.kr (Corresponding author)

DOI: 10.1587/transinf.2015EDP7252

adaptation data. When supplied with plenty of non-native speech data, Chen and co-workers [6] investigated several strategies to use native and non-native data effectively for acoustic modeling, and proposed the use of a phonetic decision tree (PDT) generated by only native speech data when constructing acoustic models using both native and non-native speech data.

In this paper, we propose a new acoustic modeling method based on a multi-space GMM [7], where multiple sub-models are trained by using the speech database (DB) under their own specific conditions and then are combined into a new acoustic model in a GMM space for the target task. In some cases, a sub-model can be a pre-existing acoustic model built for a particular target task. In case of a customer service system (e.g., a voice search service), multiple sub-models can be trained by using accumulated speech logs recorded in real environments. For each sub-model, the GMMs of all hidden Markov model (HMM) states are estimated so that the GMMs of the sub-model can be optimal elements occupying their acoustic space based on various conditions of tri-phone context, speaking style, and environment noise. All states and their GMMs from the HMMs of a sub-model are gathered into a huge pool of states. Finally, the target acoustic model is constructed from this pool of states by merging the states satisfying two criteria proposed in this paper. Experimental results show that the proposed method achieves better performance than the conventional multi-condition training without any adaptation using target domain speech DBs.

We applied the proposed method to an ASR system for both native and non-native speakers. The proposed method is different from the conventional polyphone decision tree-based extension/adaptation method [5] in that the proposed method can be extended to non-native speakers with multiple sources of first languages without target domain speech DBs. As shown in experimental results to be described later, the method proposed in [6] can be used to handle the same problem, but shows limited performance improvement compared to the proposed method. Whereas the previous paper [7] focused on the approach of building an acoustic model robust to environment noise, this paper expands target tasks to include non-native speech recognition and point-of-interest (POI) recognition. Compared with the SGMM-based approach [4], our method has a simple algorithmic structure and does not require additional complicated parameter estimation steps.

In terms of handling speaking style with multiple phonetic variations, the proposed method is closely related to the study of multidialectal speech recognition. Caballero *et al.* [8] presented research results about ASR dealing with five dialects of Spanish, in which different methods for combining data between dialects were proposed and compared. Whereas dialect information is also needed in the regression through the decision tree during the training and decoding stage for the dialect-context-dependent (DCD) acoustic modeling approach proposed in [8], our proposed method does not require any information on whether input speech is native or non-native during the decoding stage.

The remainder of this paper is organized as follows. Section 2 describes the proposed method to combine the GMMs from multiple AMs. Section 3 and Sect. 4 present experimental results in two cases to show the effectiveness of the proposed method. Finally, concluding remarks are given in Sect. 5.

2. Proposed Method

Figure 1 shows an overview of the main procedure of the proposed acoustic modeling method. In the figure, the sub-models are the pre-existing AMs that have been trained and optimized for specific tasks using the appropriate training speech DB. For example, 'sub-model 1' can be a native English AM trained by using native-spoken English speech DBs, and 'sub-model 2' can be a non-native English AM trained by using Korean-spoken English speech DBs.

2.1 Training Sub-Models

For each sub-model, mono-phone HMMs are expanded into tri-phone HMMs, where the model parameters Λ are estimated by maximizing the likelihood $P(O\,|\,\Lambda)$, where O is a sequence of speech feature vectors of a training speech DB. In LVCSR, the states occupying a similar acoustic space are tied using a decision-tree-based clustering mechanism [9], [10]. In the conventional HMM-based speech recognition, the probability of observing a data vector $\mathbf{x}$ in an HMM state j is expressed through a GMM as follows.

$$p(\mathbf{x}|j) = \sum_{i=1}^{M_j} w_{ji}\mathrm{N}\left(\mathbf{x};\boldsymbol{\mu}_{ji},\boldsymbol{\Sigma}_{ji}\right), \tag{1}$$

where M_j is the number of mixture components in state j, w_{ji} is the weight of the i-th component and $N(.;\boldsymbol{\mu}_{ji},\boldsymbol{\Sigma}_{ji})$ is a multivariate Gaussian probability density function with mean vector $\boldsymbol{\mu}_{ji}$ and covariance matrix $\boldsymbol{\Sigma}_{ji}$. In this step, each sub-model has a unique set of states and GMMs that are the optimal elements characterizing their own acoustic space based on various conditions of tri-phone distribution, noise environment, and channel distortion. Each sub-model includes a GMM and a single Gaussian for each context-dependent HMM, and includes a physical and logical HMM

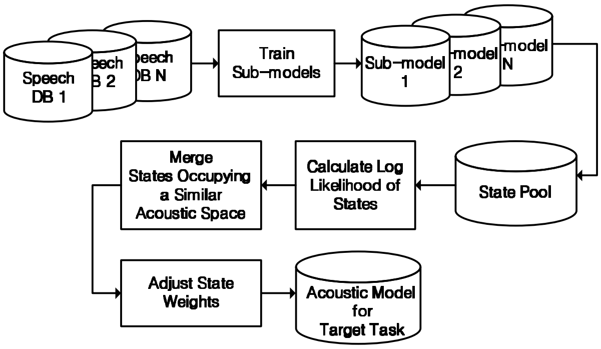

Fig. 1 Block diagram of the proposed acoustic modeling method.

mapping table generated by a PDT. The single Gaussian of each state is used only to find the pairs of states to be merged from a pool of states. Actual merging of states is done in the states with a full GMM which are used in the decoding stage.

2.2 Computing Log Likelihood of States

At first in the succeeding step, all GMMs from the multiple sub-models are gathered into a huge pool of states. Then the state occupation counts for all states are computed in the state pool using the Baum-Welch re-estimation algorithm [9]. These state occupation counts are generated requisitely in the process of training sub-models and referred to generate a PDT for each sub-model. Since the state occupation counts are saved together with the corresponding sub-models, we do not need to access the original training data in the next step of state merging. In combination with the means and variances, the state occupation counts form sufficient statistics to calculate the log likelihood for single-Gaussian distributions [9]. Assuming that state tying [10] does not change the frame alignment, the log likelihood of each state in the state pool $L(s)$ is approximately calculated as follows.

$$L(s) = \sum_{f \in F} \log\Big(P\big(\mathbf{x}_f; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\big)\Big)\, \gamma_s(\mathbf{x}_f), \tag{2}$$

where F is the set of time frames aligned to state s, $\gamma_s(\mathbf{x}_f)$ is the *a posteriori* probability of the observation data vector $\mathbf{x}_f$ being generated by state s, and $P(\mathbf{x}_f; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ is the state observation probability.

In particular, for a Gaussian distribution, the log likelihood is given by

$$L(s) = -\frac{1}{2}(\log[(2\pi)^n |\boldsymbol{\Sigma}_s|] + n) \sum_{f \in F} \gamma_s(\mathbf{x}_f), \tag{3}$$

where n is the dimensionality of the data vector $\mathbf{x}_f$ [10].

2.3 Merging States Occupying a Similar Acoustic Space

Among the pool of states, all pairs of states clustered into the same terminal node of the PDT, which has the same center-phone and the same phone state position in the HMM, are merged if one of two criteria is satisfied. A combined GMM is obtained by concatenating the Gaussian distributions of all merged GMMs as follows. Assuming that a GMM with M_j mixtures at state j has a parameter set of $\boldsymbol{\lambda}_j = \{w_{ji}, \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji}, i = 1 \cdots M_j\}$, the parameter set $\tilde{\boldsymbol{\lambda}}_c$ of the combined GMM is given as

$$\begin{aligned}\tilde{\boldsymbol{\lambda}}_c = \{&w_{1i}, \boldsymbol{\mu}_{1i}, \boldsymbol{\Sigma}_{1i}, \cdots w_{ji}, \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji},\\ &\cdots, w_{S_c i}, \boldsymbol{\mu}_{S_c i}, \boldsymbol{\Sigma}_{S_c i}, i = 1 \cdots M_j\},\end{aligned} \tag{4}$$

where $\boldsymbol{S}_c$ is the total number of states merged to the target state c. Consequently, each state of the combined model has a variable number of Gaussian distributions.

The first criterion is as follows:

$$\begin{aligned}&\Delta L_{merged} = L(state_1) + L(state_2) - L(state_{merged})\\ &\Delta L_{merged} \le Threshold,\end{aligned} \tag{5}$$

where ΔL_{merged} is the decrease in the log likelihood after state merging, $L(state_1)$ and $L(state_2)$ are the log likelihood of any pair of states in the state pool. These log likelihoods are computed using Eq. (3) with a single Gaussian model of each state. Any pair of states whose decrease in log likelihood is less than a threshold is merged. The threshold influences the total number of states of the final acoustic model. As the distribution of state likelihood depends on the speech DB used for training sub-models, the merging threshold is tuned from a small randomized subset of the training speech DB. In our experiments, the best performance in terms of the final AM size and the recognition accuracy was obtained when the threshold was set to the values ranging from 4,000 to 8,000 depending on the target tasks.

The second criterion is as follows:

$$\begin{aligned}&\{T_{sub1} \,|\, T_{sub1} \text{ is the logical tri-phone sharing state}_1\}\\ &== \{T_{sub2} \,|\, T_{sub2} \text{ is the logical tri-phone sharing state}_2\},\end{aligned} \tag{6}$$

where $\{T_{sub1}\}$ and $\{T_{sub2}\}$ are a set of logical tri-phones sharing states $state_1$ and $state_2$ of the state pool in a specific state position, respectively.

Figure 2 shows the concept of the second criterion for merging states. The cloud-like figure describes a huge pool of states in an acoustic space composed of sub-models. An arrow in the figure indicates that any state of the state pool can be a state position of several logical tri-phones of a sub-model. If a set of states shares the same logical tri-phone set from sub-models, these states are likely to occupy the same acoustic space and can be merged as a single state. For example, if *state₁* is a state position of a logical tri-phone set

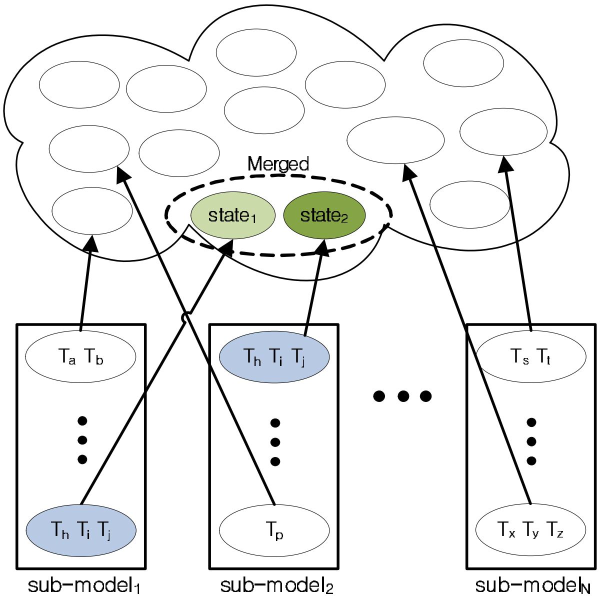

Fig. 2 Concept of the second criterion for merging states.

$\{T_h, T_i, T_j\}$ of *sub-model*$_1$ and *state*$_2$ is a state position of the same logical tri-phone set of *sub-model*$_2$, then *state*$_1$ and *state*$_2$ are merged. After the merging process using the two criteria, a common PDT for the target task is constructed based on the clustered terminal nodes.

2.4 Adjusting State Weights and GMM Parameters

Because the state pool is generated through the concatenation operation, the mixture weights of the GMMs constituting these states should be adjusted. Because these GMMs are constructed from a combination of several sub-models reflecting multiple spaces, they should be adjusted to represent an acoustic space that their states delegate for the target task. If a target domain speech DB such as user's speech log is available, the weights of the combined GMMs in each state, which are induced from the GMM spaces from multiple AMs, can be adjusted based on the maximum *a posteriori* estimation or discriminative adaptive training [11], [12]. For the state weights to influence on recognition performance, we need a sufficient amount of target domain DB from real environment conditions such as speech logs to adjust mixture weights of several or tens of thousands of states. Since we could not obtain a sufficient amount of adaptation DB from real environment conditions, we just proportionately scaled the original mixture weights of sub-models as follows,

$$\mathbf{w}_{adj} = \frac{\mathbf{w}_{orig} \text{ of each state of sub-models}}{\text{Total number of states merged into target state}}, \tag{7}$$

where $\mathbf{w}_{adj}$ is the adjusted weight vector of GMMs in the target state and $\mathbf{w}_{orig}$ is the original weight vector of GMMs in the state of a sub-model merged to the target state.

3. Combining Native and Non-Native Acoustic Models

3.1 Experimental Conditions

For many applications such as a language learning or evaluation system for a second language, an acoustic model should provide robust speech recognition performance for both native and non-native speech. For this purpose, we applied the proposed acoustic modeling method for combining native and non-native acoustic models. The speaker information on whether input speech is native or non-native is assumed to be unknown during the decoding stage in the proposed method. Instead, the kind of the target task, for example, whether an English evaluation system for Korean speakers or a car navigation system, is assumed to be known during the training and decoding stages.

The training speech DB for acoustic modeling consisted of 426 hours of native American English speech data, which were derived from the training set of several speech DBs: The Wall Street Journal database (WSJ1) [13], the native English speech corpus of the Speech Information Technology and Industry Promotion Center (SiTEC) [14], and the native English speech DB of the Electronics and Telecommunications Research Institute (ETRI). The training speech DB for non-native English AM was composed of 382 hours of Korean-spoken English speech data extracted from the training set of the Korean-spoken English speech DB of ETRI.

In our experiments, all states in the acoustic models for the baseline and sub-models were set to have a fixed number of Gaussians for simplicity. The number of states of each model was determined according to the amount of training speech DBs. In detail, the number of states was counted after the leaf nodes with the number of occurrence less than threshold are removed in the PDT. For fair comparison, we used the same parameters for the proposed method as the baseline and sub-models. Using native-spoken English AM and Korean-spoken English AM, we built the combined native and non-native AM following the procedure of the proposed acoustic modeling. We used both criteria for the state merging step described in Sect. 2.3. We adjusted the state weights by scaling proportionately the original mixture weights of GMMs from sub-models as described in Sect. 2.4.

We extracted 39 dimensional feature vectors composed of 13 mel frequency cepstral coefficients (MFCCs) including C0, and their first and second derivatives. The acoustic models consisted of left-to-right and cross-word tri-phone HMMs with three states, where each state has a mixture of 16 GMMs. The acoustic models were built by using the hidden Markov model toolkit (HTK) [9]. For each AM, the mono-phone HMMs were expanded into tri-phone HMMs, and the states of the tri-phone HMMs were then tied using the PDT clustering mechanism [9], [10]. For acoustic and pronunciation models, we used the CMU-DICT American English phone set [15], which consists of 39 phones.

For native English evaluation, we used 4,878 utterances from the evaluation set of the WSJ1 (NatEngSet). For non-native English evaluation, we used three evaluation sets. The first non-native evaluation set is made up of 3,109 utterances from the evaluation set of the Korean-spoken English speech DB of ETRI (KorEngSet1). The second one is made up of 3,000 utterances which were spoken by kids and recorded in real situations by GnB, an English education institute in Korea (KorEngSet2). The third one is made up of 400 utterances recorded by 10 Korean speakers for assessing GinieTutor [17], an English-learning application of ETRI (KorEngSet3).

Two types of language models in the form of backed-off bigrams were used for the native and non-native English evaluation sets. For the native English evaluation set, 5,593 unique words and 21,378 bigram entries were used. For the two non-native English evaluation sets, 5,059 unique words and 105,401 bigram entries were used. For the speech recognition engine, we used a finite state transducer (FST)-based large vocabulary speech recognizer developed at ETRI [16].

3.2 Experimental Results

First, we compared the performance of AMs trained with native-spoken English speech DB and non-native-spoken English speech DB.

Table 1 shows that word error rates (WERs) of native-spoken English AM are 3.2% and 38.9% for NatEngSet and KorEngSet1, respectively. Native-spoken English AM performs well for NatEngSet. However, due to the mismatch of acoustic and linguistic characteristics, it shows serious performance degradation for Korean-spoken English evaluation sets. On the contrary, Korean-spoken English AM yields WERs of 8.1% and 18.9% for NatEngSet and KorEngSet1, respectively.

Next, following the conventional method to build acoustic models for covering non-native speech, we directly combined the native English speech DB and the Korean-spoken English speech DB, and constructed native/Korean-spoken English AM with an 808 hour speech DB. As shown in the table, native/Korean-spoken English AM gives comparable performance with native-spoken English AM for NatEngSet, but gives small performance degradation compared to Korean-spoken English AM for KorEngSet1. As Chen and co-workers proposed [6], we generated a PDT by using only native-spoken speech DB and then trained GMMs for each leaf node based on native/Korean spoken speech DB. In spite of a smaller number of total Gaussian mixtures as shown in Fig. 3, native PDT & native/Korean spoken English AM gives comparable performance with native/Korean spoken English AM for NatEngSet, but results in performance degradation for Korean-spoken English evaluation sets.

Using native-spoken English AM and Korean-spoken English AM, we combined two acoustic models in the state level by the proposed method. The results show that the proposed method achieved significant error rate reduction (ERR) of 15.0%, 7.6%, and 13.9% for all non-native English evaluation sets of KorEngSet1, KorEngSet2, and KorEngSet3 compared to the conventional method, respectively.

Compared to native-spoken English AM, the proposed AM consistently achieves significant improvement for all 3 kinds of Korean-spoken English evaluation sets, but has worse performance for NatEngSet. The proposed method outperformed Korean-spoken English AM, achieving significant ERR of 42.0% and 6.9% for NatEngSet and KorEngSet1, respectively, which implies that the combined native and non-native AM gives better performance than the AM obtained by using only non-native speech data.

To evaluate the effect of the two criteria on recognition accuracy, we applied the first and second criteria separately and then compared their performance with the proposed method which used both criteria together. Table 2 shows the recognition performance of two criteria in the state merging step described in Sect. 2.3. While two criteria got similar performance for the Korean-spoken English evaluation sets, the second criterion showed somewhat more contribution to the recognition performance for NatEngSet. The proposed method which uses two criteria consistently achieved good performance compared with applying each criterion separately.

Figure 3 shows the total number of Gaussian mixtures of native-spoken English AM, Korean-spoken English AM, the conventional native/Korean spoken English AM and the proposed AM. The figure suggests that the proposed AM achieves better performance than the conventional native/Korean spoken English AM, even though the proposed method has the total number of mixtures similar to the conventional one.

Table 1 Comparison of word error rates (WERs, in %) of native English AM, Korean-spoken English AM, Native/Korean-spoken AM, Native PDT & native/Korean spoken English AM, and the proposed AM.

Evaluation DB	NatEngSet	KorEngSet1	KorEngSet2	KorEngSet3
Native-spoken English AM	3.2	38.9	38.3	21.5
Korean-spoken English AM	8.1	18.9	21.0	9.5
Native/Korean spoken English AM	3.6	20.7	22.4	10.1
Native PDT & native/Korean spoken English AM	3.4	21.5	23.3	11.1
Proposed AM	4.7	17.6	20.7	8.7

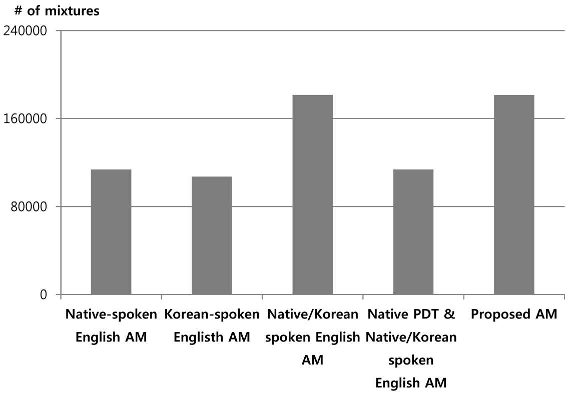

Fig. 3 Comparison of the total number of Gaussian mixtures.

Table 2 Comparison of the two state merging criteria of the proposed AM.

Evaluation DB	NatEngSet	KorEngSet1	KorEngSet2	KorEngSet3
The first criterion	6.2	18.1	21.1	9.8
The second criterion	5.8	18.0	21.3	9.9
Proposed AM	4.7	17.6	20.7	8.7

4. Combining Acoustic Models for Car Navigation Systems

4.1 Experimental Conditions

We applied the proposed method to build a noise-robust acoustic model for speech recognition in a car navigation system. Our target system has a speech recognition interface for both POI input and dialog-based information retrieval. The target domain for dialog-based information retrieval includes the recognition of POIs and addresses, location-based search items such as gas stations and parking lots, command and control for digital multimedia broadcasting (DMB), and information retrieval for traffic reports about congested areas. The vocabulary size was about 1,800k.

The training speech DB for acoustic modeling was 1,100 hours of training speech data composed of 750 hours of dictation speech and 350 hours of phonetically optimized words (POWs) [18], connected digits, words with low-frequency phoneme strings, and speech logs of a mobile search service. To build an acoustic model based on the proposed method, we used two sub-models trained with a clean speech DB and noisy speech DB, respectively. The first sub-model was trained by using speech DBs in a quiet office environment and the second sub-model was trained for noisy environments. The original training speech DB for the second sub-model was the same as the speech DB of the first sub-model. However, noise signals recorded in a car under various conditions were randomly mixed into the original training speech DB at a randomly selected SNR within the range of 5 to 15 dB. For the baseline multi-condition training, we used all speech DBs: The clean speech DB of the first sub-model and the noisy speech DB of the second sub-model. For evaluation in real car environments, we used 1,200 utterances recorded in two types of vehicles: The Hyundai Sonata (D segment sedan) and the Hyundai Santa Fe (E segment SUV). To emulate real environments of car navigation systems, three different driving conditions were considered: Idling, driving through town, and driving at a high speed.

4.2 Experimental Results

Table 3 compares the WERs between AM obtained by the proposed method and AM used in baseline multi-condition training. To check the contribution of two criteria used in the state merging step, we applied the first and second criteria separately and then compared their performance with the baseline and proposed methods. The experimental results indicate that the proposed AM has better performance than the baseline AM under various driving conditions and the AMs with each criterion applied separately.

Table 3 Comparison of WERs (%) of the baseline AM and the proposed AM.

	Sonata			Santa Fe		
	idle	town	high	idle	town	high
Baseline AM	9.7	14.6	16.7	13.9	22.2	39.4
The first criterion	8.9	13.9	16.3	13.2	22.1	40.0
The first criterion	8.4	14.1	16.0	13.6	22.1	39.7
Proposed AM	7.7	13.5	15.8	12.9	22.0	39.8

5. Conclusion

We proposed a new method for combining multiple AMs in GMM spaces for robust speech recognition of non-native and noisy speech. The proposed method is motivated by the fact that each acoustic model with an intrinsic target task has the optimal elements occupying its acoustic space reflecting specific conditions and environments. Thus, after gathering all states and their GMMs from sub-models into a huge pool, a new acoustic model for the target task can be constructed by state merging and weight adjustments.

To evaluate the proposed acoustic modeling method for robust speech recognition, we performed computer experiments for two tasks: A non-native speech recognition task for English learning systems as a second language and a noise-robust speech recognition task for car navigation systems. For the non-native speech recognition task, the proposed method of combining native and non-native models achieved an average ERR of 12.2%. For the noise-robust speech recognition task, the proposed method achieved consistent and significant error reduction under various driving conditions compared to the conventional method.

Acknowledgments

This work was supported by the ICT R&D program of MSIP/IITP. [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning]

References

[1] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Garret, and B. Strope, "Google search by voice: A case study," in Visions of Speech: Exploring New Voice Apps in Mobile Environments, Call Centers and Clinics, A. Neustein, Ed. Springer, 2010.

[2] R.P. Lippmann, E.A. Martin, and D.B. Paul, "Multi-style training for robust isolated-word speech recognition," Proc. ICASSP-1987, pp.705–708, Dallas, Texas, USA, May 1987.

[3] D. Povey, S.M. Chu, and B. Varadarajan, "Universal background model based speech recognition," Proc. ICASSP-2008, Las Vegas USA, March 2008.

[4] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N.K. Goel, M. Karafiat, A. Rastrow, R.C. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian mixture models for speech recognition," Proc. ICASSP-2010, Dallas, Texas, USA, March 2010.

[5] U. Nallasamy, F. Metze, and T. Schultz, "Enhanced polyphone decision tree adaptation for accented speech recognition," Proc. INTERSPEECH-2012, pp.1902–1905, 2012.

[6] X. Chen and J. Cheng, "Acoustic modeling for native and non-native Mandarin speech recognition," Proc. International Symposium on

Chinese Spoken Language Processing, 2012.

[7] B.O. Kang, H.Y. Jung, and O.-W. Kwon, "Noise robust spontaneous speech recognition using multi-space GMM," Proc. INTERNOISE-2013, Innsbruck, Austria, Sept. 2013.

[8] M. Caballero, A. Moreno, and A. Nogueiras, "Multidialectal Spanish acoustic modeling for speech recognition," Speech Communication, vol.51, no.3, pp.217–229, 2009.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book, Cambridge University Engineering Department, 2009.

[10] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," Proc. ARPA Human Language Technology Workshop, pp.307–312, Plainsboro, New Jersey, USA, 1994.

[11] B.O. Kang, H.Y. Jung, and Y. Lee, "Discriminative noise adaptive training approach for an environment migration," Proc. INTERSPEECH-2007, pp.2085–2089, Antwerp, Belgium, Aug. 2007.

[12] H.-Y. Jung, B.-O. Kang, and Y. Lee, "Model adaptation using discriminative noise adaptive training approach for new environments," ETRI Journal, vol.30, no.6, pp.865–867, Dec. 2008.

[13] D.B. Paul and J.M. Baker, "The design for the Wall Street Journal-based CSR corpus," Proc. ICSLP-1992, pp.899–902, Oct. 1992.

[14] Y.J. Lee, S.C. Rhee, B.W. Kim, and Y. Um, "Speech corpora development in Korea - on the activities of speech information technology & industry promotion center," Proc. International Congress on Acoustics (ICA) 2004, pp.I-397–I-400, Kyoto, Japan, 2004.

[15] CMU, Carnegie Mellon Pronunciation Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[16] H. Chung, J. Park, H. Jeon, and Y. Lee, "Fast speech recognition for voice destination entry in a car navigation system," Proc. INTERSPEECH-2009, pp.975–978, Brighton, UK, 2009.

[17] O.W. Kwon, K.Y. Lee, Y.H. Roh, J.X. Huang, S.K. Choi, Y.K. Kim, H.B. Jeon, Y.R. Oh, Y.K. Lee, B.O. Kang, E.S. Chung, G.G. Park, and Y.K. Lee, "GenieTutor: A computer assisted second-language learning system based on spoken language understanding," Sixth International Workshop on Spoken Dialog System (IWSDS 2015), Pusan, South Korea, Jan. 2015.

[18] Y. Lim and Y. Lee, "Implementation of the POW (phonetically optimized words) algorithm for speech database," Proc. ICASSP-1995, Detroit, Michigan, USA, 1995.

Byung Ok Kang received his BS and MS degree in electrical and electronics engineering from the POSTECH, Rep. of Korea, in 1997 and 1999, respectively. From 1999 to 2001, he joined S/W Center for mobile phone application of Samsung Electronics. Since 2002, he has been a researcher at the Automatic Speech Translation and Artificial Intelligence Research Center of Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. Currently, he is a candidate for a PhD in control and robot engineering from Chungbuk National University (CBNU), Cheongju, Rep. of Korea. His research interests include speech recognition, speech signal processing and pattern recognition.

Oh-Wook Kwon received his BS degree in electronics engineering from Seoul National University, Rep. of Korea in 1986, and the MS and PhD degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1988 and 1997, respectively. From 1988 he was with Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. In 2000, he joined Brain Science Research Center at KAIST as a research professor. From 2001 to 2003, he worked for Institute for Neural Computation at University of California, San Diego as a postgraduate researcher. Since 2003, he has been a professor at Chungbuk National University, Cheongju, Rep. of Korea. His research interests include speech recognition, speech and audio signal processing, and pattern recognition.