PHONE SET CONSTRUCTION BASED ON CONTEXT-SENSITIVE ARTICULATORY ATTRIBUTES FOR CODE-SWITCHING SPEECH RECOGNITION

Chung-Hsien Wu, Han-Ping Shen and Yan-Ting Yang

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, R.O.C
{chunghsienwu, hanpinsheen, s800333 }@gmai.com

ABSTRACT

Bilingual speakers are known for their ability to code-switch or mix their languages during communication. This phenomenon occurs when bilinguals substitute a word or phrase from one language with a phrase or word from another language. For code-switching speech recognition, it is essential to collect a large-scale code-switching speech database for model training. In order to ease the negative effect caused by the data sparseness problem in training code-switching speech recognizers, this study proposes a data-driven approach to phone set construction by integrating acoustic features and cross-lingual context-sensitive articulatory features into distance measure between phone units. KL-divergence and a hierarchical phone unit clustering algorithm are used in this study to cluster similar phone units to reduce the need of the training data for model construction. The experimental results show that the proposed method outperforms other traditional phone set construction methods.

Index Terms— speech recognition, code-switching, articulatory attribute, phone set construction

1. INTRODUCTION

Due to globalization, the demand for code-switching automatic speech recognition (ASR) has been increased with time. Phone set construction for code-switching ASR plays an important role in code-switching speech recognition. However, it is difficult to collect sufficient number of code-switched utterances to train the acoustic models for code-switching ASR. On the other hand, code-switched utterances usually carry accents of speakers. Acoustic model training for code-switching ASR directly using the well-collected database from the native speakers is impractical. Data sparseness is a vital problem in code-switching ASR.

In recent years, multilingual or code-switching speech recognition phone set construction can be divided into three categories [1]. First, some approaches constructed one multilingual phone set by combining different languages phone sets directly. The disadvantage of this method is that the parameters and data of each defined phone model are not shared. With the increase of the number of languages, the number of phone models increases, too. This phenomenon will lead to the degradation of speech recognition performance and recognition speed. Second, multilingual phone set was constructed by mapping different languages phone sets into the same phone set according to expert knowledge. International Phonetic Alphabet (IPA), Speech Assessment Methods Phonetic Alphabet (SAMPA) and Worldbet are well-known phone sets defined by experts. This method can share the parameters in the acoustic models among different languages. However, this method does not take spectral characteristics into consideration. Third, some methods merged similar phone units of different languages into the same phone unit according to the spectral characteristics. Phone units with similar spectral properties are combined into one phone unit according to the likelihood or distance between two phone units. For example, [2][3][4]estimated the similarities among different phones and merged similar phones using a decision tree or confusion matrix. The distances between different phones can be calculated by applying Bhattacharyya Distance or Kullback-Leibler (KL) Divergence. This method takes spectral characteristics into consideration. However, if the collected corpus is not enough, data sparseness problem will occur and lead to generating unreliable acoustic models. Nowadays, triphone model can achieve good recognition accuracy if the collected data is sufficient. Although state-tying can partially ease the data sparseness problem, this problem still exists, especially for training the acoustic models using a code-switching corpus because the number of utterances from the second language are often much fewer than the number of utterances from the first language. In this study, we propose a systematic method which integrates traditional acoustic feature and context-sensitive articulatory attributes for Chinese-English phone set construction for code-switching ASR. As we know that the preceding phone of the current phone affects the pronunciation of the current phone. Furthermore, the current phone changes its partial pronunciation according to the articulation properties of the succeeding phone. In order to deal with the data sparseness problem, the articulatory attributes (AAs) which are well-known for the cross-lingual

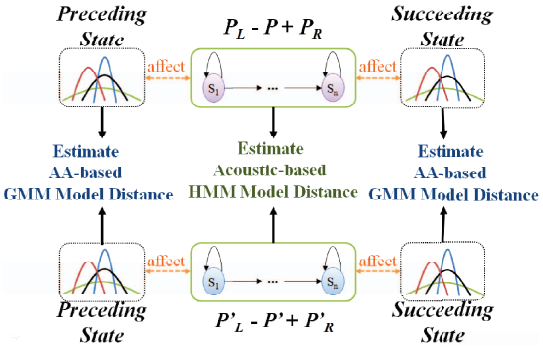

Figure 1: The idea of the proposed method

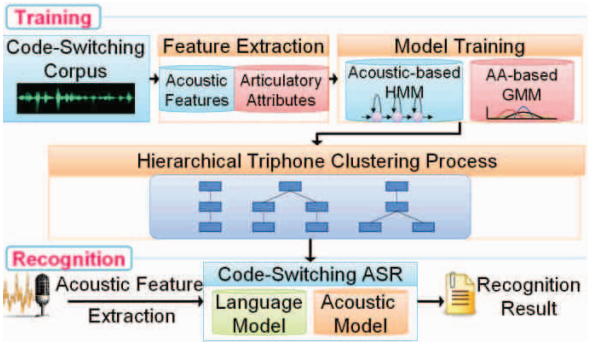

Figure 2: Code-Switching ASR System Diagram

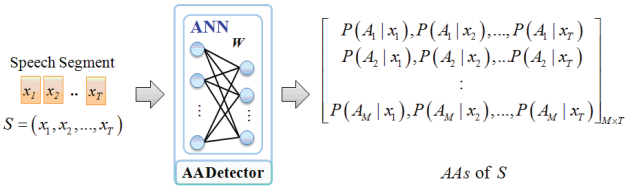

Figure 3: AA-based Detection Process

property are adopted for distance estimation between two phones. Different languages can share the same AA set. Even the state of the preceding/succeeding phones coming from different languages, if they have the same AAs, then they have similar effects on the spectrum of the current triphones. Therefore, the context-sensitive articulatory attributes of the last state of the preceding phone and the first state of the succeeding phone are taken into consideration in estimating the distances between two triphones. Figure 1 shows the idea of the proposed method. First, the distance between two triphones, P_L-P+P_R and P'_L-P'+P'_R based on acoustic-based HHMs is estimated. Second, the context-sensitive articulatory attributes are used to estimate the distances between the preceding/succeeding state AA-based GMMs of P_L-P+P_R and P'_L-P'+P'_R. KL-divergence is employed for model distance estimation. Instead of using acoustic-based distances to determine if two triphones are similar only, the AA-based GMM distances of the preceding and succeeding states are also considered to deal with the data sparseness problem and improve the reliability for distance estimation. Finally, a hierarchical phone unit clustering algorithm is used to cluster similar phone units to reduce the need of the training data for model construction.

In Section 2, the system diagram of the proposed system is presented. Feature extraction and model training are discussed in Section 3. In Section 4, a hierarchical triphone model clustering method is proposed. Experimental results are shown in Section 5. Conclusions are drawn in Section 6.

2. SYSTEM DIAGRAM

Figure 2 shows the system diagram of the proposed phone set construction method. In the training phase, the acoustic features and context-sensitive articulatory attributes from an intra-sentential code-switching corpus are extracted. Then, acoustic feature-based HMM and AA-based GMM are trained using the extracted features. A hierarchical triphone clustering process is applied to cluster similar triphone models. Finally, the resulting triphone models and code-switching language model are integrated for Chinese-English code-switching speech recognition.

3. FEATURE EXTRACTION AND MODEL TRAINING

In this study, similar phone units are clustered into a phone unit according to not only traditional acoustic features but also the articulatory attributes. 39-dimensional MFCCs are adopted as the acoustic features. These features are extracted from the current triphone speech segment. On the other hand, an AA detector [5] is adopted to extract the articulatory attributes of the speech frames in the preceding state and succeeding state of the current triphone speech segment. The AA detector, trained using a speech database with articulatory attribute labels, estimates the attribute likelihoods for each state-based speech segment. All attribute likelihoods of a state-based speech segment can be formed as an M×1 AA vector, where M is the number of AAs. In this study, 22 AAs are considered as the articulatory attributes including vowel, fricative, nasal, voiced, approximant, coronal, labial, back, stop, glottal, lo w, vocalic, dental, high, mid, continuant velar, anterior, retroflex, round, tense and silence. The values of the elements in the AA vector are the average likelihood of all the frames in a specific speech segment. Figure 3 shows the AA-based detection process.

The attribute detector is constructed using an Artificial Neural Network. Eq. (1) is used to estimate the likelihood of an articulatory attribute E for the speech frame corresponding to the state-based speech segment of the GMM.

$$p(y = E \mid X_i^S) = T^{-1} \sum_{t=1}^{T} \left(\frac{\exp(w_E^T z(t))}{\sum_{j=1}^{M} \exp(w_j^T z(t))} \right) \quad (1)$$

$$z(t) = [1 \quad x(t) \quad u(t)]^T \quad (2)$$

$$u(t+1) = (1+\exp(-v_E z(t)))^{-1} \quad (3)$$

where $z(t)$ is the input vector of a frame and composed of feature vector and the current state vector u(t) in the attribute detector. w_E and v_E are the weighting parameters for the output and the next state of the attribute detector, respectively. T is the total frame numbers in the speech segment. After feature extraction, MFCC features extracted from the triphone speech segments are used for acoustic model training and the context-sensitive attributes extracted from the preceding and succeeding states of the current triphone are used to train the AA-based GMMs.

4. HIERARCHICAL TRIPHONE MODEL CLUSTERING

After feature extraction, this study clusters similar triphones in a hierarchical manner. Distances among different triphones are calculated firstly. Kullback-Leibler Divergence (KL-Divergence) is applied to estimate the distances between different triphone acoustic models of the current triphones and AA-based GMMs for the states of their preceding and succeeding triphones. After calculating the similarity between each pair of triphones, this study merges triphone pair with the smallest distance into a triphone unit.

4.1. Distance between two triphones

The computational complexity for calculating all triphone pairs distances is very high. Thus, we only calculate distances among triphones having the same central phone mapping into the same IPA symbol. The distance between two triphones, X and Y, is estimated by Eq. (4)

$$D(X,Y) = w_{AC} D_{AC}(X,Y) + w_{AA} D_{AA}(X,Y) \quad (4)$$

where D_{AC} is the acoustic model distance between two current triphones. D_{AA} is the AA-based distance summation of the preceding and succeeding states between two triphones. w_{AC} and w_{AA} are the weighting factors for the acoustic models and articulatory attribute models. D_{AC} and D_{AA} are calculated by using KL-Divergence. Eq (5) defines the estimation of the KL-Divergence between two Gaussian distributions.

$$KL\left(N(\mu_1,\Sigma_1) \| N(\mu_2,\Sigma_2)\right) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} + Tr\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1-\mu_2)^T \Sigma_2^{-1}(\mu_1-\mu_2) - d\right) \quad (5)$$

where N_1 and N_2 are two different Gaussian distributions. μ and Σ are the mean and covariance, respectively. d is the dimension of the observed data. $D_{AC}(X,Y)$ and $D_{AA}(X,Y)$ can be estimated by using Eq. (6) and Eq. (7), respectively.

$$D_{AC}(X,Y) = \frac{1}{2}\left(\sum_{s=0}^{S} w_{x_s} KL(x_s \| y_s) + \sum_{s=0}^{S} w_{y_s} KL(y_s \| x_s)\right) \quad (6)$$

$$D_{AA}(X,Y) = \frac{1}{2}\left(D_{AAPre}(X,Y) + D_{AASuc}(X,Y)\right) \quad (7)$$

where x_s is the s-th mixture of X and y_s is the mixture of Y, which is closest to x_s. D_{AAPre} and D_{AASuc} can be denoted in Eq. (8) and Eq. (9), respectively.

$$D_{AAPre}(X,Y) = \frac{1}{2}\left(KL(X_{pre} \| Y_{pre}) + KL(Y_{pre} \| X_{pre})\right) \quad (8)$$

$$D_{AASuc}(X,Y) = \frac{1}{2}\left(KL(X_{suc} \| Y_{suc}) + KL(Y_{suc} \| X_{suc})\right) \quad (9)$$

where D_{AAPre} is the AA-based model distance between the preceding states of triphones X and Y. D_{AASuc} is the AA-based model distance between the succeeding states of X and Y.

4.2. Hierarchical Clustering

After calculating all the distances among all triphones, the closest triphone-pair will be merged. A hierarchical clustering algorithm is used to cluster the closest triphone-pairs iteratively. After clustering the closest triphone-pair in each iteration, the mean and covariance vectors of the new merged triphone will be updated by Eq. (10) and Eq. (11) instead of re-training the new triphone model directly [6].

$$\mu_{Zsi} = \frac{m_{Xs}\mu_{Xsi} + m_{Ys}\mu_{Ysj}}{m_{Xs} + m_{Ys}} \quad (10)$$

$$\Sigma_{Zsi} = \frac{m_{Xs}\left(\mu_{Xsi}^2 + \Sigma_{Xsi}\right) + m_{Ys}\left(\mu_{Ysj}^2 + \Sigma_{Ysj}\right)}{m_{Xs} + m_{Ys}} - \mu_{Zsi}^2 \quad (11)$$

where μ_{zsi} and Σ_{zsi} are the mean and covariance of the i-th mixture in the s-th state of the new merged triphone model, respectively. m_{Xs} and m_{Ys} are the observed data amount of triphone X and Y, respectively. μ_{Xsi} and μ_{Ysj} are the mean vectors of the i-th mixture in the s-th state of triphone X and Y, respectively. Σ_{Xsi} and Σ_{Ysi} are the covariances of the i-th mixture in the s-th state of triphones X and Y, respectively. The reason why we update the parameters is that re-training for each merged triphone-pair is time consuming and we have many triphone-pairs needed to be merged. Updating parameters by using this estimation can reduce the total computational cost without losing too much accuracy and increase the clustering performance in each iteration. After finishing parameter updating in each iteration, the distance between the new and the existing triphones will be calculated and the closest triphone-pair will be merged in the next iteration. The hierarchical process will be terminated until an optimal clustered triphone set has been obtained.

5. EXPERIMENTS

5.1. Corpus

In this study, the code-switching database CECOS [7] is adopted to construct the phone set and evaluate the performance of code-switching ASR. The corpus contains 6650 utterances totally. A large portion of the collected utterances is the Chinese-English code-switched utterances.

Of the 6650 utterances, 5985 and 665 sentences were selected as the training data and test data, respectively. We adopted the HTK (Hidden Markov Tool Kit) to build the code-switching speech recognition system. 39-dimensional MFCCs and 22-dimensional AAs were used as the acoustic features and articulatory attributes. Each phone was modeled by 3 states and each state contained 16 mixtures. State-tying technique was used. In language model training, TCC-300, TIMIT and CECOS, containing 16150 sentences totally, were used to train a bi-gram language model.

5.2. Experimental Results

Before triphone-pair clustering, the total number of triphone is 4171. The recognition rate of using the triphone set directly is 80.79%. When the hierarchical clustering process starts, the total number of triphones is decreased by 1 in each iteration. The recognition results of the proposed method are shown in Figure 4. The results are demonstrated every 100 iterations in the merging process. From the results, the best accuracy is about 84.27% when around 500 triphones have been clustered. After achieving the best accuracy, the recognition accuracies declined because over-clustering increases the ambiguity of the phone set. The recognition accuracy for the constructed phone set is compared with the recognition accuracy obtained by directly combining Chinese and English phone units into one phone set based on IPA definition. Table 1 shows these three results. The proposed method absolutely outperforms the other two methods by 3.48% and 2.96%, respectively. Furthermore, we compared the proposed method with the traditional data-driven method which calculates distances by using only acoustic features. The recognition accuracy of using only acoustic features is 81.53% which shows the accuracy of the proposed method is 2.74% better than the accuracy of the traditional data-driven method. Thus, considering the context-sensitive articulatory attributes of the preceding and succeeding states can help ease the data sparseness problem and increase reliability in estimating distances among triphones.

6. CONCLUSION

This study integrates traditional MFCC features and context-sensitive articulatory attributes into phone set construction for code-switching ASR. MFCC features are used for acoustic model training, while AAs of the preceding and succeeding states of a triphone are used to train the GMMs for distance estimation. A hierarchical triphone clustering process is used to cluster similar triphones into a triphone unit iteratively. KL-divergence is applied to estimate the distance between two different triphone models.

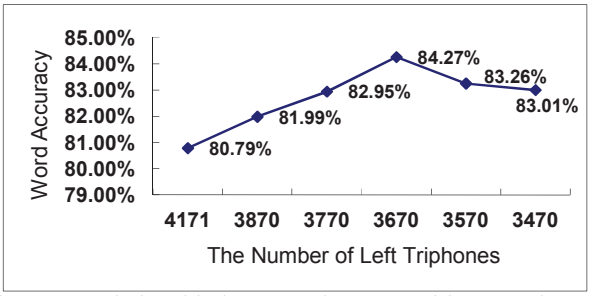

Figure 4: Relationship between the Recognition Results and the left number of triphones

Table 1: Recognition Results of using the Simply Combined Phone Set, IPA-based Phone Set and Proposed Method

Method	Directly Combined Phone Set	IPA-based Phone Set	Proposed Method
Recognition Rate	80.79%	81.31%	84.27%

The experimental results show that the proposed method outperforms other traditional methods, including directly combining Chinese and English phone models, mapping Chinese and English phone models into the same IPA phone set and estimating the triphone distances based on acoustic features, by 3.48%, 2.96% and 2.74%, respectively. The evaluation results confirm that the context-sensitive articulartory attributes can help eliminate the data sparseness problem and increase reliability in estimating distances among triphones.

7. REFERENCES

[1] S. Yu, S. Zhang, and B. Xu, "Chinese-English bilingual phone modeling for cross-language speech recognition," in *Proc. of ICASSP*, vol.1, Montreal, Canada, pp. I- 917-20, 2004.

[2] H.-P. Shen, J.-F. Yeh, and C.-H. Wu, "Speaker Clustering Using Decision Tree-based Phone Cluster Models with Multi-Space Probability Distributions," *IEEE Trans. Audio, Speech, and Language Processing,* Vol. 19, No. 5, July 2011, pp.1289~1300.

[3] C.-L. Huang and C.-H. Wu, "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis," *IEEE Trans. Computers*., Vol. 56, No. 9, September 2007, pp. 1245~1254.

[4] Y.-J. Chen, C.-H. Wu, Yu-Hsien Chiu, and Hsiang-Chuan Liao, "Generation of Robust Phonetic Set and Decision Tree for Mandarin Using Chi-square Testing," *Speech Communication,* Volume 38, Issues 3-4, November 2002, pp 349-364.

[5] S.-M. Siniscalchi, T. Svendsen and C.-H. Lee, "Toward A Detector-Based Universal Phone Recognizer," In *Proc. of ICASSP*, Las Vegas, USA, pp.4261-4264, 2008.

[6] Y. Qian and J. Liu, "Phone Modeling and Combining Discriminative Training for Mandarin-English Bilingual Speech Recognition," in *Proc. of ICASSP*, Dallas, USA, pp.4918-4921, 2010.

[7] H.-P. Shen, C.-H. Wu, Y.-T. Yang and C.-S. Hsu, "CECOS: A Chinese-English Code-Switching Speech Database," in *Proc. of Oriental-COCOSDA*, Hsinchu, Taiwan, 2011