ACOUSTIC MODELING USING AN EXTENDED PHONE SET CONSIDERING CROSS-LINGUAL PRONUNCIATION VARIATIONS

Dau-Cheng Lyu [1], Ren-Yuan Lyu [2], Ming-Tat Ko [3]

[1] Dept. of Electrical Engineering, Chang Gung University, Taiwan
[2] Dept. of Computer Science and Information Engineering, Chang Gung University, Taiwan
[3] Institute of Information Science, Academia Sinica, Taiwan
d9221003@stmail.cgu.edu.tw, renyuan.lyu@gmail.com, mtko@iis.sinica.edu.tw

ABSTRACT

To deal with the issue of data unbalanced condition among a task of multilingual speech recognition and a phenomenon of pronunciation variations across languages, we propose an approach to clustering context dependent phones from an extended phone set in an acoustic model trained on a data unbalanced bilingual corpus. First, we generate an extended phone set using pronunciation modeling by a confidence measure between Mandarin and Taiwanese. Second, we use a two-step agglomerative hierarchical clustering with delta Bayesian information criteria to automatically generate a merged extended phone set (MEPS). Third, we choose a parametric modeling technique, model complexity selection, to increase the final number of Gaussian components dependent on the available training data in a data unbalanced condition. The experimental results show that the proposed automatic extending phone clustering approach reduced relative syllable error rate by 8.3% over the best result of the decision tree based phone clustering approach.

1. INTRODUCTION

In Taiwan, over 75% of the population speak at least two languages/dialects - Mandarin and Taiwanese - in their daily conversations. Due to the different mother tongues of the speakers, the problem of pronunciation variations becomes an important issue in multilingual speech recognition [1]. Among them, allophonic variation, changing the phoneme from a base-form to the surface form, is perhaps the most common. [2]. For example, for a Chinese word "魚肉" (fish meat), all of the Mandarin native speakers would lexically pronounce it as /yu-rou/. However, some of the Taiwanese native speakers are apt to pronounce it as /i-lou/ (the same pronunciation of "遺漏" (to miss)) due to the lack of phonemes /yu/ and /r/ phones in their mother-tongue language.

Some researchers suggest extending the standard phone set to deal with phonetic confusion problem in training a robust acoustic model of an automatic speech recognition (ASR) task [3]. Taken the confusing pair of "yu-rou, i-lou" as an example, the extended phones are generated as /yu-i/ and /r-l/ for the phonetic confusion, where /yu-i/ represents an extending phone to deal with /yu/, /i/ confusion, and /r-l/, the /r/, /l/ confusion. In such an approach, we can more specifically describe the possible acoustic space, and we believe that the issue of the allophonic variations across languages will become more and more important because of the multilingual tendency in globalization.

On the other hand, in order to compensate for the allophonic variations; it is money and time consuming approach to generate extending phones in a large scale corpus via manual effort. Therefore, we choose an automatic approach, where we employed a well trained recognizer to replace the human recognition [2]. However, the accuracy of the automatic way only relies on the performance of the recognizer, and the performance of the state-of-the-art ASR still can not achieve 100% so far. Besides, even human beings have different annotations of the same sound. Therefore, the further processing after the automatic approach is indeed necessary.

To extend the phone set increases the total number of the original existing phone set, such as IPA (International Phonetic Alphabet) [4], and it results in the reduction of available training data of each phone. Besides, in our case, the training data of the minor languages, such as Taiwanese, is much less than that of the major language, Mandarin. In fact, the total quantity of Taiwanese speech data is about half that of Mandarin. In order to solve the data shortage problem, an approach, clustering phones which they are similar, should be taken. In such approach, the final total number of the acoustic units is reduced thus the parameters in the acoustic model can obtain relatively sufficient data for estimating [5].

There are two main methods of phone clustering for training acoustic models, i.e., the knowledge-based and the data-driven approach. The former is to utilize a universal phone set to build a language-independent speech recognizer based on phonetic knowledge to construct the multilingual phone inventory [6, 7]. The latter is to merge the similarity phones according to some specific distance measure between acoustic models [8, 9]. Although the researcher in [9] claims that the data-driven approach can find the better result by trying several different thresholds of the confidence measure to merge the similar phones, we can not automatically obtain an optimal phone set in the clustering process.

In this paper, we proposed an approach to automatically cluster an extended phone set via delta Bayesian information criteria (delta BIC) [8] on an unbalanced Mandarin and Taiwanese bilingual corpus. First, according to an aspect of pronunciation variations across languages, we generate an extended phone set from a confusion matrix with language dependent context independent phone (LD-CIP). Second, we re-train the acoustic models with a partial re-labeled extended phone set. Then, we rank the distance among the phones via similarity measure, Bhattacharyya distance. Third, a two-step acoustic model phone clustering criteria is

adopted. In the first step, we generate the LD-CIP clustering rules via hierarchical agglomerative clustering (HAC) [10] and delta-BIC, and the results of the first-step outcomes will be as the phonetic constraints to guide the next step clustering procedures. In the second step, we merge the extended phone set with the language dependent context dependent phone (LD-CDP) also via HAC and delta-BIC. Finally, we re-train the merged extended phones using hidden Markov model (HMM) with a model complexity selection (MCS) [11] for increasing the number of Gaussian model depending on the occurrence of the available training data.

2. GENERATING AN EXTENDED PHONE SET

In order to specifically describe the sounds which are easily confused, such as those caused by pronunciation variations, we suggest to generate an extended phone set, which could contain all pronunciation variations. The reason is that some of the easily confusing sounds are ambiguous and they could not belong to any original phonemes. Therefore, we could create a new phoneme to represent those easily confusing sounds. The procedure of generating the extended phone set is shown in Figure 1. From left to right, they can be divided into three steps: phone recognition, alignment a pair of base-form and surface form of a phone and confidence measurement.

First, a bilingual training corpus with base-form transcription, T_b, is used to build a bilingual recognizer. Then, the phone recognition on the training corpus is performed to obtain the recognized transcription, T_s. After that, we use a dynamic programming algorithm to generate T_b-T_s pair. According to the results we then try to find the candidates of pronunciation variations across the languages. Finally, we use a confidence measure along with pronunciation ranking and pruning to get an extended phone set. Table 1 shows the top 10 confusing phones which are put in the extended phone set, such as /uh_T=u_M/. It means that the created phones are new phones to represent the sounds that are very ambiguous between uh_T and u_M.

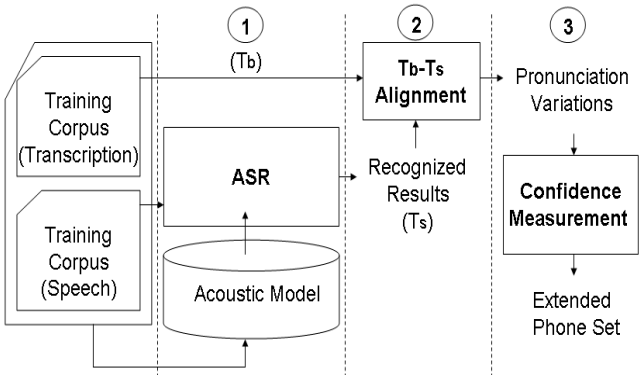

Figure 1. The procedures of generating an extended phone set.

	Baseform	Surface form		Baseform	Surface form
1.	uh_T	u_M	6.	it_T	i_T
2.	err_M	er_M	7.	ak_T	a_M
3.	ih_T	i_T	8.	gh_T	g_T
4.	zh_M	z_M	9.	ih_T	i_M
5.	oh_T	ok_T	10.	ah_T	a_M

Table 1. The top 10 outcomes of the base and surface form alignment pair according to the confidence measurement.

3. EXTENDED PHONES MERGING VIA DELTA BIC

We propose a method to automatically cluster the elements in the extended phone set according to the procedures of a phone similarity measure and a two-step phone clustering criteria. After that, we use a model complexity selection to train a HMM-based acoustic model which keeps balance between the available training data and the number of the merged extended phones. The procedures are shown in Figure 2, and the details are described in the following two subsections.

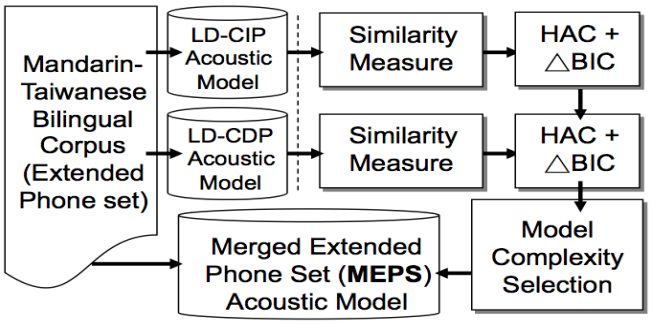

Figure 2. The procedures of training a merged extended phone set acoustic model via similarity measure and a two-step clustering criteria.

3.1. Phone Similarity Measuring

We use the Bhattacharyya distance to measure the similarity of HMM-based acoustic phone model. The Bhattacharyya distance is similar to a likelihood ratio test to evaluate a similarity of each language-dependent, context-independent phone (LD-CIP) and language-dependent context-dependent phone (LD-CDP) acoustic models. It is a theoretical distance between two Gaussian distributions, and proven to be equivalent to an upper bound on the optimal Bayesian classification error probability [5]. In this stage, the number of the Gaussian components in a Gaussian mixture in each state of a HMM acoustic model is set to be one. We gave a brief review here with the following equation and its notations.

$$D_{pqi} = \frac{1}{8}(u_{pi} - u_{qi})^T \left[\frac{\Sigma_{pi} + \Sigma_{qi}}{2}\right]^{-1} (u_{pi} - u_{qi}) + \frac{1}{2}\ln \frac{\left[\frac{\Sigma_{pi} + \Sigma_{qi}}{2}\right]}{\sqrt{|\Sigma_{pi}||\Sigma_{qi}|}} \qquad (1)$$

where D_{pqi} is the Bhattacharyya distance between p^{th} and q^{th} phonemes in i^{th} state. u_{pi} is the mean vector of p^{th} phoneme in i^{th} state, and Σ_{pi} is the covariance matrix of the p^{th} phoneme in i^{th} state.

The first term of the right side in equation (1) discriminates the class due to the difference between class means, while the second term discriminates the class due to the difference between class covariance matrices.

3.2. Extended Phone Set Merging

In order to automatically merge the extended phone set of the bilingual acoustic model, we employ a two-step clustering criteria by using the HAC and delta BIC to guide the direction of phone clustering based on a similarity matrix. The first step is to use the LD-CIP acoustic models to generate knowledge-like rules as the phonetic constraints, and the second step is to generate the merged

extended phone set from clustering the LD-CDP. Each of the merged LD-CDP shares the available training data. After the models are merged via delta BIC, the models could be probably over-merged or under-merged. Therefore, we next use model complexity selection to get a balance between the demands of resolution of acoustic models and the amount of available training data.

Because the Taiwanese speech corpus is only half of that of the Mandarin, we use a data-driven approach to obtain the knowledge-like rules to replace the knowledge sources. These rules are generated with LD-CIP level and we use them to constrain the LD-CDP clustering. In order to obtain those rules, we adopt HAC and delta BIC.

HAC is a bottom-up clustering method where the bottom nodes are the LD-CIP. There are several algorithms to evaluate the distance of each cluster in HAC, such as the centroid-linkage, average-linkage, and complete-linkage agglomerative algorithms. In this paper, due to its ability to get the best agglomerative coefficient, measuring the clustering structure of the phoneme set, we use the average-linkage agglomerative algorithm with Euclidean distance to construct the hierarchical tree from the similarity matrix.

Delta BIC is the confidence measure to cluster the similar LD-CIP and LD-CDP from the HAC results. Before we describe delta BIC, we should introduce BIC. BIC is an asymptotically optimal Bayesian model selection criterion. It is used to decide which of m parametric models best represents n data samples $x_1, \ldots x_n$, where $x_n \in R^d$. Each model M_i has a number of parameters k_i. We assume that all the samples x_n are statistically independent. According to BIC theory [12], for sufficiently large n, the best state of the data is the one which maximizes.

$$BIC_i = \log \ell_i(x_1,\ldots,x_n) - \frac{1}{2} k_i \log n \qquad (2)$$

where $l_i(x_1,\ldots,x_n)$ is the likelihood of the data under the model M_i.

In our case, according to the HAC structure, we select the nearest two nodes for model merging: choose the model M_p over M_q if delta BIC defined as $BIC_p - BIC_q$ is positive. Based on Equation (3), the formula of △BIC is written as:

$$\Delta BIC = -\frac{n_p}{2}\log|\Sigma_p| - \frac{n_q}{2}\log|\Sigma_q| + \frac{n_r}{2}\log|\Sigma_r| + \frac{1}{2}\left(d + \frac{d(d+1)}{2}\right)\log n_r \qquad (3)$$

where n_p , n_q and n_r are the number of occurrences of node p, q and r where n_r equals n_p adding n_q. Σ_p , Σ_q and Σ_r are the covariance of the model p, q and r respectively. d is a number of the dimension of the model.

3.3. Model Complexity Selection (MCS)

The number of Gaussian mixtures should be carefully controlled according to the amount of training data available for each state. We choose MCS from [10] to select the number of Gaussian components according to the number of frames of data available for that state. MCS works as follows: whenever there is a change in the amount of data assigned to a model, the number of the available training samples that are assigned to the model is used to determine the new number of mixtures in the Gaussian components in a Gaussian mixture using:

$$M_{pi} = round\left(\frac{N_{pi}}{OR}\right) \qquad (4)$$

where M_{pi} is the number of Gaussian components in a Gaussian mixture of the p^{th} acoustic model of the i^{th} state, and it is determined by the amount of the training data belonging to that model at that occurrence N_{pi} divided by the occurrence ratio (OR) where OR is a constant value across all training process. Therefore, the final number of Gaussian components in a Gaussian mixture for the p^{th} acoustic model of the i^{th} state depends on the amount of the corresponding occurrence training data.

4. EXPERIMENT

4.1. Unbalanced Bilingual Speech Corpus

Corpus is divided into two parts: training and test set. In the training set, it includes 100 speakers in Mandarin and 50 speakers in Taiwanese. The total length of the data in Taiwanese is only half of that in Mandarin, and each of the speakers recorded two sets of phonetically balanced utterances. In the test set, 20 speakers recorded the total length 0.56 hours speech data. The information of the corpus is listed in table 2.

	language	number of speakers	number of utterances	speech length (in hours)
Training Set	Mandarin	100	43,078	11.3
	Taiwanese	50	23,009	5.6
Test Set	Mandarin	10	1,000	0.28
	Taiwanese	10	1,000	0.28

Table 2. Statistics of the training and test bilingual speech corpus.

4.2. Experimental Environment Setting

For the feature extraction, each frame of short-time speech waveform is represented by a feature vector consisting of 12 mel-frequency cepstral coefficients (MFCCs), energy, their first order derivatives (delta coefficients) and second order derivatives (delta-delta coefficients). HMM is employed to train acoustic models, and each of which has three states. For the language modeling, 0-gram, a uniform distribution of base syllable is used, which implies the perplexity of the language model to be 924.

A series of the experiments to validate the proposed method are performed, so that the experiments are divided into three parts: I: baseline, II: automatic validation, and III: merged extended phone set. First, the acoustic model, LDCD, is trained to compare the performance with that of an acoustic model with MCS, LDCD-MCS. Besides, a widely used clustering algorithm, decision tree approach (DT-MCS), is also employed to be as a baseline performance compared with that of the merged extended phone set.

Second, three acoustic models, PC-3-MCS, PC-7-MCS and PC-10-MCS, using automatic ways to validate the consistency between the speech data and transcription, are trained to compare the results with those of the original transcription in the training set. Basically, each of the transcriptions of the three acoustic models is revised by the acoustic model, LDCD-MCS with a 0-gram language model. The syllable accuracy evaluated on the training data is 84.63%. After that, we choose three confidence measures to select the candidates of the pronunciation variations from the phone-based confusion matrix to re-transcribe partial changes in the training set. The total percentages of phones with label changes are 3%, 7% and 10% for PC-3-MCS, PC-7-MCS and PC-10-MCS, respectively.

The third experiment is the proposed method which automatically extracts the merged extended phone set (MEPS) from the acoustic model of the second experiment. The number of the phones in the extended phone sets of the proposed three acoustic

models, MEPS-3-MCS, MEPS-7-MCS and MEPS-10-MCS are 82, 95, and 101, respectively, where the number of the original language-dependent phones is 67. After the first step of the CIP clustering of PC-10-MCS for example, we obtain 34 phone rules to constrain the second CDP clustering, and the rules are such as: [ok_T, op_T] and [ng_M, ng_T]. In other words, the CDP containing these CIP should be put into one pool and then to automatically merge by the second step mentioned in the section 3.2.

4.3. Baseline Results

All the baseline results are shown in Table 3. First, a knowledge-based best result LDCD using IPA to transcribe the training labels gets 67.9% and 51.7% syllable accuracy for Mandarin and Taiwanese, respectively. Because of the data unbalanced condition in the training set, the gaps between these two languages we get about 16%. Second, in order to shorten the gap, we use the acoustic model with MCS, LDCD-MCS, and then we get 1.5% average syllable rate increase. As shown we can see, most of the improvements in this experiment are in Taiwanese. It points out that MCS which increases the number of Gaussian components in a Gaussian mixture depends on the available training data improves the ASR performance in the data unbalanced condition. On the other hand, we use decision tree CDP clustering to train the model, DT-MCS, and we get 62.7% syllable accuracy rate in average.

	Mandarin	Taiwanese	Average
LDCD	67.9%	51.7%	59.8%
LDCD-MCS	68.5%	54.1%	61.7%
DT-MCS	69.7%	55.6%	62.7%

Table 3. The baseline results (syllable accuracy rates) of LDCD, LDCD-MCS and DT-MCS.

4.4. Automatic Validation Results

In Table 4, we demonstrate the results of using an automatic validation with a recognizer. Comparing the performances of PC-7-MCS and PC-10-MCS with LDCD-MCS, we get better performance, but the increasing syllable accuracy rates are too small. Besides, we obtain worse result when we only change only 3% in the phone labeling (comparing PC-10-MCS with PC-7-MCS). Therefore, it is difficult for us to obtain the outstanding performance when validating the transcription of the training data in an automatic approach.

	Mandarin	Taiwanese	Average
PC-3-MCS	68.1%	51.7%	59.9%
PC-7-MCS	69.0 %	55.0 %	62.0%
PC-10-MCS	69.2 %	54.8%	62.0%

Table 4. The syllable accuracy rates of the automatic validation with different conditions.

4.5. Results of Merged Extended Phone Set (MEPS)

By using the extend phone set to automatic cluster the CDP with different experiment conditions, we show the results in Table 5. First, all of the performances of the optimal extended phone set are better than those of DT-MCS. Especially for the aspect of Taiwanese, the ASR improvements are very obviously. Taking OEPS-10-MCS as an example, we get 5.4% syllable accuracy increasing rate comparing with DT-MCS. It designates that our proposed optimal extended phone set improves the performance not only in the language with sufficient training data but also in the language with insufficient training data.

	Mandarin	Taiwanese	Average
MEPS-3-MCS	69.4%	57.6%	63.5%
MEPS-7-MCS	69.7%	60.0%	64.8%
MEPS-10-MCS	70.6%	61.0%	65.8%

Table 5. The syllable accuracy rates of the proposed extended phone set merging method with several criteria.

5. CONCLUSION

In this paper, we have achieved a better performance using an automatic phone clustering approach from an extended phone set on an unbalanced corpus. The proposed method, a two-step constrained phone clustering criteria, was used an agglomerative hierarchical clustering and delta Bayesian information criteria to automatically extract the merged extended phone set from several transcription using automatic validation with different conditions. Although, we had a slight drop performance in the automatic validation of the training data, we finally got 8.3% relative syllable reduction error rate comparing with DT-MCS.

6. REFERENCES

[1] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang and Chun-Nan Hsu, "Modeling Pronunciation Variation for Bi-Lingual Mandarin/Taiwanese Speech recognition," International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10, No. 3, Sep. 2005, pp. 363-380.

[2] Jurafsky, D. and Martin, J. H., *2000*, Speech and Language Processing, pp. 91-188. Prentice-Hall, New Jersey

[3] W. Byrne, V. Venkataramani, T. Kamm, T.F. Zheng, Z. Song, P. Fung, Y. Liu, U. Ruhi, "Automatic Generation of Pronunciation Lexicons for Mandarin Spontaneous Speech," ICASSP, 2001

[4] IPA, "Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet", Cambridge University Press, 1999.

[5] Brian Mak and Etienne Barnard, "Phone Clustering Using the Bhattacharyya Distance," ICSLP 1996.

[6] Joachim Kohler, "Multilingual Phone Model for Vocabulary-Independent Speech Recognition Task," Speech Communication, Vol. 35, Issue 1-2, 2001, pp. 21-30.

[7] C. Santhosh Kumar, V.P.Mohandas, Li Haizhou, "Multilingual Speech Recognition- A Unified Approach," InterSpeech 2005.

[8] Chien-Lin Huang and Chung-Hsien Wu, " Phone Set Generation Based On Acoustic and Contextual Analysis for Multilingual Speech Recognition," ICASSP 2006.

[9] Liu Yi and Pascale Fung, "Automatic Phone Set Extension with Confidence Measure for Spontaneous Speech," InterSpeech 2005.

[10] Fowlkes, E. B. and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," Journal of the American Statistical Association 78(383), 1983.

[11] Xavier Anguera Miro, Takahiro Shinozaki, Chuck Wooters, and Javier Hernando, "Model Complexity Selection and Cross-validation EM Training for Robust Speaker Diarization," ICASSP 2007.

[12] Schwarz, G., "Estimating the Dimension of a Model," The annals of statistics, 6(2), 1978.