OPTIMIZING THE ACOUSTIC MODELING FROM AN UNBALANCED BI-LINGUAL CORPUS

Dau-cheng Lyu [1], Ren-yuan Lyu [2]

1 Dept. of Electrical Engineering, 2 Dept. of Computer Science and Information Engineering,
Chang Gung University, Taiwan
{daucheng, renyuan.lyu}@gmail.com

ABSTRACT

Phoneme set clustering of accurate modeling is important in the task of multilingual speech recognition, especially when each of the available language training corpora is mismatched, such as is the case between a major language, like Mandarin, and a minor language, like Taiwanese. In this paper, we present a data-driven approach for not only acquiring a proper phoneme set but optimizing the acoustic modeling in this situation. In order to obtain the phoneme set that is suitable for the unbalanced corpus, we use an agglomerative hierarchical clustering with delta Bayesian information criteria. Then for training each of the acoustic models, we choose a parametric modeling technique, model complexity selection, to adjust the number of mixtures for optimizing the acoustic model between the new phoneme set and the available training data. The experimental results are very encouraging in that the proposed approach reduces relative syllable error rate by 7.8% over the best result of the knowledge-based approach.

Index Terms—phoneme set clustering, delta-BIC, multilingual speech recognition

1. INTRODUCTION

Multi-lingual speech recognition is a popular research topic in the speech recognition field [1-5]. Most multi-lingual speech recognition depends on a large-scale speech database for each language in order to train the acoustic models well. However, such abundant speech corpora are not always available for all the languages under consideration. Some major languages, such as English, Mandarin, French, Deutsch, Japanese and Spanish, may have abundant speech corpora, which could be used in training reliable acoustic models. In contrast, certain minor languages, like Taiwanese, may not have much speech corpora available [6]. Collecting a well-designed, large-scale speech corpus for every language under consideration is not feasible, so the fundamental motivation of this paper is to find an approach which can adopt the speech corpora available for the major languages to help build reliable acoustic models for the minor languages.

Several approaches which utilized a universal phoneme set for multilingual speech recognition have been proposed. One approach is to map a language-dependent phone set to a global inventory of the multilingual phonetic phone set based on phonetic knowledge to construct the multilingual phone inventory [1-4]. The advantage of this approach is that the same phonetic representation with different languages shares the training data. However, this type of approach is based only on the phonetic knowledge. It does not consider the spectral properties of the phone models. This may be the disadvantage. Another approach is to merge the language-dependent phones using a data-driven approach, such as a hierarchical phone clustering algorithm, according to some specific distance measure between acoustic models [7-9]. The advantage of this type of approach is that the distance is estimated from the statistical measure of similarity of real recognition models, which may be more appropriate. Nevertheless, most proposed approaches of this type do not consider optimizing the number of the phones in a phone set. They used several heuristic thresholds as the criteria to stop the merging or splitting process, and then chose the best one according to the local optimal performance obtained from several well trained acoustic models.

In this paper, we use a data-driven approach to choose the best phoneme set so that the bi-lingual/multilingual speech recognition system can achieve optimal recognition accuracy. For the first phase of our approach, we use a statistical distance measure, Bhattacharyya distance [9], to map phonemes across languages according to well-trained acoustic models. Then we adopted a 2-step clustering, which used the agglomerative hierarchical clustering (HAC) [10] with delta Bayesian information criteria(ΔBIC) [11] to guide phone clustering based on the distance measure mentioned above. Since the training speech data is unbalanced between the languages considered in this paper, Mandarin and Taiwanese, the goal of this step is to generate a data-driven rule set from the corpus with language independent (LI), context independent phoneme (CIP) units; we then use these rules to constrain the second step of LI context dependent phoneme (CDP) clustering. After the clustering, we generate the optimal phoneme set (OPS) which with the same unit will share the training data. Finally to optimize the OPS acoustic model, we used a model complexity selection (MCS) to adjust the number of mixtures for balance between the OPS and the available training data.

2. SOME PHONETIC INFORMATION FOR - TAIWANESE AND MANDARIN

For using a major language, like Mandarin, to assist a minor language, like Taiwanese, in acoustic modeling, basic knowledge about the phonetics, phonology and other linguistic aspects of the two languages are essential. Taiwanese (also called Min-nan in linguistic literatures) and Mandarin are two members of Sino-Tibetan language family. They are monosyllabic languages and they share the same written system, the Chinese character. Taiwanese is a language member in the Chinese language family, but it is quite

unintelligible to people speaking only Mandarin. We examined some aspects of the linguistic properties to demonstrate how they are different with each other.

In the phonemic level, there are 21 consonants and 9 vowels in Mandarin, while there are 16 consonants and 11 vowels in Taiwanese. Some of the phonemes in the two languages are labeled with the same symbols by phoneticians, meaning that they are phonetically very close. In our examination, only 18 phonemes are in common.

In the syllabic level, although both languages have similar CVC (C: Consonant, V: Vowel) structure, Taiwanese has much more abundant variations in the syllable-ending consonants, including -p, -t, -k, -h, and so on. Totally, there are 408 CVC syllables in Mandarin, and 709 CVC syllables in Taiwanese. The numbers of the union set of the syllables in both languages are 924, and only 189 syllables are in common. The information mentioned above is summarized in table 1.

	M	T	$M \cup T$	$M \cap T$
N_p	30	37	49	18
N_S	408	709	924	189

Table 1. The statistic information of all Mandarin (M) and Taiwanese (T) linguistic units in two levels: the numbers of phoneme (N_p), the numbers of syllables (N_S), where $\cap$and$\cup$ represent intersection and union of sets, respectively.

3. OPTIMIZING ACOUSTIC MODELS

We proposed a system, which is composed of 3 main steps to obtain an optimal set of bilingual acoustic models. The overall diagram is shown in figure 1. First of all, we trained a set of hidden Markov models (HMM) based on LD phoneme sets which include both CIPs, and CDPs. Then, the Bhattacharyya distance is used to evaluate a similarity of each phone model in the LD acoustic models. The Bhattacharyya distance is a theoretical distance measure between two Gaussian distributions. It is said to be equivalent to an upper bound on the optimal Bayesian classification error probability [9]. We could give a brief review here with the following equation and its notations.

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

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 $\sum_{pi}$ 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.

In order to generate the optimal phoneme set (OPS) of the bilingual corpus, we employ two steps of clustering by using the HAC and to guide the direction of phone clustering based on a similarity matrix. The first step is to use the LI-CIP acoustic models to generate the data-driven rules as the phonetic constraints, and the second step is to generate the OPS from clustering the LI-CDPs. Each of the merged LI-CDP shares the available training data. After the models were merged via ΔBIC, 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. In the following subsections, we will describe these three steps in more details.

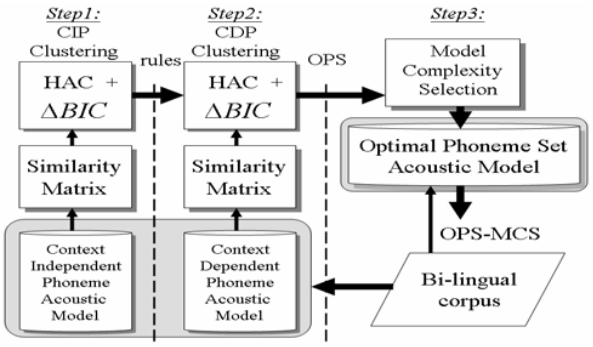

Figure 1. The overall diagram for automatically optimizing the acoustic models

3.1. CIP Clustering

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

HAC is a bottom-up clustering method where the bottom nodes are the LD-CIP. We use the single-linkage agglomerative algorithm with Euclidean distance to construct the HAC tree from the similarity matrix. Then, we employ ΔBIC as the confidence measure to cluster the "similar" CIP from the bottom nodes to the top nodes.

Before we describe ΔBIC, we introduce Bayesian information criteria (BIC). BIC is an asymtotically optimal Bayesian model-selection criterion used to decide which of m parametric models best represents n data samples $x_1,...,x_n, where\ x_t \in R^d$. Each model M_i has a number of parameters k_i. We assume that all the samples x_t are statistically independent. According to the BIC theory [12], for sufficiently large n, the best state of the data is the one which maximizes.

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

where $\ell_i(x_1,...x_n)$ is the likelihood of the data under the model M_i .λ is a constant value, and we used 1.

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 ΔBIC (defined as $BIC_p - BIC_q$) is positive. Based on equation (2), the formula of ΔBIC is written as:

$$\Delta BIC = -\frac{n_p}{2}\log|\sum_p| - \frac{n_q}{2} lpg|\sum_q| + \frac{n_r}{2}\log|\sum_r| + \frac{1}{2}\lambda(d + \frac{d(d+1)}{2})\log n_r \qquad (3)$$

where n_p, n_q and n_r are the number of occurrences of node p, q and r; $\sum_p, \sum_q$ and $\sum_r$ are the covariance of the model p, q and r respectively. The results of using to merge Mandarin and Taiwanese LD-CIPs are demonstrated in figure 2, where the squared CIPs, such as /ak_T/ and /ap_T/ (the 4th square from the right side), are merged.

3.2. CDP Clustering

We set the results of clustering CIPs as the rules to constrain the CDP clustering. We take the merging process for phonemes /*-ak_T/ and /*-ap_T/ as an example, which is shown in figure 3. First of all, we generated a hierarchical tree whose bottom nodes include all CDPs containing /ak_T/ and /ap_T/ in their phonemic transcription by using HAC algorithm based on the similarity matrix.

Then we used to find the optimum CDP set. The merged CDPs share the available training data. If is negative, the merging process stops.

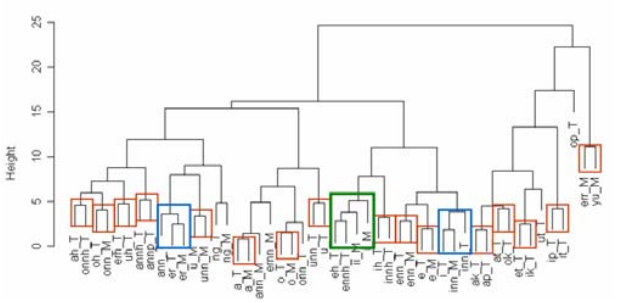

Fig. 2. The results of using HAC and of Mandarin and Taiwanese LD-CIP.

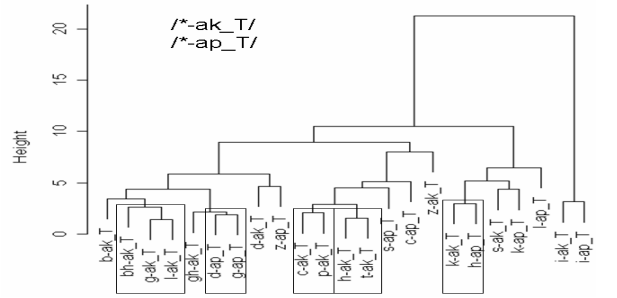

Fig. 3 The merging result of LD-CDP, /*-ak_T/ and /*-ap_T/.

3.3. Model Complexity Selection (MCS)

After we optimized the phoneme set by using the constraints of HAC and ΔBIC, we considered the balance between the resolution of the generated acoustic models and the amount of available training data. Each state of an acoustic model contains several Gaussian mixtures. According to [14], a parametric modeling technique, model complexity selection, is chosen to perform on each state with sufficient training data, and we select the number of mixtures based on the number of data frames belonging to the phoneme state in the acoustic model. Model complexity selection 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 GMM using:

$$M_{pi}^{j} = round(\frac{N_{pi}^{j}}{OR * j})$$

where M_{pi}^{j} is the number of Gaussian mixtures of the p^{th} acoustic model of the i^{th} state at iteration j , and it is determined by the amount of the training data belonging to that model at that occurrence N_{pi}^{j} divided by the j^{th} iteration multiplying the occurrence ratio (OR) where OR is a constant value across all training process.

4. EXPERIMENT

4.1. Corpora

The training corpus contained both Mandarin and Taiwanese speech, including 100 speakers in Mandarin and 50 speakers in Taiwanese. The total length of the training speech in Taiwanese was about half of that in Mandarin. Each of the speakers recorded two sets of phonetically balanced utterances, each of which contained one to four syllables. Another 20 speakers recorded the test data, and the total length was about 0.56 hours. 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 testing bi-lingual speech corpus.

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). For the acoustic modeling, HMM is used to train acoustic models, each of which has three states. For the language modeling, a uniform distribution is used, which implies the perplexity of the language model to be 924. In the configuration of the acoustic models, the numbers of LD-CDP and LI-CDP are 1503 and 1242, respectively. A set of 19 rules from the first step of CIP clustering is applied, and then a set of 1083 phonemes is obtained from the second step of CIP clustering.

A series of the experiments to validate the proposed method were performed, and the acoustic models could be divided into two parts. In the first part, we used the knowledge-based approach. In this approach, the acoustic models represent phonemes of both LD-CD and LI-CD phone set, transcribed based on IPA. Each of the models could probably use MCS or not. Next, we use the test data to evaluate the trained acoustic models then can realize the baseline knowledge-based results and also analyze the influence of the MCS on the unbalanced corpus.

In the second part, in order to analyze the influence of the rules which have been generated from the unbalanced corpus by the data-driven approach, we trained two acoustic models, CDP-MCS and OPS-MCS. The CDP-MCS only used HAC and once to generate a new CDP set. The latter OPS-MCS uses the 2-step clustering of HAC and ΔBIC to generate a new phoneme set which is also the main proposed method in this paper.

In addition, we compared our approach with the decision tree-based tri-phoneme clustering method with MCS (DT-MCS). In that method, the tri-phoneme nodes are placed in the root of the decision tree, and each node of the tree is associated with a binary question which has been selected from a set derived by linguistic experts. The best question is assigned to a node if it results in a binary splitting with minimal loss of likelihood [13]. We use the MCS for all of the acoustic models throughout all of the second part experiment, and the experiment diagram is shown in figure 4.

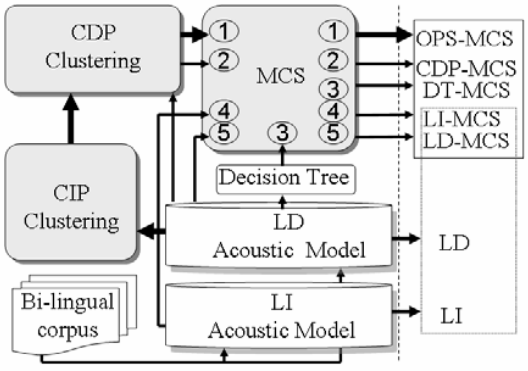

Fig. 4. The experiment diagram.

4.2. Baseline Results

The experimental results of the first part are illustrated in table 3. The best performance in terms of syllable accuracy of both the baseline LI and LD are 60.1% and 59.8%, respectively. On the other hand, using the MCS described in step 3 of the third section, we get the best results of LI and LD to be 61.7% and 60.9%. These results mean that balancing the demands of resolution in acoustic models and the amount of available training data will achieve a higher accuracy rate.

	GMM-8	GMM-16	GMM-32	GMM-64
LD-MCS	57.7%	60.9%	59.7%	59.1%
LI-MCS	59.5%	61.7%	61.3%	61%
LD	56.3%	59.8%	57.2%	56.6%
LI	58.4%	60.1%	59.5%	58.6%

Table. 3. The syllable accuracy rates of LI and LD with and without MCS in different maximum number mixtures per state.

4.3. OPS Results

In this part, we compare the performance of several other models of acoustic training with our proposed approach, and the results are shown in table 4. In this chart, it is clear that increasing the maximum number of mixture per state, only our approach, improves the syllable accuracy rate, and the best result we achieve in this paper is 64.7%. On the other hand, CDP-MCS which does not use the data-driven rules from the corpus, the performance is even lower than the baseline performance when the maximal numbers of the mixtures are below 16. This result means the second step of CDP clustering alone is neither adequate to model the data well nor generate the phoneme set that is suitable for such the unbalanced corpus. Therefore, we should use some of the data-driven approaches as the phonetic rules to constrain the clustering of CDP. After we use the knowledge sources embedded during the phoneme clustering, such as DC-MCS, the performance is better than that of CDP-MCS. However, the knowledge sources of DT-MCS does not concern with the unbalanced conditions in this corpus. Thus, the proposed OPS-MCS which uses the data-driven approach to generate the data-driven rules for concerning the conditions of the unbalanced training data achieves higher accuracy rates than DT-MCS.

To analyze the mixtures of the acoustic model, we can see the best results of the baseline when the maximum number of mixtures per state is 16. Then if we increase the number of the mixtures in a state, the performance drops. Nevertheless, we used a 2-step clustering of the HAC and in OPS-MCS, as the accuracy rate curve rises as we increase the maximum number of mixture per state. That also verifies our proposed method is a balanced approach between the choosing the suitable phonetic units and optimizing the acoustic model under an unbalanced bi-lingual corpus.

5. CONCLUSION

In this paper, we have demonstrated that a 2-step clustering approach with agglomerative hierarchical clustering and delta Bayesian information criteria is a useful data-driven method to generate data-driven rules to constrain the CDP clustering in an unbalanced bi-lingual corpus. It has been shown that the rules provide sufficient information from the unbalanced conditions of the training corpus. When a bi-lingual corpus is unbalanced, we would put more emphasis on the characteristics of the corpus identity, and the rules driven from the data could reflect the properties of the corpus completely. We also employ a model complexity selection to balance between the OPS and the available training data. The use of these steps reduces relative syllable error rate by 7.8% comparing with the best result of the knowledge-based method.

	GMM-8	GMM-16	GMM-32	GMM-64
DT-MCS	59.3%	60.4%	62.7%	62.8%
OPS-MCS	59.9%	63.1%	64.4%	64.7%
CDP-MCS	58.2%	59.5%	61.3%	61.1%

Table. 4. The syllable accuracy rates of DT, CDP and OPS with MCS in different maximum number mixtures per state.

6. REFERENCES

[1] T. Schultz and A. Waibel, "Multilingual Cross-lingual Speech Recognition," in Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] Tanja Schultz and Katrin Kirchhoff, "Multilingual Speech Processing" Elsevier, Academic Press, ISBN 13: 978-0-12-088501-5. April 2006.

[3] Ulla Uebler, "Multilingual Speech Recognition in Seven Languages," Speech Communication (35), 2001, pp. 53-69.

[4] Joachim Kohler, "Multilingual Phone Model for Vocabulary-Independent Speech Recognition Task," Speech Communication (35), 2001, pp. 21-30.

[5] C. Santhosh Kumar, V.P.Mohandas, Li Haizhou, "Multilingual Speech Recognition- A Unified Approach," in Proc. of EuroSpeech 2005.

[6] Dau-Cheng Lyu, Bo-Hou Yang, Min-Siong Liang, Ren-Yuan Lyu and Chun-Nan Hsu, "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition," in Proc. of SST 2002.

[7] Chung-Hsien Wu, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin, "Phone Set Generation Based On Acoustic and Contextual," in Proc. of ICASSP 2006

[8] Liu Yi and Pascale Fung, "Automatic Phone Set Extension with Confidence Measure for Spontaneous Speech," in Proc. of EuroSpeech 2005.

[9] Brian Mak and Etienne Barnard, "Phone Clustering Using the Bhattacharyya Distance," in Proc. of ICSLP 1996, pp. 2005-2008.

[10] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," Journal of the American Statistical Association 78 (383), 1983, pp:553-584.

[11] A. Tritschler and R. Gopinath, "Improved Speaker Segmentation And Segments Clustering Using The Bayesian Information Criterion," in Proc. EuroSpeech 1999, pp. 679-682.

[12] G. Schwarz, "Estimating the dimension of a model," The annals of statistics, vol. 6, 1978, pp 461-464

[13] Young, S. J., Odell, J. J., and Woodland, P. C., "Treebased state tying for high accuracy acoustic modelling," In Proc. of the ARPA Workshop on Human Language Technology, 1994

[14] X. Anguera, T. Shinozaki, C. Wooters, and J. Hernando, "Model Complexity Selection and Cross-validation EM Training for Robust Speaker Diarization," in Proc. of ICASSP 2007