Computational Linguistics and Chinese Language Processing
Vol. 13, No. 3, September 2008, pp. 363-386

Acoustic Model Optimization for Multilingual Speech Recognition

Dau-Cheng Lyu*, **Chun-Nan Hsu**+,

Yuang-Chin Chiang#, **and Ren-Yuan Lyu**§

Abstract

Due to abundant resources not always being available for resource-limited languages, training an acoustic model with unbalanced training data for multilingual speech recognition is an interesting research issue. In this paper, we propose a three-step data-driven phone clustering method to train a multilingual acoustic model. The first step is to obtain a clustering rule of context independent phone models driven from a well-trained acoustic model using a similarity measurement. For the second step, we further clustered the sub-phone units using hierarchical agglomerative clustering with delta Bayesian information criteria according to the clustering rules. Then, we chose a parametric modeling technique -- model complexity selection -- to adjust the number of Gaussian components in a Gaussian mixture for optimizing the acoustic model between the new phoneme set and the available training data. We used an unbalanced trilingual corpus where the percentages of the amounts of the training sets for Mandarin, Taiwanese, and Hakka are about 60%, 30%, and 10%, respectively. The experimental results show that the proposed sub-phone clustering approach reduced relative syllable error rate by 4.5% over the best result of the decision tree based approach and 13.5% over the best result of the knowledge-based approach.

Keywords: Cross-lingual Phone Set Optimization, Speech Recognition, Delta-BIC

* Department of Electrical Engineering, Chang Gung University, Taiwan
E-mail: d9221003@stmail.cgu.edu.tw

+ Institute of Information Science, Academia Sinica, Taiwan
E-mail: chunnan@iis.sinica.edu.tw

# Institute of Statistics, National Tsing Hua University, Taiwan
E-mail: jjchiang1@hotmail.com

§ Computer Science and Information Engineering, Chang Gung University, Taiwan
E-mail: renyuan.lyu@gmail.com

[Received January 3, 2008; Revised July 7, 2008; Accepted August 8, 2008]

1. Introduction

Multilingual speech recognition, integrating several language-specific recognizers into one recognizer, is one of the popular research topics in the speech recognition field [Schultz *et al.* 2006; Uebler 2001; Kohler 2001; Kumar *et al.* 2005; Lyu *et al.* 2002]. Many multilingual speech recognizers depend on a large-scale speech database for each language in order to obtain high performance. However, such a large-scale speech corpus is not always available for all of the languages to be used. For example, in Taiwanese, which is one of the most two important languages in Taiwan, the size of the available corpus of Taiwanese is not comparable to that of other majority languages, such as English, Mandarin, and Spanish. In fact, researchers who want to build a multilingual speech recognizer by collecting a well-designed, large-scale speech corpus for every language find that it is not feasible for all languages. Automatic speech recognition (ASR) systems can be prohibitively expensive to develop because of time consumption and expense as the process involves a huge amount of data collection and calls for the complete building of an acoustic model for the minor language. In this paper, we try to find an approach which uses the majority languages, defined as those languages with large-scale training data available, to help the minor languages and build a reliable multilingual acoustic model.

Several approaches are proposed for the multilingual ASR system [Schultz *et al.* 2006; Uebler 2001; Kohler *et al.* 2001; Wu *et al.* 2006; Liu *et al.* 2005; Mak *et al.* 1996]. 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 universal phone inventory. Under this approach, the same phonetic representation within different languages shares the same training data. However, the approach based on universal phone inventory only uses the phonetic knowledge, and it does not consider the spectral properties of the variations of the sounds with different speakers. Another approach is to merge the language-dependent phones using a data-driven approach to map the phones in different languages by specific distance measure between well-trained acoustic models. The advantage of this approach is that the distance is estimated from the statistical measure of similarity of speech sounds among the languages. Nevertheless, optimizing the final number of the universal phone set is an important issue in such a data-driven approach.

In this paper, we use a constrained data-driven approach to select the optimal phoneme set which directly reflects the characteristics of the unbalanced training trilingual speech data. The training corpora include the three most popular languages in Taiwan -- Mandarin, Taiwanese, and Hakka -- where the amount of Mandarin data is two times as large as that of Taiwanese and five times as large as that of Hakka. The training procedures for the trilingual acoustic model can be divided into the following steps. First, we use the Bhattacharyya distance [Mak *et al.* 1996] to measure the phonemes distance among the languages according to well-trained acoustic models. Then, we apply a 2-step clustering using the hierarchical agglomerative clustering (HAC) [Fowlkes *et al.* 1983] with delta Bayesian information criteria (△BIC) [Tristschler *et al.* 1999] to automatically select the optimal context dependent phoneme (CDP) set. The first clustering step is to generate context-independent phoneme (CIP) clustering rules. In the second clustering step, we use the rules as phonetic knowledge to constrain the CDP clustering. After the clustering procedures, we generate the new phoneme set where the sounds with the same label will share the training data. Finally, to optimize the acoustic model of the new phoneme set, we use a model complexity selection (MCS) [Anguera *et al.* 2007] to adjust the number of Gaussian components in the Gaussian mixture to balance between the new phoneme set and the unbalanced training data.

This paper is organized as follows. In Section 2, we introduce the three main languages in Taiwan -- Mandarin, Taiwanese, and Hakka -- and also discuss their linguistic and phonetic characteristics. In Section 3, we use a knowledge-based approach to build two baseline systems, a language-dependent and a language-independent recognizer, on an unbalanced speech corpus. In Section 4, we describe our proposed approach to automatically optimize a multilingual acoustic model, including three components: CIP clustering, CDP clustering, and MCS. Then, several experiments are conducted for comparison with the baseline system and the decision tree based approach. Finally, we summarize our key contributions of this paper.

2. Linguistic and Phonetic Properties of Mandarin, Taiwanese and Hakka

The goal of this paper is to use a majority language, like Mandarin, to assist minor languages, such as Taiwanese and Hakka, to train a unified acoustic model for a trilingual ASR system. In order to efficiently combine the acoustic units among the languages, some basic knowledge, such as the phonetics, phonology, and other linguistic aspects, is essential. In this section, we will introduce the linguistic and phonetic characteristics of the three languages.

2.1 Linguistic Properties

Mandarin, Taiwanese, and Hakka are the three main languages in Taiwan. Mandarin is the most important language in ethnically Chinese societies, because of its huge population and the potentially huge market. Another language, Taiwanese, is the mother tongue of more than 75% of the population in Taiwan. In addition, 10% of the local population uses Hakka as their first language. In fact, many people can use any two of these three languages. Due to this distribution, Taiwan is a bilingual, or even a trilingual society. On the other hand, the three languages, also used by many overseas Chinese living in Singapore, Malaysia, Philippine, and other areas of Southeast Asia, are the members of the Chinese language family. Some people prefer to call them "dialects" but, unlike dialects in other languages, which are usually mutually intelligible to each other, the "dialects" in spoken Chinese are almost completely mutually unintelligible. Due to their mutual unintelligibility, we prefer to call Mandarin, Taiwanese, and Hakka three languages.

2.2 Phonetic Properties

To express the three languages in phonetic properties, we applied a heuristic mapping approach based on the Formosa Phonetic Alphabet (ForPA), an easy-to-use transcribing system for the trilingual speech data of these three languages [Lyu *et al.* 2004]. Basically, the Mandarin Phonetic Alphabet (MPA, also called zhu-in-fu-hao) and Pinyin (han-yu-pin-yin) are the most widely known phonetic symbol sets to transcribe Mandarin Chinese, but both of these systems are designed for Mandarin. ForPA is designed for transcribing Mandarin, Taiwanese, and Hakka for linguistic computing. Generally speaking, ForPA might be considered as a subset of IPA [Mathews *et al.* 1975], but it is more suitable for applications to the languages used in Taiwan.

These three languages are monosyllabic languages; therefore, a syllable, the smallest meaningful unit, is the basic pronunciation unit of the Chinese languages. Although the three languages have similar CV (C: Consonant, V: Vowel) structure at a syllabic level, Taiwanese and Hakka have much more abundant variation in allowing syllable-final consonants, including -p, -t, -k, -h. Totally, there are 408 base syllables in Mandarin, 709 base syllables in Taiwanese, and 777 base syllables in Hakka. The total number of the union set of syllables among the languages is 1333, and only 141 syllables are in common.

In the phonemic level, there are 21 consonants and 9 vowels in Mandarin, while there are 18 consonants and 16 vowels in Taiwanese. In Hakka, there are 20 consonants and 25 vowels. Therefore, the total number of phonemes is 30, 34, and 45 for Mandarin, Taiwanese, and Hakka, respectively. Additionally, some of the phonemes in the three languages are labeled with the same symbols by phoneticians, meaning that they are phonetically very close. In our observation, only 26 phonemes are truly in common. The common phonemes can share acoustic data among the languages. We list the statistical information of several different phonetic units in these three languages in Table 1.

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

	M	T	H	$M \cup T \cup H$	$M \cap T \cap H$
N_p	30	34	45	73	26
N_S	408	709	777	1333	141
N_{tri}	530	1000	1049	1740	214

3. Language-Dependent and Language-Independent System

In this section, we describe two ASR systems A) language-dependent ASR and B) language-independent ASR. The first one is solely trained on a single language for each language, and we combine each recognizer into one recognizer as the language-dependent ASR system. The second one uses a universal phoneme inventory to map phonemes among the languages and build the common recognizer in one level. In the following, we first introduce the speech corpora for training the acoustic models, and then we report the performance of the two systems. Of course, some discussions are also included at the end of this section.

3.1 Trilingual Speech Corpus

In order to simulate the available speech data in Taiwan, we use an unbalanced trilingual speech corpus for all of the experiments of this paper. The details of the corpus are listed in Table 2. Three languages are included -- Mandarin Taiwanese, and Hakka -- which are recorded with a 16 kHz, 16-bit, microphone channel. There are 100 speakers, 50 speakers each for Mandarin and Taiwanese, for the training set. Each of the speakers recorded two sets of phonetically balanced utterances, and each of the utterances contained one to four syllables. In the Hakka training set, there were only two speakers; one speaker recorded two monosyllabic sets, and the other speaker recorded 4320 phonetically balanced utterances. In the two monosyllabic sets, we included the two main Hakka accents, Miaoli and Hsinzu, of Taiwan. The former contains 2269 tonal syllables, and the latter contains 2970 tonal syllables. As we mentioned before, the training sets of the corpus are unbalanced, where the total number of hours for Taiwanese is about half of that of Mandarin, and the amount of the speech data in Mandarin is five times as large as that of Hakka. For the test set, another 21 speakers recorded the test speech, with a total length of almost 1 hour.

Table 2. Statistics of training and testing the trilingual speech corpus.

	language	number of speakers	number of utterances	speech length (in hours)
Training Set	Mandarin	100	43,087	10.9
	Taiwanese	50	22,851	5.2
	Hakka	2	9559	2.1
Test Set	Mandarin	10	1,000	0.29
	Taiwanese	10	1,000	0.27
	Hakka	1	834	0.43

3.2 Language-Dependent System

For the language-dependent recognizer for the trilingual speech, we trained a monolingual acoustic model for each of the languages using individual language training speech data. After that, we combined the monolingual acoustic model of all languages into a single one. Based on the units of each monolingual acoustic model having their language tags, they do not share the training data. For example, unit /a/ for Mandarin is labeled as /a_M/, /a_T/, and /a_H/ for Taiwanese and Hakka as well. The same ForPA symbols among the languages cannot share parameters in the acoustic model. Therefore, the size of the parameters will increase proportional to the number of languages. The number of tri-phone context-dependent acoustic phone models for the language-dependent system is 2579.

The experimental setups are described as follows. 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). Each of the frames is 20 milliseconds in length with 10 milliseconds shifting, which means that the adjacent frames have a 10 milliseconds overlap. For the acoustic modeling, we used the Hidden Markov Model (HMM) for constructing a context-dependent acoustic phone model, and each of the HMMs has three states. For the language modeling of each language, a uniform distribution is used, which implies that the perplexities of the language model of Mandarin, Taiwanese, and Hakka are 408, 709, and 777, respectively. The acoustic model and the Viterbi decoding used to generate the recognized syllable sequences are implemented with HTK [Young *et al.* 2002].

The language-dependent systems are trained and evaluated only on language specific data, and the results are shown in Figure 1. Basically, the syllable error rate is positively correlated to the perplexities of the language model. For example, the highest perplexity among the three languages is Hakka, 777; therefore, the average syllable accuracy rate of Hakka is 33.9%. In contrast, the smallest perplexity is Mandarin, and the average syllable accuracy rate of Mandarin is 60.8%. We believe both the perplexity of language modeling and the amount of training data influence the results. Due to the lack of training data, the highest performance of Hakka is achieved when the number of Gaussian components in the Gaussian mixture is only 4. When we increased the number of Gaussian components in the Gaussian mixture, the performance dropped. On the other hand, Mandarin achieved the highest performance when the number of Gaussian components in the Gaussian mixture was 32, and the performance steadily increased when we increased the number components in the Gaussian mixture.

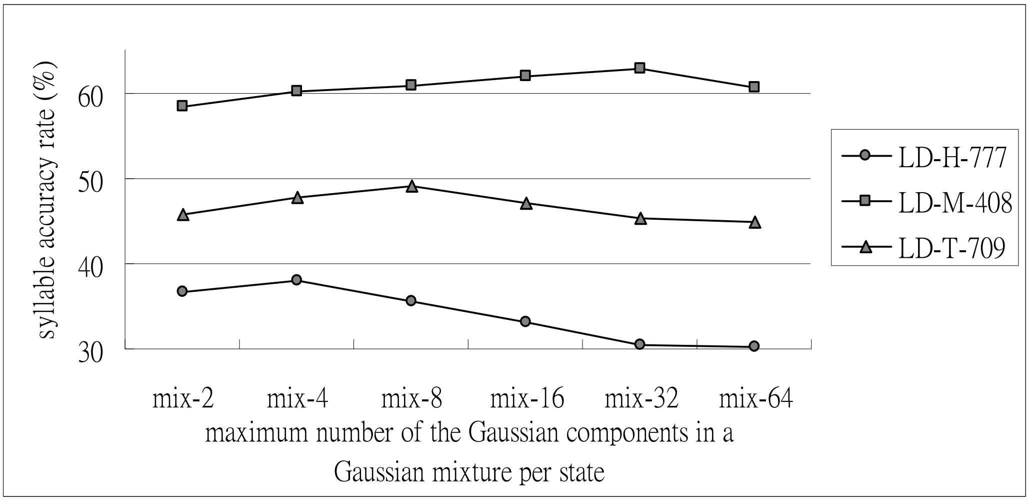

Figure. 1. The syllable accuracy rates of the language-independent system with different maximum number of the Gaussian components in a Gaussian mixture per state, where "LD-H-777" represents language-independent system tests on Hakka utterance with 777 language model perplexity.

3.3 Language Independent System

The basic concept of language-independent acoustic modeling is a knowledge-based model by sharing of the phonemes among the languages. The sound representations of phonemes are so similar among the languages that phonemes can be considered as units independent of the training set language. In this paper, we use ForPA to transcribe the sounds among the languages. In this method, the similarities of sounds are classified based on phonetic knowledge. Sounds with the same phonetic labeling that exist in different languages share the data. Like the other language-independent recognition system [Marthi *et al.* 1999], the phonemes of different languages which belong to the same ForPA units share data from different languages during training. In this way, we combine the acoustic model from the three languages, and the size of the acoustic models would not increase as large as the number of languages added. The acoustic model is unified in the language-independent speech recognition system. Due to the sharing of the training data among the languages, each parameter of a common phoneme in the acoustic model will get more available data for estimation than in language-dependent systems.

Two results are shown in Figure 2 and Figure 3. The result in Figure 2 was derived from the language-independent acoustic model with a language-specific language model, and that in Figure 3 was derived from the language-independent acoustic model with a unified language model. For example, “LI-H-777” means that we used all of the training data in the three languages to build a language-independent acoustic model, and evaluated Hakka utterances only used the Hakka language model, where the perplexity of the language model was 777. On the other hand, the system, “LI-H-1333”, used the language-independent acoustic model and tested Hakka utterances with the unified language model, and the perplexity of language model was 1333.

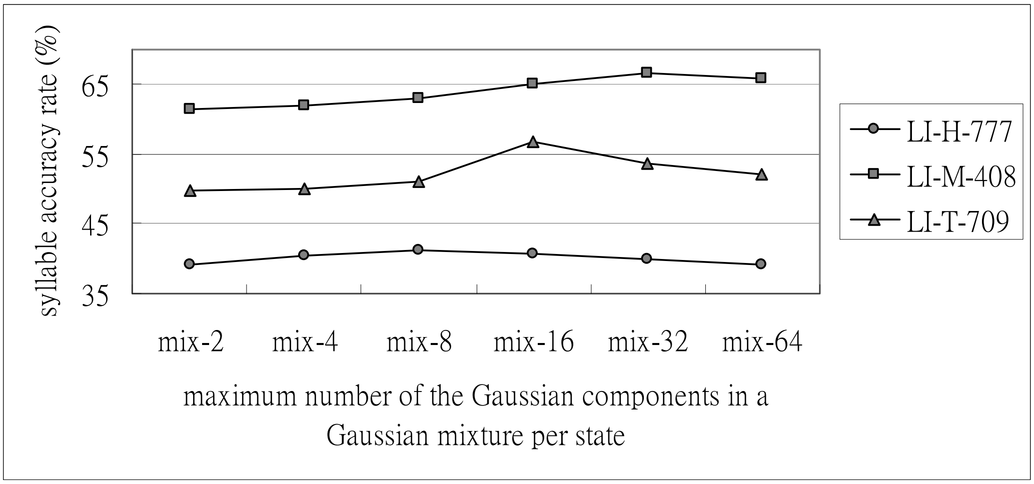

Figure 2. The syllable accuracy rates of LI system with the language-specific language model in different maximum numbers of Gaussian components in a Gaussian mixture per state.

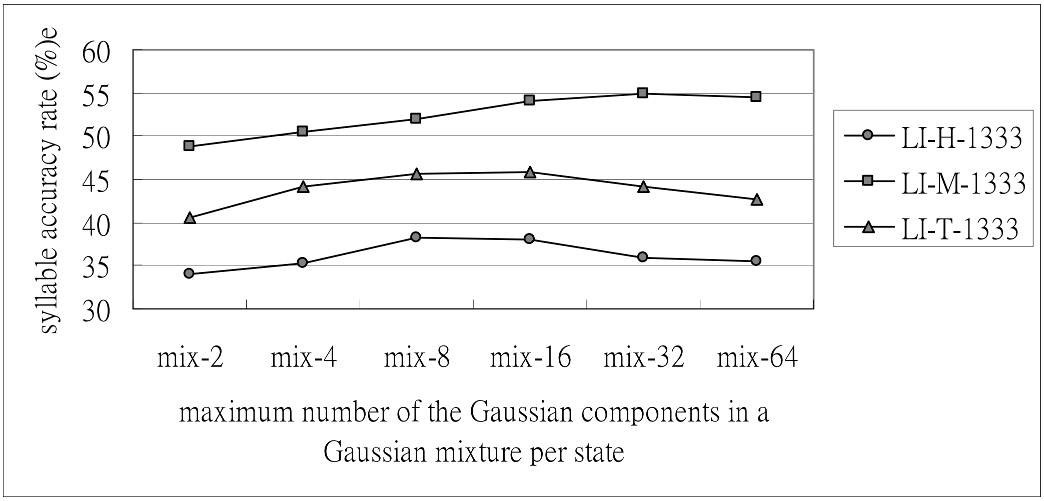

Figure 3. The syllable accuracy rates of LI system with the unified language model in different maximum numbers of Gaussian components in a Gaussian mixture per state.

Obviously, due to the sharing of training data, we obtained better performance from the language-independent acoustic model with the language-specific language model than from the language-dependent system. The average of the syllable accuracy rates for Mandarin, Taiwanese, and Hakka are 63.9%, 52.3%, and 40.1%, respectively. On the other hand, the performance dropped reasonably when we used the language-independent language model, where the language perplexity is 1333, compared with that of the language-specific case. The best performance for Hakka with the language-independent recognition system used 8 Gaussian components in a Gaussian mixture per state, but that of the language-dependent recognition system used only 4 Gaussian components in a Gaussian mixture per state. This means that the training data sharing from Mandarin helps Hakka, because some of the sounds of Hakka can get training data from the sounds of Mandarin if they share the same ForPA labels.

Based on this experiment, we had the following two conclusions: first, the number of Gaussian components in a Gaussian mixture for such a small amount of Hakka training data in the acoustic model was too large, thereby causing the poor performance. Therefore, when we used too many parameters in the number of Gaussian components in a Gaussian mixture per state, it made their estimations become unreliable because the available training data was insufficient to accurately estimate the parameters. According to this fact, second, we have to find a way to decrease the parameters in the acoustic model and give the retrenched parameters “more” training data to estimate.

3.4 Discussion

According to the performance of the baseline results of language-dependent and language-independent systems, we have found that the accuracy rate is positively correlated to the amount of training speech. This means that if we have more available training data, we may obtain higher performance. This assumption is reasonable, because, if the parameters in the acoustic model have more data for estimation, the parameters will be more reliable. However, it is time consuming and expensive to collect more training data for a new language, especially a minor one. On other words, it is not always feasible for all of the languages to get as large training data as possible. As a result, the performance of the minor language, Hakka, is not comparable with that of the majority language, Mandarin. Since the problem of training data shortage will encumber the overall performance of multilingual speech recognition, we should find a way to use the available majority language corpus to help the ASR system in the minor language to balance the total parameters in the acoustic model between the available training data.

A training data sharing approach of acoustic modeling in multilingual speech recognition helps the minor language to perform well based on the information found in the majority language. In our previous bilingual speech recognition experiments [Lyu *et al.* 2008], we clustered similar phonemes so the total numbers of the phoneme set in the acoustic model reduced and required fewer parameters to get sufficient training data for estimation. As sufficient data becomes available, parameter estimating will be more robust. Furthermore, the number of mixtures in the acoustic models is also a critical factor influencing the recognition performance when using an HMM-based recognizer. According to our Hakka experiments, when we increased the number of Gaussian components in a Gaussian mixture in the minor language, the accuracy rate dropped. Contrary to the result of the minor language, the accuracy rate of the majority language increased as the number of Gaussian components in a Gaussian mixture increased. Based on the above analysis, we identify the following two issues in order to improve the final performance using the unbalanced trilingual corpus.

- How to efficiently cluster similar phonemes in these three languages then automatically decide the final number of phonemes in the acoustic model of the trilingual recognizer. In other words, we need to redefine the acoustic unit and make the parameters in the acoustic model have sufficient training data to evaluate and then get better performance in the unbalanced speech corpus.
- The number of Gaussian components in a Gaussian mixture within a state should depend on the amount of the available training data. We should not increase the mixtures blindly.

In order to cluster the phonemes, we should first analyze the shortage condition of the training data of all the languages. To get such the knowledge, the training data should be totally “observed”. After observation, we recognize all the conditions of the training data of each language then generate the clustering rules to refine the phoneme set based on the observed data. In fact, this kind of method is a data-driven approach. According to the rules, we generate a new phoneme set which is given a consideration to balance the available training data and the training parameters. Therefore, the parameters of such a new phoneme set will have sufficient training data in an unbalanced corpus. In the next section, we propose an approach which not only automatically optimizes the number of parameters in the acoustic model but also increases the number of mixtures per state depending on model complexity selection.

4. Acoustic Model Optimization

4.1 System Overview

In this section, we propose an approach integrating three main steps: CIP clustering, CDP clustering, and MCS. The overall diagram is shown in Figure 4. First, we use a data-driven approach to estimate the similarity of the acoustic phoneme from the observed training data of all the languages. Second, based on the similarity measure, we use a clustering algorithm to

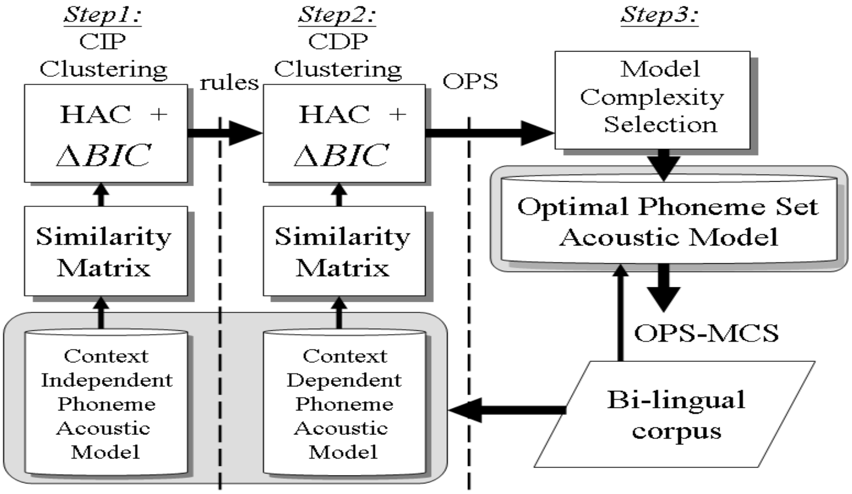
Step1:
CIP
Clustering
HAC +
ΔBIC
Similarity Matrix
Context Independent Phoneme Acoustic Model
rules
Step2:
CDP
Clustering
HAC +
ΔBIC
Similarity Matrix
Context Dependent Phoneme Acoustic Model
OPS
Step3:
Model Complexity Selection
Optimal Phoneme Set Acoustic Model
OPS-MCS
Bi-lingual corpus

Figure 4. The overall diagram for automatically optimizing the acoustic models.

group the phoneme sets that are similar. In this stage, two algorithms are used: HAC [Fowlkes *et al.* 1983] and△BIC [Schwarz 1978]. The new phoneme set is decided by the △BIC which is based on the optimal Bayesian model selection criterion; therefore, we called the new phoneme set an optimal phoneme set (OPS). Finally, we use a MCS to adjust the number of Gaussian components in a Gaussian mixture per state of a HMM-based acoustic model. By doing this, we balance the number of Gaussian components in a Gaussian mixture in the OPS and the unbalanced training data. The details of the similarity measurement are described in the following:

- Similarity Measurement:

  In order to measure the similarity of HMM-based acoustic phoneme model, we introduce a measure distance, Bhattacharyya distance, which is similar to a likelihood ratio test to evaluate a similarity of each phone model in the context independent and context dependent phone acoustic models. The Bhattacharyya distance is a theoretical distance between two Gaussian distributions [Mak *et al.* 1996], and is said to be equivalent to an upper bound on the optimal Bayesian classification error probability. In this stage, the number of the Gaussian components in a Gaussian mixture in each state of a HMM acoustic model is only one. We give a brief review here with the following equation and its notations.

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

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.

- OPS:

  The OPS is a CDP set generated by the following two steps. First, we generate CIP clustering rules, which are based on the HAC and △BIC algorithms. The goal of the rules is to phonetically constrain the next step, CDP clustering. The CDP clustering generating the OPS also uses the HAC and △BIC algorithms.

- MCS:

  Before MCS, each of the states only contains one Gaussian component. Therefore, the goal of MCS is to increase the number of Gaussian components in a Gaussian mixture in such a way where the increasing rules of the state are in accordance with the available training data corresponding to the state. In fact, MCS is an algorithm to get a balance between the number of the CDP-based acoustic models and the amount of available training data.

4.2 CIP Clustering

The goal of CIP clustering is to obtain phonetic rules to constrain CDP clustering. The procedure is initialized with context-independent phoneme models to select and merge those models that correspond to the two most similar context-independent phonemes iteratively. In order to obtain the hierarchical structure of the phoneme distances, we adopt HAC then use △BIC to decide the clustering rules of CIP clustering. In this step, if the context-independent phonemes are grouped, the same central context-independent phonemes of the context-dependent phonemes are put in a clustering pool. Then, in the next step, CDP clustering will also use HAC and △BIC to select final context dependent phoneme to cluster from each of the clustering pools. An example is shown in Figure 5 and Figure 6. A set of context-dependent phonemes with central context-independent phonemes being /c_M/ and /c_T/ (as shown in Figure 5) used HAC and △BIC to find several clustering groups, such as the group of /c_T+ng_T/ and /c_T+er_T/ in the right part of Figure 6.

There are several algorithms to evaluate the distance of each cluster in HAC, such as the centroid-linkage, average-linkage, complete-linkage, and average-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.

The similarity matrix is generated by the similarity measure and the elements of the matrix are context-independent phonemes. The average-linkage agglomerative algorithm is defined the distance between two clusters and measure the average distance between a sample in one cluster and a sample in the other cluster.

$$d_{avg}(C_p, C_q) = \frac{1}{n_p n_q} \sum_{x \in C_p, y \in C_q} d(x, y) \tag{2}$$

where n_p and n_q are the numbers of the cluster C_q and C_p respectively and x and y are two data points.

△BIC is the confidence measure to cluster the similar context independent phoneme from the HAC results. Before we describe △BIC, we should introduce 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 $x_n \in R^d$. 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 [Schwarz 1978], for sufficiently large n, the best state of the data is the one which maximizes.

$$BIC_i = \log \ell_i(x_1, \ldots, x_n) - \frac{1}{2}\lambda k_i \log n \tag{3}$$

where $\ell_i(x_1,\ldots,x_n)$ 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 △BIC defined as $BIC_p - BIC_q$ is positive. Based on Equation (3), the formula of △BIC is written as:

$$\Delta BIC = -\frac{n_p}{2}\log\left|\Sigma_p\right| - \frac{n_q}{2}\log\left|\Sigma_q\right| + \frac{n_r}{2}\log\left|\Sigma_r\right| + \frac{1}{2}\lambda(d + \frac{d(d+1)}{2})\log n_r \tag{4}$$

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, In practice, we tried to use the value of λ from 0.85 to 1.15, but the final results of △BIC did not change too much. Thus, we chose one as the value of λ. The results of consonant and vowel of using CIP clustering are demonstrated in Figure 5 and Figure 6, respectively. Table 3 shows that 12 consonants and 20 vowels are merged.

Table 3. The CIP clustering results.

Consonant	[b_K, d_K, g_K] [b_T, b_M] [d_M, d_T] [g_M, g_T] [h_K, k_M] [p_K, t_K] [z_M, z_T, zh_M] [c_M, c_T] [s_M, s_T] [l_M, l_T] [rh_M, m_M] [n_M, n_T]
Vowel	[a_K, ann_K] [ng_K, u_K,unn_K] [e_K, enn_K] [ii_K, inn_K] [a_M, a_T,] [a_M, o_T, onn_T] [er_M, err_M] [e_M, e_T] [inn_M, inn_T] [i_M, i_T] [ih_T, innh_T] [ng_M, ng_T] [uk_K, uT_K] [ok_K, op_T] [ak_T, ap_T] [ap_T, ok_T] [on_T, onnh_T, onn_M, uh_T] [et_T, ek_T] [ennt_K, et_K] [ip_T, it_T]

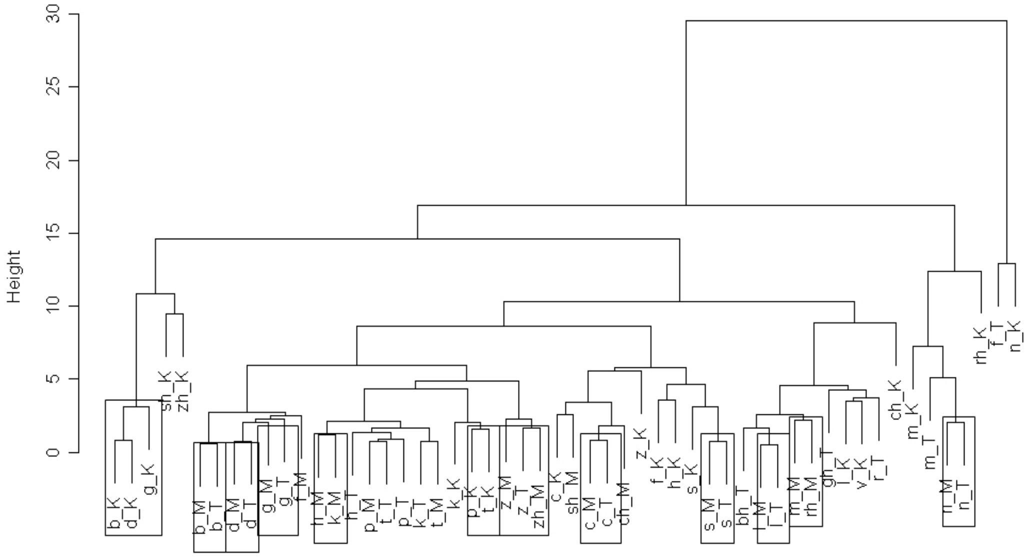

Figure 5. The consonant example of using HAC and CIP for Mandarin, Taiwanese, and Hakka, where the height is the value of the Euclidean distance among the phoneme models.

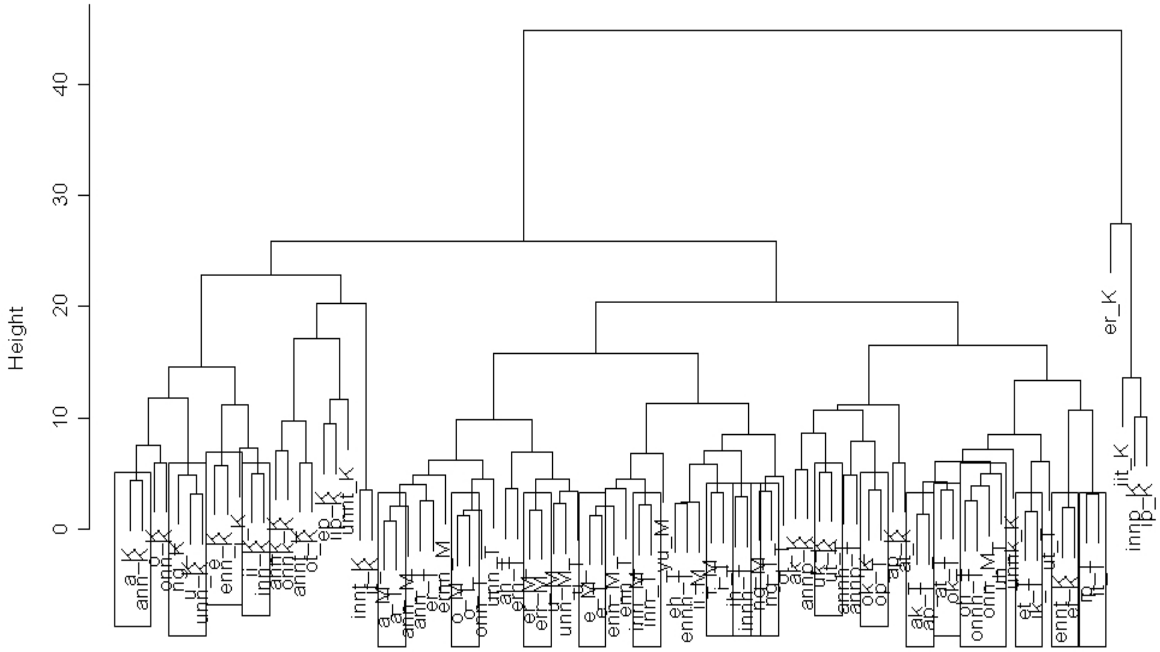

Figure 6. The vowel example of using HAC and CIP for Mandarin, Taiwanese, and Hakka.

4.3 CDP Clustering

Based on the rules derived from CIP clustering, we constrained the CDP clustering. The procedure was the same with CIP clustering, but we used context-dependent phonemes to replace context-independent phonemes. We generated 12 consonant and 20 vowel CDP clusters, and took the merging process for the phonemes /c_T+*_T/ and /c_M+*_M/ as an example, which were shown in Figure 7. First, we generated a hierarchical tree whose bottom nodes are context-dependent phonemes containing /c_T/ and /c_M/ in their phonemic transcription using the HAC algorithm based on the similarity matrix. Then, we created a new context-dependent phoneme set using △BIC. If the value of △BIC between two context-dependent phonemes or sets is bigger than zero, then the two context-dependent phonemes or the two sets are merged. The merged context-dependent phonemes or the sets are used as the new context-dependent phonemes, and the new context-dependent phonemes in the same group use common training data to train new acoustic models. On the other hand, if the value of △BIC is negative, the merging process stops.

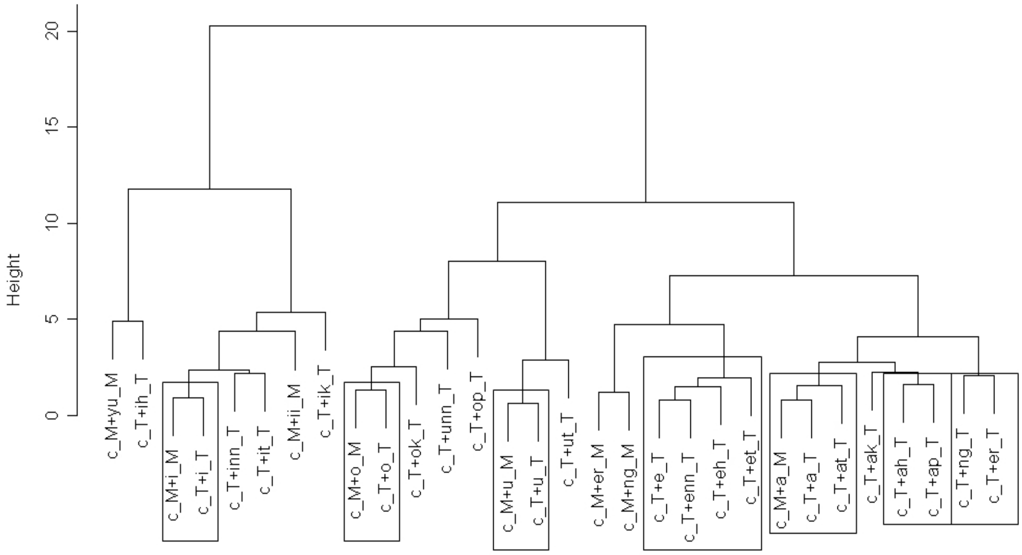

Figure 7. The merging result of LD-CDP, where the center of consonant is /c_M/ and /c_T/.

Another example, shown in Figure 5, is /b_K/, /d_K/ and /g_K/. In our case, the training data of Hakka is one-fifth of that of Mandarin. Therefore, the performance of Hakka's recognizer could not possibly be as good as that of Mandarin if we only used the language-dependent system. Therefore, the results of CIP clustering make a phonetic knowledge rule to put all of the central context-independent phoneme, such as /b_K/, /d_K/

and /g_K/, of the context-dependent phoneme, such as /b_K+*/, /d_K+*/ and /g_K+*/ in a clustering pool. Based on this rule, we used CDP clustering to generate new acoustic units. Essentially, as shown in Figure 8, there are 41 context-dependent phonemes in this clustering pool based on CDP clustering. Then, using HAC and △BIC, the final 31 new acoustic units are generated. For example, some of the sub-phone such as the first two in Figure 8, /g_K+ok_K/ and /g_K+ok_K/, are merged. It is noticed that, although the context-independent phonemes of Hakka, /b_K/, /d_K/ and /g_K/, became a confusion set, the final CDP clustering made the set of context-dependent phonemes separable. In Figure 8, we can see, basically, the sub-phones of /g_K+*/, /d_K+*/ and /b_K+*/ are not inter-merged, they are intra-merged. This means that the context-dependent phonemes merge together only when the have the same context-independent phonemes.

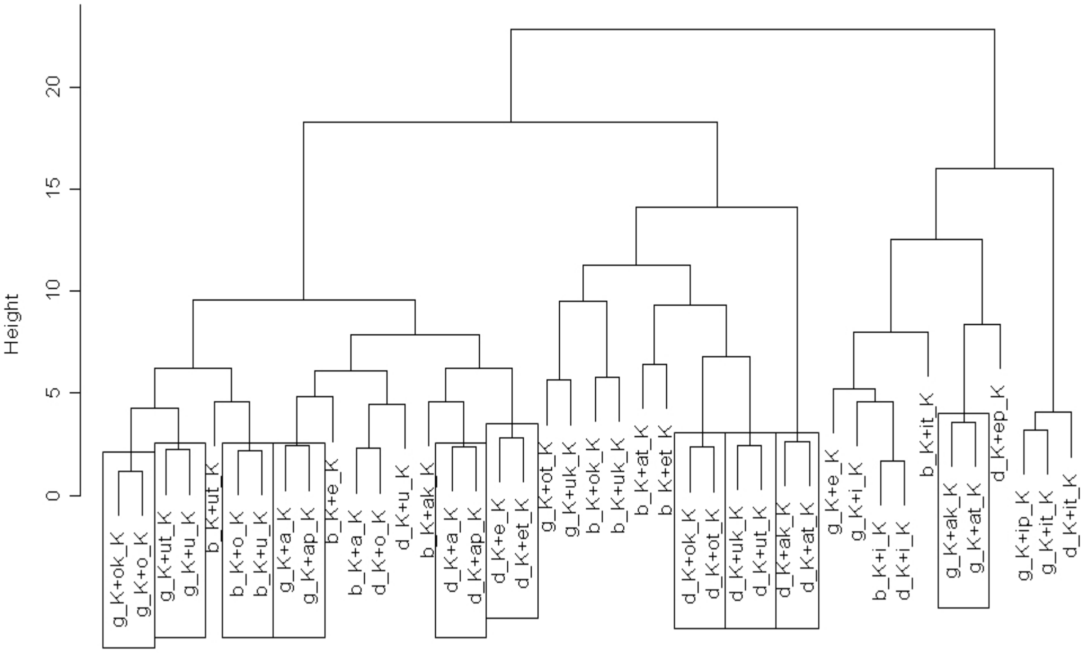

Figure 8. The merging result of LD-CDP, where the center of vowel is /ok_K/ and /op_T/.

4.4 Model Complexity Selection

After we generated the OPS using a 2-step data-driven phoneme clustering, we considered the balance between the resolution of the generated OPS and the amount of available training data. For the minor languages, Taiwanese and Hakka, the more Gaussian components in a Gaussian mixture, the lower the accuracy rate we get. Thus, the number of Gaussian components in a

Gaussian mixture for each state in the acoustic model should be carefully controlled. According to [Anguera *et al.* 2007], a parametric modeling technique, MCS, is chosen to perform on each state with sufficient training data, and we select the number of Gaussian components in a Gaussian mixture based on the amount of the data frames belonging to the state in the acoustic model. 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:

$$M_p^i = round(\frac{N_p^i}{OR}) \tag{5}$$

where M_p^i 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_p^i 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 p^{th} acoustic model of the i^{th} state depends on the amount of the corresponding occurrence training data.

5. The Experiments of OPS and MCS

In this section, we used a data-driven approach to train a trilingual acoustic model and built an OPS speech recognition system with MCS under a data unbalanced condition. Then, we performed a series of the experiments to evaluate the system on three languages, Mandarin, Taiwanese, and Hakka. The experiments were divided into two parts: OPS and MCS. In addition, we used a state-of-the-art, decision tree based tri-phoneme clustering method [Young *et al.* 1994], to compare with our proposed approach.

For the MCS part, we increased the number of Gaussian components in a Gaussian mixture per state depending on the training occurrence. This means that when we set the number of Gaussian components in a Gaussian mixture per state to be 16, not all of the states would increase the number of Gaussian components in the Gaussian mixture to 16. If the available training occurrences of the corresponding state are satisfied with Equation (5), the state will adjust to the number of Gaussian components in a Gaussian mixture to be 16. We also used the results of the LI recognition system in Figure 3 for comparison. The LI system described in Section 3 increased the number of Gaussian components in a Gaussian mixture per state in a brute-force manner, which means that when we set the number of Gaussian components in a Gaussian mixture per state to 16, then, no matter how many occurrences the particular state has, all of the states have to adjust the number of Gaussian components in the Gaussian mixture to 16. The results of MCS part are shown in Figure 9 and Figure 10.

For the OPS part, we built an OPS recognizer from the LI-CDP acoustic model based on the three procedures described in Section 4. First, we generated a set of 32 CIP clustering rules from CIP clustering; then, we generated 169 CDP rules to merge the tri-phone model in accordance with these rules. Finally, we obtained 1314 HMMs from LI-CDP's 2579 sub-phones. For each state of the HMMs, we increased the number of the Gaussian components in a Gaussian mixture by MCS approach.

We also compare the performance of a decision tree (DT) approach with our OPS recognition system. The leaf nodes in a DT ASR system are sub-phones, and each node of the tree was associated with a binary question which had been selected from a set derived by linguistic experts. The best question was assigned to a node if it results in a binary splitting with minimal loss of likelihood [Liang *et al.* 1998]. For the other parameter configuration in the experiments, we used the same setting described in Section 3.2, and we showed the results of the decision tree system and OPS in Figure 11 and Figure 12.

5.1 MCS Results

First, all of the performances of all the languages using MCS (Figure 9) are better than those without using MCS (Figure 3). The average improvements for Hakka, Mandarin, and Taiwanese are 4.5%, 2.5%, and 2.5%, respectively. The best results for LI-system for each of the languages are 42.01%, 58.45%, and 48.79% when the maximum number of Gaussian components in a Gaussian mixture per state is 32, 64, and 32 for Hakka, Mandarin, and Taiwanese, respectively. Second, unlike the results shown in Figure 3, the results in Figure 9 show that increasing of the number of Gaussian components in a Gaussian mixture caused the performance to drop in the Hakka case. This is because of increasing the number of Gaussian components in a Gaussian mixture per state based on the occurrence training data using MCS. The accuracy rate increases when the number of the Gaussian components in a Gaussian mixture is also increased.

To analyze the total number of Gaussian components in a Gaussian mixture in the acoustic model between the system with using MCS and without using MCS, we illustrated the differences in Figure 10. We can see that the average number of the Gaussian components in a Gaussian mixture in LI system without using MCS is almost the same as the maximum number of Gaussian components in a Gaussian mixture. However, for the MCS case, the average number of Gaussian components in a Gaussian mixture of mix-2, mix-4, mix-8, mix-16, mix-32, and mix-64 are 2, 3.1, 6.6, 11.3, 18.1, and 24.5, respectively. Therefore, the total numbers of Gaussian components in a Gaussian mixture using MCS are much fewer than those without using MCS system.

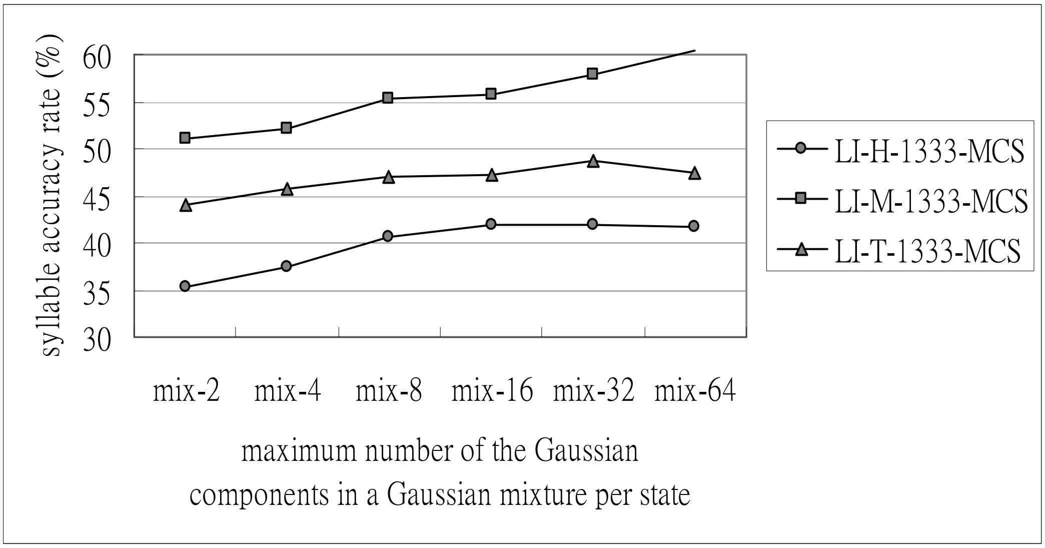

Figure 9. The syllable accuracy rates of LI system with different maximum number of the Gaussian components in a Gaussian mixture per state using MCS.

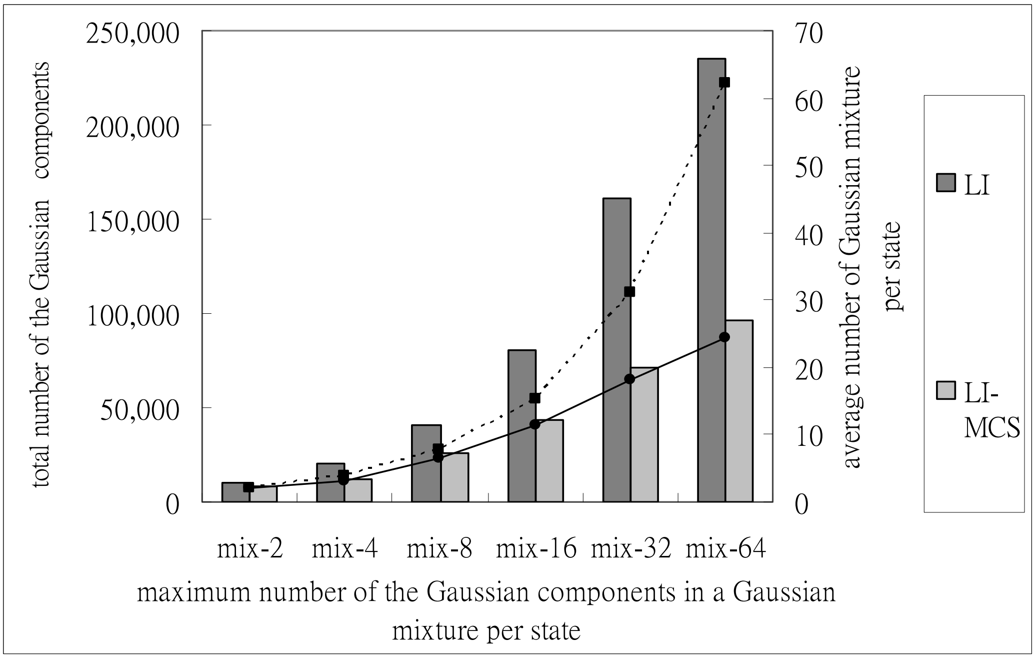

Figure 10. The total number of Gaussian components in a Gaussian mixture of the acoustic model in two different approaches.

5.2 The Results of DT and OPS

The results of using decision tree (DT) approach and OPS approach are shown in Figure 11 and Figure 12. We find that the performance of both the OPS and the DT-ASR systems are better than that of LI-MCS system. Because of the clustering approach, the similar sounds across languages are grouped, and this leads to a reduction in the total numbers of the parameters in the acoustic model. Fewer parameters get the same amount of training data in OPS and DT systems than that in LI-MCS system; therefore, we attained better results. On the other hand, the accuracy rate increased when we increased the number of Gaussian components in a Gaussian mixture per state even when the maximum number of Gaussian components in a Gaussian mixture was 64 for all languages. The results of using 64 Gaussian components in a Gaussian mixture are 45.97%, 58.94%, and 53.06% for Hakka, Mandarin, and Taiwanese, respectively. These are also the best performances achieved in this paper. Compared with the best results of LI-MCS system, we increase 3.96%, 1.49%, and 5.27% for Hakka, Mandarin, and Taiwanese, respectively.

The DT-ASR system merges similar phonemes from both likelihood scores and some phonetic knowledge, while the OPS system integrates all conditions of the unbalanced training corpus. That means both systems have their own knowledge rules to merge similar phonemes, but the OPS system generates the clustering rules after observing all of the training data. However, the rules of the DT-ASR system are generated by the phonetic expert, such as the questions which are used in splitting the nodes into the leaves. The different experts for different languages may have various rules, especially for the multilingual ASR-DT system. Furthermore, the rules for merging the similar phonemes of OPS are much easier than those in DT because we only need to follow the optimal Bayesian model selection criterion after observing all of the training data. Then, the results of CIP clustering constrain the CDP clustering. However, the DT-ASR system is not affected by the unbalanced conditions during the process of building a tree. Thus, the proposed OPS which uses the data-driven approach to generate the knowledge-like rules concerning about the conditions of the unbalanced training data achieves higher accuracy rates than that of DT-ASR system.

In our proposed OPS approach, compared with the DT-ASR, it is remarkable that the performance not only increases, but the performance gaps among the three languages are also reduced. Taking the results of Figure 12 and Figure 3 as examples, the gaps of Mandarin-Taiwanese and Mandarin-Hakka in Figure 3 are 9.26% and 16.74%, but the gaps in Figure 12 are 5.88% and 12.97%. Using the OPS approach, the performance of Taiwanese and Hakka catches up with the performance of Mandarin. This means our proposed OPS approach is able to reduce the performance gap which causes by the unbalanced condition of the available training data.

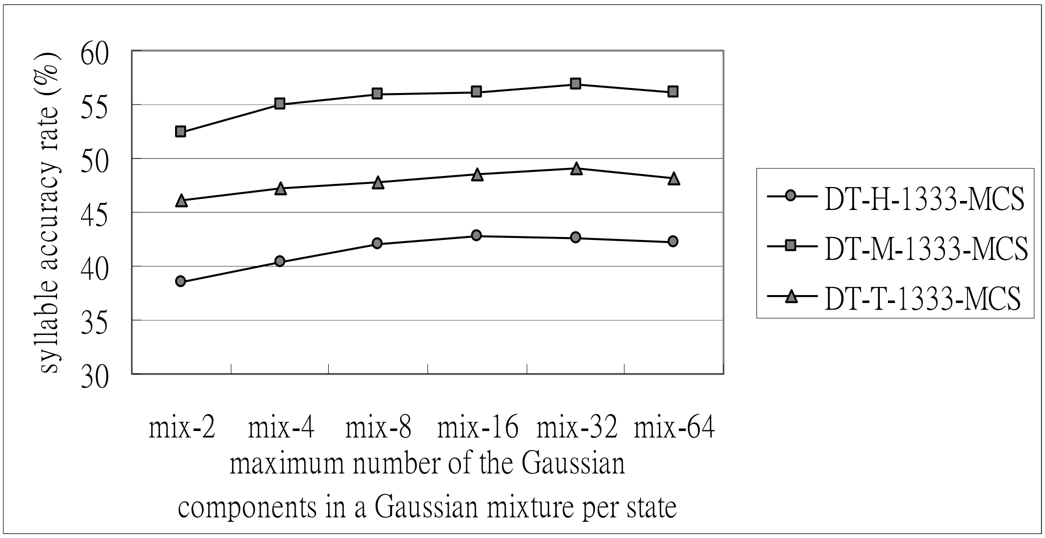

Figure 11. The syllable accuracy rates of DT with MCS in different maximum number of Gaussian components in a Gaussian mixture per state.

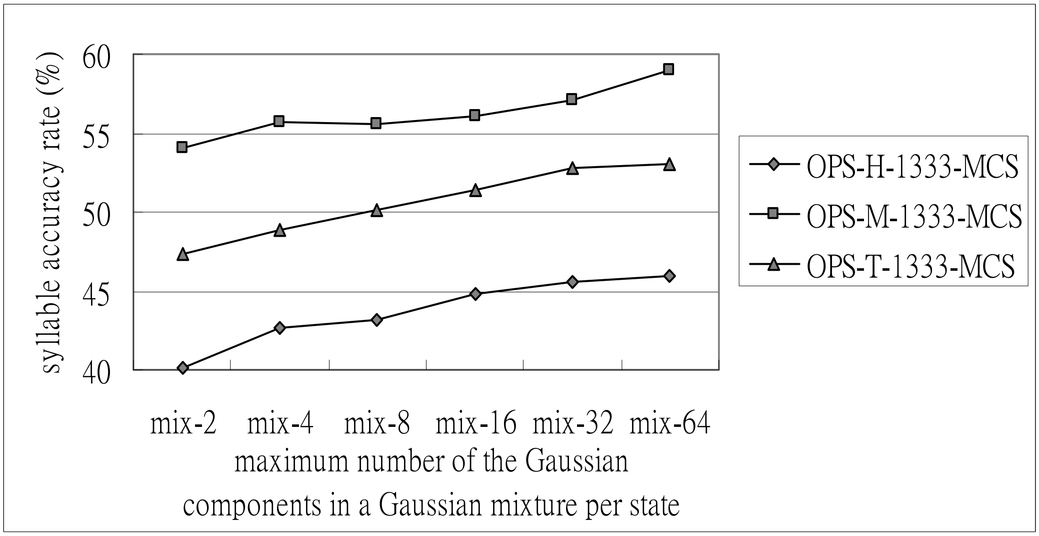

Figure 12. The syllable accuracy rates of OPS with MCS in different maximum number of Gaussian components in a Gaussian mixture per state.

6. Conclusion

In this paper, we have demonstrated a clustering approach with CIP, CDP clustering, and MCS steps with HAC and △BIC algorithm to generate the OPS of HMM-based acoustic model in an unbalanced trilingual corpus. It has been shown that the rules generated form the CIP provide sufficient information from the unbalanced conditions of the training corpus. When a trilingual corpus is unbalanced, we should put more emphasis on the characteristics of the corpus. We have shown that the rules derived from the data can reflect the properties of the corpus better and those rules are as good as the phonetic knowledge to smooth the unbalanced condition of the data. Besides, the MCS also plays an important role in balancing the OPS and the available training data. The experimental results are very encouraging in that the proposed approach reduces relative syllable error rate by 4.5% over the best result of the decision tree based approach and 13.5% over the best result of the knowledge-based approach.

Reference

Anguera, X., T. Shinozaki, C. Wooters, and J. Hernando, "Model Complexity Selection and Cross-Validation EM Training for Robust Speaker Diarization," In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing,* 2007, Honolulu, USA.

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

Kohler, J., "Multi-lingual Phone Model for Vocabulary-Independent Speech Recognition Task," *International Journal of Speech Communication*, (35), 2001, pp. 21-30.

Kumar, C. S., V. P. Mohandas, and H. Z. Li, "Multi-lingual Speech Recognition - A Unified Approach," In *Proceedings of Interspeech*, 2005, Lisbon, Portugal.

Liang, P. Y., J. L. Shen, and L. S. Lee, "Decision Tree Clustering for Acoustic Modeling in Speaker-Independent Mandarin Telephone Speech Recognition," In *Proceedings of the International Symposium on Chinese Spoken Language Processing*, 1998, Singapore, pp. 207-211.

Liu, Y., and P. Fung, "Automatic Phone Set Extension with Confidence Measure for Spontaneous Speech," In *Proceedings of Interspeech*, 2005, Lisbon, Portugal.

Lyu, D. C., B. H. Yang, M. S. Liang, R. Y. Lyu, and C.N. Hsu, "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition," In *Proceedings of SST*, 2002, Melbourne, Australia.

Lyu, D. C., and R. Y. Lyu, "Optimizing The Acoustic Modeling From An Unbalanced Bi-Lingual Corpus," In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, Las Vegas, USA.

Lyu, R. Y., M. S. Liang, and Y. C. Chiang, "Toward Constructing A Multi-lingual Speech Corpus for Taiwanese (Minnan), Hakka, and Mandarin," *International Journal of Computational Linguistics & Chinese Language Processing*, 9(2), 2004, pp. 1-12.

Mak, B., and E. Barnard, "Phone Clustering Using the Bhattacharyya Distance," In *Proceedings of International Conference on Spoken Language Processing*, 1996, Philadelphia, PA, USA, pp. 2005-2008.

Marthi, B., J. Morgan, N. Peterek, J. Picone, and W. Wang, "Towards Language Independent Acoustic Modeling," In *Proceedings of ASRU*, 1999, Keystone, USA.

Mathews, R. H., Chinese-English Dictionary, Caves, 13th printing, 1975.

Schultz, T., and K. Kirchhoff, Multi-lingual Speech Processing, Elsevier, Academic Press, 2006.

Schwarz, G., "Estimating The Dimension of A Model," *The annals of statistics*, 6(2), 1978, pp. 461-464.

Tritschler, A., and R. Gopinath, "Improved Speaker Segmentation And Segments Clustering Using The Bayesian Information Criterion," In *Proceedings of Interspeech*, 1999, pp. 679-682.

Uebler, U., "Multi-lingual Speech Recognition in Seven Languages," *International Journal of Speech Communication*, (35), 2001, pp. 53-69.

Wu, C. H., Y. H. Chiu, C. J. Shia, and C. Y. Lin, "Phone Set Generation Based On Acoustic and Contextual," In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2006, Toulouse, France.

Young, S. J., J. J. Odell, and P. C. Woodland, "Tree-based State Tying for High Accuracy Acoustic Modelling," In *Proceedings of the ARPA Workshop on Human Language Technology*, 1994, Berlin, German.

Young, S. P., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book, Version 3.2, Cambridge, 2002.