



Cross-lingual Speaker Adaptation via Gaussian Component Mapping

Houwei Cao, Tan Lee, and P. C. Ching

Department of Electronic Engineering, The Chinese University of Hong Kong
Hong Kong SAR of China

{hwcao, tanlee, pcching}@ee.cuhk.edu.hk

Abstract

This paper is focused on the use of acoustic information from an existing source language (Cantonese) to implement speaker adaptation for a new target language (English). Speaker-independent (SI) model mapping between Cantonese and English is investigated at different levels of acoustic units. Phones, states, and Gaussian mixture components are used as the mapping units respectively. With the model mapping, cross-lingual speaker adaptation can be performed. The performance of the proposed cross-lingual speaker adaptation system is determined by model mapping effectiveness and speaker adaptation effectiveness. Experimental results show that the model mapping effectiveness increased with the refinement of mapping units, and the speaker adaptation effectiveness depends on the model mapping effectiveness. Mapping between Gaussian mixture components is proved effective for various speech recognition tasks. A relative error reduction of 10.12% on English words is achieved by using a small amount of (4 minutes) Cantonese adaptation data, compared with the SI English recognizer.

Index Terms: speech recognition, speaker adaptation, cross-lingual, model mapping

1. Introduction

With the globalization of today's world, more and more people are able to speak two or more languages in their daily life. Research related to multi-lingual and cross-lingual speech has attracted much attention over the past few years, including multi-lingual speech recognition [1], mixed-language speech synthesis [2], language identification [3] and cross-lingual adaptation [4]. As one of the key research issues, cross-lingual adaptation makes it possible to make use of speech resources available in one or more languages for the recognition or synthesis of another target language. This allows fast and low-cost implementation of speech recognizers/synthesizers, which is especially useful for minority languages or dialects, in which data resources available are very limited or even not existent [5].

Hong Kong is a bilingual society, where Chinese and English are the official spoken languages. As a major working language in Hong Kong, English is widely used in commercial activities and legal matters. The usage of English, however, is much less than Cantonese in general conversational communication. In most cases, it's much easier to collect a small amount of Cantonese speech data from a specific Cantonese speaker for speaker adaptation purpose. Of course, it is also time and labour saving if we can perform adaptation on more than one languages by using only monolingual speech data captured from a desired speaker.

This paper concentrates on the use of acoustic information from an existing source language (Cantonese) to implement speaker adaptation for a new target language (English).

It is assumed English adaption data is not available for the target speaker. Speaker-independent (SI) and language-dependent acoustic models are trained for Cantonese and English respectively in the first step. Based on SI acoustic models, model mappings between Cantonese and English are established in different levels of acoustic units, viz phones, states, and Gaussian mixture components. With the model mapping, speaker adaptation on English models can be implemented by using Cantonese adaptation data from the target speaker.

The rest of the paper is organized as follows. Section 2 introduces the different model mapping schemes between Cantonese and English. In Section 3, we explain how to perform speaker adaptation via Gaussian component mapping. Experimental framework is described in Section 4. Results are discussed in Section 5. Finally, summary is drawn in Section 6.

2. Model Mapping Between Cantonese & English

In cross-lingual adaptation, one of the major problems is the model mapping between different languages. The mapping can be established for different acoustic units, such as words, syllables, phones, or others [6]. In this paper, we investigated the use of phones, states, and Gaussian mixture components for such purpose.

2.1. Phone Mapping

Phones are most widely used as the basic acoustic unit for model mapping [7]. The phone mapping table can be manually generated based on linguistic knowledge or automatically derived in a data-driven manner [8].

As an international standard of representing speech sounds of any spoken language, the International Phonetic Alphabet (IPA) is used to create mapping on the beginning. It classifies phones in terms of place and manner of articulation. Phones of different languages labeled by the same IPA symbol are considered as the same phone. However, Cantonese is a tonal language in the Sino-Tibetan family, while English is a stress-timed language in the Indo-European family. The phone sets of these two languages are significantly different. Only 17 phones can be shared according to their IPA symbols, 22 English-specific phones and 26 Cantonese-specific phones remain distinctively different. Similarity between these remaining phones can nevertheless be measured by their acoustic distributions. In this paper, Kullback-Leibler Divergence (KLD) is used. The phone mapping for the language-specific phones is created by Eq.(1). Each phone is modelled by speaker-independent, context-independent HMMs with single Gaussian distribution.

$$\hat{P}^c = \arg \min_{P^c} D_{KL}(P_i^c, P_j^e) \quad (1)$$

where, P_i^c is a phone in Cantonese phone set P^c , P_i^e is a phone in English phone set, and D_{KL} is the K-L divergence between two phones.

2.2. State-level Mapping

Since Cantonese and English are two languages highly uncorrelated phonetically, there might not exist much similarity in their phonemic counterparts. However, since the speech production is constrained by limited movement of articulators, it might be possible to find some similar acoustic units at a refined, sub-phone level. Diphthongs may be rendered by several monophthongs. Furthermore, allophones, which are highly context dependent, provide more chances for phone sharing between different languages. As a result, a tied, context-dependent state-level mapping is investigated between Cantonese and English. First, we built two speaker-independent, language-specific decision trees for Cantonese and English respectively. Each leaf node in decision tree represents a tied state, modeled by a Gaussian distribution. For each English tied states, a corresponding Cantonese tied state can be found, in the minimum KLD sense. The directional mapping from English to Cantonese can be a one-to-many mapping. Different leaf nodes in the English tree may map to the same leaf node in the Cantonese tree.

In order to achieve satisfactory recognition performance, multiple Gaussian mixture components are typically used. Similarity among states may change along with the mixture splitting. Therefore, two model mapping schemes are studied, namely the **State Mapping** and **CalState Mapping**.

In **State Mapping**, mapping is estimated based on the tied states with single Gaussian models, and the mapping would not change with the incrementation of mixture components. In **CalState Mapping**, the model distribution of each state is recalculated by Eq.(2) in multiple mixture cases,

$$CalS_i = \sum_{k=1}^K w_{ik} S_{ik} \quad (2)$$

where w_{ik} is the mixture weight of the k th mixture of state S_i , S_{ik} is the output distribution of the k th mixture in state S_i , and $CalS_i$ is the recalculated distribution of state S_i . Then, the mapping is established based on the distribution $CalS_i$. **CalState Mapping** will be created repeatedly along with the increase of mixtures, until the desired number of mixture components is reached.

2.3. Mapping among Gaussian Mixture Components

In multiple-mixture HMMs, Gaussian mixture components are the smallest elements. Furthermore, in speaker adaptation based on MLLR or MAP algorithms, adaptation is usually applied to individual mixture components in the model set. In this paper, the Gaussian component mapping, denoted by **GauMix Mapping**, is investigated between Cantonese and English. For each mixture component in English, the corresponding one in Cantonese is found by minimizing the K-L divergence,

$$\hat{M}^c = \arg \min_{M^c} D_{KL}(M_i^c, M_j^e) \quad (3)$$

where M_i^c is a mixture component in Cantonese model set M^c , M_j^e is a mixture component in English models, and D_{KL} is the K-L divergence between two mixture components. Similar to **CalState Mapping**, **GauMix Mapping** needs to be re-estimated as the number of mixtures increases.

2.4. Kullback-Leibler Divergence

KLD is an information-theoretic measure of similarity between two probability distributions. It has been used in various applications. In this study, KLD is used to measure the difference between HMMs of two speech sounds [9]. The KLD between two given distributions Q and R is defined as,

$$D_{KL}(Q||R) = \int q(x) \log \frac{q(x)}{r(x)} dx \quad (4)$$

where q and r denote the densities of Q and R. Then, the symmetric form of KLD between Q and R is,

$$D_{KL}(Q, R) = D_{KL}(Q||R) + D_{KL}(R||Q) \quad (5)$$

3. Cross-lingual Speaker Adaptation

In speech recognition, many speaker adaptation techniques have been successfully applied. HMM-based adaptation using MLLR or MAP techniques can be used to improve the recognition performance using a small amount of adaptation data from target speakers [10]. However, it is not easy to do it across different languages, especially when two languages are phonetically distant.

In this paper, cross-lingual speaker adaptation is implemented via the model mapping strategy as established in Section 2. Figure 1 gives an example of cross-lingual adaptation with **GauMix Mapping**. We want to adapt English SI models with Cantonese adaptation data from speaker A. Since the SI mapping has been created, HMM parameters at each mixture component of the English models can be replaced by the Cantonese counterpart. When standard intra-language adaptation is implemented on Cantonese SI models with Cantonese adaptation speech from speaker A, we will get Cantonese speaker adapted (SA) models for speaker A. Then, English SA models can be retrieved by Cantonese SA models via model mapping.

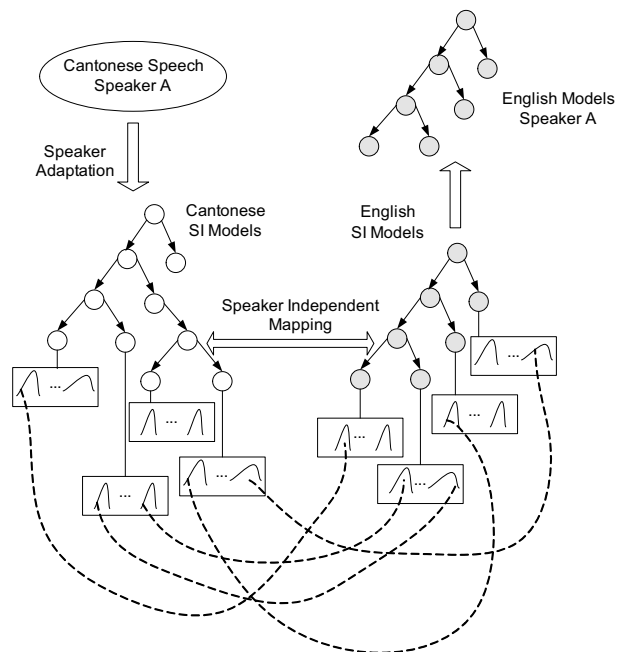


Figure 1: Cross-lingual speaker adaptation via Gaussian mixture component mapping

4. Experimental Setup

4.1. Speech Corpus

All the speech data used in our experiments are from the Cantonese-English bilingual database CUMIX [11]. It contains 16 hours of speech data from 74 native Cantonese speakers. Cantonese speech data in CUMIX are colloquial utterances. The spoken contents are mainly daily conversations or jargons by universities students in Hong Kong. English speech data in CUMIX are English words or phrases, which carry Cantonese accent to certain extent.

4.2. SI Acoustic Models

Speaker independent acoustic models are trained from the utterances of 60 speakers. 10 minutes of Cantonese and 4 minutes of English speech are available for each speaker. The acoustic feature vectors consist of 12 MFCC coefficients, log energy, and their first and second-order derivatives. The SI models are trained from context-independent monophone HMMs to context-dependent triphone HMMs, from single Gaussian to 8 mixture components.

4.3. Speaker Adaptation Setup

Cross-lingual adaptation experiment is performed on 14 different speakers, which are not included in the 60 speakers used in SI acoustic modeling. For each speaker, 4 minutes of Cantonese adaptation speech data are available. We use MLLR followed by MAP adaptation techniques in this paper. For MLLR, 32 regression classes are used in the experiments. The adapted models are evaluated in speech recognition experiments. The test speech are English words and phrases. The vocabulary size is 1.2k, and no language model is applied.

5. Results & Discussions

The performance of the proposed cross-lingual speaker adaptation system is mainly determined by two factors: model mapping performance and speaker adaptation performance.

5.1. Model Mapping Effectiveness

Model mapping performance is measured in terms of mapping effectiveness, which is evaluated by recognition performance with mapped acoustic models. Mapped SI models can be retrieved by Cantonese SI models via mapping. The recognition results of different mapped SI models are shown in Figure 2. Results from English SI models are also given as a reference. The performance degradation from English acoustic models to mapped models is defined as the mapping loss in this study. The lower the mapping loss, the higher the mapping effectiveness.

From the recognition results, it is clear that model mapping effectiveness increased with the refinement of mapping units. Due to the significant phonetic difference between Cantonese and English, it is not surprising that poor recognition results are found in the **Phone Mapping** approach. State-level mapping systems improve greatly in recognition English words due to better matched model mapping. **CalState Mapping** and **State Mapping** schemes show similar performance. However, the performance variation from single mixture to multiple mixture is not obvious. On the other hand, **English SI** models gave improved performance with increasing number of mixtures. The mapping loss found in state-level mapping increased with the mixture component splitting, model mapping effectiveness of

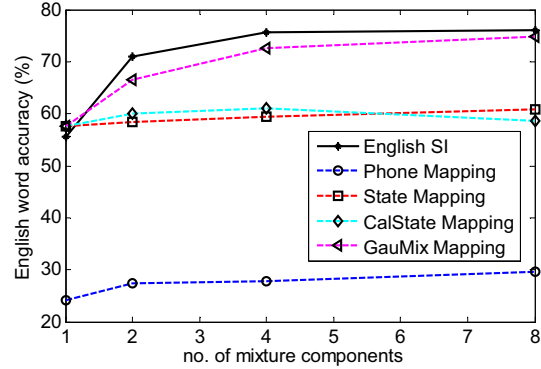


Figure 2: Recognition results from different SI models

state-level mapping is still on low side. The recognition performances of **GauMix Mapping** models however improved with mixture number increasing, and insignificant mapping loss is found between **English SI** model and **GauMix Mapping** ones. That means the acoustic space represented by gaussian components from different language are actual close to each other.

Further analysis is done by comparing the effective Cantonese states found in different model mapping strategies. There are 3248 Cantonese candidate states and 969 English states available in SI acoustic models for mapping creation. Table 1 compares the effective Cantonese states found in **CalState Mapping** and **State Mapping**. With the number of mixtures increase, effective Cantonese states in **CalState Mapping** and **State Mapping** are becoming more and more dissimilar. However, the total number of effective Cantonese states is almost unchanged with mixture splitting. This may be the reason why the recognition performance of the state-level mapped models is always the same even under different mixture component cases. The mapping details about **GauMix Mapping** are given in Table 2. It's clear that the number of effective Cantonese states increased along with mixture splitting. The resolution of **GauMix Mapping** is higher than state-level mapping. Even for fairly distant states, we may still find similar mixture components among them. This also explains why **GauMix Mapping** outperforms state-level mapping approach.

Table 1: Cantonese states found in **CalState Mapping**.

| | used states | shared states with State Mapping |
|-------|-------------|---|
| 1 mix | 566 | 566 |
| 2 mix | 571 | 409 |
| 4 mix | 590 | 356 |
| 8 mix | 591 | 272 |

Table 2: Cantonese states found in **GauMix Mapping**.

| | used mixtures | used states |
|-------|---------------|-------------|
| 1 mix | 566 | 566 |
| 2 mix | 1125 | 947 |
| 4 mix | 2266 | 1491 |
| 8 mix | 4455 | 2076 |

5.2. Cross-language Speaker Adaptation Results

Cross-lingual speaker adaptation results are summarized in Table 3. If the mapping loss exceeds the speaker adaptation improvement, the performance obtained with cross-lingual

speaker adaptation would be even worse than the monolingual SI recognizer, such as the speaker adaptation results found with **State Mapping**. Mapping between Gaussian mixture components has been proved effective in speech recognition task in Section 5.1. Via the **GauMix Mapping**, the adaptation models give an average of 78.70% word accuracy on English over all speakers. A relative 10.13% reduction on word error rate (WER) is achieved, compared with 76.30% word accuracy obtained with English SI models. Furthermore, it is found that the effectiveness of speaker adaptation is highly correlated with the mapping effectiveness. If the mapping loss is ignored, the improvement from adaptation is very limited in **State Mapping** case. However, the adaptation lead to a relative 13.67% error reduction via **GauMix Mapping**, compared with the mapped SI models. The degree of improvement from speaker adaptation increased with the improved model mapping effectiveness. Speaker adaptation experiments have also been performed on intra-language basis for reference purpose. By applying the same adaptation data, the relative error reduction on Cantonese is 19.21% on average. The experimental results show that our approach for cross-lingual speaker adaptation is promising.

Table 3: Cross-lingual Adaptation results for individual speakers (% word accuracy), 4 minutes of Cantonese adaptation speech are used.

| Speaker | SI | State Mapping | | GauMix Mapping | |
|------------|--------------|---------------|--------------|----------------|--------------|
| | | Mapped | Adapted | Mapped | Adapted |
| spkr1 | 73.42 | 58.23 | 56.96 | 73.42 | 79.75 |
| spkr2 | 72.00 | 56.00 | 54.67 | 70.67 | 76.00 |
| spkr3 | 64.38 | 56.16 | 60.27 | 69.86 | 68.49 |
| spkr4 | 79.27 | 62.20 | 65.85 | 74.39 | 84.15 |
| spkr5 | 83.95 | 67.90 | 72.84 | 85.19 | 83.95 |
| spkr6 | 89.61 | 68.83 | 67.53 | 83.12 | 89.61 |
| spkr7 | 78.31 | 55.42 | 46.99 | 81.93 | 78.31 |
| spkr8 | 79.45 | 64.38 | 67.12 | 79.45 | 83.56 |
| spkr9 | 77.50 | 62.50 | 70.00 | 80.00 | 82.50 |
| spkr10 | 80.28 | 63.38 | 69.01 | 77.46 | 81.69 |
| spkr11 | 67.95 | 55.13 | 53.85 | 60.26 | 65.38 |
| spkr12 | 74.65 | 50.70 | 50.70 | 76.06 | 81.69 |
| spkr13 | 70.00 | 51.25 | 56.25 | 67.50 | 68.75 |
| spkr14 | 76.62 | 64.94 | 64.94 | 76.62 | 77.92 |
| Ave | 76.30 | 59.81 | 61.20 | 75.46 | 78.70 |

Further analysis is done by investigating adaptation using different amounts of data. For each speaker, only 4 minutes of adaptation speech are available. We divide it into 3 adaptation subsets, which contain 2, 3 and 4 minutes of speech respectively. The adaptation results are shown in Figure 3. It is clear that noticeable improvement can be found, along with an increase of adaptation speech. It is believed that further improvement could be achieved if there is more adaptation data available. In addition, if a speaker-dependent Cantonese recognizer is existed for a particular speaker, the corresponding speaker-dependent English recognizer can be implemented via the proposed **GauMix Mapping** in a fast and low cost way.

6. Summary

Speaker adaptation techniques can be used to improve speech recognition performance if a small amount of adaptation data from the target speaker is available. However, it is not easy to do it across different languages, especially when two languages are phonetically distant. Cross-lingual speaker adaptation via

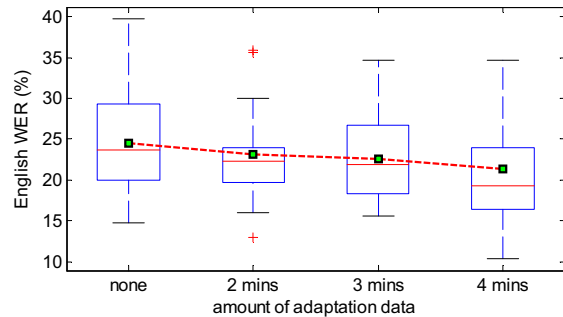


Figure 3: Boxplots for cross-lingual speaker adaptation results with different amount of adaptation data, pooling all target speakers.

model mapping is investigated in this paper. SI model mapping between Cantonese and English is established for different acoustic units. Experimental results show that the model mapping effectiveness increased with the refinement of mapping units, and the degree of improvement from speaker adaptation depends on the mapping effectiveness. Cross-lingual speaker adaptation via state level mapping is not efficient due to the significant mapping loss. Mapping between Gaussian mixture components is proved effective for various speech recognition tasks. A relative error reduction of 10.12% on English words is achieved by using a small amount of (4 minutes) Cantonese adaptation data. The experimental results show that our approach for cross-lingual speaker adaptation is promising.

7. Acknowledgements

The authors would like to thank Dr. Yao Qian for providing inspiration to this research.

8. References

- [1] T. Schultz and K. Kirchhoff (eds.), *Multilingual Speech Processing*, Elsevier Inc., 2006.
- [2] Yao Qian, Houwei Cao, Frank K. Soong, "HMM-based Mixed-language (Mandarin-English) Speech Synthesis", *Proc. ICSLSP 2008*.
- [3] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech", *IEEE Trans. Speech and Audio Processing*, Vol. 4, No. 1, pp. 31-44, 1996.
- [4] Viet Bac, Laurent Basacier, "First steps in fast acoustic modeling for a new target language: application to Vietnamese", *Proc. ICASSP 2005*.
- [5] J Latorre, K Iwano, S Furui, "Polyglot synthesis using a mixture of monolingual corpora", *Proc. ICASSP 2005*.
- [6] S. Maskey, L. Tomokiyo, A. Black, "Bootstrapping phonetic lexicons for new languages", *Proc. Interspeech 2004*.
- [7] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds", *Proc. ICSLP 1996*.
- [8] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition" *Speech Communication*, vol. 35, no. 1-2, pp. 31C51, 2001.
- [9] T. A. Myrvoll, and F. K. Soong, "Optimal Clustering of Multivariate Normal Distributions Using Divergence and Its Application to HMM Adaptation", *Proc ICASSP 2003*.
- [10] CJ Leggetter, PC Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer speech and language*, pp. 171-185, 1995.
- [11] Joyce Y. C. Chan, P. C. Ching, and Tan Lee, "Development of a Cantonese-English code-mixing speech corpus," *Proc. Interspeech 2005*.