# Acoustic Model Merging Using Acoustic Models from Multilingual Speakers for Automatic Speech Recognition

Tien Ping Tan, Laurent Besacier, Benjamin Lecouteux

# Acoustic Model Merging Using Acoustic Models from Multilingual Speakers for Automatic Speech Recognition

Tien-Ping Tan
School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia
tienping@cs.usm.my

Laurent Besacier, Benjamin Lecouteux
Equipe GETALP,
Laboratoire d'Informatique de Grenoble,
UJF BP53 38041 Grenoble Cedex 9
laurent.besacier@imag.fr, benjamin.lecouteux@imag.fr

***Abstract*—Many studies have explored on the usage of existing multilingual speech corpora to build an acoustic model for a target language. These works on multilingual acoustic modeling often use multilingual acoustic models to create an initial model. This initial model created is often suboptimal in decoding speech of the target language. Some speech of the target language is then used to adapt and improve the initial model. In this paper however, we investigate multilingual acoustic modeling in enhancing an acoustic model of the target language for automatic speech recognition system. The proposed approach employs context dependent acoustic model merging of a source language to adapt acoustic model of a target language. The source and target language speech are spoken by speakers from the same country. Our experiments on Malay and English automatic speech recognition shows relative improvement in WER from 2% to about 10% when multilingual acoustic model was employed.** *(Abstract)*

***automatic speech recognition; context dependent acoustic model merging; multilingual approach***

## I. INTRODUCTION

Training a robust acoustic model for automatic speech recognition (ASR) system requires a large speech corpus. However, a lot of time is required and high cost involves in preparing and acquiring a corpus. Thus, many studies have explored on the usage of existing multilingual speech corpora in building an acoustic model for a target language. Schultz (2001) had categorizes the works on multilingual acoustic modeling into three main categories [1]: cross-language transfer without requiring any target language data [2], limited data for adapting an initial model [3, 4] and bootstrapping approach using an initial model created from multilingual acoustic models [2], which is then adapted using target data. These works on multilingual acoustic modeling attempt to use multilingual acoustic models to create an initial model which might be adapted with some speech from target language. This initial model created is often suboptimal in decoding the target speech compared to the model created in the target language.

In this paper, we investigate multilingual acoustic modeling from a difference angle than described in [1]. Instead of using multilingual acoustic models to create an initial model, our work starts with an acoustic model in the target language, and the purpose of the multilingual resource is used to enhance the recognition capability of the acoustic model in an ASR system. The approach requires only the source acoustic model (acoustic model in source language) to adapt the target acoustic model (acoustic model in target language), without requiring the speech from the source or target language. This study was carried out specifically on acoustic models trained from speech of multilingual speakers. In other words, the acoustic models is built from speech of multilingual speakers. Our hypothesis is that multilingual speakers pronounce in a similar manner when they speak different languages. Thus, similar acoustic units in different languages can be combined to cross adapt each other for automatic speech recognition by benefiting from more acoustic context varieties and more variety of speakers.

As more and more multilingual automatic speech recognitions are developed, there is a lot of independent works being done. People are increasingly willing to share resources such as acoustic models. Thus, the possibility that other acoustic models can be jointly used to produce a better acoustic model will be an interesting one. Our study was conducted on Malay and English, which is spoken by Malaysian. Malaysia is a multicultural and multilingual society. Although the national language is Malay, other languages are also widely used such as English, Mandarin, and Tamil. Malay and English are compulsory subject in school and were taught since primary school. Thus, most people are fluent in Malay and English. This make Malay and Malaysian English suitable testing subjects. The study attempts to improve Malay and English acoustic model for automatic speech recognition by using these models to cross adaptation each other. We assume the acoustic model is model using HMM with Gaussian mixtures.

## II. BACKGROUND

We investigate using acoustic model merging to cross-adapt acoustic models. Acoustic model merging requires only acoustic models, without any raw speech data. It involves combining two or more acoustic models (Gaussian mixture model) from normally two different sources. Acoustic model merging has been used in non-native acoustic model adaptation. Often, the target acoustic model is merged with the corresponding native language acoustic model of the non-native speaker [5, 6, 7] to form a new model to decode non-native speech. The idea is that different speakers are likely to use different strategies to pronounce a sound. In this case, it is either the target language speech sound or the native speech sound of the speaker. There are two ways to merge an acoustic unit. See

Figure 1. A weight will be assigned to each of the merged model, either on the transition of each merged acoustic unit (Figure 1a) or into the mixture weights (Figure 1b). Acoustic model merging increases the number of states or Gaussians in each HMM.

Acoustic units defined in the acoustic models can be phones, graphemes, syllable, words or others. Phones is the most popular acoustic units used in the acoustic model. A phone is a distinct speech sound which is related to the phoneme. A phoneme is the smallest contrastive unit in the sound system of a language, which may change the meaning of a word. International Phonetic Table (IPA) defines a standard for phoneme and phone.

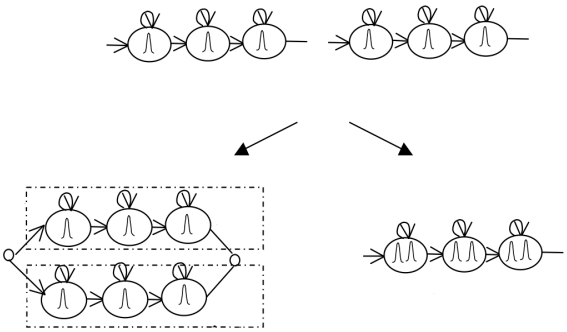

Figure 1. Acoustic model merging.

## III. CONTEXT DEPENDENT ACOUSTIC MODEL MERGING

The first step in acoustic model merging is to determine the matching acoustic units for merging between the target and source acoustic models. There are few approaches to determine the matching acoustic unit. If the acoustic units are phones based on IPA, the IPA table can be used to find the matching phones. For similar phone in two languages, such as [b] in Malay and [b] in English, this type of mapping can be set directly. As for unique phones that do not exist in both languages, mapping can be determined based on studies from acoustic phonetics or based on other approaches [3, 8]. The second approach is to conduct a perception test. For each target phone, speakers are given a list of source phones that are similar to the target phones. The speakers select zero or more source phones that are similar to target phones. The third approach is by using forced alignment. Context independent acoustic models are required in this test. A list of test sentences in the target language are forced aligned using the target acoustic model, and the acoustic scores from the alignment are recorded. The same sentences are then forced aligned with target acoustic model, which has been merged with a specific acoustic unit from source acoustic model. The source unit that gives the highest improvement in acoustic score will be chosen as the match. Other approaches such as creating a confusion matrix between the phoneme decoding by a source phoneme recognizer and a forced alignment from a target speech recognizer can also be applied [3]. There are two approach of merging. We applied the merging approach shown in Figure 1b where the Gaussians in the source state are merged into the matching target state. For context independent acoustic model merging, the same state in the matching target and source acoustic unit are merged. For example, Gaussians in state 1 of phone [b] in English is merged into state 1 of phone [b] in Malay. See Figure 2.

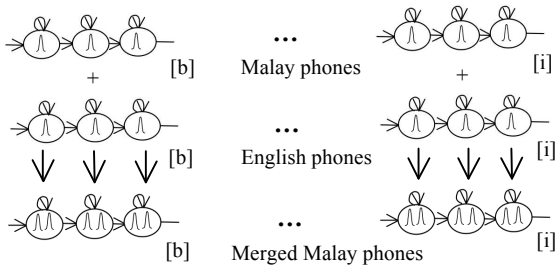


Figure 2. Merging of phone in context independent acoustic model.

However, the merging procedure of target and source context dependent acoustic model is more complex. Triphones are modelled by context dependent acoustic model. Triphone is a context dependent phone that takes into consideration the left and right context/phone of a current phone. For example, the word "phone" [f oʊ n] consists of triphone sil-f+oʊ, f-oʊ+n, oʊ-n+sil, assuming there is a silence (sil) in front and at the end of the word. The number of triphones for a language is very large and the distribution of each triphones varies, some triphones appear more frequently than others. Thus, to model each triphone separately is not possible because of rare and unseen triphones. To overcome this problem, decision tree is often used to tie similar triphones together, so that similar triphone can be modelled together. A snippet example of a decision tree used in Kaldi ASR system [9] is shown in Figure 3. At each node of a decision tree, question is asked about the context of a triphone (left phone, right phone, center phone and pdf id). Linguistic questions can also be asked. The Kaldi decision tree consists of only a single binary tree (while some other decision tree algorithms build separate tree for different phone).

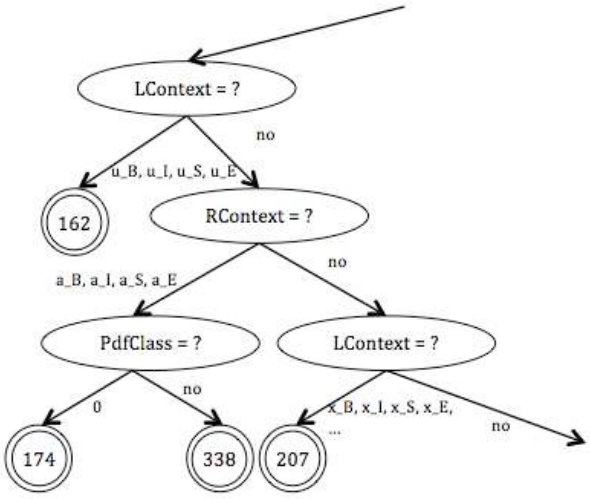


Figure 3. Decision tree in Kaldi ASR system.

All similar triphones will be classified together in the same leaf of a decision tree. Each leaf corresponds to a HMM state. Since triphones classify in a leaf is not

obvious by just looking at a decision tree, we will need to captures possible/realistic triphone sequences and their statistics from a text. The possible triphone sequences in this case can be obtained from the transcription or text corpus using a pronunciation dictionary.

The first step of context dependent merging is to add all target language triphones into the target decision tree. Similar triphones are classify together in the same leaf. Thus, each leaf may contains more than one triphone. A target triphone will be converted to a source triphone using the acoustic unit mapping found earlier. The source triphone created will be added into the source tree to find out the source state id. It is possible for a target leaf to have more than one matching source leaf because there are more than one triphone in a leaf, and different triphones in a leaf may not map to the same source state id. We tested two scenarios. The first approach selects only the source state id with the most triphone mapping for each target state id for merging. The second approach select all matching source state id for a target state id for merging. A weight is applied on the target and source mixture weights. The algorithm then traverse all leaves of the tree to find out the content of the triphones in every leaf. The merged acoustic model is then used for decoding in ASR. The approach is shown in Figure 4.

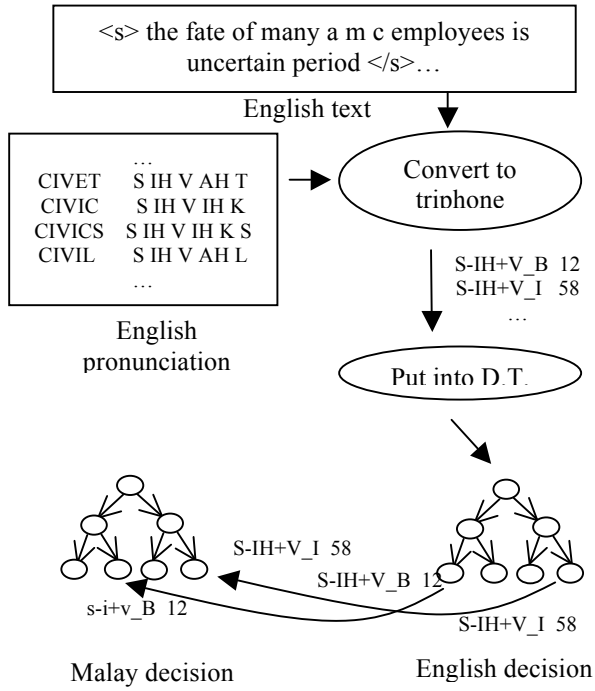


Figure 4. Merging of decision tree in context dependent acoustic model.

## IV. Experiments and Discussion

### A. Experiment Setup and Baseline ASR results

Experiments were carried out on Malay and non-native English (Malaysian English) using Kaldi automatic speech recognition system [9]. A context dependent Malay acoustic model with 4000 states and 40084 Gaussians was trained using a subset of MASS Malay read speech corpus with about 120 hours of speech [10]. A subset of 20 hours of MASS corpus was used for testing. On the other hand, the English acoustic model with 1000 states and 10017 Gaussians was trained using corpus which contains about 9 hours of speech, while the English test set consist of about 4 hours of speech. The English corpus is small because it is a non-native English corpus. Both MASS Malay and English speech corpus contain speakers of different origin, which are Malay, Chinese, Indian and others. For the pronunciation dictionary, the MASS pronunciation dictionary contains more than 60k words was used in Malay ASR, while the CMU pronunciation dictionary consists of more than 100k words was used in English ASR.

The baseline word error rate (WER) for Malay and English ASR system is given in Table 1. The baseline result using monophone, triphone (13 MFCC feature vector), and triphone (39 MFCC + delta feature vector).

Table 1. WER for baseline ASR system.

| Language | Monophone | Triphone (13) | Triphone (39) |
|---|---|---|---|
| Malay | 14.3% | 7.36% | 7.36% |
| English | 40.6% | 26.12% | 26.15% |

### B. Acoustic Model Merging Setup

The phone mapping is determined using IPA table for similar phones, while the mapping for unique phones in each language is determined using perception test. The phone mapping from Malay to English is given in Table 2, while the phone mapping from English to Malay is given in Table 3. Note that the Malay phone is in IPA character, while the English phone is in ARPAbet (because certain ARPAbet phone does not map to a single phone in IPA). All Malay phone has a corresponding English phone mapping except the glottal stop [?] which has no mapping.

Table 2. Malay-English phone mapping

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mal | ? | j | w | p | b | t | d | k | g | s | h | f |
| Eng | - | Y | W | P | B | T | D | K | G | S | HH | F |
| Mal | v | z | ʃ | x | G | tʃ | ʤ | l | r | m | n | ŋ |
| Eng | V | Z | SH | K | G | JH | CH | L | R | M | N | NG |
| Mal | ɲ | a | e | ə | i | o | u | aj | aw | oj | | |
| Eng | - | AH | EH | ER | IY | OW | UW | AY | AW | OY | | |

Table 3 shows the phone mapping from English to Malay. The mapping is similar to the mapping in Table 2 but in reverse direction.

Table 3. English-Malay phone mapping

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Eng | Y | W | P | B | T | D | K | G | S | HH | F | V | Z | SH |
| Mal | j | w | p | b | t | d | k | g | s | h | f | v | z | ʃ |
| Eng | JH | CH | L | R | M | N | NG | AH | EH | ER | IY | OW | UW | AY |
| Mal | tʃ | ʤ | l | r | m | n | ŋ | a | e | ə | i | o | u | aj |
| Eng | AW | OY | AA | DH | EY | TH | IH | AE | AO | ZH | UH | | | |
| Mal | aw | oj | a | - | - | - | i | - | - | - | - | | | |

For extracting the target triphones, we tried a few approaches, such as extracting triphone sequence from the target transcript, extracting triphone sequences from the text corpus of the target language, and extracting triphones sequence that match both target and source list. The total number of Gaussians in the merged target acoustic model do not differ very much. The WER also do not differ very much (less than ±0.1% absolute WER). The results presented below employ the matching target and source triphone list. In the acoustic model merging, a weight of 0.9 is set on the target acoustic model, while 0.1 is set to the source acoustic model. Table 3 and 4 show the WER and relative improvement in WER for Malay ASR and English ASR under different setting. Experiment were carried out by merging only the source state with the highest number of triphones and also merging of all source states for each target state.

### *C. Results*

The results shows improvement of WER under all settings. Merging all source state for every target state produces better improvement in WER than only merging a single source state. However, the total number of Gaussians in the merged model is much more. When monophone source acoustic model was merged with the monophone target acoustic model, the relative WER improvement produced in Malay ASR is 6.8%, while in English is (3.6%).

Table 3. WER for Malay ASR system using the merged acoustic model.

| Malay | Features | # GMM | WER (rel. improvement) |
|---|---|---|---|
| Monophone | 13 MFCC | 1751 | 13.33% (+6.8%) |
| Triphone (1-Best source state) | 13 MFCC | 76707 | 6.91% (+6.1%) |
| Triphone (All source states) | 13 MFCC | 96423 | 6.78% (+7.9%) |
| Triphone (1-Best source state) | 39 MFCC + delta | 77517 | 6.83% (+7.2%) |
| Triphone (All source states) | 13 MFCC + delta | 97154 | 6.72% (+9.5%) |

Table 4. WER for Malaysian English ASR system using the merged acoustic model.

| Malaysian English | Features | # GMM | WER (rel. improvement) |
|---|---|---|---|
| Monophone | 13 MFCC | 1920 | 39.15% (+3.6%) |
| Triphone (1-Best source state) | 13 MFCC | 18829 | 25.39% (+2.9%) |
| Triphone (All source states) | 13 MFCC | 47577 | 25.18% (+3.7%) |
| Triphone (1 source state) | 39 MFCC + delta | 18548 | 25.44% (+2.7%) |
| Triphone (All source states) | 13 MFCC + delta | 49546 | 24.95% (+4.5%) |

We analyse further the WER produced by analysing the result according to speaker origin. Table 5 shows a much more detail result of the decoding of the MASS test corpus (Triphone, 1-Best source state, 39 MFCC+delta). The results show that all speakers show improvement in WER.

Table 5. WER for different races of speaker in Malay ASR system using the merged acoustic model. Note: CN – Malaysian Chinese, MY – Malaysian Malay, and ID – Malaysian Indian, f – female, m – male.

| Races | CN (f) | CN (m) | MY (f) | MY (m) | IN (f) | IN (m) |
|---|---|---|---|---|---|---|
| base -line | 5.3% | 6.3% | 10.2% | 7.9% | 7.3% | 6.1% |
| tri phone | 4.9% | 6.0% | 9.7% | 7.0% | 6.5% | 5.5% |

## V. CONCLUSIONS

Experiment results show encouraging improvement in WER, from about 3% until 10% relative improvement in WER. The shortcoming of the approach is the increase in the number of Gaussians, which is about four time the original. Future work can be done on reducing the number of Gaussians by removing similar Gaussians in a state and also speaker adaptive acoustic modelling.

## REFERENCES


[1] T., Waibel, A., "Language-independent and language-adaptive acoustic modeling for speech recognition", Speech Communication Journal, 35, 31-51, 2001.

[2] Lin, H., Deng, L., Yu, D., Gong, Y.-F., Acero, A., Lee, C.-H., A study on multilingual acoustic modeling for large vocabulary ASR, ICASSP'09, 4333-4336, Taipei, 2009.

[3] Le, V. B. and Besacier, L., "First step in fast acoustic modeling for a target language: Application to Vietnamese", ICASSP'05, Philadelphia, 2005.

[4] Imseng, D., Motlicek, P., Bourlard, H., Garner, P. N., "Using out-of-language data to improve an under-resourced speech recognizer", Speech Communication Journal, 56, 142-151, 2013.

[5] Witt, S., and Young, S., Off-line acoustic modelling of non-native accents, Eurospeech'99, 1367-1370, Budapest, 1999.

[6] Morgan, J. J., Making a speech recognizer tolerate non-native speech through Gaussian mixture merging, ICALL'04:, Venice, 2004.

[7] Bouselmi, G., Fohr, D., and Haton, J.-P., Fully automated non-native speech recognition using confusion -based acoustic model integration, Eurospeech'05, Lisboa, 1369-1372, 2005.

[8] Tan, T.-P., Automatic speech recognition for non-native speakers", Dissertation, Université Joseph Fourier, Grenoble, 2008.

[9] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011, The Kaldi speech recognition toolkit, Workshop on Automatic Speech Recognition and Understanding, Hawaii.

[10] Tan, T.-P., Li, H., Tang, E.-K., Xiao, X., Chng, E.-S., 2009, MASS: A Malay language LVCSR corpus resource, Proceeding of Cocosda'09, 25-30, Beijing.