Automatic Speech Recognition for Non-Native Speakers

Tan Tien Ping

UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1

N^0 attribué par la bibliothèque

/__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1

Discipline : Informatique

présentée et soutenue publiquement

par

Tan Tien Ping

le 3 juillet 2008

Titre :

Automatic Speech Recognition for Non-Native Speakers

Directeurs de thèse : Laurent Besacier et Eric Castelli

JURY

M. Eric Gaussier	Président
M. Jean-François Bonastre	Rapporteur
Mme. Irina Illina	Rapporteur
M. Dirk Van Compernolle	Examinateur
M. Laurent Besacier	Directeur de thèse

Thèse préparée au sein du laboratoire de Groupe d'Etude en Traduction / Traitement des Langue et de la Parole, Laboratoire LIG – Université Joseph Fourier – Grenoble I

To mom and dad

Acknowledgements

Ca y est! Finally, I will be reaching the ‘finishing line’. It has been the toughest period of my life. Many people have assisted me to go as far as I am now.

First of all, I am particularly indebted to Laurent Besacier, who is also my supervisor, for accepting me to the *GETALP/ LIG* laboratory, and his guidance in the course of my study. He has not only gave many insightful advices, comments and criticisms for improving my research, even when he was at overseas, but also provided many opportunities and supports for me to participate in countless conferences, as well as opportunity to work in other laboratory, that improved my knowledge in the field. I would like to express my gratitude to Eric Castelli, for his guidance.

Special thanks to Jean-François Serignat, the leader of the speech team in *GETALP* for his kindness and helps during my stay. I would also like to thank Brigitte Bigi for her valuable ideas which help me progress in my works. Not to forget, Le Viet Bac, my office mate who has helped me in various areas, from administration paper works to research, which I’m tremendously grateful. I like to thank Professor Christian Boitet from *GETALP* for his recommendation to join this laboratory.

I would like to thank Martine Faraco from *Laboratoire Parole et Langage, Université de Provence* in Aix-en-Provence for welcoming me to her laboratory, and helping me to see my work from the phonetic perspective.

I also owe a dept of gratitude to Tang Enya Kong, Ranaivo-Malançon Bali and Professor Zaharin Yusoff from *Universiti Sains Malaysia*, Penang, who have assisted and supported me. Without them, I will not be in France to realise my dream.

I would like to take the opportunities to thank again all the volunteers that have participated in the tests. Last but not least, I would like to thank my family and friends that are always by my side, motivate me and help me, to make my work possible.

Tan Tien Ping

Preface

This dissertation has been completed in English. A short abstract in French can be found at the following section, while an extended summary in French “Résumé étendu en français” can be found near the end of the manuscript. Please refer to the table of contents.

Abstract

Automatic speech recognition technology has achieved maturity, where it has been widely integrated into many systems. However, speech recognition system for non-native speakers still suffers from high error rate, which is due to the mismatch between the non-native speech and the trained models. Recording sufficient non-native speech for training is time consuming and often difficult.

In this thesis, we propose approaches to adapt acoustic and pronunciation model under different resource constraints for non-native speakers. A preliminary work on accent identification has also been carried out.

Multilingual acoustic modeling has been proposed for modeling cross-lingual transfer of non-native speakers to overcome the difficulty in obtaining non-native speech. In cases where multilingual acoustic models are available, a hybrid approach of acoustic interpolation and merging has been proposed for adapting the target acoustic model. The proposed approach has also proven to be useful for context modeling. However, if multilingual corpora are available instead, a class of three interpolation methods has equally been introduced for adaptation. Two of them are supervised speaker adaptation methods, which can be carried out with only few non-native utterances.

In term of pronunciation modeling, two existing approaches which model pronunciation variants, one at the pronunciation dictionary and another at the rescoring module have been revisited, so that they can work under limited amount of non-native speech. We have also proposed a speaker clustering approach called “latent pronunciation analysis” for clustering non-native speakers based on pronunciation habits. This approach can also be used for pronunciation adaptation.

Finally, a text dependent accent identification method has been proposed. The approach can work with little amount of non-native speech for creating robust accent models. This is made possible with the generalizability of the decision trees and the usage of multilingual resources to increase the performance of the accent models.

Keywords: non-native speech recognition, non-native multilingual acoustic modeling, non-native pronunciation modeling, accent identification

Résumé

Les technologies de reconnaissance automatique de la parole sont désormais intégrées dans de nombreux systèmes. La performance des systèmes de reconnaissance vocale pour les locuteurs non natifs continue cependant à souffrir de taux d'erreur élevés, en raison de la différence entre la parole non native et les modèles entraînés. La réalisation d'enregistrements en grande quantité de parole non native est souvent difficile et peu réaliste pour représenter toutes les origines des locuteurs.

Dans cette thèse, nous proposons des approches pour adapter les modèles acoustiques et de prononciation sous différentes conditions de ressource pour les locuteurs non natifs. Un travail préliminaire sur l'identification d'accent a également proposé.

Ce travail de thèse repose sur le concept de modélisation acoustique translingue qui permet de représenter les locuteurs non natifs dans un espace multilingue sans utiliser (ou en utilisant très peu) de parole non native. Une approche hybride d'interpolation et de fusion est proposée pour l'adaptation des modèles en langue cible en utilisant une collection de modèles acoustiques multilingues. L'approche proposée est également utile pour la modélisation du contexte de prononciation. Si, en revanche, des corpus multilingues sont disponibles, des méthodes d'interpolation peuvent être utilisées pour l'adaptation à la parole non native. Deux d'entre elles sont proposées pour une adaptation supervisée et peuvent être employées avec seulement quelques phrases non natives.

En ce qui concerne la modélisation de la prononciation, deux approches existantes (l'une fondée sur la modification du dictionnaire de prononciation, l'autre fondée sur la définition d'un score de prononciation utilisé dans une phase de re-scoring) sont revisitées dans cette thèse et adaptées pour fonctionner sur une quantité de données limitée. Une nouvelle approche de groupement de locuteurs selon leurs habitudes de prononciation, est également présentée : nous l'appelons « analyse de prononciation latente ». Cette approche se révèle également utile pour améliorer le modèle de prononciation pour la reconnaissance automatique de la parole non native.

Enfin, une méthode d'identification d'accent est proposée. Elle nécessite une petite quantité de parole non native pour créer les modèles d'accents. Ceci est rendu possible en utilisant la capacité de généralisation des arbres de décision et en utilisant des ressources multilingues pour augmenter la performance du modèle d'accent.

Mots clés : reconnaissance automatique de la parole non native, modélisation acoustique multilingue non native, modélisation de prononciation, identification d'accent

Contents

Figures

Tables

Introduction

Most people nowadays can speak more than one language. In a world where competition becomes more and more critical, the ability to communicate in several languages gives extended advantages to the speakers. This is because language does not solely play the role of communication but also represent the identity and culture of the community who speaks that language. People who speak the same language are more capable of relating to each others.

Many people also acquire new languages to have an edge in the economy. The booming of many economies around the world has renewed interest in languages such as Arabic, Indian, Korean, Mandarin and others.

Besides that, *lingua francas* such as English, Spanish and French have long been of interest for people around the world because of their richness particularly in the domain of science and technology. Many of these languages are taught in schools and universities around the world.

People may also learn a new language when they move to a new country, since they may speak a language different from the native speakers. Human migration is becoming more common particularly for economic reasons. In the United States for example, 37.5 million of the population or nearly one in five is from foreign origin in 2006 [Ohlemacher 2007].

Tourism is a lucrative industry for many countries. In France for example, there were 78 million tourists who visited the country in 2006 [LExpansion 2007]. People who travel to other countries also often pick up some common local phrases from travel guide books or Internet to facilitate communication.

This thesis is about automatic speech recognition for non-native speakers. The popularity of using speech recognition system as a natural interface is increasing with the maturity of speech recognition technology. Nowadays, speech recognition applications are embedded in different systems such as telephony systems, computers, mobile phones, car and others. However, most people who try to use speech recognition applications to recognize their non-native speech will be discouraged by their performance. Studies show that the performance of speech recognition systems in decoding non-native speech is at least two times lower compared to native speech. This is due to the difference in characteristics between native and non-native speech.

Statistical speech recognition system uses three types of models for modeling speech at different levels, namely acoustic model, pronunciation model and language model. These models are created by using data-driven approach. Since they are usually modeled using only the native

language resources, there may be a mismatch between non-native speech and the models. As a result, the recognition rate for non-native speech is much lower compared to the native speech. The solution is to build models that better match non-native speakers by using non-native speech. However, acquiring non-native speech is time and resource consuming. There are more than six thousand languages in this world. Therefore, to record all the non-native speech for each language is difficult if not impossible. In certain cases, it is unfeasible particularly for under resourced languages.

The objective of this thesis is to propose non-native modeling methods that are flexible to be employed under different resource constraints. Multilingual resources have been proposed for adapting non-native models to overcome the difficulty to obtain non-native speech whenever possible. In situation when some non-native speech is available, it can also be taken advantage of.

In this thesis, we will look at non-native acoustic and pronunciation modeling. In acoustic modeling, we look at the usage of multilingual resources for adapting the acoustic model of the target language. The reason why multilingual resources can be used for adapting non-native speakers is because the 'cross-lingual transfer' phenomenon by non-native speakers. By using the multilingual resource and cross-lingual transfer information, a new language space that is aligned with the target language space is created. This new language space can then be used to estimate the non-native language space (see Figure Ia). Depending on the type of multilingual resources such as multilingual acoustic models or corpora that are available, different techniques are proposed. In cases where some non-native is available from the speaker, more effective approach for estimating the non-native language space is also proposed. We will also see that the multilingual approach proposed can be used for context modeling.

In pronunciation modeling, we revisit two of the conventional approaches. We have modified two of the approaches to make sure that they can be used even in situation when only small amount of non-native speech is available. We also propose a new approach of pronunciation habits clustering and pronunciation adaptation. This is done by creating a pronunciation space and the non-native speakers can then be separated into groups or cluster on the pronunciation space. For an unknown speaker, the pronunciation variants of the speaker can be estimated the position of the speaker on the pronunciation space given some non-native speech.

Besides acoustic and pronunciation modeling, a preliminary study on accent identification has also been carried out by making use of multilingual resources with non-native speech. Multilingual resources have been shown to be beneficial in improving the language identification systems. In our proposed accent identification system using phonotactic features, multilingual resources are being used to capture the pattern and degree of changes with different non-native speech (see Figure Ib). Like the approach proposed for non-native modeling before, this approach is able to take advantage of limited non-native speech.

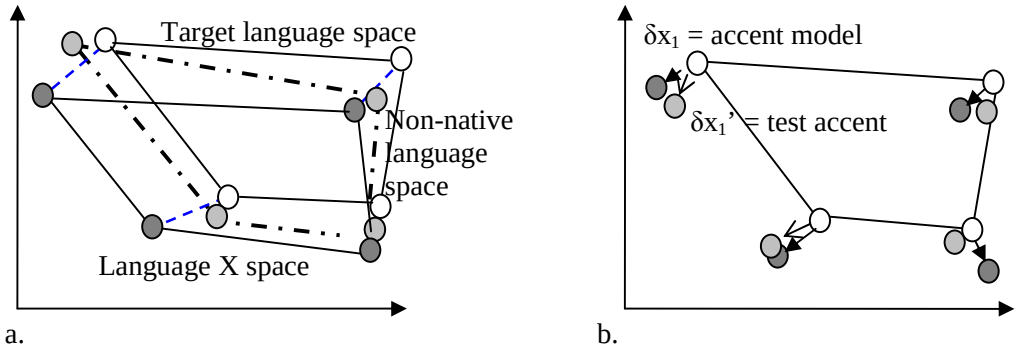


Figure I. Non-native acoustic and accent modeling using multilingual resources. a) Using target language and language X (for instance native language of the speaker) space to estimate non-native language space. b) Using model from language X and some non-native speech to create the accent model δx, and compared against the accent of the test speaker δx'

In the first chapter of this thesis, a brief introduction to the architecture and components of speech recognition is given. Subsequently, a discussion on the language acquisition of human in term of first and second language will follow. Here, we will learn why non-native speech recognition performance is significantly lower compared to native speech recognition. At the end of the chapter, recent works in the domain of non-native speech recognition will be presented.

The second chapter presents our proposed non-native acoustic modeling approaches using multilingual resources. Depending on the types of multilingual resources such as acoustic models or corpora available, different modeling approaches are proposed. If some non-native speech is available, non-native speaker adaptation techniques are also proposed. The approaches proposed are hybrid approach of interpolation and merging, and new interpolation approaches.

In the following third chapter, two pronunciation modeling approaches: pronunciation dictionary and n-best rescoring are revisited, and modified for modeling pronunciation habits. In addition, we also propose an original approach that we called "latent pronunciation analysis", which uses pronunciation eigenvectors for speaker clustering. The approach can also be employed for pronunciation adaptation. Next, we present our preliminary work in accent identification. It is a phonotactic approach which makes use of multilingual resources for creating accent models in the form of multilingual decision trees.

Chapter four presents the non-native French corpus that we have acquired for testing and adaptation purpose. The corpus is evaluated through perception and acoustic analysis. In addition, we also use data-driven approach for analysing it.

Finally in chapter five, the acoustic modeling, pronunciation modeling and accent identification approaches that have been presented in previous chapters two and three are evaluated. This chapter is followed by conclusions and future works.

CHAPTER 1

Automatic Speech Recognition for Non-Native Speakers

1.1 Introduction

Human has always been fascinated by artificial intelligence such as the ability of machine to understanding speech. However, before a speech signal can be analyzed for its meaning, it has to be first converted to a simpler form – the text transcription. Speech to text or speech recognition is an interesting but challenging domain because of its multi-disciplines nature. Among the domains involved are signal processing, pattern recognition, linguistics, information theory and others.

For more than five decades, researchers have achieved great advancement in the field of automatic speech recognition, from the earliest isolated word recognition to current large vocabulary speech recognition. Significant progress in the areas of speech recognition is achieved with the introduction of statistical based approach which uses hidden Markov model and n-gram model since 1980s for large vocabulary recognition. At the same time, the technology advancement has also propelled the progress of other areas such as biometric speaker recognition and statistical machine translation.

Nowadays, the technology used in automatic speech recognition (ASR) has matured to a level where it has been increasingly applied in services such as telephony systems, mobile phones, GPS, as well as cars and computers. However, automatic speech recognition still faces many challenges before it can be employed by everyone at anywhere. One of the problems faced by current speech recognition systems is the difficulty of recognizing non-native speech. While the word error rate of speech recognition systems for native speakers is now in the range of less

than twenty percents for a large vocabulary speech recognition system, the word error rate for non-native speakers is at least twice the rate of native speakers.

In this chapter, a brief introduction to the architecture of statistical automatic speech recognition system and its components will be presented. Subsequently, we will look at why non-native speech is poorly recognized by speech recognition system, and how the current non-native modeling techniques improve the system for recognizing non-native speech. In addition, accent identification approaches will also be discussed.

1.2 Architecture of an Automatic Speech Recognition System

An automatic speech recognition system also known as speech to text system receives an utterance as input and delivers an output text transcription. Figure 1.1 shows the main components of a speech recognition system.

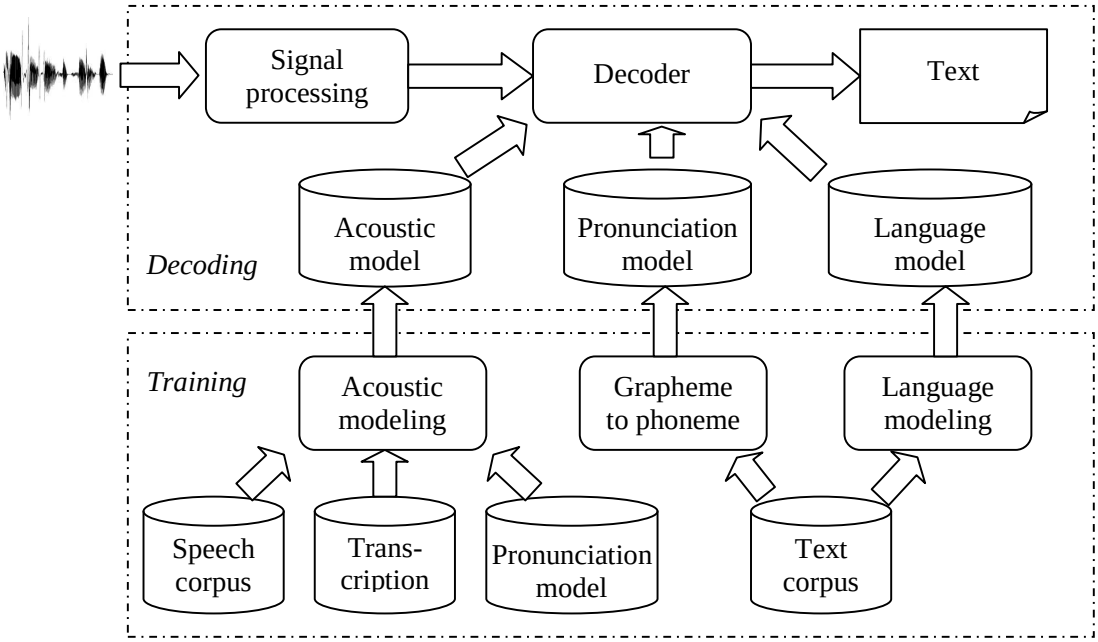


Figure 1.1 Automatic speech recognition system architecture

An automatic speech recognition system can be divided into two main processes: decoding and training. The decoding components consist of a signal processing front-end and a decoder. The purpose of signal processing front-end is to digitize analog signal and to convert it to discriminative features for recognition. The decoder is the engine of a speech recognition system that uncovers the possible word sequence from the feature vectors using the knowledge from

acoustic, pronunciation and language models. From the linguistic view point, these models have roughly the following representations in a language:

- Acoustic model – phonology of a language
- Pronunciation model – vocabulary and pronunciations
- Language model – grammar of a language

In a typical speech recognition system, acoustic model defines the elementary units of speech. It can be phones, phonemes, syllables, and words. On the other hand, pronunciation model represents language units such as word or syllable, and sometimes common word sequences using the acoustic units defined in the acoustic model. Language model in turn defines the structure and syntax of a language with the vocabulary from the pronunciation dictionary. In the following sections, we will look at the speech recognition components in more details.

1.2.1 Signal Processing Front-End

The objective of the signal processing front-end is to extract discriminative features that are perceptually important. The signal processing front-end first digitizes the analog signal to a form suitable for analysis. The process involves several stages such as pre-emphasis, filtering, sampling, and quantization [Kent 2002]. A sampling frequency of 16 kHz is sufficient to represent human speech intelligibly. Study shows that a higher sampling frequency does not give any further improvement to the speech recognition system [Huang 2001].

The digitized signal is then converted to feature vectors, a form which is more relevant for speech processing. The possible types of feature are short time spectral envelope, energy, zero crossing rates, level crossing rates, and others. Frequency-domain features such as short time spectral envelope are more accurate and descriptive compared to time-domain features for analyzing speech. Among the well known spectral analysis methods are linear predictive coding (LPC), perceptual linear prediction (PLP), and mel-frequency cepstral (MFC) spectral analysis model. The mel-frequency cepstral coefficients (MFCC) are one of the most widely used features in speech recognition. These features can be derived through the following procedures [Davis 1980; Tychtl 1999]:

- Fourier Transform is computed for each frame
- Triangular mel filter banks are applied on the power spectrum
- A log function is used to smooth the spectrum
- Discrete cosine transform (DCT) encodes the spectrum to MFCC

Studies show that 13^{th}-order MFCC contains sufficient information to represent speech [Huang 2001]. In addition to the raw MFCC features, the first and second derivatives of the MFCC features are normally also computed, because they provide temporal changes information of the spectral. For speech recognition system using hidden Markov model (HMM), these information can be useful, because the acoustic frames are assumed to be independent and

stationery. For reducing the size of the feature vectors, dimension reduction techniques such as principal component analysis (PCA) or linear discriminative analysis (LDA) can be applied on the vectors to create a more compact and discriminative feature.

1.2.2 Decoder

The word *decoder* is originated from the field of information theory, which means the conversion of a coded message to an understandable form. In speech recognition, decoder is the component that uncovers the word sequences from the speech signal or more precisely the feature vectors. The search for the most probable word sequence can be achieved by maximizing the posterior probability for the given feature vectors. It is difficult to calculate efficiently and robustly the posterior probability. Thus, instead of calculating the posterior probability directly, it can be put in another form using Bayes theorem:

$$\begin{aligned} \hat{W} &= \arg\max_{W} P(W \mid O) \\ &= \arg\max_{W} \frac{P(W)P(O \mid W)}{P(O)} \\ &= \arg\max_{W} P(W)P(O \mid W) \end{aligned} \tag{1.1}$$

where W is the word sequence w_1, w_2, …w_m which gives the maximum posterior probability P(W|O) given O, a series of observations o_1, o_2,… o_n which produce the word sequence. This means that the best word sequence can be found by combining the language probability of word sequence P(W) (prior probability) with the acoustic probability of the word sequence P(O|W) (conditional probability) from an acoustic model which gives the highest value. The state of the art acoustic model used in automatic speech recognition is hidden Markov model. It will be discussed in the coming section. Conversely, the language model provides the language probability. A widely used language model is the n-gram model.

Figure 1.2 shows an example of decoding an utterance with a single word, using a decoder with a vocabulary of only three words {A, B, C}. The connected circles represent the phoneme models except the model ‘sil’ which represents the silence at the start (<s>) and the end (</s>) of a sentence, while the dotted circles are word pronunciation models. The arrows leaving one dotted circle to another indicate the language probability, while the arrow from a connected circle to another indicates the transition from one phoneme model to another.

The complexity of the problem can be imagined from Figure 1.2. In continuous speech, where the total number of words in an utterance is unknown and the boundary of words is blurred, uncovering the most probable word sequence is not an easy task. The equation 1.1 however did not specify how to search for the best word. The most direct way is to calculate all the possible word sequences and select the one which gives the highest value from the formula. For short single word recognition as we see above, this might be possible, but when the length of the

sentence is unknown and the vocabulary gets larger, this is no longer feasible. For example, for a sentence with *m* words and with a vocabulary of *n*, there are n^m possible solutions. Search algorithms using dynamic programming strategies have been successfully applied to uncover the word sequence in a feasible and efficient manner by avoiding redundant computation. This is done by breaking up a problem to common sub-problems, finding the best solution for the sub-problems and storing the previous calculated results which are common. Another technique used to speed up the search is by pruning the unpromising path. An example of search algorithm for automatic speech recognition is time-synchronous Viterbi beam search.

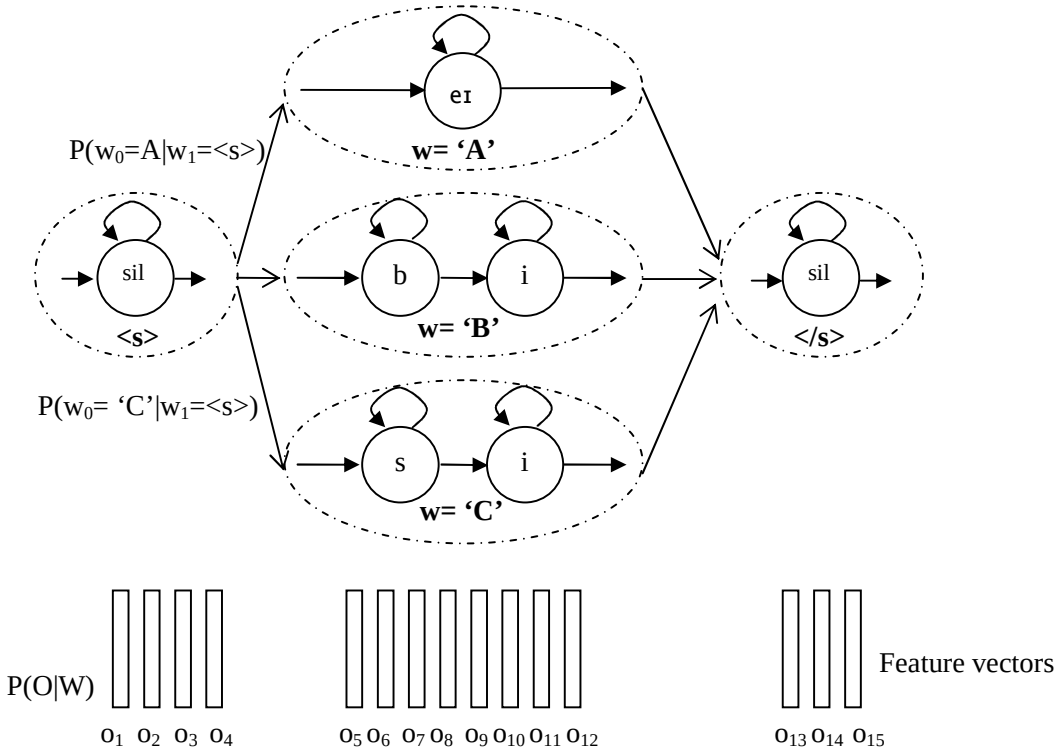


Figure 1.2 Speech recognition as a search problem. The most probable word is the word which gives the highest score of P(W)P(O|W)

The performance of a speech recognition system is typically measured by *word error rate* (WER). The word error rate is calculated from three types of errors a speech recognition system commit:

- Substitution: replacement of a correct word by another different word
- Deletion: omission of a correct word
- Insertion: addition of an extra word

For calculating the number of substitutions (subs), deletions (dels) and insertions (ins) made, the hypothesis from the system is aligned against the actual reference transcription using minimum edit distance (Levenstein distance), with the same cost given to the errors.

$$Word\ Error\ Rate = \frac{subs + dels + ins}{number\ of\ \ words\ in\ the\ correct\ sentence} \times 100\% \tag{1.2}$$

1.2.3 Acoustic Modeling

Building a robust acoustic model is one of the main challenges in the field of speech recognition. The difficulty of modeling acoustic features in a robust manner is due to the variability which exists in the speech. Context variability can happen at the sentence, word and phonetic level [Huang 2001]. In a continuous speech, words in a sentence may be connected instead of separated by silence. Variability can also exist in pronunciation when the words are pronounced in an isolate and continuous manner. At the phonetic level, variability can exist in a phoneme when it is realized under different contexts. At the speaker level, variability is most noticeable since the speech is influenced by the physical attributes such as vocal tract size, height, age, sex and also social characteristics. Environmental condition is another contribution to speech variability. The environmental noise and variation in microphone are among the factors that can influence the performance of a speech recognition system.

The selection of appropriate acoustic unit for an acoustic model is important. Among the units often used for modeling are word, syllable, and phone. From word model to phone model, the ability for each to generalize increases; therefore, the amount of speech required to train the unit robustly decreases. In small vocabulary word recognition system, word models can give better results than other type of units when sufficient speech is available for each word model. However, in continuous speech recognition, inter-word variability can happen as mentioned previously. In this case, context dependent modeling is required to record the differences to achieve high recognition performance. Thus, this means that the number of instances needed for training different word contexts will also increase. Syllable model is the next best choice if sufficient speech is available for training them. For languages with a (relatively) limited number of syllables such as Japanese and Chinese, syllable model can be an attractive option. Most speech recognition systems use phone as the unit of modeling, since it requires moderate amount of speech to robustly model and will not over generalize. However, phone model is greatly influenced by the context compared to other models. Consequently, context dependent modeling is often employed to achieve a high recognition rate.

There are many possible approaches for modeling the acoustic units, for example hidden Markov model (HMM), artificial neural network and template model etc. Hidden Markov model is one of the most widely used approaches in statistical speech recognition because of its robustness. In the following sections, we will briefly present the theory of HMM, and followed by the training procedures.

1.2.3.1 Hidden Markov Model

The theory of hidden Markov model (HMM) was developed since the late 1960s. It was used by IBM in automatic speech recognition since 1970s. A Markov chain is a stochastic process with short memory, where the current state depends only on the previous state. In a Markov chain, the observations are actually the state sequence. A hidden Markov model is an extension of a Markov chain where the observation is a function of the state; therefore, the state sequence is hidden in hidden Markov models. The probability to be at a particular state can be calculated instead given the observation.

Figure 1.3 shows an example of a continuous mixture density HMM for recognizing three phonemes /a/, /e/ and /i/. Each phoneme is represented by a state, and each state is defined by Formant 1 (F1) value with a single Gaussian. If we treat the figure as simply a Markov chain, the observation in this case will be the three possible phonemes /a/, /e/, and /i/. Given the current observed phoneme, we can know the probability of transition to the next phoneme (state). For example, if currently we observed that the speaker utters the phoneme /e/, the probability of it to transit to phoneme /a/ will be 0.1. For a hidden Markov model, the observation is a series of F1 value instead. From the F1 value, we do not know the actual state it is in, but we know that if the F1 value is near to the mean of one of the phoneme, it has a higher probability to be at that phoneme. Thus, if we observed an F1 value of 525Hz, it has a higher probability to be /e/, than /a/ or /i/.

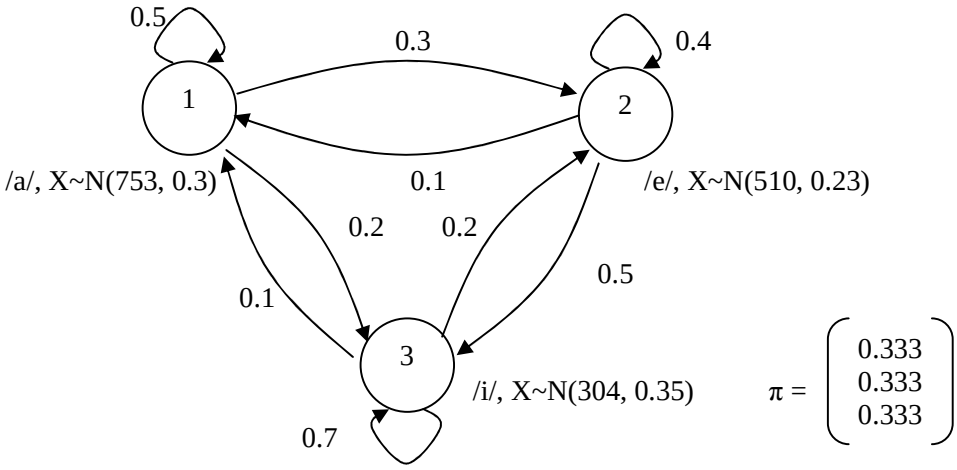


Figure 1.3 A continuous hidden Markov model for modeling three phonemes /a/, /e/ and /i/

A hidden Markov model is defined by the following parameters:

- $O=o_1,o_2,\ldots o_m$ – A sequence of observations. In the given example, the observations are a series of F1 values. For a period of time t, from an utterance of a speaker, the possible series of F1 observations are 558 Hz, 561 Hz, 562 Hz…
- $\Omega=\{1,2,\ldots N\}$ – A set of states in the model. In the given example, there are three states, which represent the phoneme /a/, /e/, and /i/.
- $A=\{a_{ij}\}$ – A state transition probability matrix. Among the values of the state transition for the given example are $a_{12}=0.3$ and $a_{31}=0.1$.
- $B=\{b_i(t) = P(O_t=o_t| s_t=i)\}$ – An output probability matrix, where $b_i(k)$ is the probability of observing the value o_t at state i. In the given example, it is a continuous mixture density HMM with single Gaussian. The conditional probability can be calculated by using the Bayes Gaussian classifier. Given the observation o with dimension n (in our case n = 1), the probability of it to be emitted by distribution in state i with mean vector μ_i and covariance matrix C_i is:

$$b_i(o) = \frac{1}{(2\pi)^{n/2} \mid C_i \mid^{1/2}} e^{-1/2(o-\mu_i)' C_i^{-1}(o-u_i)} \tag{1.3}$$

- $\pi=\{\pi_i = P(s_0=i)\}$ – an initial state transition probability, where $1\leq i \leq N$. In the case above, all three states have the equal chance to start.

There are two assumptions made in the hidden Markov model: first-order Markov assumption and output-independence assumption. The reason is to simplify the calculation and to make the system more efficient and feasible. The first-order Markov assumption states that the probability at state s and time t, depends only on the preceding state at time t-1.

$$P(s_t \mid s_1, s_2, \ldots s_{t-1}) \approx P(s_t \mid s_{t-1}) \tag{1.4}$$

The output-independence assumption states that the probability that a particular observation at time t depends only on the state s_t and is conditionally independent of the past observations.

$$P(O_t \mid O_1, O_2, \ldots O_{t-1}, s_1, s_2, \ldots s_t) = P(O_t \mid s_t) \tag{1.5}$$

In a typical large vocabulary speech recognition system, an acoustic model normally consists of multiple HMMs. Each HMM models an acoustic unit for example a phone, using a left to right architecture with three to five states, see Figure 1.4.

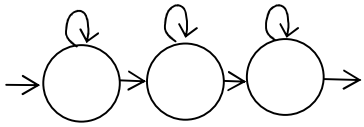


Figure 1.4 A three state left to right HMM topology

The HMM in Figure 1.3 is a continuous HMM. In fact, there are three types of HMM depending on how the feature space of the model is defined: continuous, discrete and semi-continuous (tied-mixture), see Figure 1.5. In a continuous HMM, the feature space of a model is represented using Gaussian mixtures. In a discrete HMM, the feature space is divided into clusters of speech sounds, normally using vector quantization (VQ) algorithm such as k-means. For a discrete HMM, the continuous features observed are mapped to a finite set of discrete observations. Thus, the feature space of each model is defined using the discrete features. Semi-continuous space as its name suggests, is an intermediate between discrete and continuous model. The feature space of each model is defined using a common set of Gaussian densities.

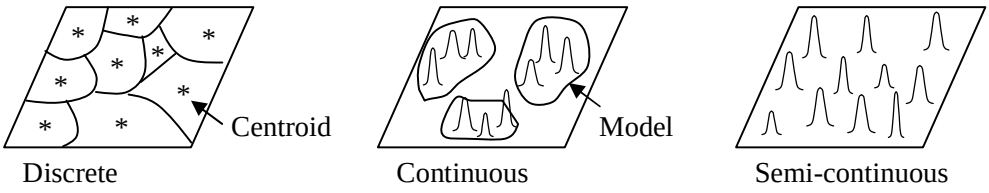


Figure 1.5 Representation of feature space using discrete, continuous and semi-continuous models [Rabiner 1993]

1.2.3.2 Training a Continuous Mixture Density HMM

Given the hidden Markov model, how do we learn the optimized model parameters or patterns for speech recognition? In HMM training, the observations from the training utterances are used to model the corresponding acoustic units in the transcriptions with the assistance of a pronunciation model through iterative re-estimation. There are many different strategies to train a continuous HMM acoustic model. Figure 1.6 shows one of the possible ways for training an acoustic model, which is applied in Sphinx and HTK systems [Woodland 1993, Woodland 1994, CMU 2000].

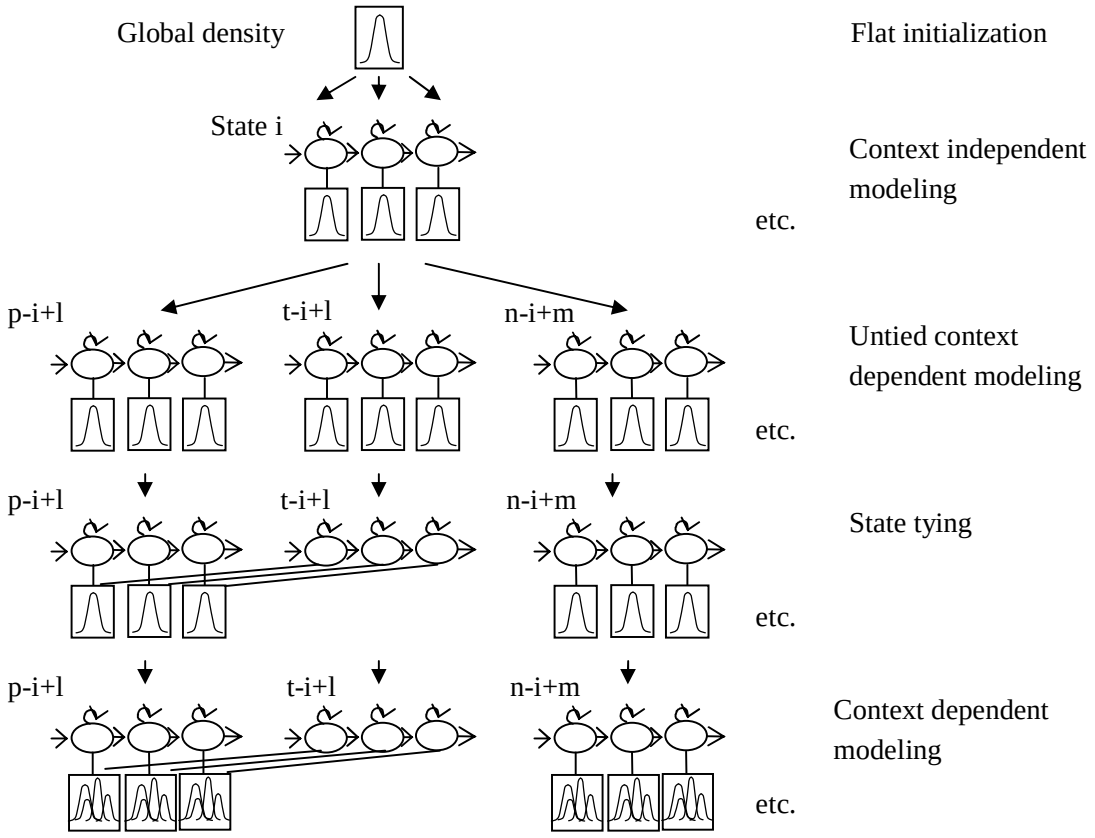


Figure 1.6 Procedure for building a context dependent continuous HMM acoustic model

Flat initialization is used here to initialize the HMM parameters by calculating a global means and variances from the features files. Note that the same front-end module used in recognition stage will be used here to convert the speech to the same type of feature, for example MFCC. Equally probable transition matrix and mixture weights are also initialized in the process. These values are then copied to the context independent (CI) models to initialize the parameters. Subsequently, the parameters are re-estimated using Baum Welch algorithm. Baum Welch algorithm is an expectation maximization (EM) method that iteratively maximizes the log

likelihood from the observed training data. The new HMM parameters will update the previous estimated values. Following are the re-estimation formulas for the coefficients of the mixture density, mixture weight *c*, mean μ, and covariance matrix *C* [Rabiner 1993].

$$\bar{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{1.6}$$

$$\bar{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k).o_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{1.7}$$

$$\bar{C}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k).(o_t - \mu_{jk})(o_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{1.8}$$

Note that the equation for the weight of the k^{th} Gaussian mixture component in state *j*, w_{jk} has the same form as the state transition from state *i* to state *j*, a_{ij}. In fact, a HMM state with a mixture density has shown to be equivalent to a multi-state single-mixture density model. $\gamma_t(j,k)$ is the probability of observing o_t in state *j*, k^{th} Gaussian mixture at time *t*.

For creating a multi-Gaussians context independent model, the number of Gaussians will be increased by splitting each Gaussian in the context independent model normally to two when it is converged by perturbing the means slightly. Context independent model is created by modeling the defined acoustic unit without taking into consideration the context (surrounding acoustic units and possibly its position) of the acoustic units during training. Context independent model is especially useful in situation when little speech is available. It is also useful as a bootstrap model in multilingual acoustic modeling. The re-estimation step is again repeated until the total number of Gaussians is reached. If the acoustic units defined are phonemes, then there will be one HMM for representing each phoneme. Note that phoneme is the smallest sound unit that distinguishes meaning. The acoustic units can also be trained by considering the context of the acoustic unit through context dependent modeling. The context dependent acoustic unit which is trained by taking into consideration the left context of current phoneme is known as biphone. The acoustic unit trained by considering its left and right context is known as triphone model. Thus, for the vowel /ɛ/ in the context of the word "yes" which is pronounced as /j ɛ s/, the monophone /ɛ/ will be modeled, but for a biphone it will be /ɛ/$_{j}$ (written also as j-ɛ), and the triphone will be /ɛ/$_{j,s}$ (also can be represented as j-ɛ+s). In this manuscript, we will consider phoneme as speech sound independent of the context (monophone), while phone (or more precisely allophone) is used to refer to the speech sound which takes into consideration of its context.

In general, context dependent model outperforms context independent model for speech recognition. However, it requires more data to train robustly compared to context independent model. The data sparsity problem implies that precise context modeling cannot be used in most cases. A possible solution is to use a combination of different contexts, where a sharper phone model is trained if there are sufficient data. Another approach to overcome this problem is by state tying [Young 1993]. The idea is to model 'similar' states together based on the contexts they are in. This is carried out using decision trees for clustering the untied context dependent states by asking linguistic questions about the context of the untied states such as the type of the phoneme, the place of articulation, the position of tongue etc., from the root node to the leaf node. Normally, a decision tree is used for classifying states from the same base unit. Consequently, for a triphone model, the phonemes realised in different contexts are clustered using the same phoneme tree. All untied states to be clustered are placed on the root node of the tree, and the log likelihood of the training data calculated on the assumption that all of the states in that node are tied. The node is then split based on the linguistic questions which give rise to the maximum increase in log likelihood. The process is repeated until the increase in log likelihood fall below a threshold or a minimum occupation count is reached. Besides using decision tree, other way of tying the states is by using neural networks [Li 1998].

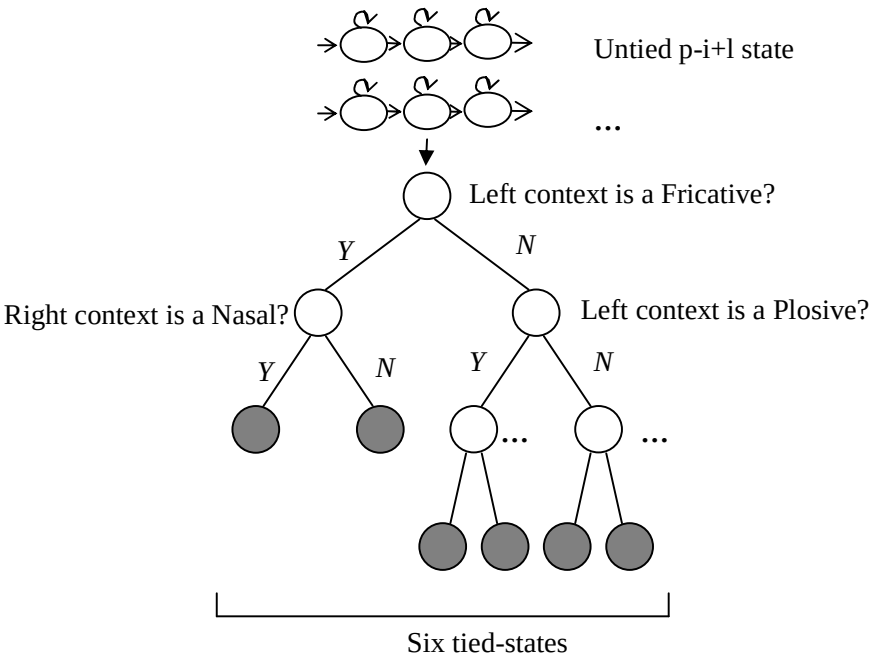


Figure 1.7 Phonetic decision tree for the phoneme /i/

Figure 1.7 shows an example of decision tree for phoneme /i/ built from triphone [i]. At each branch, phonetic question is asked about the context of the untied triphone [i] state. For example at the root node, the question is "Is the phoneme at the left of the triphone [i] a

fricative? ". For an untied triphone p-i+l (a triphone [i] with left context /p/ and right context /l/), the answer is "no" because /p/ is a plosive, not a fricative. All the untied states which reach the same leaves will be tied together as the same state. Thus, the untied state such as f-i+m and s-i+n will be tied together at the same state in the previous example.

1.2.4 Pronunciation Modeling

Acoustic model defines elementary speech units using fine phonetic features which are related to mouth, tongue, vocal tract and others from speech. Pronunciation modeling on the other hand consists of creating the bigger word or syllable models using the acoustic units defined in acoustic model. Since in most cases, phoneme or phone is the acoustic unit employed in the acoustic model, a pronunciation dictionary (lexicon) can be built from a typical dictionary, since most of them contain descriptions of how words should be pronounced using International Phonetic Alphabet.

If there is no description on the manner of pronunciation, then rules for converting the graphemes to phonemes have to be created. However, this requires an understanding of the language involved. An automatic grapheme to phoneme tool can be created for generating the 'standard' pronunciation models using linguistic rules. A manual verification of the generated pronunciation models is often required to correct words which are exception to the rules. In cases where rules for converting graphemes to phonemes do not exist, and there is limited understanding of the language involved, studies found that using graphemes (context dependent) as the acoustic units for modeling pronunciation model can produce acceptable speech recognition performance, where it is only slightly worse compared to word modeled using phonemes [Killer 2003a, Killer 2003b]. Note that, this also means that the grapheme units have to be trained in the acoustic model. Using a phoneme based acoustic model, the words can be modeled in the pronunciation dictionary as follows (English pronunciations):

ABANDON	ə b æ n d ə n
CARPET	k ɑ r p ə t

If there are few possible pronunciations for the same word, pronunciation variants can also be added into the pronunciation dictionary as follows:

VOYAGE	v ɔɪ ə ʤ
VOYAGE(2)	v ɔɪ i ʤ

Often in a word pronunciation dictionary, frequent found phrases or word sequences can also be modeled here, for example:

DRAG_AND_DROP	d r æ g æ n d r ɒ p

THERE_ARE ð e r *a* r

We will look into more details about pronunciation modeling in Section 1.5.2, particularly on different approaches to find pronunciation variants and model them, since they are also related to the problem of non-native speakers.

1.2.5 Language Modeling

The language model represents the grammar of a language. It defines rules that govern the proper use of a language such as morphology and syntax. There are two very different ways to represent the grammar of a language: formal language model and stochastic language model. Formal language modeling is a knowledge-based approach to represent a language model using linguistic knowledge, while stochastic language modeling is a statistical data-driven approach that uses text corpora for generating the rules automatically.

1.2.5.1 Formal Language Model

A formal language model corresponds to knowledge and rule based language modeling. There are two important components in a formal language model: grammars and parser. Grammars are described using a set of rewrite rules, and they define the permissible structures of a language, and the parser decomposes a sentence to smaller lexical categories (lexical class, part of speech, word class) such as noun, verb, adjective, pronoun, preposition, adverb, and conjunction. With formal language model, a sentence is analyzed to determine whether its grammatical structure is allowed by the grammars. In this case, a sentence is transformed normally to a tree structure with words at the leaves nodes. The shortcoming of this rule based approach is that the grammars need to be defined by someone with the linguistic knowledge, and the grammars defined here typically have the standard structures which are permissible by the language. However, in conversational speech, this is often not the case. Furthermore, the structure can also be ambiguous in some situations. For example the sentence “time flies” has two possible representations with the given rewrite rules in Figure 1.8.

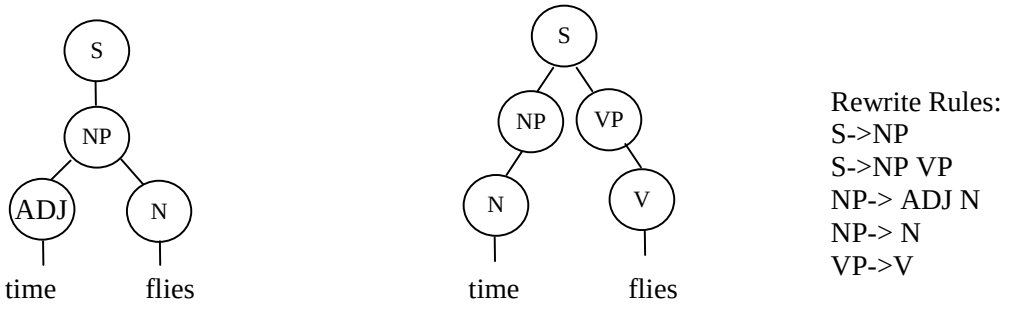


Figure 1.8 Two different tree representations of the sentence “time flies”

1.2.5.2 N-Gram Model

Stochastic language model is a statistical approach to represent grammar. In the formal language model, a sentence is classified as either an grammatically acceptable or unacceptable sentence. On the other hand, stochastic language model uses soft classification by estimating the probability of a sentence *P(W)* where *W* is a series of words $W=w_1,w_2,w_3... w_n$. *P(W)* can be calculated by decomposing it using the chain rule of probability as below:

$$\begin{aligned} P(W) &= P(w_1, w_2, w_3 ... w_n) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)...P(w_n \mid w_1, ...w_n) \\ &= \prod_{i=1}^{n} P(w_i \mid w_1, w_2 ... w_n) \end{aligned} \tag{1.9}$$

The problem of calculating the sentence probability using the above formula is in estimating w_i which depends on *i-1* previous words. Even for a moderately long sentence, this probability can be impossible to calculate robustly. The workaround to this problem is to approximate the history by using several previous words (Markov assumption), instead of using all the previous words. This approach is known as n-gram language modeling. Depending on the constraint in the amount of text available, n-gram model can be created by using different values of *n* or order. When *n* is one, the probability of a word in the sentence depends only on its frequency of occurrence in the text, it is known as unigram. Bigram model assumes that the probability of a word depends on the immediately preceding word. One of the most often used n-gram model is trigram model, since most words in a language have a strong dependency on the previous two words, and it can be built using a reasonable size corpus. The trigram probability can be estimated from total observation counts of word pair $C(w_{i-2}, w_{i-1})$ and triplet $C(w_{i-2}, w_{i-1}, w_i)$ in a training corpus using the maximum likelihood approach as follow:

$$P_{ML}(w_i \mid w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \tag{1.10}$$

The chain rule of probability in (1.4) can be represented in the following form for trigram:

$$P_{ML}(W) = P(w_1) \times P(w_2 \mid w_1) \prod_{i=3}^{n} P(w_i \mid w_{i-2}, w_{i-1}) \tag{1.11}$$

Even with the usage of n-gram to estimate sentence probability, calculating a robust trigram model can still pose problem because of data sparseness. This is because some words are more rare than others, for example proper name compared to pronouns in a text. As a result, in a trigram model, most words exist only in a few instances. Furthermore, the training corpus may also not contain certain words because of differences in training and actual environment. Smoothing is used in this case to produce a more robust estimation. Often, lower order n-gram is

used in some way for compensating the estimated trigram probability. For example, smoothing can be carried out using deleted interpolation smoothing as follow:

$$P_{DI}(w_i \mid w_{i-n+1}...w_{i-1}) = \lambda P(w_i \mid w_{i-n+1}...w_{i-1}) + (1-\lambda)P(w_i \mid w_{i-n+2}...w_{i-1}) \tag{1.12}$$

where λ is interpolation weight that depends on $w_{i-n+1}\ldots\ w_{i-1}$. Another commonly used approach is backoff approach. Backoff smoothing is first introduced by Katz, where smoothing is carried out by recursively backing off to lower n-gram until some counts are found. The Katz's trigram backoff approach is defined as below [Jelinek 2001]:

$$P_{katz}(w_i \mid w_{i-2}, w_{i-1}) = \begin{cases} P_{ML}(w_i \mid w_{i-2}, w_{i-1}) & if\ r > k \\ \alpha Q_T(w_i \mid w_{i-2}, w_{i-1}) & if\ k \geq r > 0 \\ \beta(w_{i-2}, w_{i-1}) P_{katz}(w_i \mid w_{i-1}) & if\ r = 0 \end{cases} \tag{1.13}$$

$$P_{katz}(w_i \mid w_{i-1}) = \begin{cases} P_{ML}(w_i \mid w_{i-1}) & if\ r > k \\ \alpha Q_T(w_i \mid w_{i-1}) & if\ k \geq r > 0 \\ \beta(w_{i-1}) P_{ML}(w_i) & if\ r = 0 \end{cases} \tag{1.14}$$

where α and ß can be considered as weights. Q_T is a Good-Turing type function. r is the occurrences in the training data and k is a threshold.

Another approach for reducing data sparseness when using trigram model is to use class n-gram. In this approach, words with similar semantic or grammatical behavior can be grouped together and represented using a particular class. For example for a tourism domain language model, we may group together proper name of places such as Paris, London, New York, and Tokyo into city class. This will allow the n-gram model to generalize better because proper names are rarely found in the data. Furthermore, new words for example in this case new city names can be added and associate to the same class. As a result, they will inherit the same relationship found from the previous observed training data, without using any additional training data.

1.3 Language Acquisition

The studies on language acquisition can generally be divided into first and second language acquisitions. First language acquisition deals with the development of the first or native language on children, while second language acquisition looks at the process of learning a language other than the native. First and second languages are acquired differently and often in different stages of life, which can affect the language capability of the speakers. Before we look at non-native speech recognition, it is important to first understand how language capability developed in an

infant compared to an adult who learns a new language, and how these differences will affect the speech recognition system, so that approaches can be developed to take into consideration the differences to improve the system for non-native speakers.

1.3.1 First Language (L1) Acquisition

Research in the area of language acquisition studies the developmental process of children in learning a language. For years, researchers have been fascinated by how children are able to master language effortlessly in a short period of time. An understanding of this area may help educator in proposing a better approach for learning second language.

Study shows that the ability of children to distinguish speech sounds is well developed before their speech production ability [O'Grady 2000]. Early research has shown that infants demonstrate the ability to respond to phonetic units of all languages. This discriminative ability could be accounted to the general auditory processing mechanism, which has also been shown for animals such as monkeys. The acquisition of a particular language on the other hand, involves the specialization of this general auditory processing mechanism, where specialized auditory features are exploited [Kuhl 2000]. For example, for speakers of tonal languages, they appeared to activate different regions of brain compared to those of non-tonal languages [ScienceDaily 2008]; therefore, the ability for infants to discriminate non-native consonant and vowel contrasts, and musical rhythms deteriorates across first year of life, increasing the sensitivity toward their native languages [Weikum 2007]. According to one of the theory [Kuhl 2000], infants detect patterns in language input, and exploit the statistical properties of the input, altering their perception to enhance specific language perception, see Figure 1.9. In this case, Japanese speakers have shown to have lost the ability to distinguish the /l/ from /r/ because their perceptual space for /r/ has been expanded to the area of /l/ to enhance their recognizing of the speech sound /r/.

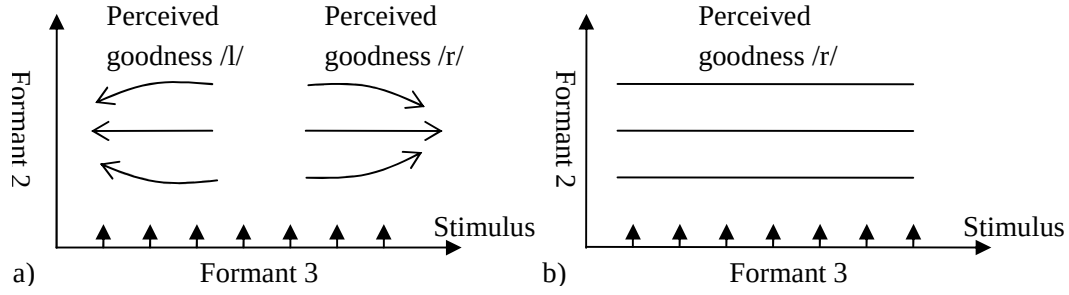


Figure 1.9 a) Physical stimulus given at different intervals equally spaced on the mel scale and the perceived category goodness by American listeners between /r a/ and /l a/ syllables. b) the perceived goodness of /r/ by Japanese listeners [Kuhl 2000, Iverson 2003].

Infants begin babbling starting from around six months of age. Studies show that there are similarities among the babbling sounds of infants, even though they are from different origin.

Among the sounds, plosive consonants are more frequent compared to fricatives [O'Grady 2000]. Understandable words are pronounced starting at around twelve months old, with one-word utterances. Syllable deletion is common in the words pronounced. There are also systematic deletions of certain sounds to simplify the syllable, for example the word *stop* is pronounced as [t *a* p] [O'Grady 2000]. Other widespread observation is the substitution of one sound by another. Among the vocabulary acquired, nouns are more frequent. It is followed by words associated with common daily expressions and interactions. Children also often express themselves using two-word utterances, few months after their first one-word utterances. For example, *baby chair* with the meaning of 'the baby is sitting on the chair' [O'Grady 2000]. More complex and longer sentences are expressed several months after this. Furthermore, studies also found that simply exposing infants with recorded material has shown to have no benefits. On the other hand, active exposure with feedback and recast from multiple speakers is important [Kuhl 1997, O'Grady 2000].

1.3.2 Second Language (L2) Acquisition

There is a general agreement that age of learning (AOL) a language plays an important role in determining how well one will master the language, whether it is the first or second language. Results from functional magnet resonance imaging (fMRI) shows that the early adult bilinguals activate overlapping regions of the Brodmann's area in brain, whereas the late bilingual subjects who acquire the language in adulthood activate two distinct regions of the area for processing the two languages [Kim 1997]. However, there are some differences on how age influences language acquisition. Early studies suggest that there is a critical period (Critical Period Hypothesis) to learn first and also second language. After this period, the ability to acquire the language successfully will decline and compromise. One of the main reasons stated is that neurology maturation will reduce the ability for a person to learn a language after the period has passed. For second language acquisition, some works show that there is a linear relationship instead of a threshold between age of learning and the perceived accent [Flege 1995]. Inaccurate perception is claimed as one of the main reason why non-native speakers are unable to articulate like the native speakers [Flege 1995, Rochet 1995, Kuhl 2000]. A further research on the topic shows that the frequent usage of L1 by a non-native speaker can also increase the perceived accent, even though the speaker learns the foreign language at a young age [Flege 1997, Flege 2004].

Besides age, the learner himself is also a factor which determines how successful he learns a new language [O'Grady 2000]. The degree of motivation on learning a language is one determining factor. The cognitive factor of a person is also important. There are two cognitive styles: field independent and field dependent. The field independent person focuses on the details of a particular subject, while those of field dependent focus more on the overall picture. Thus, field independent and dependent learners do better in different tasks in second language learning. Field independent learners are better for example in grammatical task, while field dependent learners are better in synthesizing knowledge. In addition, different people have different learning strategies which may determine their rate of success in second language learning.

There are two types of errors made by second language learners, namely transfer errors and development errors according to Ontogeny model [O'Grady 2000]. Transfer errors are mistakes made based on the knowledge of the speaker on his L1. On the other hand, development errors are mistakes made in the process of acquiring the first language. According to the model, the amount of transfer errors go down in a linear fashion over time, while development errors increase and then go down over a period of time, see Figure 1.10. Second language acquisition involves four areas: phonology, pronunciation, vocabulary and grammars.

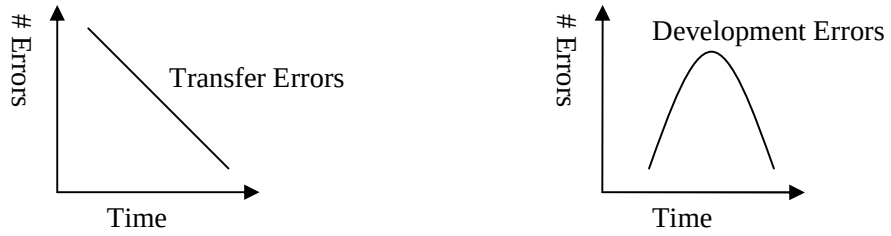


Figure 1.10 L2 errors predicted by Ontogeny model

1.3.2.1 Phonology

Mastering the L2 phonology is one of the most difficult lessons for language learners. It involves the phonetic segments (speech sounds) and prosody.

Phonetic Segments

The correct acquisition of speech sound is important, because it distinguishes one word from another. The interference from the L1 of the speaker may lead to incorrect perception, where the speakers will interpret L2 sounds based on their L1 phonology [Flege 1995], which resulted in incorrect speech production. For instance, from the previous example, Japanese speakers will interpret English /l/ as Japanese /r/. However, studies show that this is not irreversible. With appropriate training, it has shown that Japanese speakers can be trained to perceive correctly these two sounds [Logan 1991]. A general model about second language acquisition known as speech learning model (SLM) has been proposed to describe the changes in the perception of the speaker, which takes into account of age-related limits on the ability to produce L2 sounds [Flege 1995]. According to SLM's hypotheses:

i. Initially L2 sounds are linked perceptually to the closest L1 sounds (the linked L1 and L2 sounds also known as "diaphones).

ii. However, when L2 learners gain experience in L2, they may gradually recognize the phonetic difference between certain L2 sounds and the closest L1 sounds and new category can be established.

iii. The greater the perceived dissimilarity between diaphones, the higher the phonetic differences between sounds will be recognized and new phonetic category will be formed for the L2 sound.

iv. When AOL increases, the likelihood for distinguishing L1 and L2 sounds will decrease.

v. Similar L2 sounds will be classified as L1 sound (equivalent classification), and both sounds may interfere each other to achieve a 'tolerant' value to be used in both L1 and L2.

vi. If a new sound category is created, the new L2 category may not necessary be the same as the native, it can be based on different features or phonetically different from the native to maintain phonetic contrast between the phones using the common phonology space.

vii. Eventually, speaker produces L2 sounds that correspond to the phonetic category.

In term of vowels, some observations that support the model were presented by Flege [Flege 1987]. In his studies, he found that the type of perception errors one might make on L2 vowels depends on the perceived similarity of it with L1. Vowels that are perceptually far from each other are easier to distinguish. Sometimes, when listeners seem to correctly identify the type of vowel, they may in fact use a different feature than the one normally used by the natives for distinguishing the sounds. For example, Spanish speakers of English incorrectly use duration for distinguishing /I/ from /i/, rather than spectral cues [Bohn 1995]. This means that during production, they may use the wrong feature also. For 'new' L2 sound test (perceptually far from L1), an articulation experiment of French /y/ by native English speakers shows that their ability to articulate /y/ improves when experience increases, while initial learner will articulate it as /u/.

On the other hand, for consonants, AOL affects the production of consonants just like on vowels. Native Italians who learn English by about the age of 10 are capable to produce /θ/ and /ð/ consonants that are perceived to be native like. After that, their performance reduces dramatically according to AOL, and the L1 counterpart is often produced instead of the L2 sound [Flege 1995]. Experiments on "similar" L2 plosive /t/ articulated by inexperienced non-native speakers showed that they used their L1's voice onset time (VOT) in L2 [Flege 1987]. Note that VOT is the duration between a burst and the beginning of the vibration of vocal cords [Kent 2002]. On the other hand, experienced native English speakers of French and experienced native French speakers of English have a compromise VOT value which is between VOT of L1 and L2. Consonant final stops /t/ and /d/ produced by Spanish native speakers are less often identified by native English speakers. The analysis shows that the less experienced Spanish speakers have smaller closure voicing differences between /t/ and /d/, while the experienced Spanish speakers have the same closure voicing as the native English speakers. However, the failure for native English speakers to detect the final consonant uttered by the experience Spanish speakers has shown that the feature learned is not the significant one.

The speech learning model provides a general insight into the changes in the perception of language learners at different stage of learning and also on the type of speech productions that are

foreseen for different type of L2 sounds compared to the native sounds of the speaker. However, they do not specify quantitatively the effects from the L1 phoneme set of the speakers on a particular L2 phoneme set.

Prosody

In general, prosody comprises of intonation, stress and rhythm. L2 Intonation is one of the widely studied topics. The intonation shows the variation of pitch (fundamental frequency or f0) over the speech. Intonation conveys linguistic information, for example whether the utterance is a statement, command or question, and also the state of a person for instance emotion, physical, sociolinguistic and others. Hence, an inappropriate intonation may cause misunderstanding in communication. One of the widely used intonation model is by [Pierrehumbert 1980]. It separates the intonation to phonological and phonetic component. The phonology component consists of four types of tone at the higher suprasegmental level: the pitch accent (tone associates with stress), tone at the phrase level and tone at the boundary of the sentence. The tone at the phonetic component studies the changes of tone along phonetic level, for example from one phone to another. Each of these tones is represented by a pair of high and low tone.

The studies conducted on L2 intonation acquisition showed that there are similarities with L2 speech sound acquisition discussed earlier. L1 interferes with L2 intonation is apparent. Hence, speakers from different L1 commit different type of errors in intonation according to their L1. For example native English speakers who learn Italian have a local (tone) peak at the later position of a syllable, which is native English like compared to native Italian speakers which have a peak earlier in a syllable [Mennen 2006], see Figure 1.11. However, not all L1 features will affect L2 intonation. There are also different degrees of success in L2 intonation acquisition which depends on the experience of the speakers [Ueyama 1996, Menne 2004].

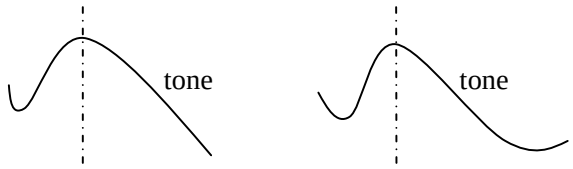


Figure 1.11 The Italian word ‘Mantova’ uttered by a) non-native Italian (native English speaker) compared to b) native Italian speaker [Mennen 2006]

Different languages have different levels of tone and pitch range. For example, English has a higher mean tone than German. Studies show that most non-native speakers are able to adapt to different level of tone when they learn a new language, but their pitch range is narrower compared to the native speakers [Mennen 1998, Ullakonoja 2007].

In term of stress, three factors have been determined to affect the placement of stress in English by L2 learners: syllabic structure, lexical class and stress patterns of phonologically similar words [Wayland 2006]. However, speakers from different L1 (e.g. stress versus tonal language) shows different degree of these effects. For example, non-native speakers where their native language is a stressed language, they tend to transfer their stressed pattern from their L1 to L2, while those from tone language do not show a specific pattern. Interestingly, studies also found that non-native English speakers from a stressed native language perceive English stress poorly compared to speakers from a non-stressed language because of L1 interference. However, the same speakers with stressed native languages performed better in the tests involving the production of L2 stress compared to speakers from non-stressed native languages [Altmann 2006].

1.3.2.2 Pronunciation

In term of pronunciation, when language beginners encounter unfamiliar L2 syllables, they may resort to use their L1 pronunciation rules. For example Spanish does not allow /s/ to be followed by consonant sequences word initial. So, it is possible that Spanish speakers pronounce the word 'Spanish' as 'Espanish' [O'Grady 2000]. Development errors which are observed in L1 acquisition by children can also happen on second language learners. For complex pronunciations, they may resort to simplify the pronunciation by inserting, deleting or substituting certain sounds with another sound. This is the same as what happened in children acquiring L1. Other possibility is to pronounce the word according to the grapheme sequence, which is not necessarily correct.

1.3.2.3 Vocabulary

Non-native speakers are likely to use the wrong vocabulary for expressing themselves. In general, the errors committed come from two sources: L1 vocabulary transfer and wrong association. L1 vocabulary transfer happens for L1 words with similar orthography to L2. For example, a Spanish who speaks English may say "my wife is embarrassed", where the actual intended meaning is "my wife is pregnant" because the Spanish word "embarasado" means pregnant [O'Grady 2000]. Another example from French and English is the words "avertissement" and "advertisement" which may seem to be similar have different meaning. The French word means warning, but the English word means an announcement. In the worst case, the speakers may also introduce words in their native language. For example, in French, there exists words that end with "-ment" like in English, for example the French word "département" with the corresponding English word "department". An English language beginner may end up introducing the English word "experiment" in French. Furthermore, "expérience" is a valid word in French, but not "experiment".

Wrong association happens because of influence by the relationship between the native language vocabulary and the actual meaning. This occurs for certain words which have one-to-n relationship between the meaning of the native word of the speaker and the foreign word. For instance, in Malay there are two possible words which can translate the English word 'rice'.

'Beras' /b ə r a s/ is the uncooked rice and 'nasi' /n a s i/ is the cooked rice. Thus, a Malay language beginner with English native background may associate one of the Malay words (e.g. beras) with rice for all contexts, and end up using it wrongly in certain context. For the sentence 'I eat rice', the speaker may translate it as below, which is incorrect.

*Saya makan beras /s a y ə m a k a n b ə r a s/

The probability of incorrect usage by a native French learning Malay is even higher. See Figure 1.12 for the relation between the words.

In term of perplexity, study shows that non-native speech has a lower perplexity compared to native speech [Tomokiyo 2000]. This may means that non-native speakers tend to use more common words in their speech, while native speakers that are expert in their language are capable of using more specific words to form sentences.

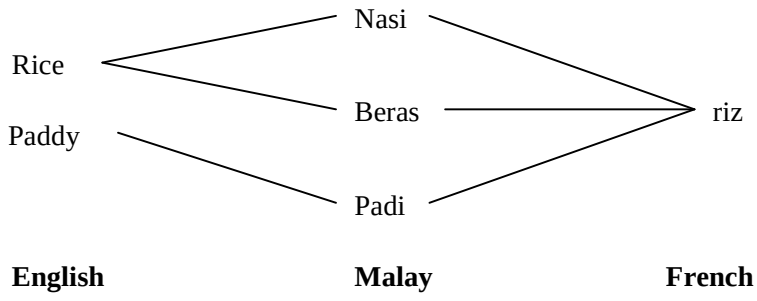


Figure 1.12 Different associations between words from different languages

1.3.2.4 Grammars

Non-native speakers also make more grammatical errors compared to native speakers, because of the unfamiliarity with the target language. They are likely to use the sentence structure of their native language in the target language. For example in some languages like Spanish, it is possible to drop the subject in a sentence. Speakers having a native language with this characteristic have also shown to commit certain grammatical errors associated with them [O'Grady 2000]. Another example is the placement of adverb by French speakers. In French, the adverb is put after the verb. Studies show that English beginners of French origin show the same tendency when they speak English. For example, they tend to say "Marie watches often television", because of the influence from the French sentence "Marie regarde souvent la télé" [O'Grady 2000].

1.4 Speech Recognition for Non-Native Speakers

The advancement in communication and transport has made cultural interactions between different parts of the world easier and more frequent. Although European languages such as English, French, Spanish and others have long been the international languages learned by people around the world at schools and universities, with the development of countries in Asia and other continents, more and more people around the world are embracing languages such as Mandarin, Indian, Arabic, Korean etc. Nowadays, apart from the native language, most people can speak at least one foreign language. Furthermore, people are more and more likely to travel to foreign countries for vacation or business. They also often pick up some common phrases with the help of the Internet and travel guides to make the communication easier with the locals.

Speech recognition technology has achieved tremendous advancement in the past decades. However, most of the works in speech recognition in the past focus on native speakers. Non-native speech as we see in previous section is different from native speech in term of phonology, pronunciation, vocabulary and grammars. These differences give rise to what is known as accent of a particular group of non-native speakers. What is the difference between non-native speech and dialects? For dialect speakers, there is no transfer of L1 like what happens for non-native speakers, because the dialect is often the first language of the speakers. However, variation from the 'standard' language can still happen in the areas of phonology, pronunciation, vocabulary and grammars. However, unlike non-native speech, dialect has commonly accepted phonology, pronunciation, vocabulary and grammars rules or standard among the speakers. Conversely, there is different degree of accent in non-native speech. The difficulty of non-native speech recognition is worsening by the number of languages available, and the limited amount of non-native resources. How non-native speech is going to affect an automatic speech recognition system? Three important components in speech recognition system are affected. They are the acoustic model, pronunciation model and language model.

The mismatch which is caused by the negative transfer of the L1 phonology of the speaker to L2 will affect the performance of the target acoustic model and pronunciation model for recognizing their speech. For similar phonemes, for example French and English /t/, they have different VOT. Inexperienced non-native speakers may use their native /t/ in the foreign language. As for prosody, although the differences do not affect the meaning of an utterance for a speaker, it may also affect the performance of a speech recognition system. The speaking rate is one obvious example. For new phonemes which are perceived to be different from L1 phonemes of the speakers, inexperienced non-native speakers may still have trouble articulating them even though they can perceive the differences. In some cases, they may even use the wrong features for differentiating the sounds, which may result in articulation that varies from the native variants which is modeled in the acoustic model. Context dependent modeling which is used to improve speech recognition performance for native speakers may not be useful for non-native speakers [Compernolle 2001]. On the other hand, a context independent model may end up performing better. As a result, the model which is built for native speakers is not fully compatible with non-native speakers. From previous section, we also learn that non-native speakers may also simplify some pronunciations which are not familiar for them through insertion, substitution and deletion.

We have seen that it is possible that Spanish speakers speaking English pronounce the word 'Spanish' as 'Espanish'. This pronunciation variant has to be added to the pronunciation dictionary.

Incorrect usage of vocabulary and grammars may also hurt the language model. Similar words, even if they exist in the speaker native language and the second language, may have different semantic and form of usage. Occasionally, vocabularies which exist only in the native language of the speaker but not in the foreign language may also be used. These will affect the n-gram probability of the words. The incorrect usage of grammars will show up as different n-gram probability of the words involved. Table 1.1 shows the difference in the accuracy of speech recognition for native and non-native speakers from some studies. The results show that the WER for non-native speakers is about twice or higher than the rate of native speakers.

Table 1.1 Comparison of the performance of automatic speech recognition (ASR) on different non-native speakers

	Target Language (L2)	Native Language (L1)	WER (native)	WER (non-native)
[Oh 2006]	English	Korean	4.2	39.2
[Liu 2006]	Mandarin	Cantonese	7.9	20.0
[Steidl 2004]	German English	50 countries German	18.5 35.0	34.0 65.6
[Wang 2003b]	English	German	16.2	49.3
[Witt 1999a]	English	Spanish and Japanese	-	28.2

1.5 Non-native Modeling in Speech Recognition

As mentioned in previous section, non-native speech has different characteristics compared to native speech. Hence, specific non-native models tailored to different non-native speaker groups have to be created to achieve better recognition speech performance. However, the lack of non-native resources implies that many of the conventional techniques proposed for native speakers are unable to be used effectively. Over the past decade, creative approaches have been developed for modeling non-native speech under the constraint of resources, by taking advantage of existing resources.

Automatic speech recognition system for non-native speakers has the same architecture as the conventional system at Figure 1.1. However, it may have an additional component which determines the accent of the speaker either manually or automatically. With this information, matching models which correspond to the accent of the speaker can be selected for decoding the speech. In this section, we will look at approaches for building acoustic, pronunciation, and language model for non-native speakers. In addition, current state of the art accent identification approaches will also be investigated.

1.5.1 Non-native Acoustic Modeling

Non-native speakers do not articulate the speech sounds like the native speakers, because their speech is often influenced by their native phonology. Non-native speakers from different origin therefore often have pronunciation habits which are related to their native language. Getting enough non-native speech to create non-native acoustic model is time consuming and sometimes unfeasible. Consequently, the existing approaches proposed for adapting acoustic model for non-native speakers make use of the native language of the speaker or a little non-native speech for adaptation. Generally, these approaches can be divided into four main categories. They are acoustic model reconstruction, acoustic model interpolation, acoustic model merging and the more general adaptation algorithms.

1.5.1.1 Acoustic Model Reconstruction

The most direct way of creating a non-native acoustic model is through acoustic model training. However, it is not easy to get enough non-native speech to create a non-native model. Thus, instead of creating it from scratch, existing target language acoustic model is employed as the bootstrapping model, which will be subsequently adapted using some acquired non-native speech. Studies have also found that native speech from the non-native speaker can also be useful for adapting the target language acoustic model in situations when non-native speech is not available [Uebler 1999, Tomokiyo 2001]. However, non-native speech has found to be a better adaptation source than the native language of the speaker.

Alternative state tying methods using phonetic decision trees to create tied-states during training have also been proposed to improve the performance of accented and non-native speech recognition systems. The idea is to initialize non-native acoustic features at state tying. These approaches have shown to reduce the recognition errors of non-native speech and at the same time cause little or no reduction in the performance of native speakers. This means that the same model can be applied for both groups of speakers at the same time. [Oh 2006] has proposed to tie all the confused target language phones together by using the same decision tree for the confused phones. This is done by using non-native speech to find the target language phones confusion. The phones with confusion probability exceeding a certain threshold will be selected and collected together. A phonetic decision tree will be constructed and tied together the states for these phones (see Figure 1.13a) using the standard decision tree approach discussed in Section 1.2.3. On the other hand, [Liu 2003a, Liu 2003b, Liu 2006] has proposed a slightly different approach for tying states earlier for accented speech by using decision tree. There are two types of tree which being used here. A standard phonetic decision trees built using the target language, and auxiliary tree which is also a phonetic decision trees but built with non-native data, where a particular phoneme is pronounced or realised as another phoneme. In another word, that phoneme is confused as another phoneme by the speaker. For example /d/ is realised or confused as /t/. The leaves of the auxiliary trees with single Gaussian density each will be merged to the nearest leaf nodes of the standard target language phonetic decision tree by applying weights (see Figure

1.13b). The purpose is to initialize the densities that define non-native speech. After that, standard training procedures follows, where only native speech is used for creating the acoustic model.

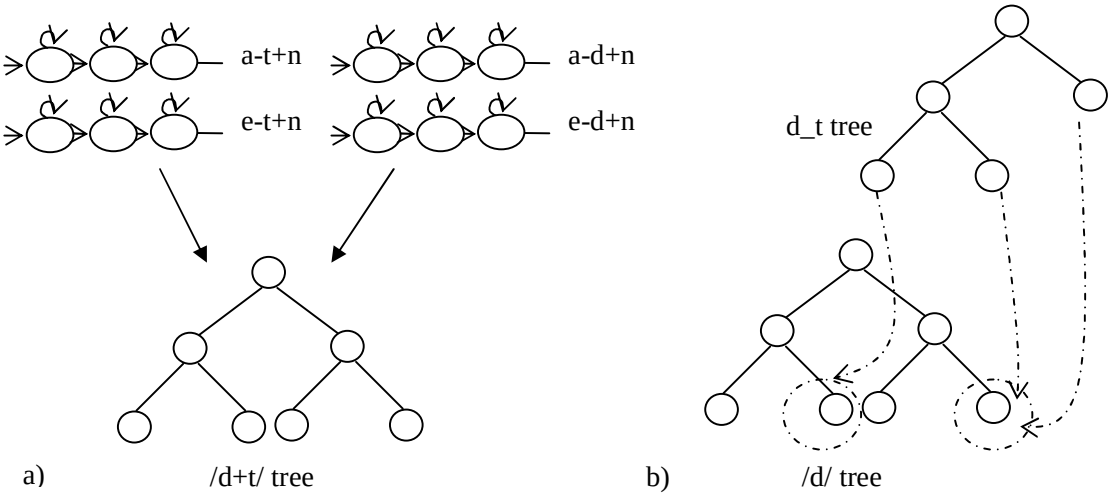


Figure 1.13 Comparing two states tying approaches for acoustic model reconstruction. a) The confusion of phoneme /t/ and /d/ is above the specified threshold. A decision tree /d+t/ is built to tie all triphones [t] and [d]. b) The confusion of /t/ and /d/ is used to build an auxiliary d_t tree, which will be merged to the leaves of the target language /d/ tree.

1.5.1.2 Acoustic Model Interpolation

Acoustic model reconstruction requires the raw corpus for modeling, but acoustic model interpolation on the other hand can be carried out even when the resources are in the form of acoustic models. It is carried out normally between two acoustic models by applying a-priori weight to the models. A target language acoustic model may be interpolated with the native language acoustic model of the speaker [Witt 1999a, Witt 1999b], see Figure 1.14. In this case, it is based on the hypothesis that the pronunciations of non-native speakers are intermediate between the two languages. For finding the target and source language model mapping, Witt has proposed three approaches. The first approach made used of linguistic knowledge for mapping the target and source language sounds. Another possibility is to conduct perception analysis by phonetician and the third approach is to use some non-native speech to create confusion matrix to find the phoneme confusion. Besides interpolating between target and the native language acoustic model, it is also possible to interpolate the target language acoustic model with the non-native acoustic model [Tomokiyo 2001, Wang 2003]. In this case, the model is created with only limited amount of non-native speech. Alternatively, instead of using it to adapt the target acoustic model, the non-native speech can also be used to find the phoneme confusion between the target language phonemes, and interpolation can be performed on these confused phonemes [Steidl 2004]. In a continuous HMM acoustic model, the matching Gaussians in the corresponding states will be interpolated. If the source acoustic model for interpolation is not derived from the target

language acoustic model, the Gaussians in the target and source are mismatched, so they have to be matched first before interpolation can be carried out using distance measure. Contrary, for a semi-continuous HMM, non-native speech is used to find the target language phoneme confusion, and interpolation can subsequently be carried out by interpolating the mixture weights [Steidl 2004]. The benefits of acoustic model interpolation are that the approach is simple to carry out and the interpolated model has the same number of components in term of number of states and Gaussians. In situation where some non-native speech is available from the speaker, [Witt 1999a] has proposed a non-native speaker adaptation approach which is able to estimate the interpolation weights automatically which is known as linear model combination.

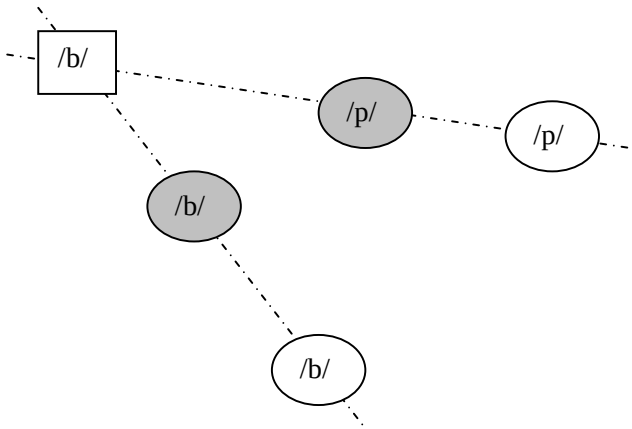


Figure 1.14 Acoustic space. Interpolation between target language phonemes in cycles and source language phoneme in square. The shaded cycles represent the interpolated non-native phonemes.

1.5.1.3 Acoustic Model Merging

Like acoustic model interpolation, acoustic model merging requires only the acoustic models, without any raw speech data. It involves combining two or more acoustic models from normally two sources. Often, the target language acoustic model will be merged with the corresponding native language acoustic model of the non-native speaker [Witt 1999a, Witt 1999b, Morgan 2004, Bouselmi 2005, Bouselmi 2006] to form a new model. 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 is also a work which merges the native and the non-native models [Minematsu 2003]. A weight will be assigned to each of the merged model, either on the transition of each model (Figure 1.15a) or into the mixture weights (Figure 1.15b). The weights can be assigned manually or estimated automatically using some non-native speech [Bouselmi 2005]. The disadvantage of acoustic merging is that it increases the number of states or Gaussians in each HMM and it may create some redundant distributions, which therefore increase the memory and computation time.

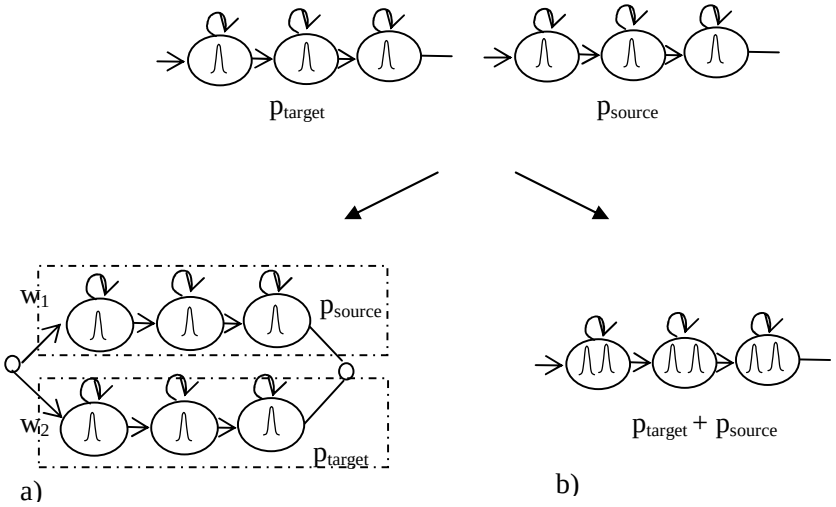


Figure 1.15 Two variants of acoustic model merging created from a target model /p_{target} / and a source model /p_{source}/. a) Two models are merged to form a new model /p/ with six states. Transition weights w_1 and w_2 are assigned. b) The mixtures from source model are merged to the corresponding state in the target state to form a new model /p/ with only 3 states. Different weights are assigned to the mixture weights of target and source.

1.5.1.4 General Adaptation Algorithms

General adaptation algorithms have proven to be effective for creating speaker specific model. By using a few utterances from a speaker, a speaker independent model can be adapted. Adaptation algorithms have also been used for adapting the environment conditions. The flexibility of adaptation algorithms, which are capable to work under limited resource constrains makes them an ideal choice to be employed for creating non-native models.

Two of the most popular adaptation algorithms in automatic speech recognition are Maximum Likelihood Linear Regression (MLLR) [Leggetter 1995] and Maximum a Posteriori Estimation (MAP) [Gauvain 1994]. [Tomokiyo 2001] found that adapting the target language acoustic model using MLLR or MAP with native speech of the speakers does not produce any improvement. Contrary with this result, the acoustic models created from merging of the target language acoustic model with the target language acoustic model adapted with the native language of the speakers, have shown to be beneficial [Bartkova 2004]. On the other hand, [Tomokiyo 2001] found that significant improvement can be obtained by adapting the target language acoustic model using small amount of non-native speech with MLLR or MAP. [Wang 2003a, Wang 2003b] proposes to apply non-native speech with MAP adaptation and Polyphone Decision Trees Specialization (PDTS). PDTS [Schultz 2000] is a decision tree adaptation algorithm which is used to grow specialized non-native branches from a target language trees by pruning to the point where it can be inserted. The adapted tree represents contexts of the non-

native speech data. Other general adaptation algorithm which has been tested on non-native speakers is [Deng 2006]. It is an unsupervised speaker adaptation algorithm using incremental singular value decomposition (SVD) adaptation technique.

1.5.2 Pronunciation Modeling

We see in Section 1.2.4 the typical ways to create and add pronunciation models in the pronunciation dictionary. Although pronunciation variants are also typically added in the pronunciation dictionary, it is not the only choice. The pronunciation variants or surface forms can also be modeled in other components. This does not only affect the memory and computation time for instance, but can affect the ways how the pronunciation variants are generated. In general, there are four possible places to model the pronunciation: pronunciation dictionary, language model, acoustic model and rescoring module [Strik 1998, Strik 1999].

1.5.2.1 Pronunciation Dictionary

Typically a speech recognition system has a pronunciation dictionary which stores at least the baseform representations or standard way for pronunciation of words or syllables. Hence, it is also natural to add the surface form or the variant pronunciation which maybe different from the baseform into the pronunciation dictionary as another possible realization of the word. For example in the previous example, the word *voyage* has the standard pronunciation /v ɔɪ ə ʤ/. It also has another possible variant pronunciation /v ɔɪ i ʤ/.

One possible way to add pronunciation variants is through listening to the utterances, and to write down their pronunciations. However, this is time consuming and not necessarily produces better result than the automatic approach. A study shows that manual pronunciation modeling do not necessary outperforms automatic approach [Goronzy 2001a]. Automatic variants generation can be performed using data-driven approaches. The general procedure for finding pronunciation variants is by aligning the hypotheses obtained from non-native speech against the corresponding reference transcriptions to create phone confusion matrix. Pronunciation variants can be observed from the phone confusion matrix. The unobserved variants can be found by generalizing the variants found according to context by using decision trees, and optionally adding the variant probability from the decision trees for each word into the dictionary [Humpries 1997].

The procedure described above requires the usage of non-native speech. However, in many situations non-native speech is hard to acquire. [Goronzy 2002] has attempted to generate pronunciation variants using the native phoneme set of the speaker. It is based on the hypothesis of cross-lingual transfer, where non-native speakers substitute target language phonemes with their native phonemes. The procedures are the same as described before for finding pronunciation variants using non-native speech. The only difference is that the target language speech is decoded by a phoneme recognition system of the source language (native language of the speaker). The phone confusions created from the alignment are then used to create the decision trees. The pronunciation variants can then be subsequently retrieved from the trees. However, the

results show that the new dictionary does not produce a significant improvement. Improvement is only obvious when the dictionary is used in conjunction with MLLR applied to the acoustic model with some speech from the speaker.

Different non-native speakers have different pronunciation habits which are specific to that group. [Raux 2004] has proposed an automatic speaker clustering method for non-native speakers based on a list of manually defined vocalic substitutions. A vector is used to represent a dialogue session from a speaker. It contains the number of times a variant appears. Clustering is carried out using model-based k-means and the vectors are randomly assigned to one of the cluster initially. In the subsequent iteration, the vector is assigned to the cluster which gives the highest likelihood to the pronunciation variants observed. This step is executed until it converges. The pronunciation dictionary created for each cluster has shown to be able to reduce the WER of a speech recognition system for each group.

1.5.2.2 Language Model

Language model is another place where pronunciation variants can be represented. There are two ways to calculate the best word sequence by taking into account the pronunciation variants. First method is to treat each variant as a separate word. An n-gram model will be built which includes all the variants. For this to be carried out, a transcribed speech corpus with the possible variants is required. Recall that in Section 1.2.2, the best word sequence can be formulated by argmax(P(O|W) P(W)). So, in this case the formula will be:

$$\hat{V} = \arg\max_{v}(P(O|V)P(V)) \tag{1.15}$$

where V is the sequence of pronunciation variant. The second method is to calculate the probability through an intermediate conditional probability as follows

$$\hat{V} = \arg\max_{v}(P(O|V)P(V|W)P(W)) \tag{1.16}$$

The probability of the variant P(V|W) can be estimated from the unigram. The second approach requires less data compared to the first approach. However, the shortcoming is that in the second approach the context of the variant cannot be modeled.

1.5.2.3 Acoustic Model

Modeling pronunciation variants at the acoustic level has blurred the difference between acoustic and pronunciation modeling. Acoustic model merging, which is described in Section 1.6.1.3 can also be used for modeling pronunciation variants [Bouselmi 2005, Bouselmi 2006]. The pronunciation dictionary in this case may store only the baseform pronunciations, and the variants

or the surface representations are stored in the acoustic model. The procedures to find the pronunciation variants are the same as discussed for pronunciation dictionary, the only difference is that the pronunciation variations are stored in the acoustic model. For example, if there is a lot of confusion between the phoneme /ə/ and /e/, then pronunciation variants can be put in the acoustic model like Figure 1.16.

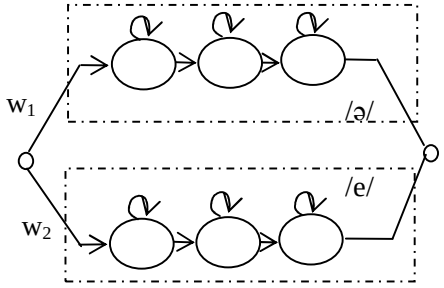


Figure 1.16 Acoustic model merging to represent pronunciation variation

Other approaches of modeling pronunciation variants at the acoustic level can be considered more as a combination with other approaches, particularly with the pronunciation dictionary. One of the most common techniques to optimize the trained acoustic model is to go through a procedure known as iterative transcribing [Nock 1998]. The idea is to use the baseform pronunciation as the initial model for training the acoustic model. Subsequently, pronunciation variants are found and added into the pronunciation dictionary. By forced-aligning the transcription using the new dictionary, the most suitable pronunciation variants for the words in the transcription will be chosen by the speech recognition system. The transcription is updated with the right variants and will be used for training a new acoustic model. The procedure is repeated iteratively.

Another issue related to modeling pronunciation variants is the selection of the type of acoustic unit to model. Most speech recognition systems model phone, but it is also possible to model units bigger than phone for example word, syllable, demi-syllable. Some speech recognition systems use a combination of these different units. The training of bigger units is only possible when the unit occur frequently enough. Smaller sub-phonemic units are also possible. We did not found any works which model units other than phones for non-native speech recognition, however indirect results from the work in accent identification [Arslan 1996] indicate that using whole word models for accent identification shows better performance. If this is true, then modeling whole word models for automatic speech recognition should also give a better result, provided that there is sufficient speech for modeling them.

1.5.2.4 Rescoring Module

Rescoring module is an optional component used in multipass search. In a multipass search strategy, decoding produces a word lattice or an n-best list, instead of a one-best word sequence, which will be re-evaluated in the rescoring module. The idea is to use a more general and simpler knowledge source (e.g. language model) during decoding to prune away unlikely hypotheses, and subsequently increase the complexity of the knowledge source in the rescoring module for finding the best match. This will make sure the search process is done in a manageable and feasible level.

The same idea can also be extended to knowledge source like pronunciation model, where during decoding a pronunciation dictionary which contains the standard pronunciations is used to produce word lattice or n-best list, while in the rescoring process the output will then be rescored using all the possible pronunciation variants. [Gruhn 2004] proposes to model the variants by using the hypotheses from the phoneme recognition system to align against the references to build a discrete HMM for the word pronunciation model. In the rescoring process, the decoder will produce n-best sentences, which will be rescored using a different pronunciation model. We will look more detail into this approach in Section 3.2.3.

1.5.3 Language Modeling

Obtaining sufficient non-native speech to create non-native acoustic model is not an easily achievable task. Acquiring enough non-native speech to model the grammars of non-native speakers is even more difficult. One possible solution for this is to collect instead the writing materials such as homework from the related non-native speakers, since the speaking habits such as common used sentence form and often made grammatical errors may also show up in their writing.

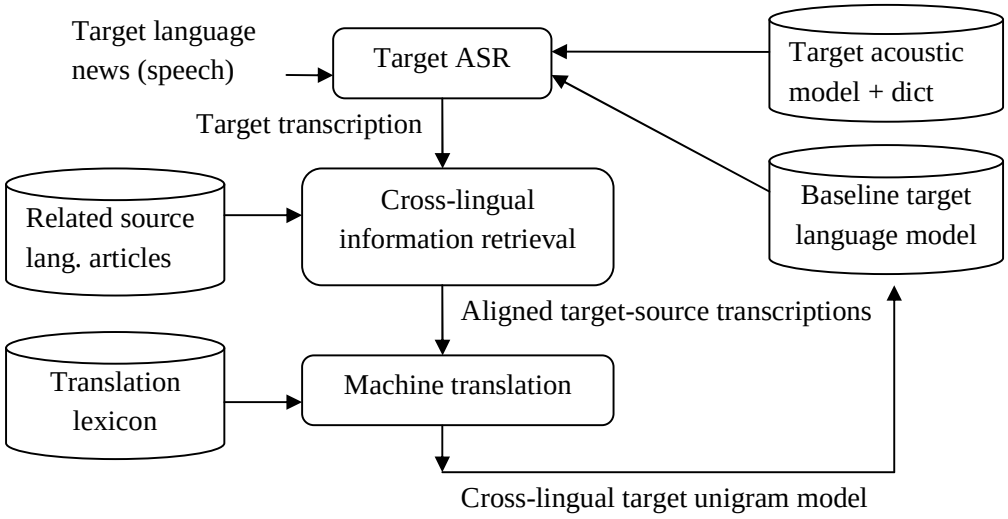


Figure 1.17 Cross-lingual adaptation of a language model

There are not many studies done in language modeling for non-native speakers, since obtaining enough dialogs for testing is not easy. One related work can be found in cross-lingual language model adaptation for multilingual speech recognition [Kim 2003], which may have the potential to be ported to non-native language modeling (see Figure 1.1.7 above). The approach tries to solve the lack of data for constructing a target language model by utilizing texts in resource rich source language, which have similarity to the target language to adapt the target language model. The related texts are identified using cross-lingual information retrieval technique. Subsequently, machine translation approach is used to estimate the target language unigram language model.

1.5.4 Accent Identification

Accent can be defined as a way of pronouncing a language that indicates the origin and social background of the speaker. Accent can generally be divided into two types: dialect (local accent) and foreign accent. Accent identification approaches can be grouped according to the type of features they treated. The two main categories are acoustic or phonotactic features.

1.5.4.1 Acoustic Features

One of the earliest works in accent identification employs f0 for dialect identification [Itahashi 1992]. The relative starting frequency and the changes in f0 are judged sufficient for Japanese dialect identification. On the other hand, [Blackburn 1993] has segmented incoming speech to voiced and unvoiced, stop and energy dip, before classifying it using neural network. Other acoustic features like formants, duration and bandwidth have also found their way into the accent identification domain [Liu 1999, Ghesquiere 2002]. The more general MFCC features were also employed for accent identification since the features can serve for dual purposes: speech recognition and accent identification. [Arslan 1996] proposes to use the HMM speech recognition system for accent identification task by employing MFCC features, see Figure 1.18.

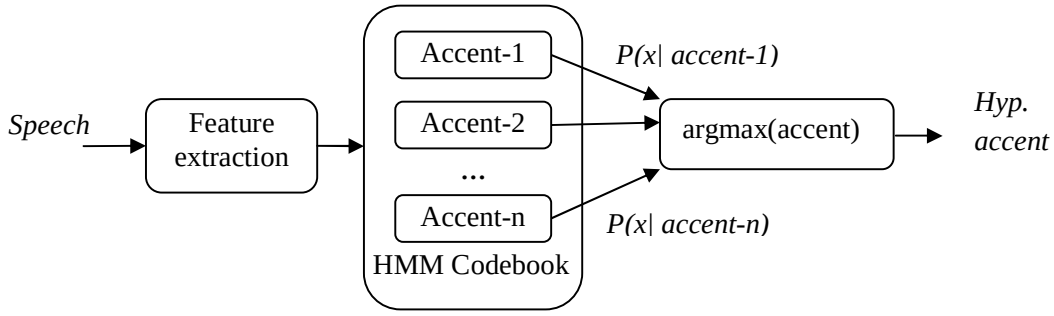


Figure 1.18 Accent identification using acoustic features

The system applies a text dependent strategy, which means the speakers were required to read specific words, and the scores derived from the acoustic models are used to identify the accent of the speakers. The accented HMM model which has the highest score for the input speech will be selected as the hypothesized accent. The study found that word based accent models perform better than phone based accent models in accent identification experiments. This shows that word based models are more capable of capturing coarticulation of non-native speech. However, phone models have the benefit of being vocabulary independent. On the other hand, in term of context dependent (CD) and context independent (CI) phones, context dependent phone models give a better result than the context independent phone models in accent identification tasks.

On the other hand, [Kumpf 1996] employed a slightly different strategy, where parallel phoneme recognizers with accented acoustic and phoneme language models (which we will see in next section) were used to evaluate the accented speech in a text independent manner. The accented speech recognizer which produces the highest score for the input speech is selected as the hypothesized accent. Besides the typical HMM, Gaussian mixture model has also proven to be useful for accent identification task [Chen 2001]. For identifying accent, Chen created accent specific models which are gender based using unsupervised approach, without the need for any transcriptions. This approach is similar to approaches in speaker identification. Approaches proposed for language identification have also found their way into accent identification. Heteroscedastic linear discriminant analysis (HLDA) and maximum mutual information (MMI) training have been combined for creating accent models based on Gaussian mixture models (GMM) [Choueiter 2008]. In addition, index language models have also been created by using the sequence of indices, which indicates the mixture component with the highest likelihood, generated by the Gaussian tokenizer.

In addition, there are also approaches that take into consideration all the possible types of acoustic features and subsequently use multivariate analysis to extract discriminative features for accent identification tasks. For example [Kumpf 1997] proposed to use Linear Discriminant Analysis (LDA) to reduce the feature vectors which consist of MFCC and other acoustic features, while [Ghesquiere 2002] proposed a one-way ANOVA to select the best discriminant features.

1.5.4.2 Phonotactic Features

Accent identification systems using phonotactic features make use of the hypothesized phonemes sequence uttered by the speaker to distinguish the accents. There are many ways to model the phonotactic features and one of it is using phoneme language model. [Zissman 1996b] has proposed a text independent system which used accent models from phoneme language models for evaluating non-native speech. Separate accented phoneme language models are built by decoding corresponding non-native speech using the target language phoneme recognizer, and the transcriptions generated are used to create phoneme bigram and unigram models. The bigram model will then be smoothed by interpolating it with the unigram phoneme language model. An utterance with unknown accent is decoded using the target language phoneme recognizer (PR).

The language model score is calculated by evaluating each accented phoneme language models using the following formula:

$$P'(w|w_{t-1}) = \alpha P(w|w_{t-1}) + ß P(w) + \gamma P_0 \quad (1.17)$$

where w is the word after w_{t-1}, and α, $ß$ and γ are interpolation weights which sum to 1.0, P_0 is the reciprocal of the number of speech sounds. The accented phoneme language model that produces the highest score will be selected as the hypothesized accent (see Figure 1.19). In fact, a similar approach can also be found in language identification [Zissman 1996a]. In language identification, instead of accented models, language models of different languages are created using the native languages. Besides the phoneme sequence, the position of the phoneme realised by a non-native speaker in a particular syllable is also important for predicting the origin of a speaker [Berkling 1998]. However, this means that the reference phoneme sequence has to be known, so that it can be compared against the hypothesis.

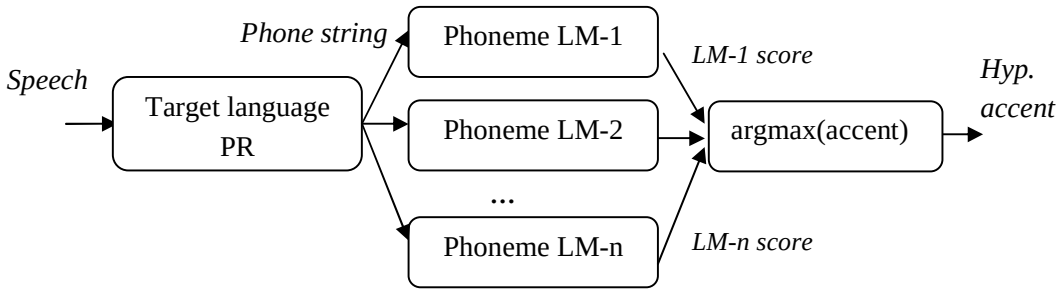


Figure 1.19 Accent identification using accented phoneme language model (LM)

Besides using phoneme language model to represent the phonotactic features of the accent, other approaches of phonotactic modeling using support vector machine (SVM) are also possible options. Support vector machine is a supervised learning approach for classification [Duda 2000]. The procedure to create SVM accent models is similar to text categorization, where instead of calculating the word distributions in a text during training, phoneme distributions from the decoding of the utterances are used for modeling. They are calculated by first decoding the non-native utterances with a phoneme recognizer. Then, each utterance is described by a vector which contains the hypothesized phoneme distributions of that utterance. Each utterance is an instance for training the SVM accent models. For improving the accuracy, several phoneme recognizers of different languages can be used at the same time. Using several phoneme recognizers of different languages have shown to improve accuracy in language identification tasks [Zissman 1996a]. Different phoneme distributions are calculated from each phoneme recognizer and they are later appended to form a single vector, which is subsequently used for modeling. During classification, an unknown utterance is decoded by the same multiple phoneme recognizers and its hypothesized

phoneme distributions are calculated. The accent of the utterance is then determined by using the SVM classifier (see Figure 1.20).

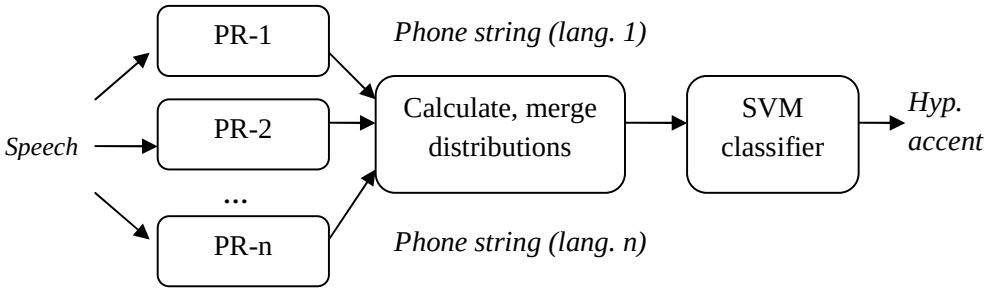


Figure 1.20 Accent identification using phoneme distribution features

Another similar approach is to use distance measure such as Kullback–Leibler divergence (KL distance) for classifying the accent of the utterance by calculating the phoneme distributions of the utterance compared to the phoneme distributions of different accent models. This is carried out using a phoneme recognizer for estimating the phoneme distribution of each accent. During identification, the phoneme distribution for test utterance is calculated. The accented model which gives the shortest distance for that utterance will be selected as the hypothesized accent.

1.6 Conclusions

In this chapter, we have presented the architecture for a statistical speech recognition system. Speech recognition can be considered as a pattern recognition and search problem. The acoustic model defines the most basic unit in speech, possibly with a phone, syllable or word unit. Phone is often selected as the preferred acoustic units because of the better generalizability of it compared to others. HMM is used to represent these acoustic units. An iterative re-estimation procedure is used for training the models until converged. For improving the precision of the model trained, the speech sounds can be modeled by taking into consideration the context. On the other hand, the pronunciation model defines words and phrases by using the acoustic units as the basis. Normally, the pronunciation entries in the dictionary which is based on IPA are used. However, in cases where this is not available, a grapheme based pronunciation dictionary can also be used instead. Given the defined vocabulary, the language model in turn defines the rules that govern the usage of words and phrases to construct sentences. N-gram is a widely used approach for this purpose. During decoding, the search process finds the best combination of acoustic and pronunciation units which is governed by the language rules, producing the most probable word sequence.

The performance of speech recognition system for non-native speakers is at least twice lower than for the native speakers. The reason is because there is a mismatch between the models trained (normally only using native speech), and the actual non-native speech that one tries to

recognize. In fact, non-native speakers do not speak like native speakers. The characteristics of non-native speech are different compared to the native. Non-native speech is influenced particularly by the native language of the speaker in phonology, pronunciation, vocabulary and grammar. The negative transfer of L1 knowledge to L2 may happen. Besides that, different types of development errors may also occur in these areas. In term of phonology, language beginners may interpret L2 speech sounds based on their native language. They may also use the wrong features to distinguish the L2 speech sounds, which will be reflected in wrong production. The influence of the native language of the speaker in phonology is not only limited to speech segments, it has also shown up in suprasegmental level such as intonation and stress. In the area of L2 pronunciation, non-native speakers have the tendency to use their native pronunciation rules in the target language. They will also simplify complex syllable that they are not familiar with by insertion, deletion and substitution. On the other hand, in term of vocabulary, they may even transfer their native vocabulary to the target language, or wrongly associate the vocabulary in their native language to the target language, which resulted in incorrect usage. Finally at the grammars level, non-native speakers may unconsciously employ the grammatical structure of their native language on the target language. As a result, the acoustic, pronunciation and language model that are trained do not match the characteristics of non-native speech.

Non-native modeling techniques have been proposed for reducing or compensating the mismatch between models trained for the native speech and the non-native speech, by inserting non-native speech characteristics into the models. The problem of modeling non-native speech lies in the difficulty of acquiring non-native data. Thus, the existing approaches proposed attempt to use a small amount of non-native speech or the native language of the speakers for improving the target models. Non-native acoustic modeling approaches can be divided into acoustic model reconstruction, acoustic model interpolation, acoustic model merging and the general adaptations. Acoustic model reconstruction and the general adaptation approaches are interesting methods which can be used when the non-native or the native language corpus of the speaker is available, while acoustic model interpolation and merging can be applied when the resources are in the form of acoustic models. However, there remain some unanswered questions. First, there are not many studies that look into how context modeling might affect non-native speakers. Second, how the multilingual acoustic modeling which has been applied for modeling new languages can be adopted for non-native adaptation. Third, it would also be interesting to know in what ways the existing linguistic studies can be taken advantage of.

Pronunciation modeling approaches can be divided based on the component where the pronunciation variants are modeled. The possible places are pronunciation dictionary, language model, acoustic model and rescoring module. Existing studies show that using native language alone for pronunciation modeling does not produce very encouraging results. Thus, the interesting questions are how the existing approaches can be modified, to make use of limited non-native speech for modeling, and which approach is better in modeling pronunciation variants. An equally interesting question is whether non-native speakers can be classified into groups based on their pronunciation habits, so that pronunciation dictionary based on group or speaker can be created.

In term of language modeling, there is not much works on non-native language modeling. We will also not be analyzing non-native language modeling here because of the lack of non-native textual data for testing.

Finally, accent identification approaches will be investigated. Accent identification approach can generally be divided according to the type of feature used, which are acoustic and phonotactics. We are interested to know how these approaches perform when only limited amount of non-native speech is available for creating the accent models. In addition, is it possible to propose an approach which is capable of modeling the accent models robustly using limited non-native resources?

CHAPTER 2

Multilingual Acoustic Modeling for Non-Native Speech Recognition

2.1 Introduction

The speech from non-native speakers has a different characteristic compared to the one from native speakers as we have mentioned in Chapter 1. Non-native speakers do not articulate the target language phonemes just like the natives do, because their perception pattern is influenced by their native language phonology. To articulate the 'new' phonemes, which do not exist in the mother tongue of the speaker, is a challenge that language beginners have to face. For example native English speaker who starts learning French may pronounce /y/ as /u/. On the other hand, for 'similar' phonemes which exist in both the target language and the native language of the speakers, non-native speakers may have trouble changing certain articulation habits which are specific to their mother tongue. They may also discriminate them using the wrong features which resulted in a different articulation. For experienced non-native speakers who acquire the language at an older age, they may not be able to articulate like the native speakers.

Current speech recognition systems achieve high recognition rate by taking advantage of very precise context clues such as triphone and even pentaphone. For example, context dependent model such as triphone model produces 25% relative error reduction compared to context independent model [Huang 2001]. However, non-native speakers are not capable of pronouncing the target language phoneme precisely like native speakers. Furthermore, for complex syllables, they may decide to simplify the pronunciations. As a result, the system performs very well on native speakers but at the expense of the non-native speakers.

Getting enough non-native speech to create a non-native acoustic model is time consuming and sometimes unfeasible especially when it comes to under-resource native languages such as Vietnamese or Khmer. Furthermore, there are more than six thousand languages in this world [WilFord 2007]. It would be an enormous task to collect all the non-native speech for each language. Although currently many efforts have been put in collecting non-native speech, the number of corpora available in language resource distributors such as LDC and ELRA is still small. According to a review [Raab 2007], only about one third of about forty corpora appeared in the research publications are actually available to others. In addition, the corpora are normally small compared to the native corpora.

The works in non-native acoustic modeling make use of a little non-native speech or the native language of the speaker for adaptation. Generally, non-native acoustic modeling approaches can be divided into four main categories. They are acoustic model reconstruction, acoustic model interpolation, acoustic model merging and general adaptation approaches. Creating non-native acoustic model using some non-native speech or the native language of the speaker by training is normally not as effective compared to other non-native acoustic adaptation approaches. However, appropriate state tying seems to be able to improve non-native speech recognition considerably. However, this can only be done if some non-native speech is available. Acoustic model interpolation and merging are interesting as they can be performed easily with the native language acoustic model of the speaker, without any non-native speech, while speaker adaptation approaches are also very useful for non-native adaptation when some non-native speech or the native speech of the speaker is available.

Existing works on acoustic modeling for non-native speakers however do not address some important issues. Firstly, multilingual acoustic modeling has been used for some time for constructing acoustic model for new languages. Many existing works employ the native language of the speaker for adaptation. It will be interesting to look at how the existing multilingual resources can be further utilized for adapting acoustic model for non-native speakers, since getting the target language non-native speech for adaptation is not always practical. Secondly, is to look at how the existing linguistic studies on non-native speakers can be used advantageously for acoustic model adaptation. Thirdly, there is no study which compares the effects of context (CI and CD) modeling on non-native speakers, although it is understandable that very precise context dependent models may not work in favor of non-native speakers. It will be useful to be able to employ context dependent model since it is beneficial for the native speakers and at the same time without causing any reduction in performance for non-native speakers.

Non-native multilingual acoustic modeling approaches have been proposed in our works by taking into consideration these issues. They are *hybrid of interpolation and merging* approach and a new *interpolation* approach. Section 2.2 presents the overview of multilingual acoustic modeling for non-native speakers. Section 2.3 describes the first step in non-native acoustic modeling, which is determining the matching phonemes between the languages. The remaining section 2.4 and 2.5 describes our proposed approaches, hybrid of interpolation and merging approach and new interpolation approaches to multilingual acoustic modeling for non-native speakers. We define the following terms to make the chapters easy to follow: the target language

is the spoken language, or the language for recognition by speech recognition syste,, while the source language is the language used for adapting the target model.

2.2 Non-Native Multilingual Acoustic Modeling

Multilingual resources such as acoustic models and corpora have shown to be particularly useful for creating acoustic models for new languages [Schultz 1998, Le 2005]. The idea is to use multilingual resources to overcome the difficulty in acquiring speech corpora especially for those rare languages. The general strategy is to construct a global phone set using the multilingual acoustic model. A new acoustic model for a particular language can be constructed by matching as much as possible the polyphone context. If some target language speech is available, it can be equally used to adapt the acoustic model created with the multilingual acoustic resources. Multilingual resources can be potentially useful for adapting the target language acoustic model to better suit the non-native speakers, since non-native speakers show cross-lingual transfers in their speech. Unlike multilingual acoustic modeling for new languages which creates a new model out of the existing multilingual resources, the multilingual acoustic modeling for non-native speakers uses the multilingual resources to adapt the target language acoustic model for non-native speakers. However, not all language resources are suitable to be used as source language for a particular group of non-native speaker adaptation. We identified three types of multilingual resources which can be used to adapt the target acoustic model:

- The native language of the speaker (L1)
- Any non-native language spoken by the same native group (L2)
- Language close to the native language of the non-native speaker (L3[1]).

These are the possible candidates for adapting the target language acoustic model. For instance, if we consider French as the target language for automatic speech recognition system, and if the task is to recognize non-native speech from Vietnamese speakers, the resources considered will be Vietnamese speech (L1), any non-native speech uttered by Vietnamese for example non-native English by Vietnamese (L2) and a language close to Vietnamese (L3), respectively.

In general, the approach consists of first, determining the cross-lingual phoneme transfer (as described in the subsequent section) of the non-native speakers. Next, with this information, the non-native adaptation can be carried out depending on the type of resources available. The idea is to create an intermediate model between the target and multilingual model mentioned above, which better suit non-native speakers. We propose two different methods of adaptation for modeling cross-lingual transfer of non-native speaker depending on whether we have a multilingual acoustic model or a speech corpus. In cases where the suitable multilingual acoustic model is available, a hybrid approach of interpolation and merging is useful for offline adaptation. However, when the original multilingual corpora can be accessed, it is possible to use it for

[1] Non-standard abbreviation

adaptation directly. Three types of cross-lingual transfer modeling are evaluated. They are manual interpolation which can also be applied offline, weighted least square and eigenvoices for cross-lingual transfer speaker adaptation. Finally, we will also see that the hybrid approach introduced for modeling cross-lingual transfer can also be used for modeling context variation for non-native speech recognition.

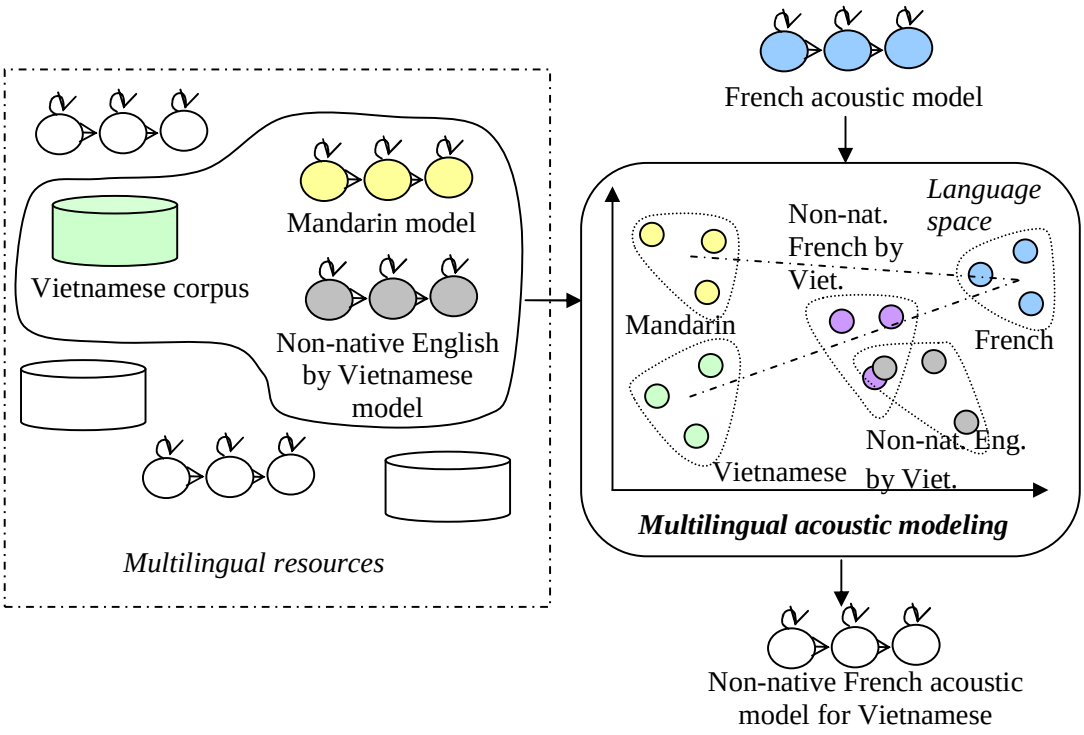


Figure 2.1 Creating a non-native French acoustic model for Vietnamese speakers using multilingual resources

2.3 Cross-Lingual Phoneme Transfer

As described in the previous section, there are three types of speech which can be used as source language to adapt the target acoustic model. They are the native languages (L1) of the non-native speakers, any non-native languages from the same native speakers (L2) and languages close to the native languages (L3). The hypothesis is that non-native speakers show cross-lingual transfer when they learn a new language (L2), where their native language (L1) sounds or phonemes are transferred to the new target language. So, the term "transfer" here carries the meaning of "applying the familiar to the unfamiliar" [Bohn 1995]. When finding the source-target phonemes transfer, the aim is to find the corresponding target language phoneme which is perceptually similar, according to the non-native speakers, to the one in the source language. This is important because it provides the necessary information for adapting the target language acoustic model using the available source language resources in the subsequent stage.

The approaches to measure the similarity of phonemes can generally be grouped into knowledge-based and data-driven approaches. In knowledge-based approaches, the probable source phonemes transfer for non-native speakers can be obtained from existing linguistic studies [Flege 1995], perception tests, and acoustic phonetic analysis or simply through the analysis of the International Phonetic Alphabet (IPA) table of both the target and the source languages. On the other hand, data-driven methods, which are used in multilingual acoustic modeling, can be carried out by using phoneme distances such as Euclidean distance, Kullback-Leibler, HMM distance [Juang 1985], phoneme confusion matrix and others.

We have adapted two popular approaches often used in multilingual acoustic modeling on new languages for measuring phoneme distance, so that they can be used to find the source language phoneme transfer, one using phoneme confusion matrix and another using IPA table.

2.3.1 Phoneme Confusion Matrix

Phoneme confusion matrix is created by aligning the hypothesis from the phoneme recognition system against the corresponding reference phoneme sequence from the forced alignment of a speech recognition system. The alignment will shows the hypothesized phoneme actually realized at the position of the actual phoneme. Although the current phoneme recognizer is not perfect (often with accuracy in the range of 50%), with sufficient amount of data available, the confusion matrix result can give insight to the actual substitutions that occur.

The alignment can be performed by using for example time alignment or Levenshtein distance. In our case, we use a variant of time alignment. Examples of the ways the alignment is carried out are shown in Figure 2.2. Each hypothesized phoneme is aligned to a reference phoneme according to the time alignment. Except in the case of Figure 2.2c, where deletion happens to the reference phoneme, a hypothesized phoneme can be assigned to more than one reference phoneme. However, it is better to introduce a deletion label. The confusion probability can be calculated in two pass, where the score from the first pass can be used to determine whether the deletion occurred at the first or second phoneme like in the case of Figure 2.2c. If this probability is not available, we assume the second or beyond phonemes are deleted.

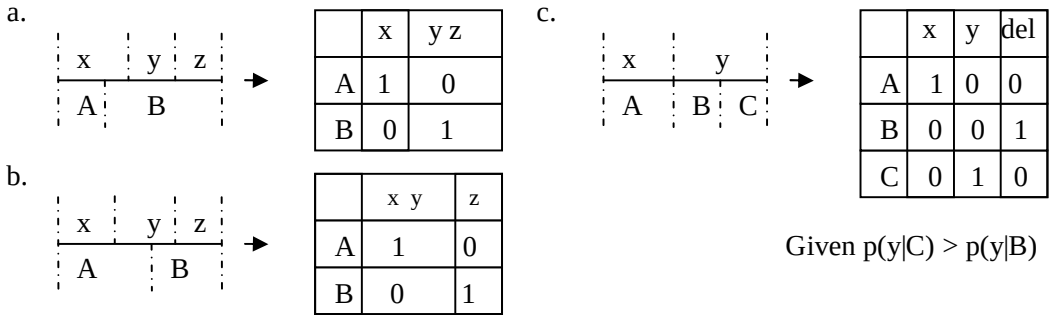


Figure 2.2 Examples of phoneme alignment. On the left are three pairs of hypothesis (top) and reference (bottom) phoneme string. On the right are the monophone confusion matrices

To find the matching source phoneme transfer for every target phoneme, without using any non-native speech, one possibility is to use a source language phoneme recognition system to decode the target language speech. The target language speech will also be forced-aligned using the target language acoustic model. The source phoneme with the most probable alignment for a particular target phoneme will be selected (see Figure 2.3).

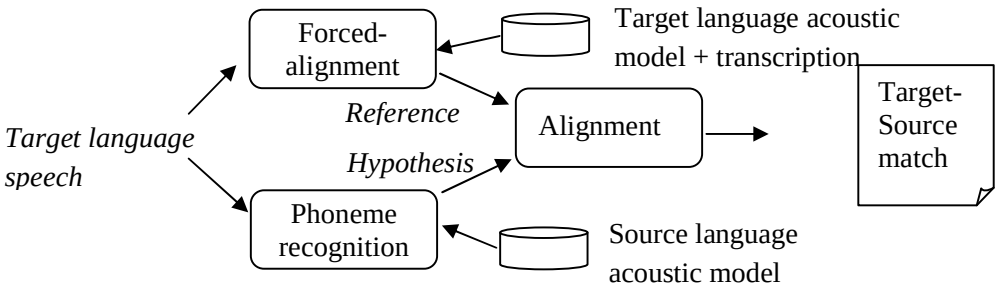


Figure 2.3 Determining phoneme match by using phoneme confusion matrix

In certain cases where we have access to the non-native speech, even though it is from another target language, it can also be used to create a phoneme confusion matrix. This secondary information derived from the phoneme confusion matrix can then be applied especially for those 'similar' phonemes (according to the International Phonetic Alphabet-IPA) which also exist in the target language of our interest. Thus, for example if the phoneme confusion matrix of non-native English gives the matching Vietnamese phoneme /s/ for the English phoneme /ʃ/, we can also use the same match for non-native French (see figure 2.4).

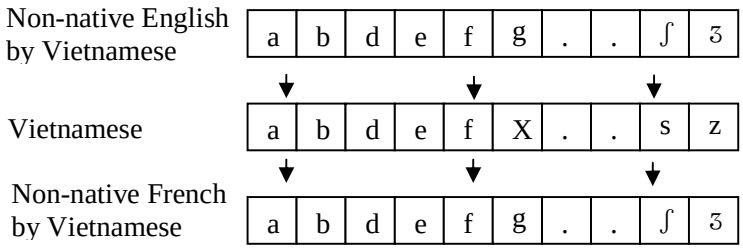


Figure 2.4 Using the knowledge from other non-native language for determining the phoneme match

2.3.2 International Phonetic Alphabet (IPA) Table

Existing linguistic studies [Flege 1987, Bohn 1995, Flege 1995] can also provide information on the kind of substitution to apply, especially for new phonemes. It is based on the idea that non-native speakers substitute target language phoneme with their native language phoneme. For

instance, it is well known that native Japanese speakers of English have difficulty with /ɹ/ and /l/, because they are perceived to be similar to Japanese /r/. Italians tend to hear word initial /ð/ as /d/. This information can help us in deciding the source-target phonemes transfer, either as primary or secondary information source. However, it is important to note that the substitution choice is not always easy to predict. Some studies showed that it is dependent on the mother tongue and education level. For example, Russians tend to substitute /t/ for English /θ/, whereas Japanese beginners use /s/. However when primary result is unavailable, this secondary information can provide clues on the type of substitution to apply. Table 2.1 shows the corresponding L1 phoneme transfers for the respective target English phoneme.

Table 2.1 Common observed source (L1) phoneme transfer from speakers of different origins for various target English (L2) phonemes [Flege 1995]

Target	Source	Description
/æ/	/a/ (Spanish), /ɛ/ (Korean)	/æ/ (front, open-mid_open), /a/ (front, open), /ɛ/ (front, open-mid)
/ɑ/	/a/ (Spanish)	/ɑ/ (back, open), /a/ (front, open)
/ɛ/	/e/ (Spanish)	/ɛ/ (front, open-mid), /e/ (front, close-mid)
/ɪ/	/i/ (Spanish), /i/ (Korean), /i/ (Chinese), /i/ (Italian)	/ɪ/ (front, close_close-mid), /i/ (front, close)
/ʊ/	/u/ (Italian)	/ʊ/ (central_back, back_close-mid), /u/ (back, close)
/θ/	/s/ (French), /s/ (Japanese), /t/ (Russian), /t/ (Italian)	/θ/ (fricative, dental), /s/ (fricative, alveolar), /t/ (plosive, alveolar)
/l/	/r/ (Japanese)	/l/ (lateral, alveolar), /r/ (trill, alveolar)
/ð/	/d/ (Italian)	/ð/ (fricative, dental), /d/ (plosive, alveolar)
/p/	/b/ (Arabic)	/p/ (plosive, bilabial), /b/ (plosive, bilabial)

The IPA table is constructed based on the linguistic study; therefore we can also take advantage of that knowledge. It consists of two main parts, a consonant table and a vowel chart. For consonants, the results from our tests in Chapter 4 show that non-native speakers often transfer the nearest source language phoneme (according to IPA) to the target language phoneme. It means that similar phonemes in the target language will be substituted for the same phoneme in the native language of the speaker. For new consonants which do not exist in the native language of the speaker, the nearest native phoneme can often be found in the same row (manner of articulation) for example Vietnamese speaker speaking English may replace /ʃ/ and /ʒ/ for the phoneme /s/ and /z/ respectively, or in the same column (place of articulation). However, it is not always like that. In some cases, we will see the non-native speaker substitutes a target language phoneme for a native phoneme at a nearby column in the IPA, although there is one native phoneme at the same column.

As for vowels, the vowel distances can be observed and compared using the vowel formant chart instead of using the vowel chart from IPA. As a result, a vowel formant chart has to be constructed to determine the similarity between the target language and native language phonemes. Figure 2.5 shows an example of vowel formant chart for target and source phonemes. Using this approach, we assume that a vowel in the target language will be substituted for a similar vowel in the source language (vowel which exists in both the language according to IPA). We can project these source language vowels to the corresponding target language vowels. The other remaining source vowels (new vowels which do not have a corresponding vowel in the native language according to the IPA) will then be projected by taking into consideration the projection of all other similar phonemes.

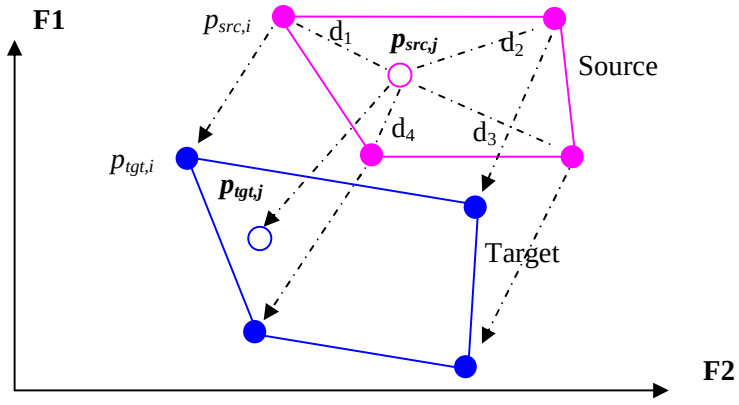


Figure 2.5 An example of vowel formant chart of target and source phonemes

The estimation is carried out simply using equation 2.1 and 2.2 below. We want to know the transformation of the vowel $p_{src,j}$, which is the source language (L1 of speaker) vowel which does not have a corresponding vowel in the target language, to the new target point $p_{tgt,j}$. $p_{src,i}$ is the source language vowel with a corresponding vowel in the target language, $p_{tgt,i}$; *d()* is the Euclidean distance, and $w_{j,i}$ is the weight for the vowel $p_{src,j}$, contributed by the similar vowel $p_{src,i}$. The weight is depend on the distance of the vowel $p_{src,j}$ to the vowel $p_{src,i}$. After all the source vowels are projected, the matching source language vowels for the vowel in the target language can then be determined by using Euclidean distance. The source vowel which is nearest to the target vowel is selected.

$$w_{j,i} = \frac{\sum_{i=1}^{n} d(p_{src,j}, p_{src,i})}{d(p_{src,j}, p_{src,i})} \tag{2.1}$$

$$p_{tgt,j} = p_{src,j} + \sum_{i=1}^{n} \left[\frac{w_{j,i}}{\sum_{k=1}^{n} w_{j,k}} * (p_{src,i} - p_{tgt,i}) \right] \tag{2.2}$$

Besides the method describe above, other possible method is using weighted least square (see equation 2.12) for estimating the transformation of the source vowels.

Since there are few possible sources of phoneme transfer information, we present here a general ranking of the confidence from the highest to the lowest:

- Perception test results and literature knowledge
- Phoneme confusion matrix using non-native speech
- IPA Table
- Phoneme confusion matrix using native speech of the speaker

In a situation when several sources of information can be obtained, it is also a good idea to compare their results and pick the most suitable one. In next section, we will look at how the information found here is exploited for modeling cross-lingual transfer by non-native speakers.

2.4 Cross-Lingual Transfer Acoustic Modeling

This section presents the proposed methods for modeling cross-lingual transfer of non-native speakers by using multilingual resources under different constraints. Two offline adaptations and two online speaker adaptations are proposed. The hybrid of acoustic interpolation and merging can be used for adapting target acoustic model using appropriate multilingual acoustic models, while interpolation approaches employ multilingual corpora for adapting the target acoustic model.

2.4.1 Hybrid of Acoustic Model Interpolation and Merging Approach for Offline Adaptation

Acoustic model interpolation is a promising approach to create a model which is intermediate between two languages using only the target and source language acoustic model. However in some cases, non-native speakers may also introduce sounds which do not have correspondence in the target language or vice versa. The idea is similar to the one described in Section 1.3.2.1, where non-native speakers tend to achieve an intermediate level for similar target and source language sounds, while for two very different speech sounds, speakers will use the one or another. An approach which therefore incorporates interpolation and merging seems appropriate.

The general approach of interpolation is to select the nearest Gaussian from the corresponding source state for every Gaussian in the target state using certain distance measure. Instead, we propose to carry out the interpolation in a different manner, where every Gaussian in the target state is treated like the ‘centroid’ for the Gaussians in the source state. The next step is to find the nearest target Gaussian for all source Gaussians using distance measure like Euclidean distance or approximated divergence distance.

$$Euclidean\ \ distance = \sqrt{\sum (\mu_i - \mu_h)^2} \tag{2.3}$$

$$Approximated\ \ divergence\ distance = \sqrt{\sum \frac{(\mu_i - \mu_j)^2}{\sigma_i \sigma_j}} \tag{2.4}$$

Every source Gaussian will be associated with only one target Gaussian. Certain target Gaussians will be instead associated with zero or more source Gaussians. When the distance between the associated target Gaussian and the source Gaussian is below a threshold, their means, variances and mixture weights will be interpolated (equation 2.5). Otherwise, merging is performed: for those target Gaussians without any associated source Gaussian (equation 2.7) or for the source Gaussian that are far (more than the threshold) from their associated target Gaussians (equation 2.6). In equation 2.6 and 2.7, their mixture weights will be reduced by the interpolation weight. The threshold can be calculated for example by measuring the average distance among the Gaussians, and then multiplying it with a constant. The resulted model is a hybrid model of interpolation and merging. Let p_{src} = { $p_{Src,1}$, ..., $p_{Scr,j}$, $p_{Src,n}$ } where p_{Src} is the set of source Gaussian associate with $p_{Tgt,i}$ the target Gaussian. $p_{Adp,k}$ is the adapted model with the weight α, while *d()* is the distance function and ω is the mixture weight.

$$p_{Adp,k} = \alpha.p_{Tgt,i} + (1-\alpha).p_{Src,j},\ p_{Src} \neq \phi,\ d(p_{Tgt,i}, p_{Src,j}) \leq threshold \tag{2.5}$$

$$p_{Adp,k} = p_{Scr,j},\ \omega_{Adp,k} = (1-\alpha).\omega_{Scr,j},\ p_{Scr} \neq \phi,\ d(p_{Tgt,i}, p_{Src,j}) > threshold \tag{2.6}$$

$$p_{Adp,k} = p_{Tgt,i},\ w_{Adp,k} = (\alpha).\omega_{Tgt,i},\ p_{Scr} = \phi \tag{2.7}$$

Using the information from the target and source phonemes, we can model non-native speaker cross-lingual transfer using the method proposed above with the target and source language acoustic model. The target language in this case is the new acquired language of the non-native speakers. The possible source languages can be any of the three types of languages we mentioned in the earlier of Section 2.2 (L1, L2 and L3).

The target and source acoustic models may have different configuration in term of number of states and number of Gaussians. Consequently, before the modeling can be carried out, the states and Gaussians of the target and source acoustic model have to be matched. In our current implementation, we use a simple context matching. This means that in cases where the models used for modeling are context dependent (CD) models, the matching triphone in the source model will be looked upon. If there is no matching triphone, a backoff strategy is applied where the context independent (CI) monophone in the CD acoustic model will be used instead. Another possibility is to use decision trees to select the best matching context.

Figure 2.6 shows an example of what will take place in two dimensions acoustic space for two target language Gaussians (French) and three source language Gaussians (Vietnamese) from

the matching state. Two Vietnamese Gaussians $p_{VN,\ s1g1}$ and $p_{VN,\ s1g2}$ will be associated with the French Gaussian $p_{FR,\ s1g1}$. Both will be interpolated with $p_{FR,\ s1g1}$, while $p_{FR,\ s1g2}$ which although associates with $p_{VN,s1g3}$ is far from the French Gaussian (more than the threshold), so both of them will be merged into the state, and their mixture weight will be recalculated with the weight given. The new state created will have four Gaussians.

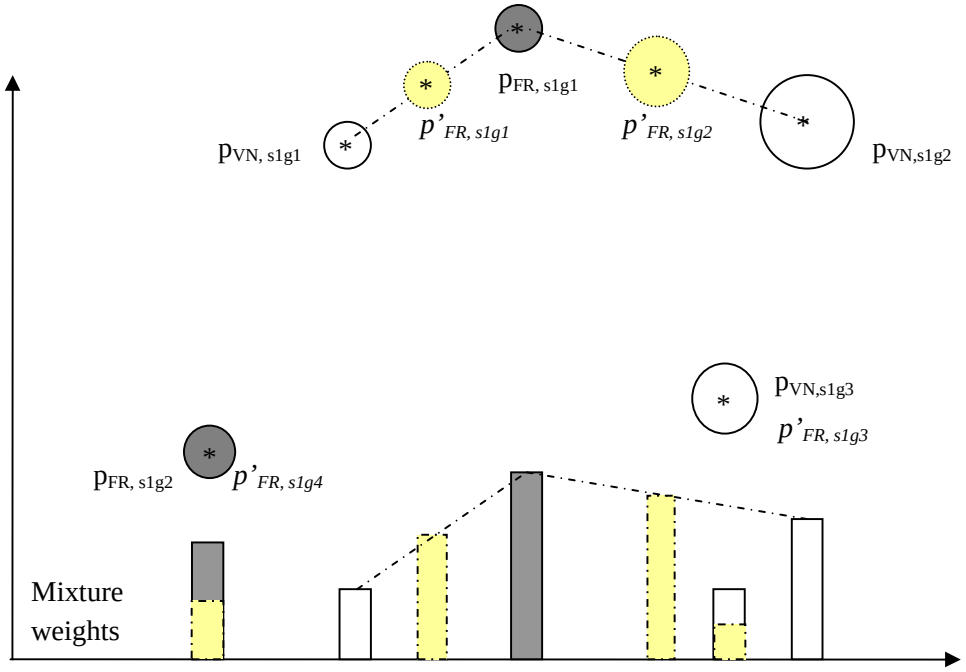


Figure 2.6 Interpolating and merging of the target model p_{FR} (French) and the corresponding source model p_{VN} (Vietnamese) to create the new model p'_{FR} in a two dimension acoustic space by setting the weight at 0.5. The points and circles indicate the means and variances. The newly created Gaussians are in dotted circles. The histogram presents the Gaussian mixture weights

2.4.2 Acoustic Model Interpolation for Offline Adaptation

Non-native cross-lingual transfer can be modeled using hybrid approach with source language acoustic model, as we have shown in the previous section. However, if we have access to the original source corpus, it is better to use the source language speech directly to create a new source language acoustic model which has the same configuration as the target acoustic model, in terms of the number of states and Gaussians, so that we can carry out the interpolation directly with the target acoustic model. This will avoid the use of distance measure from matching Gaussians between target and source states. Furthermore, using an adaptation algorithm will allow predicting unobserved means and the total number of Gaussians will stay the same, with no addition of Gaussian like in the case of the hybrid approach.

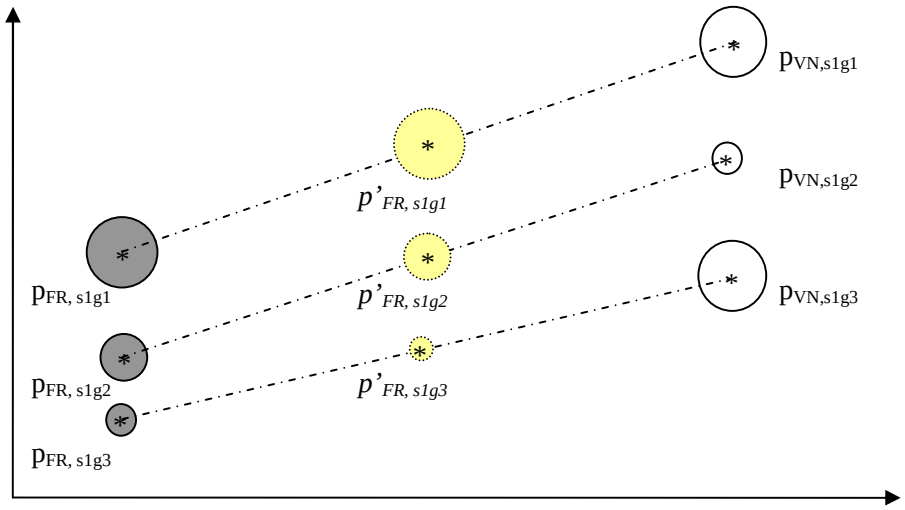


Figure 2.7 Interpolation of the target state p_{FR} (French) and the corresponding source state p_{VN} (Vietnamese) in a two dimension acoustic space by setting the weight at 0.5. The points and the circles indicate means and variances. The newly created Gaussians are in dotted circles.

This is done by adapting the target acoustic model using the source corpus. The first step is to map each phoneme in the pronunciation dictionary of the source language to the phoneme of the target language using the phoneme matching information that we had found previously. It must be noted, however, that several source phonemes may be mapped to the same target phoneme. For source phonemes which associate to multiple target phonemes, there will be several possible pronunciations (source language words with target language phonemes). One of the possibilities is to create all the possible pronunciation variants, and randomly associate one of the pronunciation variants in the pronunciation dictionary to the word in transcription. It is also possible that there are some source phonemes which do not have any associated target phoneme. We can either associate the source phonemes without any partners to one of the nearest target phonemes, or copy these source HMMs to the target acoustic model, and remove them later in the process. Our test shows that the results do not differ much. Now that the phonemes in the source language pronunciation dictionary have been converted to target language phonemes, the next step is to adapt the target acoustic model using the speech in the source corpus. Instead of using only Baum Welch algorithm to recalculate the Gaussians, we adapt the target language acoustic model (in our case MLLR and then MAP) using the source language speech and the modified pronunciation dictionary for a few iterations. This will create a source acoustic model which has the same amount of Gaussians, and at the same time, matches the target acoustic model. A weight is then predicted and assigned to the target and source acoustic model, and a new model is created with the following interpolation formula (see Figure 2.7),

$$\mu_{Adp} = w.\mu_{Tgt} + (1-w)\mu_{Src}, 0 \le w \le 1 \tag{2.8}$$

where μ_{Adp} is the interpolated means using the weight *w*, from μ_{Tgt}, the target language means, and μ_{Src} the source language means.

2.4.3 Acoustic Model Interpolation for Online Speaker Adaptation: Weighted Least Square

Manual interpolation is useful for adaptation without the need for any non-native speech. However, in certain situations when we are able to obtain some speech from the non-native speakers involved, we may want to predict the weights to apply on the acoustic models. Here we attempt to use weighted least square (WLS) to predict the weights to assign to the target and source acoustic model (created using the procedure describe in previous section). Since only two variables are estimated, the approach is suitable for adaptation even when small amount of non-native speech is available. The idea is to use some speech to predict the mean values of the speaker. Nevertheless, because the mean values are created using only few utterances, the vector contains a lot of missing values. Equation 2.8 can be rewritten in a matrix formulation,

$$Ax = b \tag{2.9}$$

$$A = \begin{bmatrix} \mu_{Tgt} & \mu_{Src} \end{bmatrix}, x = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, b = \begin{bmatrix} \mu_{Spk} \end{bmatrix} \tag{2.10}$$

where μ_{Tgt} and μ_{Src} are the target and source language means, while b is the mean values of the speaker. We want to find *x* that minimizes the value of $|Ax-b|$, which can be viewed as a measure of error. If there is an exact solution for *x*, then the error will be zero. We can solve the above equation and find *x*, given the least square errors by using the least square formula in equation 2.11. Variances *C* are used as weight to the least square formula [Montgomery, 2001].

$$A^T Ax = A^T b \tag{2.11}$$

$$A^T C^{-1} Ax = A^T C^{-1} b \tag{2.12}$$

The mean vector of the speaker, *b*, can be estimated by force aligning some speech from the non-native speaker using the target language acoustic model. However this method is not that accurate because the weight will tend to be close to the mean vector of the target acoustic model that we had used. A better approach is to create a merged acoustic model by using both the target and source acoustic models, and use it for the forced alignment instead. The merged model has the same number of states, but with the combination of the target and source Gaussians. Assume that we create the merged model (Figure 1.15b) by appending the source Gaussians of every state to the corresponding target state. The mean vector of the speaker can then be estimated using the formula below,

$$b_{j,k} = \frac{\sum \gamma(j,k)\theta + \sum \gamma(j,k+n)\theta}{\sum \gamma(j,k) + \sum \gamma(j,k+n)}, \quad \sum \gamma(j,k) + \sum \gamma(j,k+n) > 0 \tag{2.13}$$

where $b_{j,k}$ is the mean vector of the speaker at state j and Gaussian k, and n is the number of Gaussians for state j for the target language model. Since only a few utterances are used, many of the values in the vector b will be zeros. Only the values in the vector b which is non zero are used for estimating the weights by using equation 2.12.

2.4.4 Eigenvectors Interpolation for Online Speaker Adaptation: Eigenvoices

Eigenvoices method has been successfully applied in speaker adaptation [Kuhn 1998a, Kuhn 1998b, Kuhn 1999]. The works in eigenvoices are inspired by the works in eigenfaces for face recognition [Turk 1991]. These works are made possible through a pattern analysis approach known as principal component analysis (PCA) for finding the vectors that best characterize or describe the pattern of a set of feature vectors. This is achieved by analysing the covariance structure of a data set and finding directions of different variability. The vector (axis or linear equation) which forms the largest variance from the data is the principal component [Flury 1988]. These generalized vectors are known as eigenvectors and their relative descriptive power of the data are indicated by their eigenvalues, see Figure 2.8. The eigenvectors can then be used to describe or normalize any feature vectors including those unknown relevant vectors by finding the eigenvalues for them.

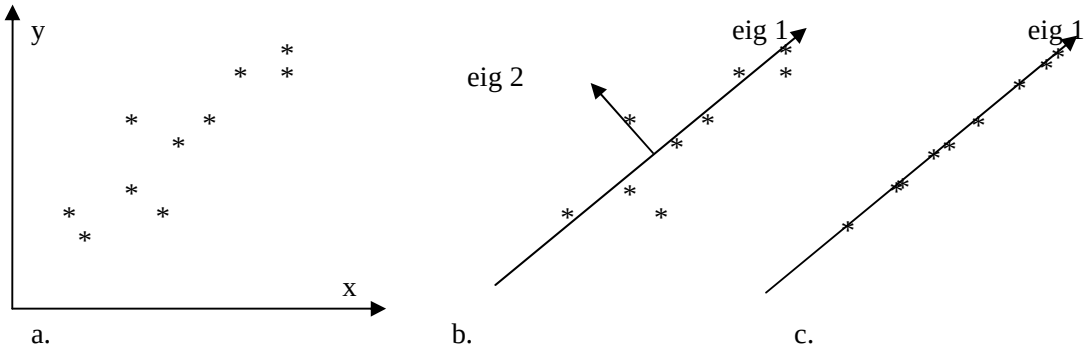


Figure 2.8 a) Original plot of data. b) Plotting of data using eigenvector one and two. c) Normalization of data using eigenvector one, the new axis with the largest variance

In the previous two methods, interpolation is carried out on the acoustic models. For eigenvoices, on the other hand, we can see it as an interpolation of eigenvectors. The standard

eigenvoices technique is applied here. However, we attempt to insert language space into the eigenspace by adding source language supervectors using the source language.

Speaker adaptation in eigenvoices is carried out by creating a speaker space and subsequently finding the speaker we want to adapt on that space. The first step to create a speaker space is to create a speaker dependent acoustic model for each speaker. For each target language speaker dependent acoustic model, the process is the usual one in which we first create a speaker independent acoustic model. Subsequently, speaker dependent models for each speaker are derived, by adapting the speaker independent model using a few iterations of combined supervised MLLR and MAP adaptations with the speech from each speaker. Next, we create the components for language space by going through similar steps we used to create the source language acoustic model for the previous interpolation methods. The only difference is that in eigenvoices, we have to use MLLR and MAP to adapt the target speaker independent acoustic model to speaker dependent acoustic models using the source language speech from each source speaker (see Figure 2.9).

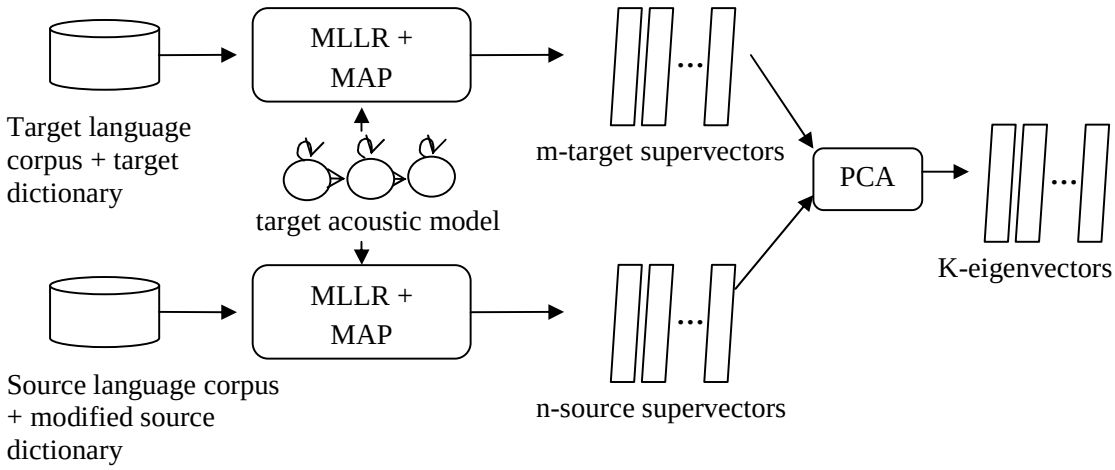


Figure 2.9 Steps to create eigenvectors using target and source language corpus

Once the speaker dependent acoustic models for target and source languages are created, the means of the acoustic models are written out, each as a sequential vector which is known as supervector. A total of K supervectors, each with a dimension of D will be created. Next, principal components analysis (PCA) or singular value decomposition (SVD) can be used to find the eigenvectors E=e(1)…e(K), from the supervectors which define the eigenspace (Howard, 2000). Not all eigenvectors will be used. A subset of eigenvectors, k, which have among the highest eigenvalues (principal components), will be selected for interpolation, where k is less than K, and K<<D. The projection methods in PCA, MLED [Nguyen 1998] or other methods [Westwood 1999] can be used to find the interpolation weights by using some speech from the

speaker. We have chosen to use MLED approach to estimate the new means. The weights for the eigenvectors can be calculated with the equations below:

$$v = Qw \tag{2.14}$$

$$v_e = \sum_{e=1}^{E} \mu_e(j,k)^T C(j,k)^{-1} \sum_{t=1}^{\tau} \gamma_t(j,k) o_t \tag{2.15}$$

$$q_{e,j} = \sum_{e=1}^{E} \sum_{f=1}^{F} \mu_f(j,k)^T C(j,k)^{-1} \mu_e(j,k) \sum_{t=1}^{\tau} \gamma_t(j,k) \tag{2.16}$$

where v is an E-dimension vector, Q is an (E x E) matrix, and w is the E-dimension eigenvalues or weights. μ_e is the means of the eigenvector and C is covariance matrix of the speaker independent acoustic model; j and k are the state and Gaussian mixture component respectively. The weights w can be estimated using Gaussian elimination and the new means can be estimated as follow:

$$\hat{\mu} = Pw \tag{2.17}$$

where P is the eigenvectors $[\mu_1^T, \mu_2^T, ...\mu_k^T]$.

2.5 Context Variation Modeling

Precise context dependent modeling as mentioned earlier is not suitable for non-native speakers. One possibility is to use a smaller tied state or even context independent acoustic model for non-native speakers. However, this means that different acoustic model has to be used for native and non-native speakers, because native speakers can benefit from a precise context modeling. Here, we propose to use the hybrid of interpolation and merging proposed earlier for modeling cross-lingual transfer for non-native speakers for modeling context variation.

2.5.1 Hybrid of Acoustic Model Interpolation and Merging Approach for Offline Adaptation

The idea applied for modeling context variation is similar to modeling cross-lingual transfer, where the hybrid approach proposed is used to create an acoustic model which is intermediate between a very precise context dependent model and a very flat context independent model. When modeling context variation, the model with a smaller number of states (e.g. context independent model) will be treated as the target model while the other one will be considered as the source model. In context modeling, since all models with bigger number of states are also part of the model with smaller number of states (both are from the same language), all source Gaussians are assumed to have a target Gaussian interpolation partner. Thus, no threshold needs

to be set. This is the difference compared to cross-lingual transfer modeling.

For example, if we employ a context independent model (target model) and a context dependent model (source model) for context variation modeling, context dependent (CD) triphones are matched to their corresponding context independent (CI) monophones. Next, the corresponding CI Gaussian for every CD Gaussian is found using a particular distance measure (Equation 2.3 or 2.4). Interpolation is then performed on the CI Gaussians with their associated CD Gaussians, while the CI Gaussians without any interpolation partner will be merged. See Figure 2.10 below.

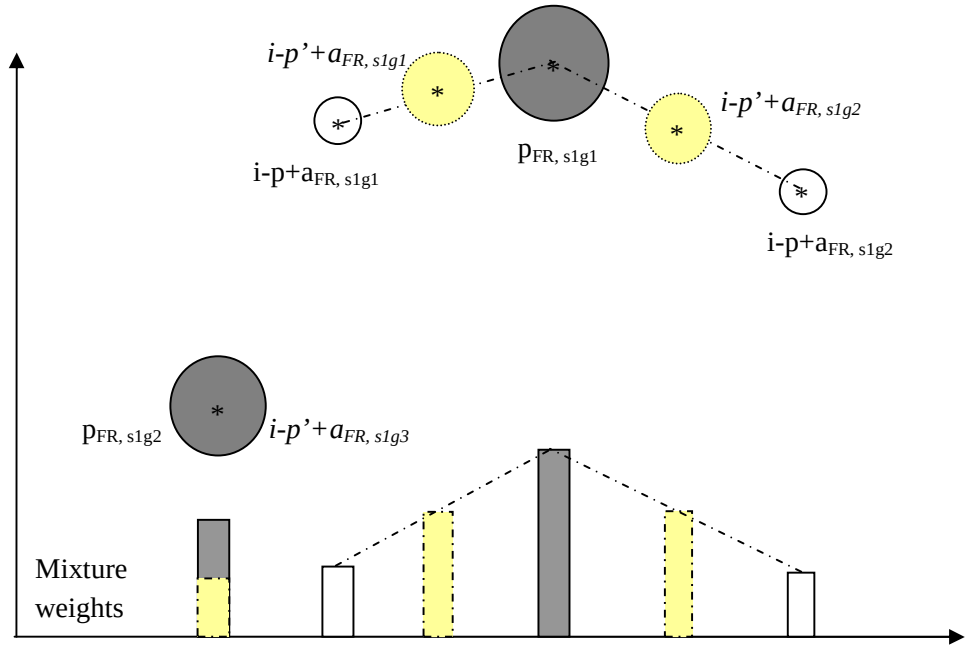


Figure 2.10 Interpolating and merging of context independent model p_{FR} and the corresponding context dependent model i-p+a_{FR} to create a new *i-p'+*a_{FR} model in a two dimension acoustic space by setting the weight at 0.5. The point and circle indicate mean and variance respectively. The newly created Gaussians are in dotted circles. The histogram presents the Gaussian mixture weights

2.6 Conclusions

We have presented multilingual acoustic modeling approaches to adapt target language acoustic model for non-native speakers without requiring any non-native resources from the target language. The multilingual acoustic modeling approaches proposed can be used for modeling cross-lingual transfer and context variation to improve non-native speech recognition. Multilingual resources from three types of speech can be used for non-native adaptation when the

target non-native language is not available. They are the native language of the speaker, any non-native speech from the same native speakers and languages close to the speaker native language.

Non-native speakers often transfer their native language phonology to the target language. The mismatch of speech sounds and the acoustic model will degrade the system recognition capability. In addition, unlike the native speakers, non-native speakers are not able to pronounce the speech sounds of the target language precisely because of unfamiliarity. Hence, context dependent modeling which is beneficial for improving the speech recognition performance for native speakers may not be useful for non-native speakers.

For modeling cross-lingual transfer by non-native speakers, two approaches have been proposed for treating multilingual acoustic models or corpora. However, before the modeling can be carried out, the target and source language speech sounds have to be matched. This can be done by using phoneme confusion matrix or IPA table. Phoneme confusion matrix is a data-driven method, which can be employed by using the native language speech with a source language phoneme recognizer and a target language speech recognition system. The IPA approach on the other hand makes use of linguistic and IPA information for finding the match. Depending on the type of resources that are available, acoustic modeling can then be performed. If the source resource is in the form of acoustic model, the hybrid approach used for modeling context variation can also be applied for modeling cross-lingual transfer. However, if the resource is in the form of corpus, interpolation can be carried out. In certain situation when some non-native speech is available, the speech can be used to estimate the interpolation weights by using weighted least square method. This approach is attractive because there are only two parameters to measure. This means that we do not need a lot of speech from the speaker to estimate the weights. Eigenvoices approach which is coined to fast (limited speech) adaptation has also been proposed for non-native acoustic modeling. It uses the source language for creating a bi-lingual space in the eigenspace, and subsequently finding the position of the speaker on the eigenspace for adaptation. For context variation modeling, the hybrid of interpolation and merging approach has been proposed for creating a model which is intermediate between a very flat context independent model and a very precise context dependent model. With appropriate weight, the new context dependent model created can be applied not only for improving speech recognition system for non-native speakers, but also be used for native speakers without causing huge decrease in word error rate. All the proposed approaches are experimented in Chapter 5.

CHAPTER 3

Non-Native Pronunciation Modeling and Accent Identification

3.1 Introduction

In the previous chapter, we have looked at acoustic modeling without using any non-native speech for adaptation. However, we learned from previous study that using only the native acoustic units of the speaker to model pronunciation variants is not effective. Thus, in the coming section, we look at pronunciation modeling approaches for modeling pronunciation variants by using a little amount of non-native speech. Following that, a preliminary work on accent identification has also been proposed. The new approach can work even with limited amount of non-native speech for creating the accent models.

3.2 Non-Native Pronunciation Modeling

As discussed in Chapter 1, non-native speakers have difficulty to pronounce words or syllables like the native speakers. For complex and unfamiliar syllables, non-native speakers tend to simplify them, just like the children learning their first language by insertion, deletion or substitution of speech sounds. On the other hand, for target language syllables which are similar to the native syllables of the speaker, they may tend to articulate them by employing their native manner of articulation. The differences in the pronunciation strategies and the pronunciation model used result in lower speech recognition accuracy.

Pronunciation modeling approaches can be divided based on the component in the speech recognition system where the pronunciation variants are modeled. There are four possible locations, namely the pronunciation dictionary, language model, acoustic model and rescoring module. Studies in pronunciation modeling found that modeling non-native pronunciations by generating the pronunciation variants using the native language phonemes of the speaker alone into the pronunciation dictionary, do not seem to be effective for modeling the pronunciation behaviour of the speaker, neither is applying linguistics rules blindly to all speakers. On the contrary, some non-native speech seems to be prerequisite for modeling non-native pronunciation correctly.

We have experimented with some modifications to the *pronunciation dictionary* and *n-best rescoring* approach, so that with little amount of non-native speech, it is possible to estimate the non-native pronunciation variants of the speakers. We have also tested the possibility of clustering non-native speakers according to their pronunciation habits. For this, we propose an original speaker clustering approach which group speakers based on their pronunciation habits and use this information for pronunciation adaptation. We call this approach *latent pronunciation analysis*.

3.2.1 Pronunciation Dictionary: Decision Trees

There are few possibilities to derive pronunciation variants, and one of it is through the use of decision trees [Humpries 1996, Humpries 1997]. The procedure used here is the general one proposed, except that we derive the pronunciation variants by going through two passes, since we only have a little amount of non-native speech, and the phoneme recognizer employed produces around 50% recognition errors. It is thus important to have the hypotheses as accurate as possible. In the first pass, the observed variants are extracted using confusion matrix. Only variants that are observed more than the given threshold are selected from the confusion matrix. The possible pronunciations are then generated into the temporary dictionary. Then in the second pass, from the observed variants, the pronunciation variants will be generalized according to the features of the pronunciation context using decision trees to predict unobserved variants. Figure 3.1 shows the steps for deriving the pronunciation variants.

The objective of the first pass is to retain the more likely observed pronunciation variants and to remove those less likely. The hypothesis phoneme strings generated by phoneme recognizer from the decoding of non-native speech are aligned against the corresponding reference phoneme strings from forced alignment using the modified time-alignment approach presented earlier in Section 2.3.1. A triphone confusion matrix is then created from the alignment. A low threshold is set to the triphone confusion matrix, so that the pronunciation substitutions or variants which appear more than the threshold are selected. All possible (word) pronunciation variant combinations will be generated and added into a temporary pronunciation dictionary. This will produce a pronunciation dictionary which is very big compared to the original.

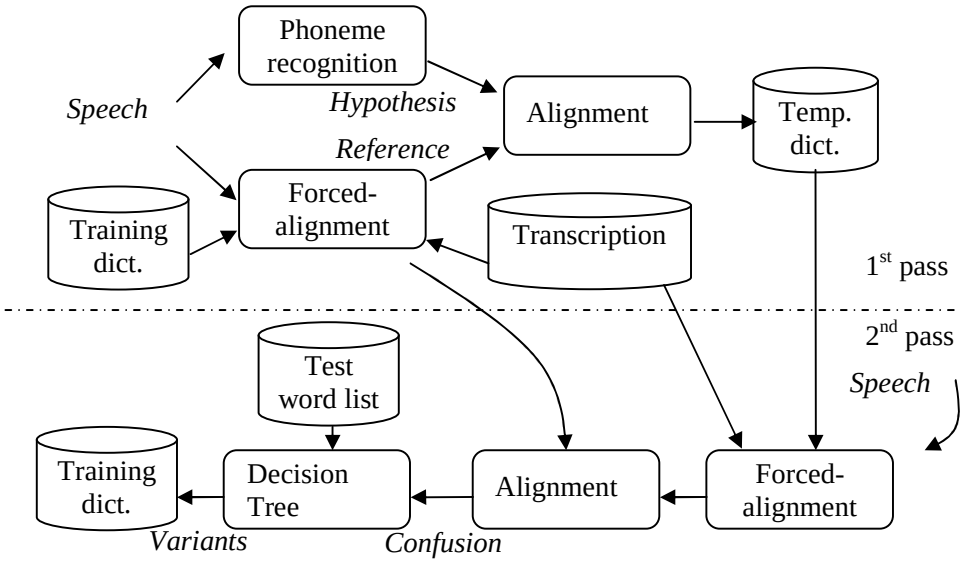


Figure 3.1 Generating pronunciation variants using decision tree

In the second pass, pronunciation variants will be generalized by using decision trees according to the context features. The first step is to re-estimate the hypotheses of non-native utterances, this time by force aligning the previous speech using the new (temporary) pronunciation dictionary created in the first pass. The new hypothesis phoneme time stamps will then be aligned against the corresponding reference phoneme time stamps estimated in the first pass. The triphone confusions with the same base phoneme will be collected together, and a tree will be built for each base phoneme except silence. The left and right phoneme contexts need to be translated to the corresponding feature vector (see Figure 3.2b). One possibility is to convert the context to phonetic feature vector using IPA based features, so that phonemes can be classified according to similar phonetic context. A decision tree algorithm such as CART or C4.5 [Quinlan 1993] will then classifies the triphone confusion according to the features defined. The idea is to classify triphones with similar pattern of substitutions together by searching for feature or attribute with high information gain. This will allow the unobserved pronunciation contexts to be predicted from the decision trees. This is similar to the usage of decision tree in state tying in Figure 1.7. After the decision trees for all the phonemes are built, a probability threshold is set to extract pronunciation variants which are observed more than the given probability from the leaves of the decision trees. The pronunciation variants will subsequently be added into the pronunciation dictionary by generating all the possible pronunciation combinations.

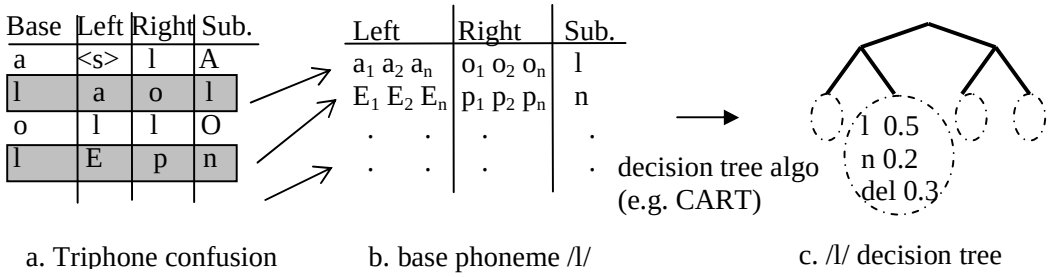


Figure 3.2 Sub-steps to create the decision trees in the decision tree process

3.2.2 N-Best List Rescoring

In the previous approach, pronunciation variants are added into the pronunciation dictionary, and the speech recognition system will then selects the best pronunciation during decoding. On the contrary, it is also possible to evaluate the pronunciation variants at the word lattice or n-best list stage after decoding [Gruhn 2004]. Figure 3.3 shows the architecture of the n-best rescoring system, where a phoneme recognizer is employed to decode non-native speech to produce hypothesis that will be used to re-evaluate or re-rank the n-best sentences from the speech recognition system. The approach applied here has the same architecture as the one suggested in [Gruhn 2004], except that we attempt to use a triphone model to represent the variants instead of a word model. The main reason is to reduce data sparseness because of limited data.

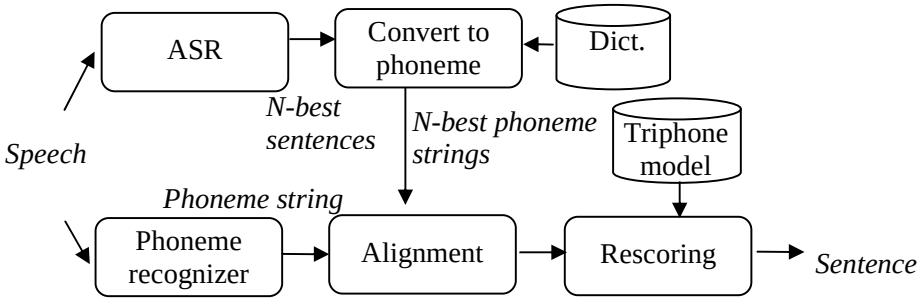


Figure 3.3 Pronunciation modeling using n-best rescoring

Before pronunciation rescoring can be carried out, the triphone model has to be created. The triphone model actually contains the triphone confusions of non-native speakers. In previous pronunciation dictionary approach, triphone confusions are also used to find pronunciation variants. However, only those variants from the decision trees that exceed the threshold are added into the pronunciation dictionary. In this approach, all the variants are used and the confusion probabilities are also made used of during evaluation. Some non-native speech is required for training the triphone model. This is done by first decoding non-native speech with the phoneme recognizer and the hypotheses produced are then aligned against the corresponding reference

phoneme strings. For smoothing the triphone confusion matrix, the triphone confusion values are interpolated with the corresponding monophone confusion probability. A floor value is used if both the confusion values are zero.

$$P'(sub \mid base, left, right) = w\, P(sub \mid base, left, right) + (1 - w)\, P(sub \mid base), \quad P(sub \mid base) > 0,\ 0 \le w \le 1 \tag{3.1}$$

$$P'(sub \mid base, left, right) = \ floor\ \ probability,\ P(sub \mid base) = 0 \tag{3.2}$$

where *w* is weight, *left* is the phoneme to the left of the base phoneme, *right* is the phoneme to the right of base phoneme, and *sub* is the hypothesis phoneme(s).

During evaluation, the non-native speech is decoded by the speech recognition system and the phoneme recognizer. Note that the pronunciation dictionary used during decoding contains only the baseform representations or standard pronunciations of the words. The speech recognition system produces n-best sentences with the n-highest P(W)P(O|W) score, where W is the word sequence and O is the observation. These sentences will then be converted to the corresponding (reference) phonemes strings from the word strings using the pronunciation dictionary. The hypothesis phoneme string from the phoneme recognizer will then be aligned against each of the reference phoneme string. The pronunciation score for each sentence from the n-best list is calculated by considering the triphone confusions of the hypothesis phoneme string against the reference string using the triphone model.

$$\text{Pr}onunciation\ score = \prod_{i=1}^{n} P'(sub_j \mid base_i, left_i, right_i) \tag{3.3}$$

where *i* is the reference base phoneme, and *j* is the hypothesized substitution. The pronunciation score for each sentence is then included in the log-linear model that calculate the final speech recognition composite score using acoustic and language score. The sentence from the n-best list with the highest composite score will be selected.

Figure 3.4 below shows a toy example of pronunciation rescoring using n-best list. Given a non-native utterance, the speech recognition system in this case produces two most probable sentences: "*ah bon*" and "*allô*", which are converted to the corresponding phoneme strings. The same utterance is also decoded by a phoneme recognizer producing the hypothesis phoneme string /a n o/. The hypothesis is then aligned against the references from the speech recognition system, and the pronunciation scores are calculated by considering the triphone confusion from triphone model. For example, the confusion of triphone a-b+ɔ~ as phoneme /n/ is 0.1. The total pronunciation score is calculated and the log probability is added to the total score (acoustic and language model). The word with the highest score is selected.

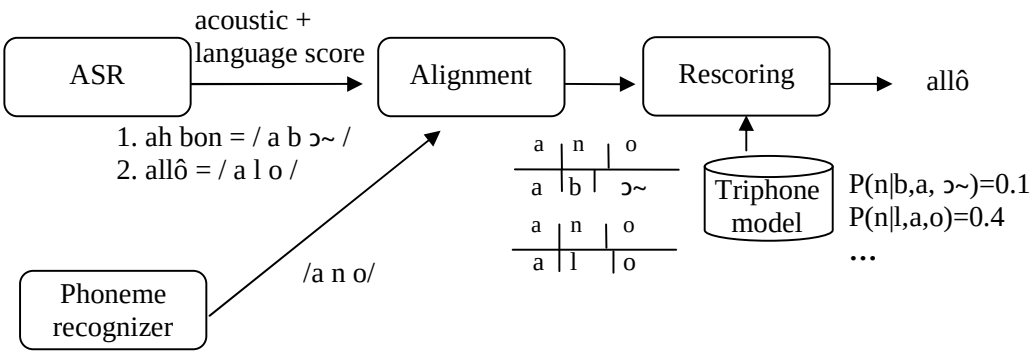


Figure 3.4 An example of pronunciation modeling using a 2-best list rescoring

3.2.3 Latent Pronunciation Analysis

Non-native speakers from different origin differ in their pronunciation habits. Sometimes, even non-native speakers from the same native language may have different pronunciation habits which are influenced by education, social-economy and other factors. Thus, it would be interesting to know whether it is possible to cluster speakers to different groups. Earlier work in speaker clustering using pronunciation habits has been proposed before [Raux 2004]. The work is not fully automatic since the possible vowel substitutions are manually defined. In this approach, we attempt to cluster non-native speakers automatically into groups based on their pronunciation habits, and subsequently use this information for adaptation. An unsupervised speaker clustering method based on pronunciation habits is proposed here. The approach is inspired by eigenfaces and eigenvoices approaches and also from idea given (but not experimented) in [Goronzy 2002]. We call it "latent pronunciation analysis" by analogy with the 'latent semantic analysis' used in natural language processing. The idea is to create a pronunciation eigenspace and use it for speaker clustering and pronunciation adaptation. This can be done by creating a set of speaker dependent pronunciation confusion vectors, which are used to derive pronunciation confusion eigenvectors. The eigenvectors can subsequently be used for clustering and estimating the pronunciation confusion of the test speakers. From the estimated pronunciation confusion, a speaker dependent pronunciation dictionary can be constructed and included in the speech recognition system.

Like the previous eigenvoices approach proposed for acoustic modeling, the first step in creating an eigenspace is to build speaker dependent models. In this case for each speaker, a speaker dependent pronunciation confusion supervector will be created. The supervector is in fact the triphone confusion matrix being laid out in a vector format. Table 3.1 shows an example of K supervectors created for the non-native training speakers (note that in actual case, all operations are done in column vector, instead of a row vector). The supervectors are built from their corresponding speaker dependent pronunciation decision trees. The procedure to create speaker dependent decision trees is the same as to create speaker independent decision trees described earlier in Section 3.2, except that the speech from each speaker is separated to build the speaker

dependent trees. The next step is to create the supervector *context* structure by extracting all pronunciation contexts from the test pronunciation dictionary. The possible *substitutions* for each context and their values are extracted from speaker dependent decision trees from every speaker and put into a pronunciation confusion vector. Since every speaker may have a different set of substitution, a standard pronunciation confusion vector (supervector) must contain all the possible substitutions for every speaker and in the same order. For each context, the total probability of the substitution for each context will then be normalized to 1.0.

Table 3.1 K supervectors of pronunciation confusion. The context row shows the base/ target phoneme followed by its left and right context

Contexts	b-a+n			d-ə+p		
Substitutions	a	*a~*	*a*	ə	DEL	
speaker 1	0.9	0.1	0.0	0.9	0.1	…
speaker 2	0.7	0.2	0.1	1.0	0.0	…
.						
.						
speaker K	0.5	0.05	0.45	0.4	0.6	…

D

The pronunciation models or pronunciation eigenvectors E=e(1)…e(k) are derived from the covariance matrix of K supervectors V with dimension D, where k is less than K, and K<<D by using principal component analysis (PCA) or singular value decomposition (SVD). Table 3.2 shows an excerpt of the actual eigenvectors created.

Table 3.2 An excerpt (feature 1-9) of the actual pronunciation confusion eigenvector 1 and 2 for non-native English speakers

Ctx.	aɪ−*a*+l				b−*a*+d				
Sub.	*a*	ɔ	DEL	əʊ	*a*	ʌ	ɔ	aɪ	l
Eig1	0.00580	0.00337	-0.0036	-0.00559	-0.00709	0.00994	-0.01280	0.01918	-0.00923
Eig2	-0.0132	0.00233	0.01552	-0.00463	-0.01389	-0.01586	0.00248	0.03498	-0.00771

For clustering the speakers, the eigenvalues (weights) of the speaker is found using equation 3.4 and plotted to the k-space of the eigenspace. The speakers can then be separated to groups manually or automatically using clustering approach.

$$w = E^T x V \tag{3.4}$$

For pronunciation adaptation, some adaptation speech with transcription from the test speaker is required. This can be done in some speech recognition applications, where the speaker can be asked to read some sentences. If this is not possible, the initial decoding of the non-native speech from the speech recognition system can probably be used as the transcription for the speech, although the accuracy will be lower with this unsupervised approach[2]. The idea is to use some non-native speech to get the 'partial picture' of the pronunciation habits of the speaker and then project it to the eigenspace to estimate the 'complete picture'. The speech is forced aligned using the standard dictionary to get the reference phoneme string. It is also forced aligned using another dictionary which contains all the variants from the supervectors to obtain the hypothesis phonemes for the pronunciations. The corresponding hypotheses and reference phoneme strings are compared to create a confusion matrix. A supervector is constructed for each test speaker by finding its triphone and monophone confusion matrix, and subsequently interpolating them, and filling the supervector. The weights of the test speaker are first calculated using equation 3.4, and subsequently the weights are used to reconstruct the supervector by using the eigenvectors:

$$V' = E \, x \, w \tag{3.5}$$

Recall that each vector is in fact the pronunciation confusion of each speaker. Consequently, a threshold is set to extract the speaker specific pronunciation variants from the vector *V'*. The variants are subsequently added into the pronunciation dictionary by generating the possible combinations. The new dictionary is then ready to be employed on the utterances of the particular speaker.

3.3 Accent Identification

The accent of the speaker is a factor that affects greatly the performance of speech recognition systems. By knowing the accent information, suitable models that match the speaker can be selected for speech recognition tasks. Although accent information can be manually given by the speaker, automatic accent identification could be useful in situation when this is not possible, or when the purpose is to provide user with a friendlier system. Thus, accent identification is sometimes an important component in speech technology. Besides that, another area where accent identification has the potential to be applied is in global security, where one tries to identify the origin of a non-native speaker.

The type of accent can be classified as either dialectal or non-native. Although dialectal and non-native speeches are variants of the 'standard' spoken language, both are quite different. The most obvious difference between dialectal and non-native speech is the fact that dialect is often acquired as first language, while non-native speech is considered as the second language of the

[2] However, we did not test this unsupervised scenario, and in our experiments, pronunciation adaptation is made using manually transcribed adaptation data for each speaker.

speaker. Thus, the differences between a language and its dialects are variant rules in phonology, pronunciation, vocabulary and possible grammars are learned since infancy, while for non-native speakers, the accent is caused particularly by interference from the native language rules of the speaker. Hence, there is an acceptable norm among the dialect speakers, while there is no acceptable standard among non-native speakers, where different degrees of variability exist. As a result, dialect and foreign accent may require different strategies for identification. Non-native speech may also be harder to collect compared to dialectal speech. As a result, dialect identification may use methods that are more data intensive for creating the accent models, while non-native accent identification will not have such a privilege. Accent identification approaches can be divided according to the features used for classifying the accent: acoustic or phonotactic features. Acoustic features that have been studied in accent identification are pitch, energy, formants, MFCC and others, while for phonotactic features phoneme sequence and position are important. We are particularly interested in approaches that are capable to generalize the accent features even though data from only a few speakers is given, because in many cases, only speech from a few speakers is available. In this preliminary work, we propose an approach using phonotactic features. Multilingual decision trees are used to model the phonotactic features. For the moment, the approach is text dependent which requires the transcription of the input speech. Although this strong hypothesis, such method can be used in situation where speakers can be asked to read a particular sentence, or when the speech of the user can be predicted accurately.

3.3.1 Multilingual Decision Tree for Accent Identification

Works in language identification have received much more attention than accent identification. One of the propositions using parallel multilingual phoneme recognizers for language identification task has shown promising results [Zissman 1996a, Schultz 2002]. The idea is to use multiple phoneme recognizers (PRs) of different languages to generate phoneme strings and subsequently score them using the corresponding language model. Since phoneme recognizers are not perfect and the type of errors made by each of them maybe different, the use of multiple phoneme recognizers for generating several phoneme sequences can enhance the performance of a single language system, although it is more computational intensive.

The same idea can be applied for non-native speech recognition with some modifications, so that the accented models can be trained for recognizing the accent, even with small amount of non-native speech. It has a similar architecture compared to parallel multilingual phoneme recognizers for language identification, the difference is that it also uses speech recognition system in the known target language (for instance French) to force align the same utterance at the same time. Another difference is that the accent models are made up of decision trees of different languages instead of language models. The reason for using accent models created from decision trees is to take advantage of the context of the phoneme and the generalizability of decision trees for classifying the accent. Furthermore, decision trees have been proven to be useful in state-tying and pronunciation modeling to classify similar context together, and to predict unobserved or missing data. However, this means that we need to have the transcription of the utterance during accent identification. Figure 3.5 shows our proposed accent identification system using

multilingual decision trees. Before accent identification can be carried out, the accent models have to be trained. In the following section, the training procedure is first presented, followed by the identification step.

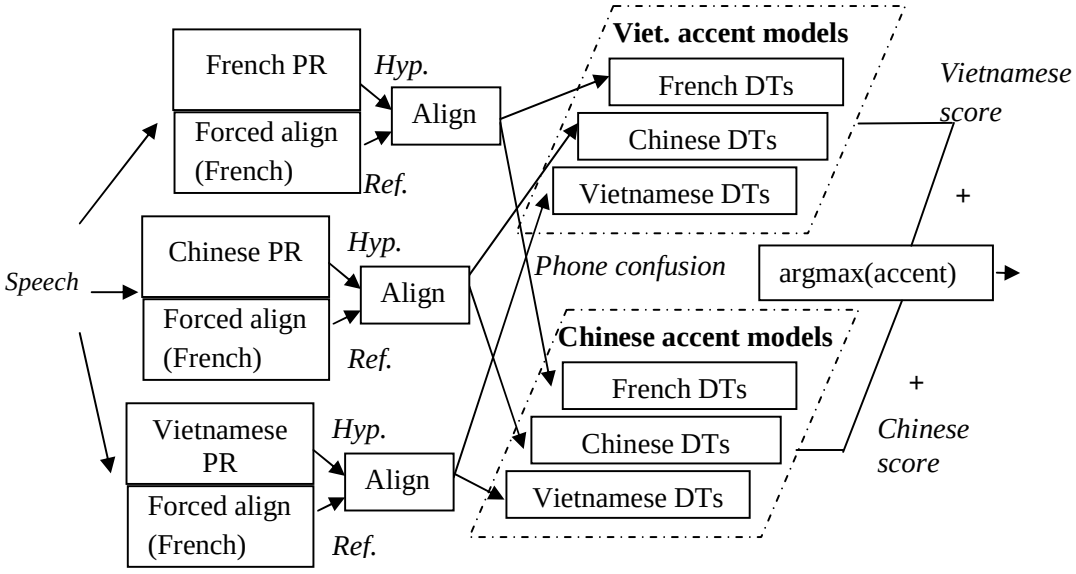


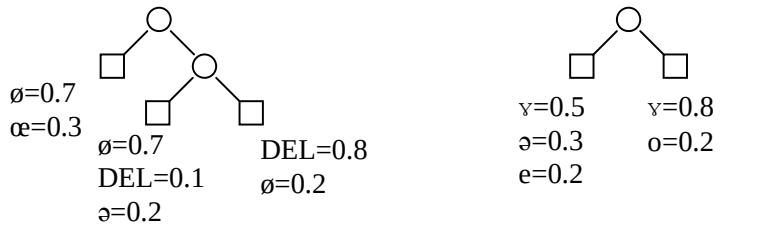
Figure 3.5 The usage of multilingual decision trees (DTs) for non-native French accent (Chinese and Vietnamese) identification

3.3.1.1 Training

In the training step, accent models which are made up of multilingual decision trees will be built. The steps to create the decision trees are similar to the one presented for pronunciation modeling in Section 3.2. Instead of going through two passes to create the decision trees, here only one pass is applied because all phone confusions whether it is significant or not should be taken into consideration for building the accent models. The hypothesis phoneme strings from the phoneme recognizer are aligned against the corresponding reference phoneme time stamps from the forced alignment to create the triphone confusions. The triphone confusions with the same base phoneme are gathered to build the base phoneme decision tree, and their contexts are converted to the corresponding phonetic feature vectors. They are subsequently passed to the decision tree algorithm to create the decision trees.

The procedure is repeated by decoding the same non-native speech with phoneme recognizer of different languages, and aligning the hypotheses against the corresponding reference phoneme strings of the target language. As a result, for each accent to be identified, *n* set of decision trees will be created, where *n* is the number of languages available for the

phoneme recognizers. Each set contains *x* phonemes of the target language. The hypotheses generated from each phoneme recognizer are aligned against the corresponding reference phoneme strings from the forced alignment producing triphone confusions. Recall from Section 3.2.1 that the triphone confusion of the same base (target language) reference phoneme will be collected together for each language. The left and right context phoneme will then be converted to feature vectors, for example by using IPA articulation features. Decision tree algorithm can finally be applied to build the trees. Since the trees are created by using the hypotheses generated by phoneme recognizers of different languages and aligned against the corresponding target language references, they are actually decision trees of the target language phoneme set. An interesting remark is that these trees have the leaves of different languages depending on the phoneme recognizer used. Thus, the language of the decision trees actually refers to the language of the phoneme recognizer used. Figure 3.6 shows two decision trees for the French phoneme /ø/, which is created by aligning the hypotheses from French and Mandarin phoneme recognizers respectively against the French reference phoneme strings. In this case, the triphone confusions of the base phoneme /ø/ with French and Mandarin phonemes are each gathered to create the decision trees below.



Phoneme Tree /ø/ — *French Decision Tree*

Phoneme Tree /ø/ — *Mandarin Decision Tree*

Figure 3.6 Example of accent models for Vietnamese in the form of decision trees created using French and Mandarin phoneme recognizers

3.3.1.2 Identification

During identification, the utterance from a non-native speaker will be sent to the phoneme recognizers of different languages and at the same time forced aligned using the transcription of the utterance, supposed known in advance, see Figure 3.5. Only one speech recognition system of the target language is actually needed for force alignment. The triphone confusions generated from the alignment of hypotheses and references are scored using the accent models. The phone confusion probability is retrieved from the decision trees based on the reference phoneme context and substitution information. A small floor probability is assigned to avoid getting a zero confusion probability. The accent score for a particular language is calculated by taking into

consideration all the phone confusion given that decision trees. This is done by multiplying all the phone confusion probabilities given the decision trees (equation 3.6). The total accent score (for the decision trees for L language) is calculated by multiplying the accent score from L language trees (equation 3.7). The accent model with the highest score will be selected.

$$accent\ \ score_l = \prod_{i=1}^{n} P'(sub_j \mid base_i, left_i, right_i) + \beta \tag{3.6}$$

$$total\ accent\ \ score = \prod_{l=1}^{L} accent\ \ score_l \tag{3.7}$$

where l is the language of the decision tree, *sub* is the hypothesis substitution for the *base* phoneme, *left* and *right* indicates the left and right context of the base phoneme, and *ß* is a small floor probability.

3.4 Conclusions

In this chapter, we have presented two modified pronunciation modeling approaches using a limited amount of non-native speech, and an original approach called "latent pronunciation analysis" that can be used for pronunciation clustering and adaptation. The first approach is the conventional approach of pronunciation modeling, which models the variants in the pronunciation dictionary. Variants are estimated by using decision trees. Two passes are applied for finding the variants because the phoneme recognizer is not accurate and there are only limited amount of non-native speech available for modeling. Hence, it is important to generate the hypotheses as accurate as possible for estimating the unobserved variants using decision trees. The second approach models the pronunciation variants in the rescoring module. The rescoring module employs a triphone model to rescore the n-best list produced by the decoder. A unigram model is used for smoothing the triphone model. The pronunciation score obtained is included in the log-linear model to compute a composite score and re-rank the n-best hypotheses. The third approach called *latent pronunciation analysis* is in fact a new pronunciation clustering method, which clusters the speakers according to their pronunciation habits. This approach is motivated by eigenfaces and eigenvoices, where it uses eigenvectors derived from speaker dependent decision trees. Thus, *latent pronunciation analysis* attempts to use the pronunciation eigenvectors derived for pronunciation adaptation. However, this method requires more non-native data to carry out, but the benefit is that the knowledge about the accent of the speaker is not required to be known a-priori, unlike the typical non-native pronunciation modeling approaches.

Preliminary work in accent identification has also been proposed. A new accent identification approach using multilingual decision trees has been presented. This is a text dependent accent identification approach which requires the transcription of the test utterance. The approach can be used in situation when speakers can be asked to read a particular sentence, or when the utterance of the user can be predicted accurately. The advantage of the approach is

that the accent models require a little amount of non-native speech to create. The benefits come from the generalizability of the decision trees to model phonotactic features. The identification capability is further improved with the usage of parallel phoneme recognizers to create decision trees of different languages for each accent model. Multilingual phoneme recognizers have proven to be beneficial in language identification, where they are able to improve the language identification rate. This approach will be experimented in Chapter 5, and compared to the existing baseline accent identification approaches discussed in earlier.

CHAPTER 4

Non-native Corpus Acquisition and Evaluation

4.1 Introduction

Speech corpus is required for testing the approaches that have been proposed. For this purpose, a non-native French speech corpus in the tourism domain has been recorded. In the next chapter, the corpus will be employed for testing. This chapter presents the procedure used for recording the corpus, follows by some analyses and evaluation tests. Among the analyses that have been carried out are intelligible test, phonetic analysis and data-driven analysis.

4.2 Acquisition of a Non-Native French Corpus

A typical speech corpus for speech recognition development consists of training, testing and development parts. This is generally true for native speech corpora. However, for non-native automatic speech recognition, collecting sufficient samples of non-native speech for training non-native models is difficult. Furthermore, there are simply too many possible groups of non-native speakers that may involve. Thus, our non-native speech corpus is recorded only for testing and adaptation purposes.

This corpus has been developed for testing, adaptation and research in mind. For testing, we would like to test the non-native speakers in the tourism domain, which might be a realistic case, where non-native speakers are likely to stumble upon. Although this is a read non-native corpus, there is a dialog part where speakers are asked to read and simulate the sentences in real situation. The sentences are also designed to contain proper names of places, person names and others.

4.2.1 Text Corpus Acquisition

The corpus is divided into two parts. The testing part consists of common dialog and article sentences from the tourism domain. The adaptation text comprises of sentences from the ESTER corpus [Gravier 2004].

4.2.1.1 Read Sentences that Simulate Dialog

For the first part, the common dialog phrases in the tourism domain were selected (for example from dialogs in hotel, restaurant, transport and other related areas). They were collected from web resources, travel books and elementary French language books. After the sentences were collected, we extracted the vocabularies out and used a script to generate their pronunciations. In the first step, the script simply searched for words that were defined in the existing pronunciation dictionaries. If the words were found, they were copied from the pronunciation dictionary and added to the new pronunciation dictionary. For words that were not found, the LIA grapheme to phoneme application [Béchet 2005] was used to generate the possible pronunciations for each word. After the pronunciation dictionary was generated, sentences were selected to be read by speakers from the text pool. The sentences were selected such that those with the most number of unique unseen triphones were selected, so that we can evaluate non-native speaker in as many context as possible. For each speaker in the same group, a hundred unique sentences were selected.

4.2.1.2 Read Articles

The texts in this second part are also from tourism domain, but instead of dialog, they are sentences from tourism articles on the web. The texts were first gathered from tourism websites using a web crawler. Subsequently, the texts were extracted from the HTML files. Next, the sentences were filtered and normalized by removing the punctuations, changing the digits to text numbers, lowering the case of the text, changing paragraph to sentences, limiting the size of sentences etc. After having manually verified that the sentences were suitable, the same approach described in previous section was applied to select sentences to be uttered by speakers. The total number of unseen triphones found over time is showed in Figure 4.1. The graph shows that the number of unique triphones found drops dramatically for the first hundred sentences. This shows that frequent triphones are repeatedly found, which is something desirable, because they should

be tested more frequently compared to rare triphones. A hundred unique sentences were assigned to every speaker in the same group.

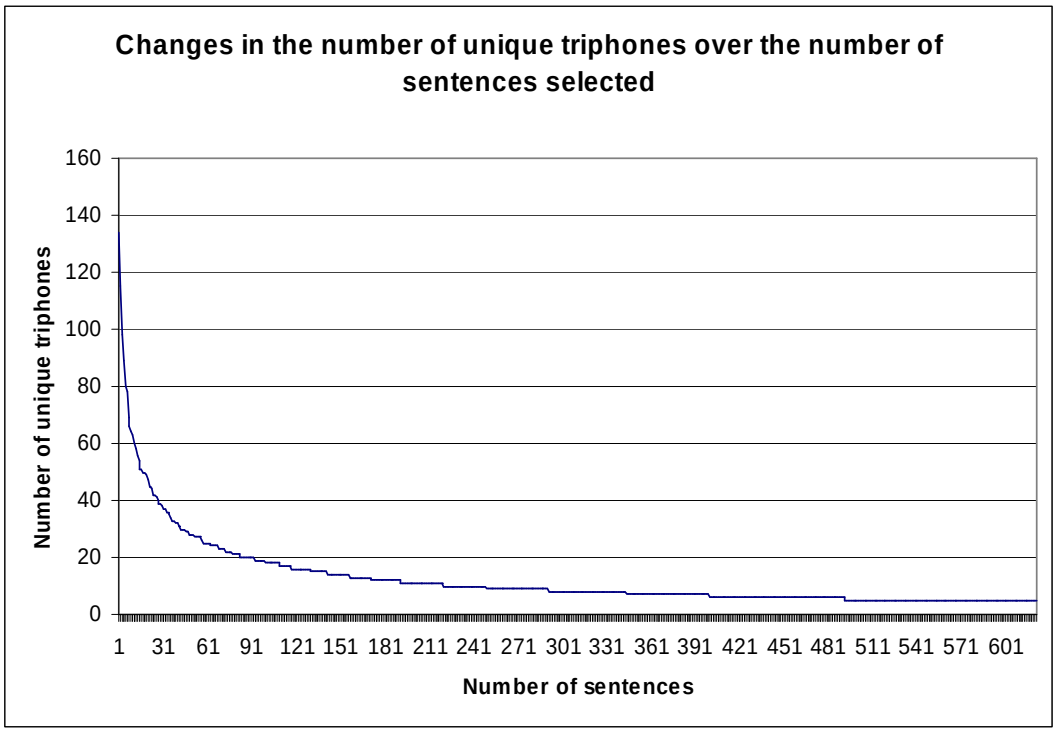


Figure 4.1 Changes in the number of unique triphones found in the sentences over the number of sentences selected from text corpus

4.2.1.3 Adaptation Text

Adaptation sentences were selected from ESTER corpus. The ESTER corpus is a broadcast new corpus. It was recorded and transcribed as part of campaign for the evaluation of Broadcast News enriched transcription systems using French data. Text from ESTER corpus was used since it contains about sixty hours of transcribed speech to choose from. At the same time, we can also take advantage of other resources that are readily available together with the corpus such as the pronunciation dictionary. It is also possible to compare the adaptation speech and the speech recorded in the ESTER corpus if necessary. The same procedure discussed before was used for collecting the adaptation text. A hundred sentences are selected for each speaker. All speakers were assigned sixty sentences with the most number of unique triphones. The other forty sentences selected for each speaker were unique. The purpose is to adapt as many triphones as possible, while making sure those frequent triphones are adapted with more data.

4.2.2 Text Corpus Evaluation

We calculated the correlation coefficient of our corpus compared to the general phoneme distribution in French according to [Vaufreydaz 2000], to have an idea of the phoneme distribution in our corpus by using equation 4.1. The result in Table 4.1 shows that our corpus has a correlation coefficient of about 0.9 for all its three parts, which means that it is phonetically well balanced. Note that the phoneme distribution gives only a general idea of the speech corpus, because we only select one possible pronunciation for each word. In addition, we assume 'liaison' occurred. This is why there is a high percentage of /z/. See Figure 4.2.

$$Corr(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma x \sigma y} \tag{4.1}$$

where $\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $\sigma x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Table 4.1 Correlation coefficients for dialog, read article and adaptation parts of our corpus

Type	Correlation Coefficient
Dialog	0.910
Article	0.893
Adapt	0.920

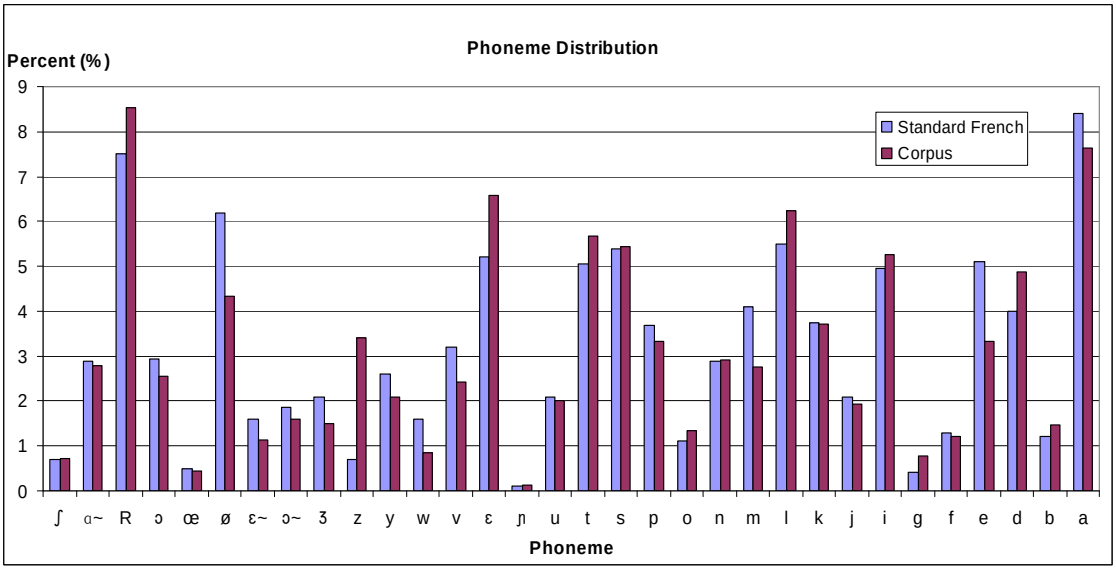


Figure 4.2 Phoneme distribution of Standard French (numbers taken from [Vaufreydaz 2000]) compared to our non-native corpus

4.2.3 Recording

A total of seven native Chinese speakers and eight native Vietnamese speakers with a comfortable degree of experience in the target language (French) were recruited. They consist of seven males and eight females (see Table 4.2). Chinese speakers who took part in the recording have previously taken 500 hours of French course in China before they came to France, and they were attending French courses at a local language school, at the time of the recording. All of them have been in France for less than a year. Most of the speakers are from Beijing. The Vietnamese speakers are students from local universities. Most of them are from Hanoi. All of them have been in France for more than a year and have learned French for more than three years. For baseline comparison, three native French speakers were also selected for recording the same test part.

Table 4.2 Number of native and non-native French speakers involved in test and adaptation

Speakers	French		Vietnamese		Chinese	
	Test	Adaptation	Test	Adaptation	Test	Adaptation
Male	1	0	3	2	2	0
Female	2	0	2	1	3	2
Total	3	0	5	3	5	2

Recording was done in a sound proof room, using a headset microphone, with sampling frequency set at 16 kHz. EMACOP (Multimedia Environment for Acquiring and Managing Speech Corpora) was used for recording and managing the speech corpus [Vaufreydaz 2000]. A supervisor was assigned to monitor and facilitate the recording. Table 4.3 shows the average duration of the utterances spoken. The results show that the average duration of the non-native utterances is longer compared to the utterances from native speakers. The Chinese speakers read the slowest. This might be due to the lower experience of Chinese speakers compared to Vietnamese speakers.

Table 4.3 Average duration of a sentence and total duration (in parenthesis) of sentences read by different native groups

	French	Vietnamese	Chinese
Read Dialog	2.84s (852s)	3.64s (1822s)	4.09s (2047s)
Read Article	6,27s (1843s)	10.2s (4694s)	11.72s (5740s)
Adaptation	-	12.54s (3687s)	17.9s (3509s)

4.3 Intelligibility Test

To investigate the intelligibility of the speech read by the non-native speakers, we invited native French speakers from CLIPS/IMAG laboratory for a perception test study. The test was conducted through Internet, and the volunteers were given eight recorded files to listen. Thirteen persons took part in the test. They were allowed to listen to the files for an unlimited number of times at their own place and pace. Subsequently, they were required to transcribe the utterances they heard on the specified textbox, and the system stored their answer in a database. The results from Table 4.4 shows that the pronunciations of non-native speakers are not clear even for the human native French speakers. Refer to appendix at Figure A1 and A2 for the web based interface used in the test.

Table 4.4 Average human WER from the Intelligibility test

Speakers	Vietnamese	Chinese
WER	12.1	11.3

4.4 Phonetic Analysis

Perception test was conducted for analysing in more detail the pronunciation of the non-native speakers. For the perception test, we had spent a month at *Aix-en-Provence* with Dr. Martine Faraco from *laboratoire parole et langage* at *Université de Provence* to analyze some of the non-native speech. To complement the perception test, we have also conducted some simple acoustic analysis on the non-native speech using Praat [Boersma 2007]. Table 4.5 below shows the summary of the analysis results. Only frequent occurred errors found for more than one non-native speaker are presented.

Table 4.5 Perception test and acoustic analysis results of non-native French speakers

French Phoneme	Perception Test	Acoustic Observation	Speakers
/b/	/p/	No voiced feature	Chinese
/d/	/t/	No voiced feature	Chinese
/g/	/k/	No voiced feature	Chinese
/ø/	/o/	-	Chinese
/g/	–	Have the features of a fricative, instead of a plosive. When /g/ is followed by /R/, /g/ seems to be deleted	Vietnamese
/ʃ/	/s/	More energy at higher frequency range instead of lower frequency	Vietnamese

		below 4000 Hz [Ladefoged 2000, Kent 2002]	
Final plosive in a syllable, e.g. /p/, /k/	deletion	No burst was found	Vietnamese
/ʒ/	/z/	More energy at higher frequency range instead of lower frequency below 4000Hz [Kent 2002]	Chinese and Vietnamese
/R/	too strong	-	Chinese and Vietnamese

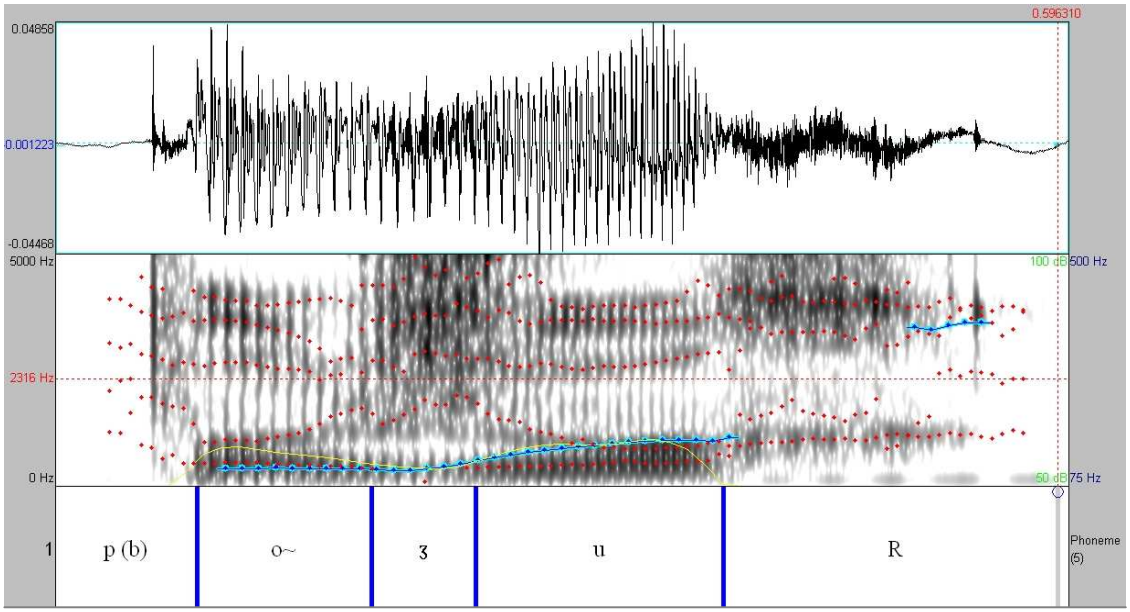


Figure 4.3 The word *bonjour* pronounced by a non-native French speaker of Chinese origin. Voiced feature are shown with blue line. Notice that there is no voiced feature on the first phoneme. This indicates that it is a /p/ instead of a /b/

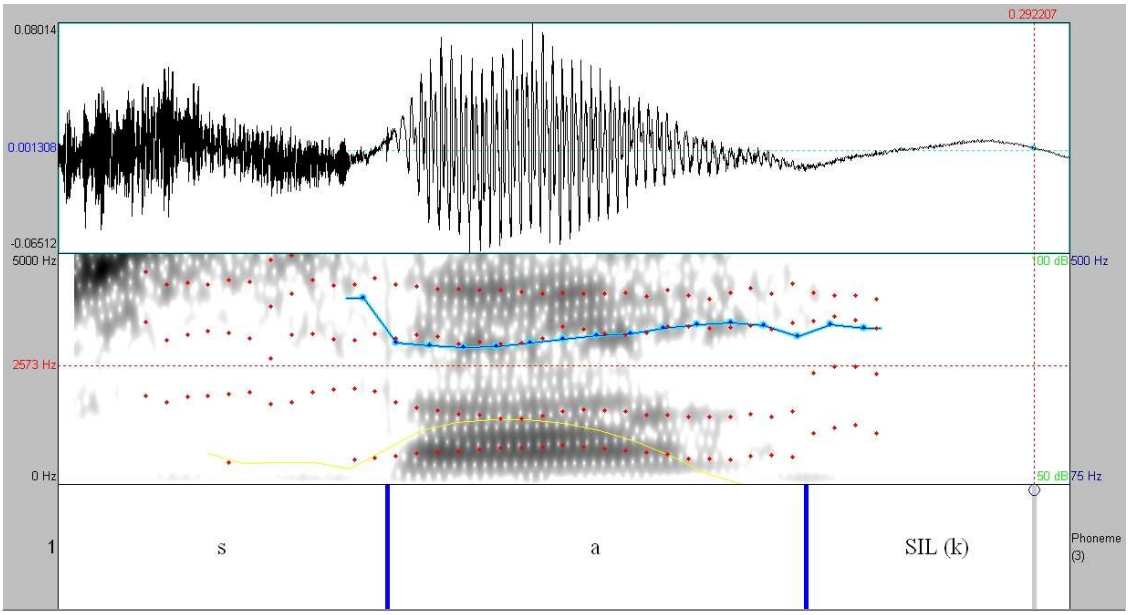


Figure 4.4 The word *sac* pronounced by a non-native French speaker of Vietnamese origin. Notice that the phoneme /k/ is not visible. It is either deleted or it is an unrelased /k/ common in Vietnamese

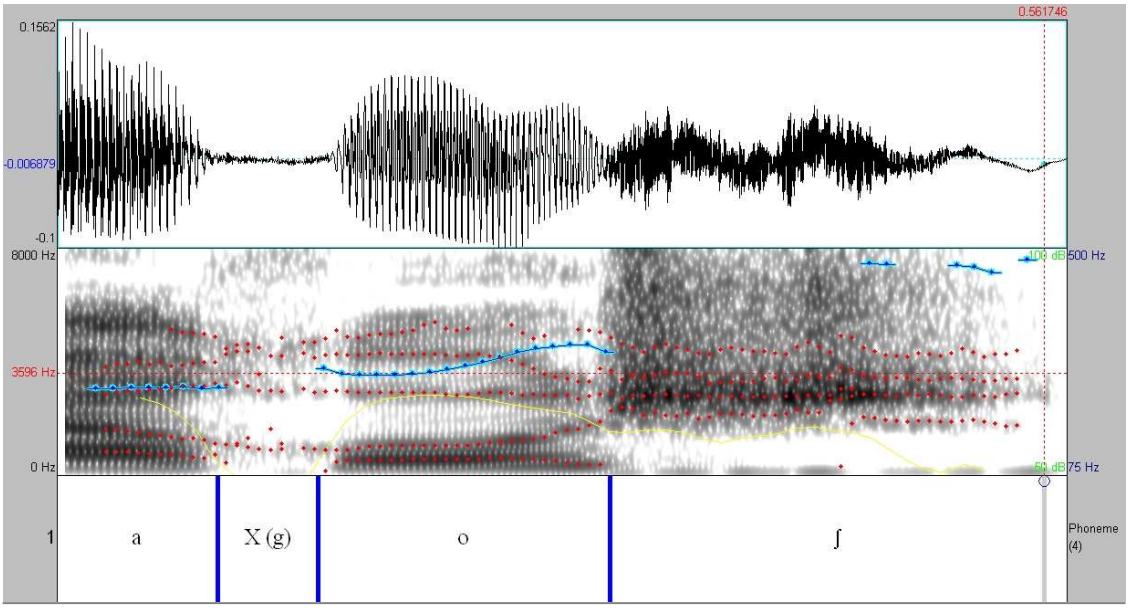


Figure 4.5 The non-native speaker of Vietnamese origin (female) read the words *à gauche*. Notice that instead of a /g/ which is a voiced plosive, the phoneme looks more like a fricative.

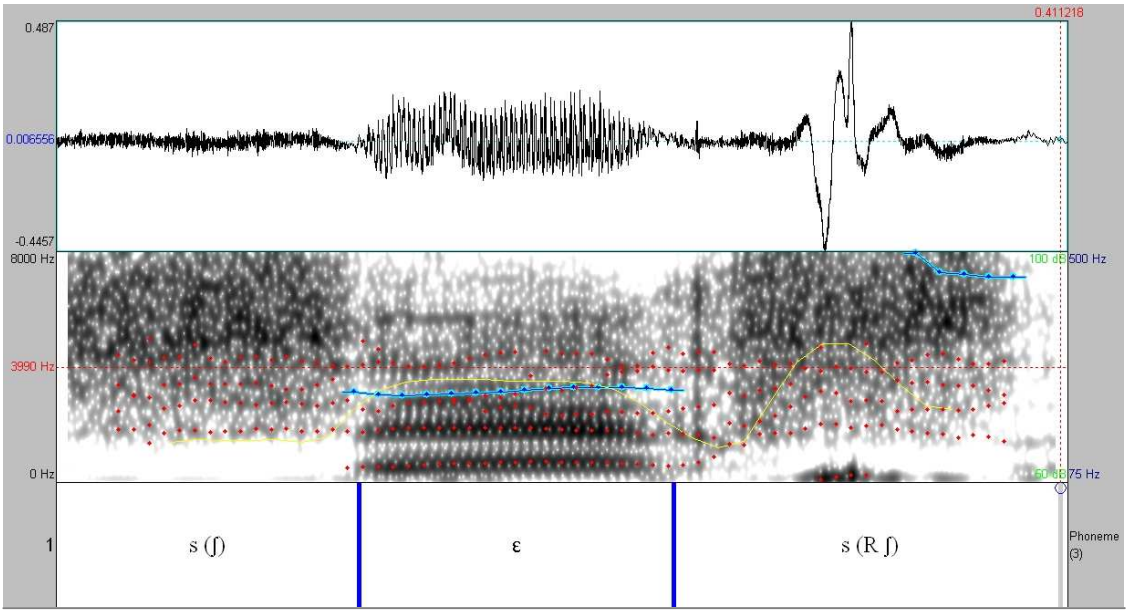


Figure 4.6 The word *cherche* pronounced by a non-native French speaker of Vietnamese origin (male). Notice that the speaker pronounced /s/ instead of /ʃ/ at the start and final phonemes. Compare it with the accurate /ʃ/ at Figure 4.5, which has more noise energy below 4000Hz.

The finding from the acoustic tests confirms the results from perception test. The results show that different non-native groups have a different tendency of substituting certain phoneme with another phoneme. The phenomena can be explained by cross-lingual phoneme transfer, where non-native speakers substitute the target language phonemes with their native language phonemes, and pronunciation simplification that was discussed in Section 1.3. Before going more detail into this, we will look at the finding from data-driven analyses on non-native speech.

4.5 Data-driven Evaluation with Phoneme Confusion Matrix

Evaluating the speech corpus through knowledge-based approach is time and resource consuming. Furthermore, it requires person with specialized linguistic knowledge in the field. Getting phoneticians to agree upon the same transcription is another difficulty. In addition, the phoneticians involved have to possess the knowledge of the languages involved. Some phonemes can also be difficult to analyse, for example the French /R/. It is also difficult to determine whether a particular phoneme realised is near the native form. Data-driven approach has the benefit to be fast, standardized and can be quite accurate if the models used are robustly built. It can therefore provide researchers with some insights into the data. For analysing the cross lingual transfer of non-native speakers, we employed the technique of phoneme confusion matrix as described in Section 2.3.1. This verifies further the phoneme confusion approach used in many of the next experiments.

The idea is to find the confusion or mismatch between the phonemes pronounced by the non-native speakers and the actual pronunciations of the words. Phoneme confusion matrix is created by aligning the hypotheses from a phoneme recognizer against the corresponding reference phoneme strings from the forced alignment of a speech recognition system (see Figure 4.7). For analysing the pronunciation habits of non-native speakers, we want to know the type of speech sounds that non-native speakers have mastered and the type of errors they are likely to commit. Since non-native speakers are influenced by their native language greatly when they learn a new language, the phoneme recognizer used must be able to recognize both the target and the native phoneme set of the speaker. Thus, the acoustic model of the phoneme recognizer has both the target and the native language acoustic units of the speaker.

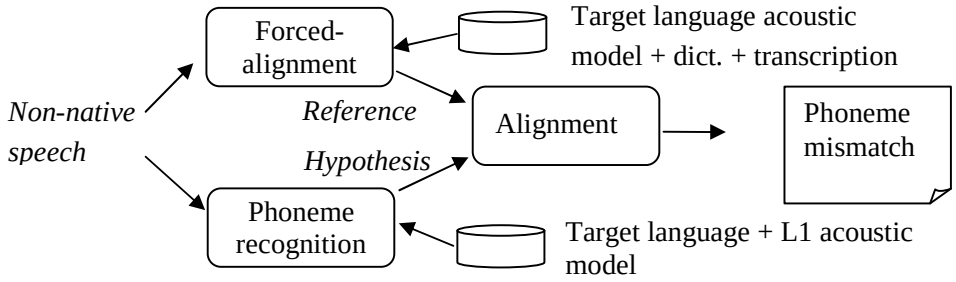


Figure 4.7 Finding the pronunciation habits of non-native speaker by using phoneme confusion matrix

For our data-driven analyses, the French acoustic model was the target acoustic model, while the source acoustic models were Vietnamese or Mandarin. The acoustic model used by the phoneme recognizer was created by merging French acoustic model with Vietnamese and Mandarin acoustic model to enable it to recognize both target and source phonemes. For more information about the system and corpora used, refer to Section 5.2 ahead. Table 4.6 and 4.7 below show the results.

For Vietnamese speakers, similar phonemes which exist in both the target and the native language of the speaker (according to IPA) were recognized as both the variants. They also have problem pronouncing /p/ even though it exists in both the languages. Probably this is because Vietnamese /p/ exists only as an unreleased plosive at the final position of a syllable. In French however, it can appear in the front of a syllable and it is a released form. There are few interesting sightings about new phonemes which exist only in the target language French (according to IPA) but not in Vietnamese. The French vowels /ø/, /œ/ and /ə/ were recognized with a high probability as the Vietnamese phoneme /ɤ/. The French /ɑ/ on the other hand was recognized as /a/. These are some new sighting from the data-driven analysis. Similar to our results from perception and acoustic analysis, /ʃ/ and /ʒ/ were replaced by phoneme /s/ and /z/. /R/ was recognized also as Vietnamese phoneme /X/. It is possible that Vietnamese speakers pronounce

/ɡ/ as /R/, since it has the same top two substitutions for /R/. There is also confusion of the French semivowel /ɥ/ with the vowel /y/. This is understandable because they are quite similar and both are articulated by rounding the lips, and /ɥ/ can be considered as the semivocalic counterpart of /y/.

Like Vietnamese speakers, similar French and Mandarin phonemes (according to IPA) were also recognized as both French and Chinese variants most of the time in the test. The result shows that the accuracy of recognizing /ə/ is low, even though the phoneme also exists in Mandarin. For new French plosives such as /b/, /d/ and /ɡ/, the results from data-driven test are comparable with the perception and acoustic analysis. However, the results from the analysis show that /z/ is pronounced as /s/, which is something we did not expect. It is possible that the /z/ pronounced by Chinese speakers is more similar to the model /s/, even though the voiced feature is articulated. Different from Vietnamese speakers, Chinese speakers have learned to pronounce the French post-alveolar fricatives /ʃ/ and /ʒ/ rather well, with a high accuracy rate. For new vowels only found in French, Chinese speakers have the tendency of substituting /ø/ and /œ/ with back vowels /o/ and /ɤ/ respectively. They are able to grasp nasal vowels rather well with a high accuracy, particularly the nasal vowels /ɛ~/ and /œ~/, although there seem to be some confusions between these two phonemes.

Table 4.6 Top two phoneme confusions for every French consonant uttered by Vietnamese and Chinese speakers.

French Consonants	Vietnamese speakers		Chinese speakers	
p	p (vn, 0.081)	t (vn, 0.075)	p (cn, 0.209)	p^h (cn, 0.12)
b	b (vn, 0.284)	b (fr, 0.148)	p (cn, 0.199)	b (fr, 0.118)
t	t (vn, 0.174)	t (fr, 0.121)	t (fr, 0.118)	t^h (cn, 0.094)
d	d (vn, 0.232)	d (fr, 0.198)	t (fr, 0.151)	d (fr, 0.123)
k	k (vn, 0.205)	k (fr, 0.158)	k (fr, 0.157)	k^h (cn, 0.155)
g	R (fr, 0.129)	X (vn, 0.097)	k (cn, 0.283)	k^h (cn, 0.13)
m	m (fr, 0.444)	m (vn, 0.209)	m (fr, 0.503)	m (cn, 0.129)
n	n (fr, 0.293)	n (vn, 0.125)	n (fr, 0.419)	n (cn, 0.074)
ɲ	ɲ (fr, 0.467)	ɲ (vn, 0.2)	ɲ (fr, 0.643)	n ɛ (cn, 0.071)
R	R (fr, 0.189)	X (vn, 0.048)	R (fr, 0.294)	x (cn, 0.25)
f	f (vn, 0.467)	f (fr, 0.383)	f (fr, 0.394)	f (cn, 0.129)
v	v (fr, 0.335)	v (vn, 0.161)	v (fr, 0.175)	u (cn, 0.072)
s	s (fr, 0.377)	s (vn, 0.302)	s (fr, 0.605)	s (cn, 0.159)
z	z (fr, 0.258)	z (vn, 0.2)	z (fr, 0.331)	s (cn, 0.11)
ʃ	ʃ (fr, 0.494)	s (fr, 0.187)	ʃ (fr, 0.734)	ç (cn, 0.086)
ʒ	ʒ (fr, 0.259)	z (vn, 0.176)	ʒ (fr, 0.594)	s (cn, 0.058)

j	j (fr, 0.234)	ie (vn, 0.286)	j (fr, 0.203)	j (cn, 0.194)
ɥ	y (fr, 0.246)	i (fr, 0.154)	e (fr, 0.155)	y (cn, 0.138)
w	w (fr, 0.222)	a (vn, 0.089)	w (fr, 0.381)	*a* (fr, 0.074)
l	l (vn, 0.314)	l (fr, 0.109)	l (fr, 0.199)	l (cn, 0.124)

Within the parenthesis is the language information of the phoneme (French-fr, Vietnamese-vn and Chinese-cn) and its confusion probability (0-1). Note that the phoneme /ŋ/ is not included in the table because it does not occur sufficiently in the data to calculate a reliable confusion value.

Table 4.7 Top two phoneme confusions for every French vowel uttered by Vietnamese and Chinese speakers.

French Vowels	Vietnamese speakers		Chinese speakers	
i	i (fr, 0.309)	i (vn, 0.272)	i (fr, 0.24)	e (fr, 0.229)
y	y (fr, 0.267)	i (vn, 0.131)	y (fr, 0.34)	e (fr, 0.202)
u	u (fr, 0.293)	u (fr, 0.168)	u (fr, 0.265)	o (fr, 0.177)
e	e (vn, 0.29)	e (fr, 0.204)	e (fr, 0.335)	ɛ (fr, 0.196)
ø	ɤ (vn, 0.288)	ø (fr, 0.096)	ø (fr, 0.189)	o (fr, 0.108)
o	o (fr, 0.332)	o (vn, 0.205)	o (fr, 0.234)	*a*~ (fr, 0.173)
ə	ɤ (vn, 0.162)	ɯ (vn, 0.065)	ɤ (cn. 0.084)	œ (fr, 0.053)
ɛ	ɛ (vn, 0.262)	ɛ (fr, 0.257)	ɛ (fr, 0.323)	a (fr, 0.08)
œ	ɤ (vn, 0.332)	œ (fr, 0.153)	œ (fr, 0.242)	ɤ (cn, 0.194)
ɔ	ɔ (vn, 0.201)	*a*~ (fr, 0.097)	ɔ (fr, 0.172)	*a*~ (fr, 0.172)
a	a (vn, 0.322)	a (fr, 0.178)	a (fr, 0.233)	a (cn, 0.132)
a	a (vn, 0.328)	ɛ~ (fn, 0.131)	*a* (fr, 0.424)	a (cn, 0.153)
ɛ~	ɛ~ (fr, 0.128)	ɛˇ (vn, 0.128)	ɛ~ (fr, 0.272)	œ~ (fr, 0.141)
œ~	œ~ (fr, 0.198)	ɔˇ (vn, 0.121)	œ~ (fr, 0.333)	ɛ~ (fr, 0.144)
ɔ~	*a*~ (fr, 0.276)	o (vn, 0.088)	ɔ~ (fr, 0.333)	*a*~ (fr, 0.188)
a~	*a*~ (fr, 0.259)	ɔ (vn, 0.13)	*a*~ (fr, 0.361)	ŋ (cn, 0.102)

4.6 Cross-lingual Transfer in Non-native Speakers

To understand why certain target language phoneme is replaced by another phoneme by a particular group of non-native speakers, it is important to understand cross-lingual transfer of non-native speakers. This can be done by comparing the target and native phoneme set of non-native speakers. Table 4.8 and 4.9 show the consonant and vowel tables respectively for Vietnamese, while Table 4.10 and 4.11 present the Mandarin consonant and vowel tables. Note that the IPA tables are not the standard one, the approximants and affricates have been combined to give a compact presentation.

Here, we attempt to generalize the finding from the perception, acoustic and also data-driven analysis, so that it can be applied for other cases. Notice that similar phonemes are often replaced by the same phoneme in the target language or the native language of the speaker. The more interesting observations are the new phonemes which do not exist in the native language of the speaker. The new phonemes are often replaced by the nearest native phonemes according to the IPA table. For new consonants which do not exist in the native language of the speaker, the nearest native phoneme can often be found in the same row (manner of articulation) or in the same column (place of articulation). Notice an interesting fact that the place of articulation in IPA table is in fact arranged in order from the lips to the glottal (refer to the vocal tract Figure A3 in the appendix). For example, Chinese speakers will replace /b/, /d/ and /g/ with /p/, /t/ and /k/, while Vietnamese speakers will substitute /ʃ/ and /ʒ/, with /s/ and /z/. However, this is not always true as we see for the phoneme /g/. It is substituted by /X/ instead of /k/ or /ɣ/, which is nearer to /g/ according to the IPA table.

As for vowels, it is harder to explain using the vowel table. Instead, it seems to be more obvious from the vowel formant graph derived from the speech. Figure 4.8 and 4.9 show the original vowel formant graph plotted using the raw formant values from Vietnamese and Mandarin compared to French. The formant values were extracted automatically using Praat [Boersma 2007]. The phoneme alignment is obtained from the forced alignment of the speech. For understanding why certain new vowels in the target language are substituted by native language vowels, we assume that similar vowels in the target language will be substituted by similar vowels in the source language. We can project these source language vowels to the corresponding target language vowels. The other source vowels are projected by taking into consideration the projection of all other similar vowels, using the equation 2.1 and 2.2. For a new target language vowel, the nearest native vowel from the speaker is chosen. For example in our case, after projecting all the Vietnamese vowels, we found that the nearest source language vowel for /ə/, /œ/, and /ø/ is /ɤ/, because the Vietnamese /ɤ/ will be projected to somewhere between /ə/ and /œ/, which corresponds to our results obtained from phoneme confusion matrix in previous section. This can explain most of the substitutions observed.

French and Vietnamese share a lot of similarities in term of consonants and vowels. There are 23 similar phonemes between them, most of them exist as consonants. Vietnamese has relatively more fricatives than French. In term of vowel, there are short vowels and dipthongs in Vietnamese, but not in French. The diphthongs in Vietnamese are /ie/, /ɯɤ/and /uo/. On the other hand, French has more types of vowels and also nasal vowels.

Table 4.8 Comparison of French and Vietnamese consonants

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Glotal
Plosive	p b			t tʰ tˋ d		ʈ	c	k g		
Nasal	m			n			ɲ	ŋ		
Trill									R	
Tap or Flap										
Fricative		f v		s z	ʃ ʒ	ʂ ʐ		ɣ	χ	h
Affricate										
Lateral fricative										
Approximant							j ɥ*	w*		
Lateral approximant				l						

Consonants in black are common in both languages. Consonants in green and dotted square are found only in French, while consonants in blue and circled are available only in Vietnamese [Le 2006]. * - rounded

Table 4.9 Comparison of French and Vietnamese vowels

	Front	Central	Back
Close	i y		ɯ u
Close-mid	e ø		ɤ ɤ˜ o
		ə	
Open-mid	ɛ ɛ~ ɛ˜ œ œ~		ɔ ɔ~ ɔ˜
Open	a a˜		ɑ ɑ~

Vowels in black are common in both languages. Vowels in green and dotted square are found only in French, while vowels in blue and circled are available only in Vietnamese

Mandarin do not shares a lot of similarity with French in terms of consonants and vowels. Only eighteen 'similar' phonemes are shared between them, and seven of them are vowels. In Mandarin there is no voiced plosive and fricative compared to French. On the other hand, affricate does not exist in French but can be found in Mandarin.

Table 4.10 Comparison of French and Mandarin consonants

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Glotal
Plosive	p b pʰ			t d tʰ				k g kʰ		
Nasal	m			n			ɲ	ŋ		
Trill				r					R	
Tap or Flap										
Fricative		f v		s z	ʃ ʒ	ʂ	ç	x		
Affricate				ʦ ʦʰ		ʈʂ ʈʂʰ	ʨ ʨʰ			
Lateral fricative										
Approximant							j ɥ*	w*		
Lateral approximant				l						

Consonants in black are common in both languages. Consonants in green and dotted square are found only in French, while consonants in blue and circled are available only in Mandarin [Duanmu 2002]. * - rounded

Table 4.11 Comparison of French vowels and Mandarin vowels

	Front	Central	Back
Close	i y		u
Close-mid	e ø		ɤ o
		ə	
Open-mid	ɛ ɛ~ œ œ~		ɔ ɔ~
Open	a		ɑ ɑ~

Vowels in black are common in both languages. Vowels in green and dotted square are found only in French, while vowel in blue and circled are available only in Mandarin [Duanmu 2002]

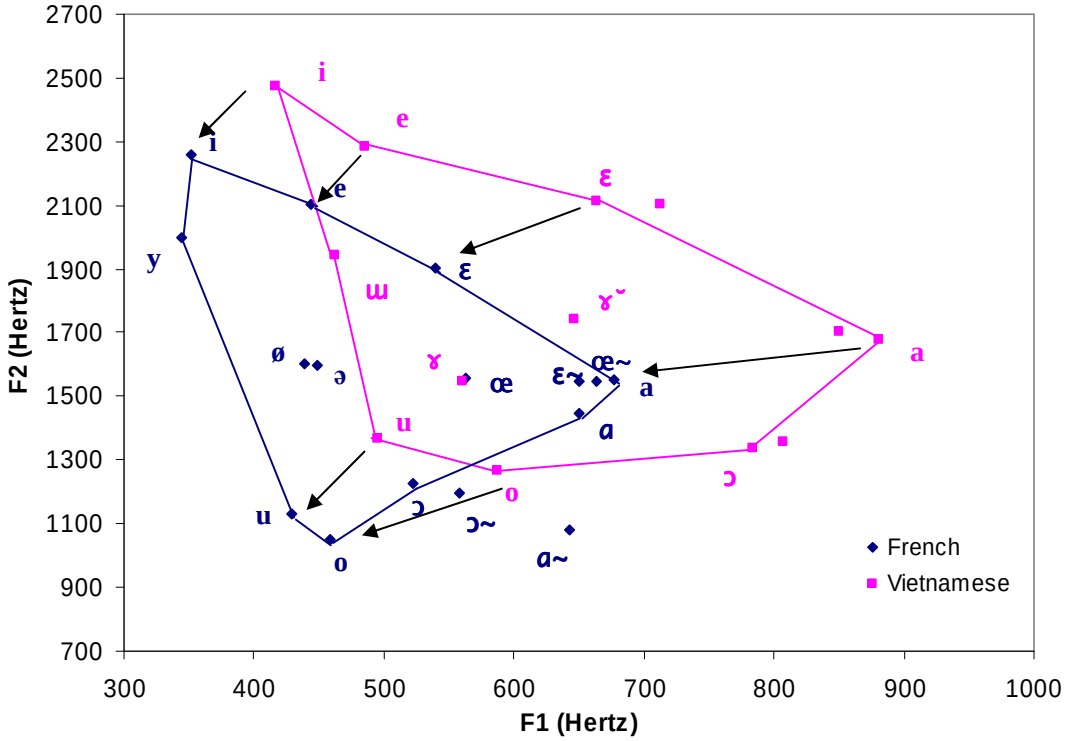


Figure 4.8 French and Vietnamese vowel charts

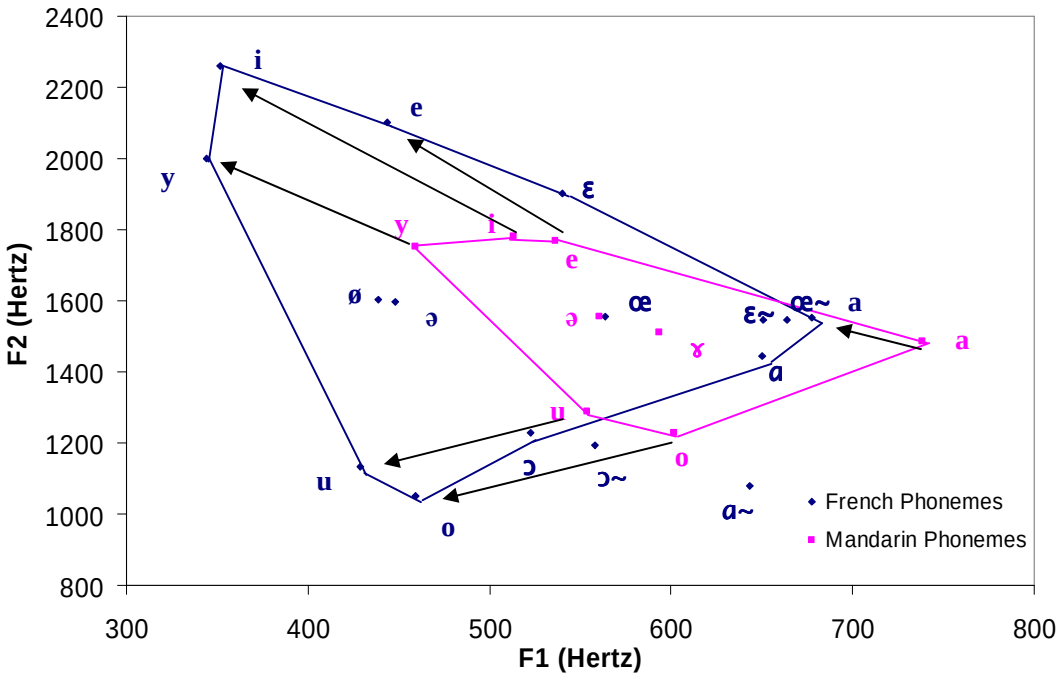


Figure 4.9 French and Mandarin vowel charts

4.7 Conclusions

In this chapter, we have presented our non-native French corpus in the tourism domain, which consists of read speech from speakers of Chinese and Vietnamese origin. In addition, few native French speakers have also taken part in the recording for comparison purpose. The corpus is used for testing and adaptation in the coming chapter. Four types of analysis, namely intelligible, perception, acoustic and data-driven tests have been carried out on the corpus for analyzing the speech of non-native speakers. The influence from the native language phonology of the speaker on the target language is very obvious from the non-native speech. Among the frequent observations are cross-lingual transfer and pronunciation simplification mentioned in Section 1.3. The finding shows that the non-native speech recorded has a high degree of accent, which may affect dramatically the speech recognition performance.

CHAPTER 5

Evaluations of Non-Native Modeling Approaches

5.1 Introduction

In the previous chapters, acoustic and pronunciation modeling for non-native speakers have been proposed together with accent identification approach. In this chapter, we will present the experiments that had been conducted to examine and verify the performance of the proposed techniques. In the coming section, a general description of the experimental setup will be given. Then in Section 5.3, we will examine the multilingual acoustic modeling approach proposed (in Chapter 2), follows by tests on pronunciation modeling in Section 5.4 (proposed in Chapter 3), and finally in Section 5.5, the proposed accent identification approach (in Chapter 3) is compared with some baseline accent identification approaches.

5.2 Experimental Setup

First, this section gives a general description of the speech recognition system employed for both speech recognition and accent identification tasks. Next, the corpora used for testing, training and adaptation through out this chapter are presented.

5.2.1 Automatic Speech Recognizer: Sphinx

Sphinx speech recognition system [CMU 2000][Ravishankar 2006] from Carnegie Mellon University (CMU) was selected for the speech recognition tasks. It consists of several speech applications. The two main applications are the trainer and the decoder. The trainer, known as SphinxTrain, can be used for training continuous HMM acoustic models. The original SphinxTrain application provides only context dependent modeling, and we have modified it for producing context independent model. The difference between context independent and context dependent models lies in the modeling of the context of the speech sounds. The iterative re-estimation procedure described in Chapter 1 is employed by using Baum Welch algorithm. In addition, for creating robust triphone context dependent models, states are being tied together (senones). On the other hand, Sphinx3 is the application for decoding speech. Sphinx3 is a fast decoder, capable of decoding speech at real time. This is achieved using conventional Viterbi search strategy and beam heuristics. In addition, it has a lexicon-tree search structure. Sphinx3 uses the acoustic model created by SphinxTrain. For language model, it accepts n-gram model in binary format, which is converted from a standard ARPA n-gram model. Speaker adaptation utilities such as MLLR and MAP are also part of the Sphinx speech recognition package.

In the following experiments, the front-end module was used to pre-process the raw speech at 16 bits sample with sampling frequency of 16 kHz to cepstral feature vectors together with its first and second derivative. This produces feature vectors with a total of 39 dimensions. SphinxTrain makes use of the feature vectors to create a continuous HMM acoustic model. Phoneme or phone were used as the unit of HMM, and each has three states, with a left-to-right topology. Conversely, the n-gram language model was created using CMU statistical language modeling toolkit [Clarkson 1997].

5.2.2 Speech Corpora

Almost all the experiments were carried out on non-native French speakers, and further tests were conducted on non-native English speakers to reinforce the results whenever possible. The following section presents the general description of the French and English corpora used for training and testing the non-native French and English speakers. For creating non-native acoustic models, multilingual corpora were also employed.

5.2.2.1 French

The non-native French corpus (NNF) described in the previous chapter was tested. It contains speakers of Chinese and Vietnamese origin, each non-native group is made up of 5 speakers who read about 100 common sentences related to dialogs in the tourism domain. The non-native speech recognition experiments made use of the data from BREF120 corpus [Lamel 1991] for creating the target French acoustic model (See Table 5.1). A general domain trigram language model was first created by using the texts from *Le Monde* newspaper. The text corpus contains about 2.8 Gigabyte of texts, from the year 1992 to 2003. The generic language model was then

interpolated with a tourism domain language model from NESPOLE project [Besacier 2001]. Conversely, for the French test pronunciation dictionary, it contains more than two thousand entries.

Table 5.1 Summary of the Corpora Used for Training and Testing for French

Task	Corpus	Description	Spk.	Hours
Training	BREF120	French - training	120	100+
Test	NNF	Non-native French	10	1

5.2.2.2 English

The tests on non-native English speakers were carried out on the ISLE corpus [Menzel 2000], which contains native German and Italian speakers. The English acoustic model was created from TIMIT corpus[3] [Fisher 1986], while the general domain language model and pronunciation dictionary were originated from Sphinx speech recognition system. However, the entries in the test pronunciation dictionary have been reduced to about one thousand for testing. Table 5.2 summarizes the corpora usage.

Table 5.2 Summary of the Corpora Used for Training and Testing for English

Task	Corpus	Description	Spk.	Hours
Training	TIMIT	English - training	630	4
Test	ISLE	Non-native English	46	18

5.2.2.3 Multilingual

The multilingual corpora used in the experiments consist of multiple independent corpora from different sources, namely a Vietnamese (VN) corpus [Le 2004], a Mandarin CADCC corpus [CCC 2005], a small non-native English corpus with Chinese and Vietnamese speakers from a public archive (GMU) [Weinberger 2007], and a Malay speech corpus courtesy of Universiti Sains Malaysia. The general information of the corpora is presented in Table 5.3. These multilingual corpora were used in non-native acoustic modeling for adapting target acoustic model, while they were employed in accent identification tasks to create accent models.

[3] We are aware that this corpus maybe small to train an English acoustic model; but at the time of these experiments, we did not have the WSJ corpus yet; moreover, the experiences on non-native English must be seen only as a validation of what is done for non-native French speech recognition.

Table 5.3 Summary of the multilingual corpora used

Task	Corpus	Description	Spk.	Hours
Adaptation and/ or identification	VN	Vietnamese	29	15
	CADCC	Chinese	20	5
	GMU	Non-native English (Vietnamese and Chinese)	17	0.14
	MSC	Malay	18	5

5.3 Non-Native Multilingual Acoustic Modeling

This section describes the experimental performance of the multilingual acoustic modeling proposed for non-native speakers in Chapter 2. Two types of modeling are being examined here: cross-lingual transfer and context. For modeling cross-lingual transfer in these experiments, non-native speech of the target language was assumed to be unavailable. Three types of multilingual resources were examined for non-native adaptation, namely the native language of the speaker (L1), any non-native language spoken by the same native group (L2), and languages close to the native language of the non-native speaker (L3). These are the possible candidates for adapting the target language acoustic model. On the other hand, for context modeling, it is sufficient to have the target language acoustic models to carry out the adaptation.

5.3.1 Cross-Lingual Transfer Modeling

In Section 1.3.2, we mentioned that non-native speakers are influenced by their native language when they learn to speak a new language, where they often transfer their native language speech sounds to the corresponding target language speech sounds. This mismatch between non-native speech and the trained models causes major reduction in speech recognition performance.

We will start by describing the experiments conducted to determine the source phoneme transfer of non-native speakers without using any non-native speech. Next, two baseline approaches for modeling cross-lingual transfers were tested. They are acoustic model interpolation and merging. Subsequently, the proposed hybrid approach of acoustic interpolation and merging are examined for modeling cross-lingual transfer by using multilingual acoustic models. We will also verify our proposed interpolation approaches for modeling cross-lingual transfer by employing multilingual corpora. Two speaker adaptation approaches using interpolation were tested. They are weighted least square and eigenvoices.

The proposed approaches were experimented using our multilingual resources. Three types of multilingual resources will be examined. They are the native language of the speaker (L1), any non-native language spoken by the same native group (L2), and languages close to the native language of the non-native speaker (L3). For instance, if we consider French as the target language for speech recognition system, and if the task is to recognize non-native speech from Vietnamese speakers, the resources considered will be Vietnamese speech, any non-native speech uttered by Vietnamese, and a language close to Vietnamese respectively. For non-native acoustic

modeling, non-native French speakers from our NNF corpus were tested, and BREF120 was used for training the target French acoustic model. For modeling cross-lingual transfer by non-native speakers, the Vietnamese (VN) corpus, the Mandarin CADCC corpus and the TIMIT corpus were used. We have also made used of non-native English speech from a public archive (GMU). The GMU corpus was either used separately or together with TIMIT corpus. Since the non-native English speech itself is not sufficient for creating non-native models, it is used to adapt the TIMIT acoustic model to create non-native English models for Chinese and Vietnamese speakers. Table 5.4 describes the ways the multlingual corpora were employed for adapting Chinese and Vietnamese speakers.

Notice that, Mandarin and Vietnamese are assigned as "close" languages even though both are belonging to different language families. Mandarin belongs to Sino-Tibetan family in the Sinitic branch, while Vietnamese is classified in the branch of Mon-Khmer in Austroasiatic family [O'Grady 2000]. However, both are Asian tone languages, where Mandarin has four tones while Vietnamese has six. In addition, a recent study [Ou 2007] found that both group of language learners share similar usage of stress in English. This suggests that there are similar characteristics between the two languages which can be exploited. In the coming section, we will examine our proposed approaches for modeling different context variation and cross-lingual transfer. All experiments were carried out using context independent acoustic models with 16 Gaussians mixture except specifies otherwise.

Table 5.4 Description of multilingual corpora used for adapting French acoustic model (BREF120)

Speaker	Corpus	Description
Vietnamese	VN	L1
	CADCC	L3 (Mandarin)
	GMU	L2 (English by Vietnamese)
	TIMIT + GMU	L2 (English by Vietnamese)
Chinese	VN	L3 (Vietnamese)
	CADCC	L1
	GMU	L2 (English by Chinese)
	TIMIT + GMU	L2 (English by Chinese)

5.3.1.1 Cross-Lingual Phoneme Transfer

Two approaches to determine non-native cross-lingual phoneme transfer by non-native speakers from a source to a target language without using any non-native speech from the target language has been presented in Section 2.3. This can be achieved by using phoneme confusion matrix or by using existing linguistic knowledge and IPA.

Recall that phoneme confusion matrix can be created by aligning the hypotheses from a phoneme recognizer against the corresponding reference phoneme strings from force alignment. Since we assume that non-native speech is not available, the source language phoneme recognizer and the target language speech recognition system are employed for decoding the target language

speech. Hence, for determining the phoneme transfer for Vietnamese and Chinese speakers, the source language will be Vietnamese and Mandarin respectively, while the target language will be French. Creating the phoneme confusion matrix does not need a lot of target language speech. Consequently, only 20 utterances were selected randomly from each speaker (total 120 speakers) from the BREF120 corpus.

For determining the possible phoneme transfer from IPA table, the corresponding source phoneme for a particular target phoneme is determined by referring to the IPA and existing linguistic knowledge. Refer to Vietnamese and Chinese IPA at Table 4.8, 4.9, 4.10, 4.11 and vowel chart in Figure 4.8 and 4.9. For similar phonemes (according to IPA) which exist in the source and also in the target language, these phonemes are assumed to transfer from the source to the target. For example, in Vietnamese and French, /p/ and /b/ exist in both languages. Thus, we assume that Vietnamese /p/ and /b/ are transferred to French /p/ and /b/ respectively. For new target language phonemes which do not exist in target language, the nearest source phonemes in the IPA will be chosen by taking into consideration existing linguistic studies. For example for Vietnamese speakers, the possible source phoneme transfer for /ʃ/ for Vietnamese is /s/ and /ʂ/ since they are the nearest phonemes. In some cases, we also take into consideration linguistic knowledge, for example /ɛ~/ is near to /a/ in vowel chart for Vietnamese speakers, but this phoneme is a nasal version of /ɛ/. For French vowel /y/, the nearest Vietnamese vowels are /i/ and /ɯ/. On the other hand, linguistic studies also suggest that American speakers replace /y/ with /u/. Consequently, the same transfer may also happen to Vietnamese speakers. Thus, the possible Vietnamese vowels that may substitute /y/ are /i/, /ɯ/ and /u/.

The possible source vowel and consonant transfer for Vietnamese, Chinese and non-native English speakers (of Vietnamese and Chinese origin) are presented in Table 5.5 and 5.6 respectively. For both Vietnamese and Chinese speakers, the possible source phoneme transfer using confusion matrix and IPA are shown, except for English speakers (native Vietnamese and Chinese origin) where solely the IPA choices are used since both French and English shares a lot of similar phonemes. The general guidelines used here for selecting the final source phonemes are:

- Source phonemes which also exist in the target language are assumed to transfer to the corresponding target phonemes. e.g. Vietnamese /p/ to French /p/.
- Existing linguistic study that shows a particular source phoneme is transferred to another target phoneme is applied. e.g. Vietnamese /u/ to French /y/.
- Compare the IPA and confusion matrix choices and select the best option.

It is also interesting to compare our choice of source phonemes set with the results from our earlier corpus analysis results in Table 4.6 and 4.7.

Table 5.5 Determining the transfers of source consonants (Vietnamese, Mandarin and English) using confusion matrix and IPA to target language

French Consonants	Vietnamese Speakers			Chinese Speakers			English Speakers
	Conf. Mat.	IPA	**Selected**	Conf. Mat.	IPA	**Selected**	**Selected**
p	v, t	p	p	p, p^h	p, p^h	p	p
b	b, v	b	b	p, t	p, p^h	p	b
t	t^h, t	t, t^h	t	t, t^h	t, t^h	t	t
d	d, t	d	d	t, p	t, t^h	t	d
k	k, g	k	k	k, k^h	k, k^h	k	k
g	k, d	k, ɣ	ɣ	k, t	k, k^h	k	g
m	m, l	m	m	m, n	m	m	m
n	n, ɲ	n	n	n, m	n	n	n
ɲ	ɲ, ie	ɲ	ɲ	j, m	n, j	j	n
ŋ	ŋ, ɲ	ŋ	ŋ	ŋ, c	ŋ	ŋ	ŋ
R	X, ɔ	X	X	a, w	x	x	r
f	f, v	f	f	f, s	f	f	f
v	v, i	v	v	f, i	f	f	v
s	s, ʂ	s	s	s, ʂ	s	s	s
z	z, zr	z	z	s, j	s, ʂ	s	z
ʃ	ʂ, ʐ	s, ʂ	s	s, t	s, ʂ	ʂ	ʃ
ʒ	ʐ, ʂ	z, ʐ	z	j, s	s, ʂ	ʂ	ʒ
j	ie, s	j	j	j, s	j	j	j
ɥ	i, l	w, y	w	i, j	w, y	w	w
w	uo, w	w	w	w, o	w	w	w
l	l, m	l	l	l, m	l	l	l

Note: Confusion matrix result shows the top two phoneme confusions in descending order, and IPA shows the likely phoneme transfer

Table 5.6 Determining the transfers of source vowels (Vietnamese, Mandarin and English) using confusion matrix and IPA to target language

French Vowels	Vietnamese Speakers			Chinese Speakers			English Speakers
	Conf. Mat.	IPA	**Selected**	Conf. Mat.	IPA	**Selected**	**Selected**
i	i, c	i	i	i, j	i	i	i
y	l, c	i, ɯ, u	u	i, j	y	y	u
u	u, o	u	u	w, u	u	u	u
e	e, ie	e	e	i, j	e	e	e
ø	ɯ, ɯɤ	ɤ	ɤ	ɤ, i	ə, ɤ	ɤ	ə
o	o, u	o	o	w, u	o	o	o
ə	ɯ, l	ɤ	ɤ	ɤ, t	ə	ə	ə

ɛ	ɛ, e	ɛ	ɛ	ɤ, e	e	e	ɛ
œ	ɤ, ɯ	ɤ	ɤ	ɤ, a	ə, ɤ	ɤ	ə
ɔ	o, ɔ	ɔ	ɔ	u, ɤ	o	o	ɔ
a	ɤ̆, ɔ	a	a	a, ɤ	a	a	ɑ
ɑ	ɔ, a	a	a	a, ɤ	a	a	ɑ
ɛ~	a, ɤ̆	ɛ, a	a	a, ɤ	ɛ, a	a	ɑ
œ~	a, ɤ̆	a	a	a, ɤ	a	a	ɑ
ɔ~	o, w	ɔ	ɔ	u, w	o	o	ɔ
ɑ~	ɔ, o	ɔ, a	ɔ	u, a	ɔ, a	a	ɑ

Note: Confusion matrix result shows the top two phoneme confusions in descending order, and IPA shows the likely phoneme transfer

5.3.1.2 Baseline Non-native Acoustic Modeling

Baseline non-native acoustic modeling approaches, acoustic model interpolation and merging were tested by using the target-source phoneme matching that we found in Table 5.5 and 5.6. Acoustic interpolation and merging were performed by using French and the corresponding context independent L1 acoustic model with 16 Gaussian mixtures. For acoustic model merging, the merging variant in Figure 1.15b was applied, and for acoustic model interpolation, Euclidean distance was used for measuring the distance between the Gaussians. Figure 5.1 shows the word error rate (WER) of non-native French speakers of Chinese and Vietnamese origin by employing acoustic model interpolation and merging across varied weights.

Overall, the results show that acoustic model merging performs better than acoustic model interpolation, although it creates a model with twice the number of Gaussian mixtures compared to the acoustic model interpolation. Note that, when French weight is equal to 1.0, it is the baseline result. When French weight equals to 0.0, the French acoustic model is replaced by the corresponding phonemes from the L1 acoustic model of the speaker.

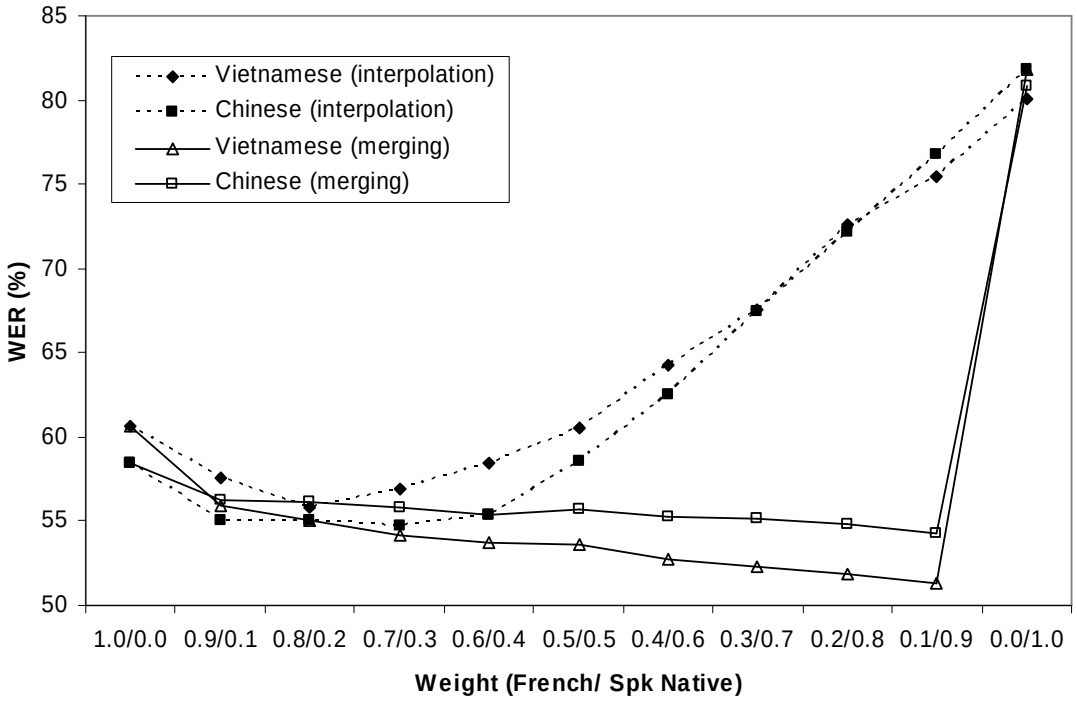


Figure 5.1 WER on non-native French speakers of Chinese and Vietnamese origin by interpolating and merging acoustic models, which are created from a 16 Gaussian CI target (French) and source (Chinese/ Vietnamese) acoustic models across different weights

5.3.1.3 Hybrid of Interpolation and Merging Approach

In this section, the hybrid of interpolation and merging approach proposed in Section 2.4.1 is applied for modeling cross-lingual transfer. Recall that the approach interpolates source and target Gaussian in the matching states that are near to each other. For Gaussians that are far from each other, the source or target Gaussians will be merged. Thus, before this can be carried out, the source-target phoneme transfer information found previously will first have to be used to map the states of the target and source acoustic model. The target language in this case is the newly acquired language of the non-native speakers, which is French. The possible source languages can be any of the three types of languages (L1, L2 and L3) proposed. For Vietnamese speakers, L1, L2 and L3 will be Vietnamese, non-native English by Vietnamese and Mandarin respectively. As for Chinese speakers, L1, L2 and L3 are Mandarin, non-native English by Chinese and Vietnamese respectively. The non-native English acoustic models for Vietnamese and Chinese speakers were created by adapting TIMIT acoustic model with non-native English speech from GMU corpus using MLLR algorithm with one regression class.

The results from non-native cross-lingual transfer modeling using L1, L2 and L3 are presented in Figures 5.2 and Figure 5.3. Euclidean distance was used as the distance measure. The initial context independent acoustic models for French, Vietnamese, Chinese and non-native English were employed for the experiments. The resulted models have an average of 26 Gaussians per state. A threshold was set at about two times the average Gaussian distance. In fact, currently Sphinx speech recognition system is not capable of handling varied number of Gaussians per state. To model this, we set all states to the maximum number of Gaussian mixtures possible. As a result, the means, variances and mixture weights for the empty Gaussians are set to zero.

The results of adapting the target acoustic model with L1 and L2 (English) acoustic model are very promising. On average, using L1 for adaptation with the hybrid approach performs better than acoustic model merging in Figure 5.1. In most cases, it scores an average relative WER improvement of 6.61% and 12.78% for Chinese and Vietnamese speakers respectively, while the acoustic merging approach has an average of 5.86% and 11.86% of improvement for Chinese and Vietnamese speakers respectively. Furthermore, the proposed approach uses smaller number of Gaussians. As for L2 adaptation, surprisingly for Chinese speakers, the results show that the non-native acoustic model created from only about 5 minutes of non-native English speech is equally effective compared to Mandarin acoustic model created from 5 hours of CADCC corpus to adapt the French acoustic model (by using the procedure mentioned). The improvement from L2 adaptation for Vietnamese is lower. This is understandable since there are only 7 speakers and only slightly more than 3 minutes of non-native English speech is available for adaptation. Another interesting result shown in the graph is that L3 can be useful for adaptation. By giving appropriate weight, Vietnamese acoustic model seems to be able to adapt French acoustic models for Chinese speakers and vice versa. A 3% reduction in WER for non-native speech recognition of native Chinese and Vietnamese speakers is recorded when French weight is equal to 0.8.

Overall, the results from the modeling are very promising, not only for L1 but also for L2 and L3 acoustic models. The results from our experiments suggest that L2 resource, even though

from a different target language may produce adaptation results which is equal or better than the L1 acoustic model, when there are sufficient amount of data available. However, more tests should be carried out to verify the same happens using other adaptation algorithms and languages. As for L3, using a low weight for modeling has shown to be beneficial for adaptation.

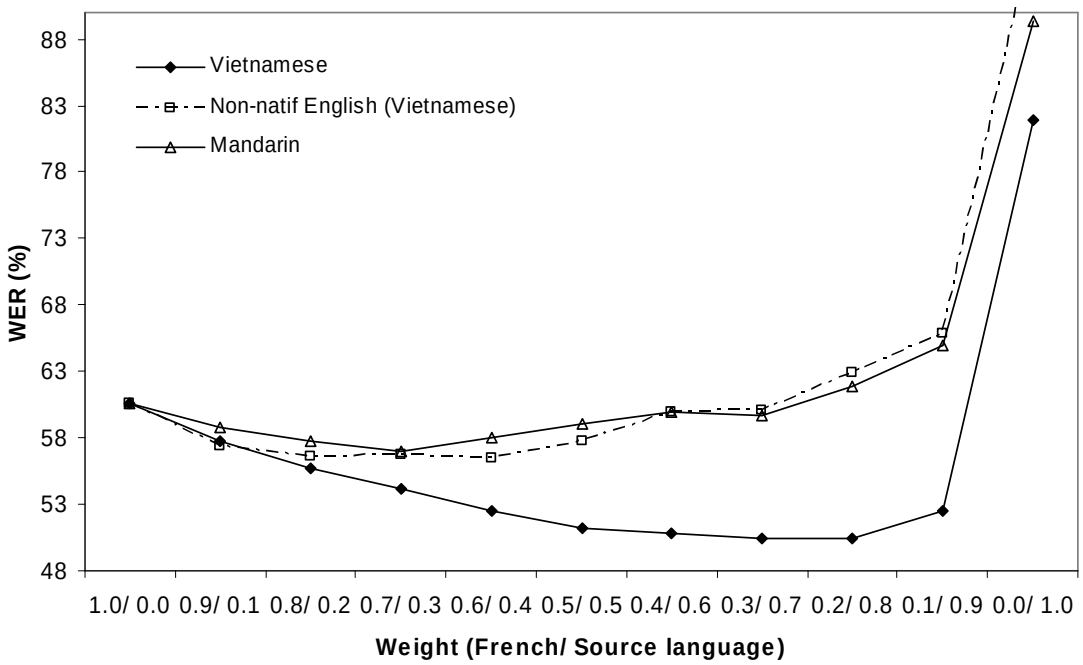


Figure 5.2 WER on non-native French speakers of Vietnamese origin using hybrid models created from a 16 Gaussian CI French and different source acoustic models with varied weights

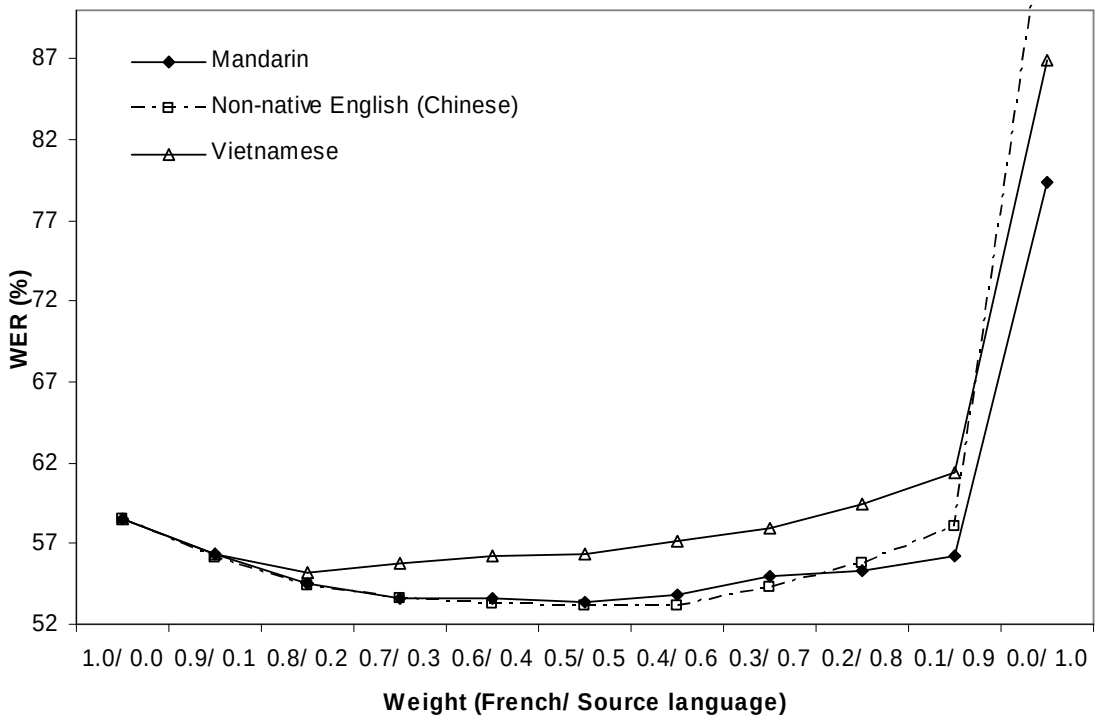


Figure 5.3 WER on non-native French speakers of Chinese origin using hybrid models created from a 16 Gaussian CI French and different source acoustic models with varied weights

5.3.1.4 Manual Interpolation

The hybrid approach works by interpolating and merging of Gaussian mixtures of target and source acoustic models. A small shortcoming is that the resulting model will have a higher number of Gaussian mixtures than the original target model. However, the advantage is that it is capable of treating acoustic model resources. However, when the original source language speech corpus is available, it may be more beneficial to use the corpus directly to create a source acoustic model that corresponds to the target language acoustic model for modeling purpose by adapting the target acoustic model with source language. This will also avoid using distance measure for matching the Gaussians. This can be achieved with our proposed interpolation approach in Section 2.4.2 for modeling cross-lingual transfer using multilingual corpora.

MLLR (with single regression class) and MAP were used to adapt the target acoustic model with the same three types of languages, experimented in the previous hybrid approach. Only two iterations of MLLR and an iteration of MAP were applied to avoid the transformations go too far until deteriorate the recognition results. Figures 5.4 and 5.5 present the speech recognition performance using the acoustic models created from our proposed interpolation approach across different weights for non-native French speakers of Vietnamese and Chinese origin.

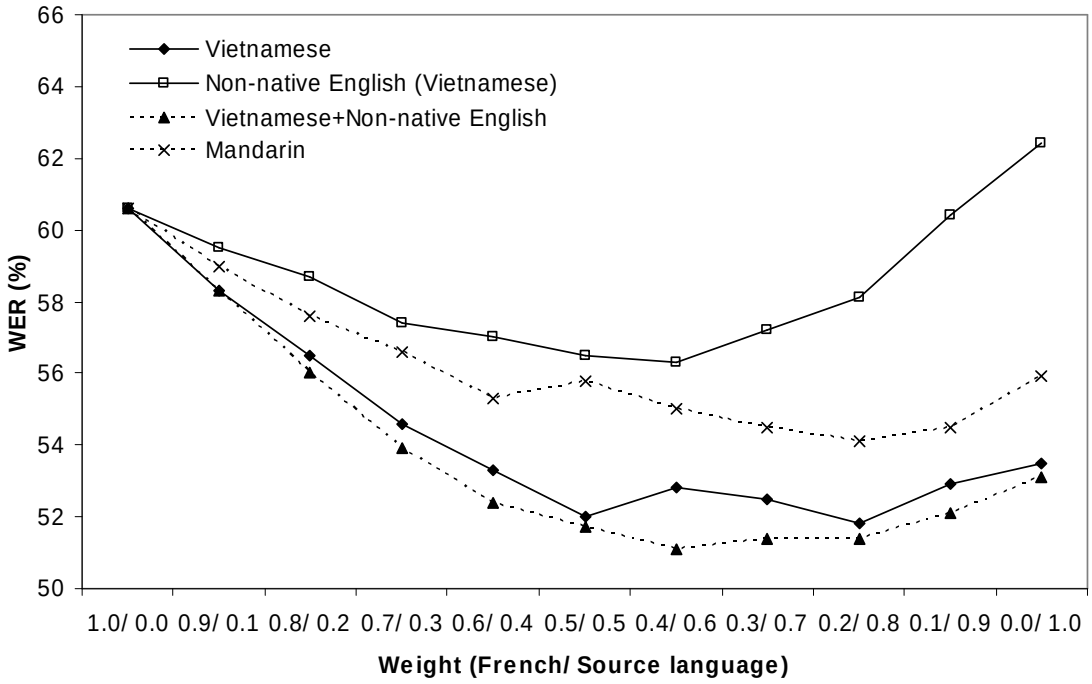


Figure 5.4 WER on non-native French speakers of Vietnamese origin using the proposed interpolated models which are created from a 16 Gaussian CI French and different source acoustic models with varied weights

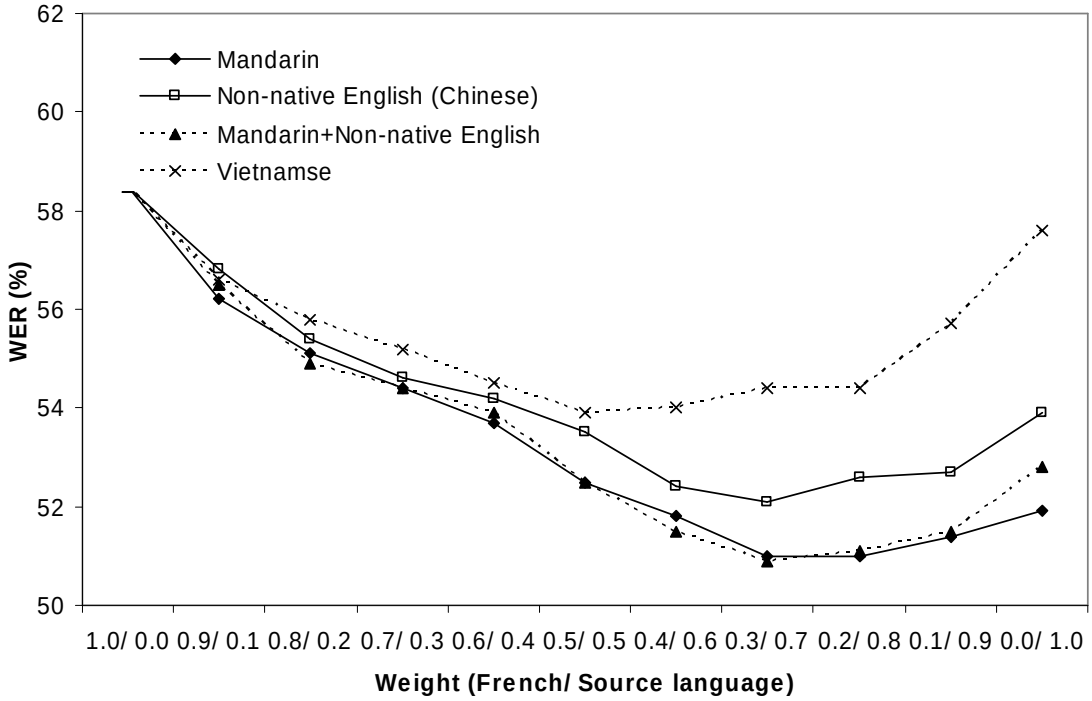


Figure 5.5 WER on non-native French speakers of Chinese origin using the proposed interpolated models which are created from a 16 Gaussian CI French and different source acoustic models with different weights

The modeling results show that L1, L2 and L3 are able to improve the non-native French speakers. Again, the results show that there is high level of accent in the non-native speech. Even without doing any interpolation, all three languages except non-native English for Vietnamese are able improve the target acoustic model. The improvements from using L1 and L2 (English) with our proposed interpolation are comparable with the hybrid approach, but without any increase in the amount of Gaussian mixtures. The average relative WER improvements of using the corresponding L1 as source language for Vietnamese and Chinese speakers are 11.13% and 9.38% respectively. Similar to the L2 adaptation results from the hybrid modeling, 5 minutes of non-native English speech seems to perform rather well compared to 5 hours of Mandarin speech for adaptation. We compared the performances of both the L1 and L2 in adaptation further by randomly selecting about the same amount of speech from the corresponding L1 corpus compared to the L2 speech to adapt the French acoustic model and subsequently tested it on the non-native speakers. The tests were not very conclusive. The results which are not shown in the figures here indicate that non-native acoustic models created from adapting French acoustic model using the non-native English speech produced lower WER across different weights for Chinese speakers compared to the one created by adapting with Mandarin speech (L1), resulting in as many as 3% reduction in error rate in the best case. However, for Vietnamese speakers, Vietnamese speech (L1) is slightly better than non-native English speech from Vietnamese speakers for adapting French acoustic model. We have also tested the combination of L1 and the L2 for adaptation. The

combined model was built by simply adapting the target model with the non-native English speech and then adapting it again with the native language of the speaker. From the figures, they show that combining both speech sources for adapting the French acoustic model produced models which are only slightly better for Vietnamese speakers, while there are no big differences for Chinese speakers. Finally, French acoustic models interpolated with L3 produce a more significant reduction in error rate compared to the models created using the hybrid approach on average, where word error rate reduced as much as 6% for native Vietnamese and 4% for native Chinese speakers. Surprisingly for Vietnamese speakers, even without any interpolation, Mandarin speech is able to improve the target acoustic model, but only slightly for Chinese speakers with Vietnamese speech. One possible explanation for this is that the transformation carried out with Mandarin speech and the (original non-native Chinese) phoneme mapping suits Vietnamese more than the opposite for Chinese speakers. Furthermore, only few iterations of MLLR and MAP were employed.

These results from our proposed interpolation approach are promising. The acoustic model created from modeling with L1 is comparable with hybrid approach, but at the same time there is no increase in the number of Gaussian mixtures in the models. On the other hand, the modeling with L2 and L3 produces acoustic models that are better than the corresponding one adapted with hybrid approach for speech recognition on average.

5.3.1.5 Weighted Least Square

The proposed interpolation approach above can produce very good recognition results if suitable weight is assigned for modeling. However, in our experiment, we evaluated different weights as a priori. In fact, when some non-native speech is available from the speaker, the speech can be used to estimate the weights. In this section, weighted least square (WLS) proposed in Section 2.4.3 will be tested for its effectiveness in estimating the weights.

Table 5.7 Comparing WER from manual interpolation and WLS

Native speaker	Source language	FR=0.5, VN/CN=0.5	Manual Int. (best result)	WLS
Vietnamese	VN (L1)	52.0	51.8	52.3
	VN+GMU (L1+L2)	51.7	51.1	50.2
	CADCC (L3)	55.8	54.1	54.9
Chinese	CADCC (L1)	52.5	51.0	51.8
	CADCC+GMU (L1+L2)	52.5	50.9	51.1
	VN (L3)	53.9	53.9	52.7

Baseline results for Vietnamese and Chinese speakers are 60.6% and 58.5% respectively

Table 5.7 presents the performance improvement of applying WLS for estimating the weights for different source languages compared to using manual interpolation. Three utterances were selected from each speaker for estimating the weights for each speaker. The results are very

encouraging, where the average improvements for L1, L2 and L3 are equal or better than the best results found in the manual interpolation. This is because in manual interpolation, the same weight is applied for all the speakers, and only one weight is used. Using WLS, speaker specific weights can be estimated. The improvement can therefore be higher if the estimation is accurate. This shows that the weights automatically estimated for modeling are relevant.

5.3.1.6 Eigenvoices in Bilingual Space

In this section, eigenvoices approach presented in Section 2.4.4 is examined for non-native acoustic modeling. The idea is to use the source language corpus to create a language space in the eigenspace to improve non-native speaker adaptation.

We did not create speaker dependent models from non-native English speech (for creating supervectors) since only about 30 seconds of speech are available from each speaker. Hence, only language resources of L1 and L3 were verified. The French supervectors were created from BREF120 corpus, producing 120 supervectors (one for each speaker from BREF120). On the other hand, the 'non-native' supervectors were created using VN and CADCC corpora, but using the same French speaker independent context independent model as the initial model. This is done by using MLLR and MAP to adapt the French speaker independent acoustic model using the Vietnamese and Mandarin corpora. This produces 29 and 20 speaker dependent models for Vietnamese and Chinese respectively. Subsequently, eigenvectors were derived from both the target and source supervectors (bi-lingual space). Figure 5.6 shows the positions of ten of the French and Vietnamese speakers used for creating the eigenvectors on eigenspace. This example shows that the second dimension of the eigenspace may correspond to the bi-lingual space.

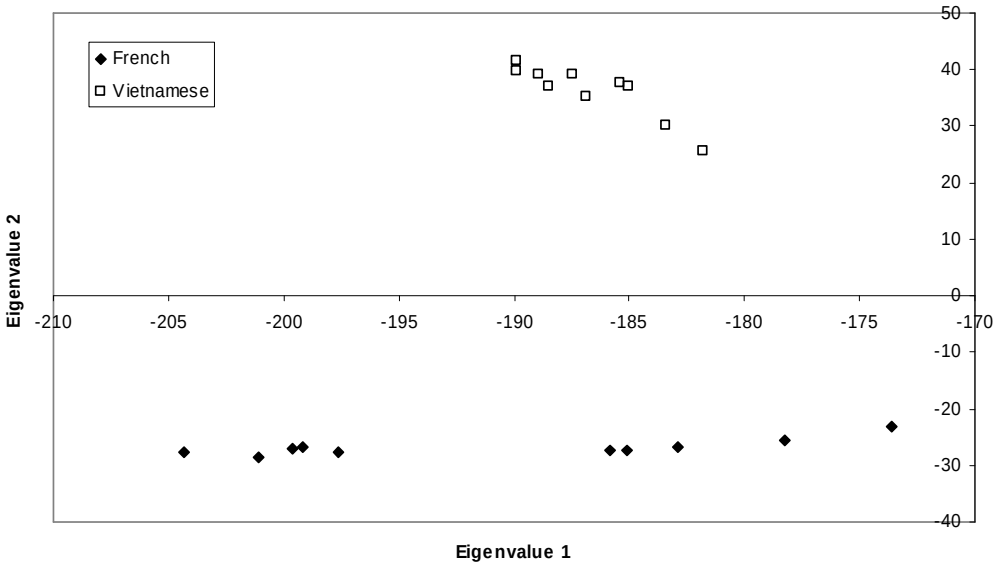


Figure 5.6 Position of ten French and ten Vietnamese training speakers in eigenspace created from French and Vietnamese supervectors

Table 5.8 presents the improvement in WER using 20 principal components with MLED eigenvoices approach. The results show that using only French supervectors for creating the eigenvectors produces 5.9% and 5.8% reduction in error rate for Vietnamese and Chinese speakers. This improvement corresponds to a conventional speaker adaptation using eigenvoices approach. When French supervectors are combined with the L1 supervectors of the speakers, to form a bi-lingual space, this can further improve the performance of eigenvoices. However, L3 resources do not produce any significant improvement.

Table 5.8 Average WER of Eigenvoices using 20 components

Native Speaker	Baseline	French supervectors	French + Vietnamese supervectors	French + Chinese supervectors
Vietnamese	60.6	54.7	**51.9**	54.4
Chinese	58.5	52.7	52.6	**51.5**

We went further to verify the performance of using the source language supervectors by verifying whether adding more target language supervectors will actually lead to the same improvement. For this, we varied the number of French supervectors used for creating the eigenvectors and setting the number of principal components used at constant. The results in Figure 5.7 show that when the number of French supervectors reaches 40, the adaptation has already reached an optimum state for native Vietnamese and native Chinese, where subsequent results do not differ much after that. The addition of source language supervectors after 120 French supervectors produces a significant drop in WER.

We have also evaluated the performance of our proposed approaches with the conventional (supervised) MLLR used in speaker adaptation, and subsequently combined the proposed approaches with MLLR (using one regression class). The same amount of adaptation speech was used in all the tests. The results are presented in Table 5.9. Surprisingly, the results show that WLS and eigenvoices perform better than MLLR. In this case, WLS produces better results compared to eigenvoices. In addition, WLS is simpler to carry out. It is also able to take advantage of small amount of L3 speech, but not the eigenvoices. When the non-native adaptation approaches are combined with MLLR, for both Chinese and Vietnamese speakers, the proposed approaches give more than 10% absolute WER reduction.

We did not test unsupervised adaptation with the non-native data. However, we suppose that for both the WLS and eigenvoices approaches are able to do well, since eigenvoices are famous for its fast adaptation (using a little amount of speech), while for WLS, there are only two parameters (weights) to estimate.

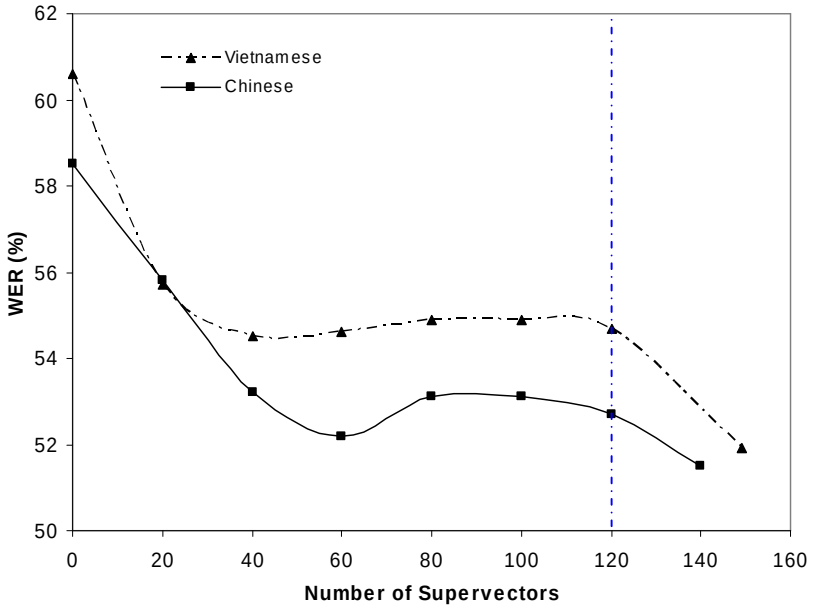


Figure 5.7 WER on non-native French speakers using different number of supervectors to create eigenvectors and maintaining the number of principal components at 20 (the number of supervectors beyond 120 corresponds to the addition of the source language speakers in the eigenspace).

Table 5.9 Comparing WER of different approaches for non-native speaker adaptation

Native speaker	Baseline	MLLR	WLS	Egv.	WLS + MLLR	Egv + MLLR
Vietnamese	60.6	53.4	50.2	51.9	48.9	50.1
Chinese	58.5	51.5	51.1	51.5	49.1	48.9

5.3.2 Context Variation Modeling

In the previous section, the hybrid of interpolation and merging has proven to be effective for modeling cross-lingual transfer by non-native speakers. In this section, we compare the effects of context modeling using the common state-tying approach during acoustic model training against our proposed context modeling in Section 2.5.1 using the hybrid approach of interpolation and merging. Unlike the native speakers, the non-native speakers are not capable of articulating accurate second language sounds. Their speech is influenced by their native language and there are possibly incorrect pronunciations. Hence, using very precise context dependent modeling such as triphones during training may not be a good choice. The hybrid of interpolation and merging is applied on two acoustic models with different contexts to achieve an intermediate level between context independent and context dependent modeling.

5.3.2.1 Baseline Context Modeling

Baseline speech recognition experiments were carried out for testing acoustic models created using different number of tied-states (triphones). The results for native French and non-native French speakers are shown in Table 5.10. In general, the average WER for Vietnamese and Chinese speakers is high, about twice the rate of native French speakers. The results from non-native French speakers also show that speech recognition system performed better using acoustic model trained with a low number of tied-states. Vietnamese speakers who are more experienced in this case show a slight reduction in word error rate when there are a small number of tied-states. On the other hand, Chinese speakers do not seem to benefit at all from context dependent modeling. These results confirm our expectation that very precise context modeling does not improve and even degrades the performance of non-native speakers. On the opposite, context dependent model works well with the native French speakers, with a reduction of about twelve percent absolute word error rate by changing from context independent to context dependent with 8129 states.

Table 5.11 shows the results of training context independent model using varied amount of Gaussian mixture. Adding more Gaussian reduces the word error rate quite substantially for non-native speakers, but the improvement for native French speakers was significantly less than the improvement gained from using context dependent model. This shows that adding more Gaussians seems to be more effective for improving the recognition of non-native speech than using very precise context dependent modeling.

Table 5.10 WER of native (French) and non-natives (Vietnamese and Chinese) using CI and CD acoustic models at different number of tied-state, with 16 Gaussians per state

State	CI: 129	429	629	4129	8129
French	36.9	30.0	27.9	23.9	24.1
Vietnamese	60.6	59.7	59.6	66.5	70.2
Chinese	58.5	63.8	67.5	78.0	83.0

Table 5.11 WER of native (French) and non-natives (Vietnamese and Chinese) using CI acoustic models with different number of Gaussians mixture per state

Gaussian	16	32	64	128	256
French	36.9	34.0	32.7	31.4	32.2
Vietnamese	60.6	57.4	56.4	55.8	54.9
Chinese	58.5	56.6	56.1	55.6	56.0

5.3.2.2 Hybrid of Acoustic Model Interpolation and Merging for Context Modeling

Context dependent modeling as mentioned in previous section is not beneficial for non-native speakers with a strong accent in general. In this section, we attempt to create a model which is intermediate between a flat context independent model and a very precise triphone context

dependent model by using the hybrid approach of interpolation and merging. For the experiments, a context independent (CI) model and an 8129 states context dependent (CD) model were used. Each contains 16 Gaussians per state.

Recall that in the hybrid modeling, each CD Gaussian will be associated with one CI Gaussian, and subsequently interpolated with the given weight, while the CI Gaussian without any associated CD Gaussian will be merged instead. Table 5.12 shows WERs of non-native French speech recognition using acoustic models created by modeling the context across different CI and CD weights. Approximated divergence distance was used as the distance measurement. The interpolation-merging produced CD models with 8129 states, where each state has an average of 25 Gaussians (except for CI weight=0 and CI weight=1.0).

We noticed that there was a slight decrease in WER when the CI weight is at 1.0, compared to the original CI result. Note that when CI weight is equal to 1.0, the algorithm produces a model with 8129 states, where all triphones are replaced by their respective monophones. The best WERs for native Vietnamese and Chinese speakers of French were achieved when CI weight was 0.7. The WERs were 51.5% and 54.0% for Vietnamese and Chinese respectively. The results are even better than context independent modeling using 256 Gaussian mixtures. The results show that when appropriate weight is used, the hybrid method produces encouraging results. The weight to apply seems to correspond to the experience of the speaker in the language. The Vietnamese speakers who are more experienced show higher improvements in WER compared to the Chinese speakers. We also found that the WER of native French speakers only showed slight increase of 2% compared to the baseline CD model when the CI weight is equal to 0.5.

The results from the experiments are very encouraging, since the approach is very easy to apply and produces a better result than context independent model with a lot of Gaussians. Furthermore, native French speakers only show a slight decrease in performance using our proposed modeling approach.

Table 5.12 Interpolation-merging of a 16 Gaussian CI (129 States) and a CD (8129 States) model

CI weight	1.0	0.9	0.7	0.5	0.3	0.1	0.0
French	36.9	33.2	28.6	26.1	25.0	22.7	24.1
Vietnamese	58.0	53.5	51.5	51.5	53.4	56.4	70.2
Chinese	57.3	55.2	54.0	55.2	58.3	63.5	83.0

CI weight at 1.0 denotes CI model, and CI weight at 0.0 denotes CD results

Finally, it is also interesting to verify whether combining context and cross-lingual transfer modeling will produce an additive effect. We used the new French CD acoustic model created from our proposed hybrid approach of context modeling at CI weight equals to 0.5, followed by cross-lingual transfer adaptation using the corresponding L1 acoustic model of the non-native speakers. The results are presented in Table 5.13. When French weight equals to 0.5, the results showed an overall improvement in WER compared to the baseline presented in Table 5.12 from

60.6% to 44.1% for Vietnamese speakers and from 58.5% to 52.1% for Chinese speakers. On the other hand, the WER for native French speakers (not shown here) increased only about 3% with the non-native acoustic models, compared to the baseline CD acoustic model.

Table 5.13 WER of non-native French using combination of context and cross-lingual modeling (using the 0.5/0.5 hybrid model of the previous table)

French weight	1.0	0.9	0.7	0.5	0.3	0.1
Vietnamese	51.5	49.1	46.6	44.1	45.0	45.9
Chinese	55.2	54.4	52.5	52.1	53.0	53.0

5.3.3 Conclusions from Non-Native Acoustic Modeling

We have examined the proposed multilingual acoustic modeling approaches to adapt French acoustic model for non-native speakers without using any non-native French speech. Three types of speech were experimented for adaptation. For Vietnamese speakers, they were Vietnamese (L1), non-native English by Vietnamese (L2) and Mandarin (L3). On the other hand for Chinese speakers, the languages tested were Mandarin (L1), non-native English by Chinese (L2) and Vietnamese (L3). Among these three types of speech, the non-native speech can be more or equally effective as the native language of the speaker for adaptation, even though the language pronounced by non-native speakers is different from the target language of speech recognition system. Interestingly, with appropriate (smaller) weight to the source model, native Vietnamese speech seems to be useful for adapting non-native French speaker of Chinese origin and vice versa. This shows that native language close to the mother tongue of the speaker can be useful.

Two approaches have been proposed for adapting the acoustic models for non-native speakers, depending on whether the resource available is in the form of acoustic model or corpus. The hybrid approach performs better than the baseline acoustic model interpolation and merging approach given the source L1 acoustic model. It has also shown to be beneficial for adaptation with L2 and L3 acoustic models. If L1, L2 or L3 corpus is available, our proposed manual interpolation approach performs nearly as effective as the hybrid approach, and at the same time maintains the number of Gaussians in each state the same as the initial target acoustic model. In addition, two speaker adaptation techniques have also been proposed for non-native speaker adaptation, they are weighted least square (WLS) and eigenvoices. In term of language resources, WLS has shown to be capable of taking advantage of L2 and L3 resources even though in small quantifies. However, eigenvoices approach does not seem to be able to work with L3 resources. Both WLS and eigenvoices when combined with MLLR have shown additional reduction in WER, producing more than 10% reduction in absolute WER.

In term of context variation modeling, using context dependent modeling for inexperienced non-native speakers will end up producing more errors. On the other hand, increasing the number of Gaussian mixtures gives a better improvement in speech recognition accuracy. However, this means that the benefit of context dependent for native speakers have to be sacrificed. The hybrid approach proposed for multilingual acoustic modeling has proven to be also useful for modeling

between two different contexts to achieve an intermediate state, where the resulting model reduces the errors of non-native speakers and at the same time causes only a slight increase in the recognition errors for the native speakers.

5.4 Pronunciation Modeling

This section examines the pronunciation modeling methods proposed in Chapter 3 for non-native speakers. In these experiments, we assume that some non-native speech is available for finding the pronunciation variants. Three pronunciation modeling approaches were tested. Two of the approaches are modifications of earlier works, where we attempt to use small amount of non-native speech available for modeling pronunciation variants with pronunciation dictionary and n-best list rescoring approach. We have also tested a new approach called latent pronunciation analysis for pronunciation habit clustering and non-native pronunciation adaptation. The non-native French (NNF) and non-native English corpus (ISLE) are tested in all the approaches, except the latent pronunciation analysis, which is only tested with non-native English speakers, because there is no sufficient data in French for executing the test.

5.4.1 Pronunciation Dictionary: Decision Tree

This section presents the experiments carried out for testing pronunciation modeling using pronunciation dictionary in Section 3.2.1. The pronunciation variants are derived from decision trees, and then added into the pronunciation dictionary. Table 5.14 below shows the data used for testing and training the accent models for non-native French and English speakers.

Table 5.14 Number of speakers involved in test and pronunciation modeling

Corpus	Description	# Speaker
NNF	Test	10
	Pronunciation modeling (Vietnamese)	3
	Pronunciation modeling (Chinese)	2
ISLE	Test	34
	Pronunciation modeling (German)	6
	Pronunciation modeling (Italian)	6

In the first pass, triphone confusion matrix was created by aligning the hypotheses and the references from the phoneme recognizer and the forced alignment system respectively. The threshold for the confusion matrix was set very low at 0.15, which created on average about 10 variants per pronunciation. The variants were then added into a temporary dictionary.

In the second pass, the dictionary created in the first pass was used for forced-alignment. The phoneme time stamps produced from the force alignment were aligned again with the corresponding references from the first pass. This creates triphone confusions which will be used to create phoneme decision trees. Wagon utility employing CART algorithm from the Festival speech synthesis system [Taylor 1998] was used to create the phonetic decision trees. Before the decision trees can be created, the left and right contexts of the phoneme confusions from the

second pass have to be converted to feature vectors. IPA based articulation features were used to represent the phoneme for classification, see Table 5.15.

Table 5.15 Articulation feature vector (complete) used for building decision trees for four French phonemes

Phoneme	ɔ	*a*~	t	m
vowel	vowel	vowel	consonant	consonant
Tongue	back	front	NA	NA
Opening	open-mid	open	NA	NA
Lips	rounded	unrounded	NA	NA
Duration	normal	normal	NA	NA
Nasal	non-nasal	nasal	NA	NA
Manner	NA	NA	plosive	nasal
Place	NA	NA	alveolar	bilabial
Voiced	NA	NA	unvoiced	unvoiced
Aspirated	NA	NA	unaspirated	unaspirated

The tests for non-native French and English speakers were carried out using context independent models with 16 Gaussians per state. Phoneme decision trees were subsequently created, and the threshold for the decision trees was set at 0.3, producing about one extra variant per pronunciation. Table 5.16 shows an excerpt of the non-native variants added into pronunciation dictionary. From the example variants, it is obvious that substitution, deletion and insertion have happened. Some of the pronunciation variants have been seen in previous chapter, for instance the deletion of final plosive by Vietnamese speakers, while others are new observations. As discussed in Chapter 1, the reasons for this may be caused by wrong perception, native pronunciation norm or simplification of the pronunciation that occur on non-native speakers.

Table 5.16 Examples of non-native pronunciation variants (Vietnamese and Chinese) derived from decision trees.

French Word	French (standard)	Vietnamese variant	Chinese variant
août	u t	u	–
brochure	b R ɔ ʃ y R	–	b R ɔ ʃ i R
désirez	d ɛ z i R e	–	d ɛ s i R e
excusez	ɛ k s k y z e	ɛ s k y z e	–
fixé	f i k s e	f i s e	–
monnaie	m ɔ n ɛ	m o n ɛ	m *a*~ n ɛ

Table 5.17 below shows the performance improvement of the speech recognition system after adding the pronunciation variants into the baseline dictionary and for testing. Context independent model with 16 Gaussians mixture was used in the tests. Non-native French speakers from Chinese and Vietnamese origin shows a reduction in WER of 1.3% and 1.8% respectively, while German and Italian speakers speaking English show a reduction of 2.6% and 4.6%.

Table 5.17 The improvement in WER by modeling pronunciation variants using decision trees for non-native French and English speakers

Approach	Non-native French		Non-native English	
	Chinese	Vietnamese	German	Italian
Baseline	56.2	58.1	58.7	81.5
Decision Tree	54.9	56.3	56.1	76.9

Dependent T-tests were calculated to verify whether the improvement was significant (95% confidence). The t-values for Chinese and Vietnamese speakers (non-native French) are 4.089 and 3.182 respectively, while the values for German and Italian speakers (non-native English) are 7.975 and 4.877 respectively. The results for non-native English speakers show that the reductions in WER are significant, but the results from non-native French speakers are not. This is due to the small sample of number of speakers that we used for testing non-native French speakers. The calculations for statistical significant are done for reference purpose, and the statistical insignificant of the non-native French results do not necessary indicate that the approach is flaws, or the difference is not important.

5.4.2 N-Best List Rescoring

In the n-best list rescoring approach, the pronunciation models are re-evaluated after decoding in the rescoring module. The same speakers used in the previous section were used for creating the triphone pronunciation model and also for testing (see Table 5.14). To create the triphone model, the triphone confusion is interpolated with the monophone confusion, using the weight 0.8 and 0.2 respectively. The number of n-best sentences is set at 100. Table 5.18 below shows the improvement after rescoring.

Table 5.18 The improvement in WER by modeling pronunciation variants using n-best rescoring for non-native French and English speakers

Approach	Non-native French		Non-native English	
	Chinese	Vietnamese	German	Italian
Baseline	56.2	58.1	58.7	81.5
N-Best rescoring	55.5	56.8	56.9	79.6

The results show that n-best rescoring is able to reduce the error made. However, compared to the pronunciation dictionary results, n-best rescoring performance is slightly lower in all the tests. Furthermore, n-best list rescoring requires much more processing compared to the traditional dictionary approach.

5.4.3 Latent Pronunciation Analysis

In latent pronunciation analysis, pronunciation confusion supervectors are created to derive eigenvectors, which will be used for clustering speakers according to their pronunciation habits, and estimating the pronunciation model of an unknown speaker.

In this test, only the non-native English corpus (ISLE) was evaluated, because there was no sufficient adaptation data to create French pronunciation models for testing. Eighteen non-native English speakers, each with about 15 minutes of transcribed speech from native German and Italian were selected as the training set to create the speaker dependent pronunciation confusion supervectors. For creating the 'training' supervectors, speaker dependent decision trees have to be grown. Wagon utility was again applied for this purpose. The next step is to create the supervectors. First, triphone contexts have to be extracted from the test dictionary to create the bare bones of the supervectors. In total, 1464 triphone contexts were extracted from the test pronunciation dictionary. Subsequently, the threshold for the decision trees was set at 0.5, pronunciation variants for the triphone contexts were extracted from the speaker dependent decision trees, and this creates supervectors with 4023 dimensions (features) each. This means that for each context, there were about 4 possible variants. With the 36 pronunciation confusion supervectors, a covariance matrix was calculated, and subsequently the eigenvectors were derived.

First, we examined the approach in term of clustering the speakers according to pronunciation habits. We will look only at the effectiveness of first and second eigenvectors in clustering the speakers. This is done by plotting the eigenvalues for the 36 training speakers. The results in Figure 5.8 show that the two groups of non-native speakers can be indeed separated by using the first eigenvector. Without any knowledge about the accent of the speakers, one way is to manually divide the speakers to two groups using eigenvector 1, at the value zero.

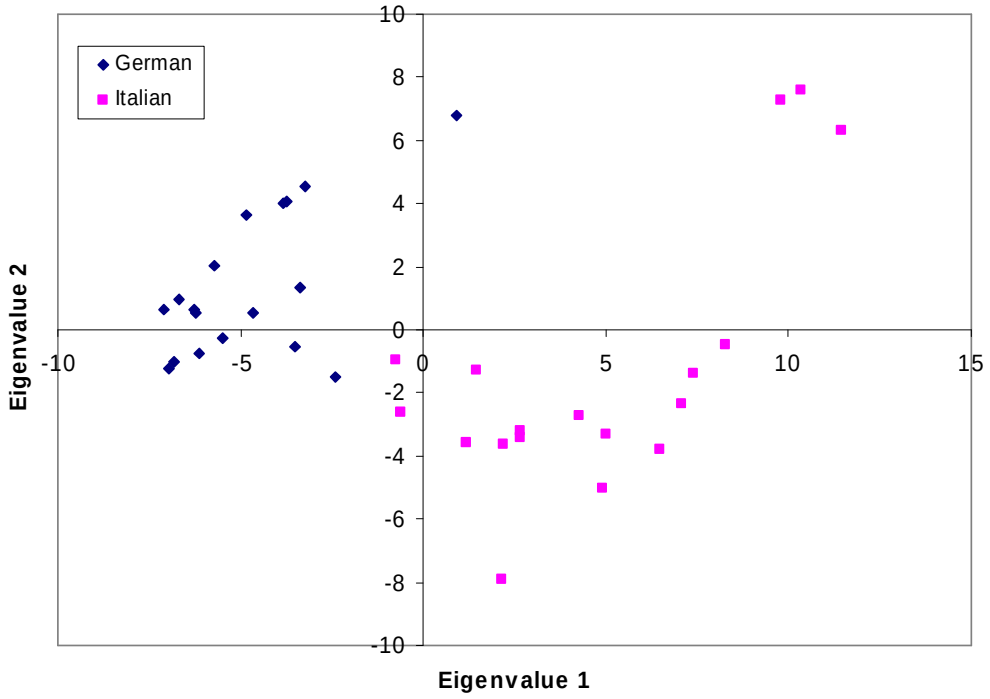


Figure 5.8 Plotting 36 non-native English (18 German/ 18 Italian) speakers on eigenspace (dimension one and two)

For pronunciation adaptation, ten principal components with the highest eigenvalues were used. The remaining ten speakers were tested. Two minutes of transcribed speech from each test speaker was used for estimating the test pronunciation confusion supervectors. Subsequently, the weights of the supervector on the eigenspace were estimated. The weights are then used for re-estimating the test supervector by projecting using the eigenvectors. The threshold was set at 0.4 and 0.3 to extract the pronunciation variants from the supervector and added into the pronunciation dictionary. Table 5.19 shows the results compared to the conventional pronunciation dictionary approach. Since only ten speakers were involved in the test, the baseline results in this experiment are different from the previous two tests. The results show that the latent pronunciation analysis approach is able to predict the pronunciation variants rather well with reduction in WER. However, the improvement is less compared to the general pronunciation dictionary (decision tree) approach. The benefit of this approach is that the accent of the speaker does not need to be known in advance for selecting the right dictionary. We have also carried out a simple comparison of the variants generated by randomly selecting the variants estimated for an Italian speaker, and compared them against the variants for other Italian and German speakers. The average difference (deletion and addition) between the number of variants for that Italian speaker and the other Italian speakers is 209, while the average difference between the variants for the Italian speaker and the German speakers is 411. This shows that the variants generated are speaker specific, and at the same time more related to the accent group of the speaker.

Table 5.19 Comparing the result of latent pronunciation analysis with the normal decision tree approach for modeling pronunciation variants

Speaker	Baseline (≈1k words)	Latent Pronunciation Adaptation		Decision Tree	
		200 variants	400 variants	200 variants	400 variants
Italian	75.5	72.6	72.2	73.0	71.2
German	59.0	57.6	57.2	56.3	56.0

5.4.4 Conclusions from Pronunciation Modeling

We have examined three different approaches of pronunciation modeling. The use of decision trees remains the best way to model pronunciation variants. Adding variants created from decision trees into pronunciation dictionary reduce the WER more than the rescoring n-best list. Furthermore, the n-best list rescoring requires much more processing compared to the traditional dictionary approach. As for the latent pronunciation analysis, it has shown to be promising for clustering speakers using pronunciation habits. It is also shown to be beneficial for pronunciation adaptation given some speech. The method can be used in situation when we do not know in advance the accent of the speakers. If the accent of the speaker is known, it is better to use the accent specific pronunciation dictionary, since it produces better results.

5.5 Accent Identification

In this section, accent identification using multilingual decision tree which has been proposed in Chapter 4 will be examined. Some baseline accent identification approaches mentioned in Chapter 1 were also tested by comparing them with the proposed approach. In all the experiments, non-native French from our NNF corpus and non-native English speakers from ISLE corpus were examined.

5.5.1 Baseline Approaches

Four baseline accent identification approaches described in Section 1.5.4 were tested. Two of the approaches employ acoustic features, and the other two use phonotactic features for accent identification. Table 5.20 shows the amount of non-native speech used to create and test the accent models. The NNF corpus contains about a thousand test sentences (Chinese and Vietnamese), while the ISLE corpus has about two thousand eight hundred sentences (German and Italian).

Table 5.20 Number of speakers involved in test and accent modeling

Corpus	Description	# Speaker
NNF	Test	10
	Accent modeling (non-native French by Vietnamese)	3
	Accent modeling (non-native French by Chinese)	2
ISLE	Test	34
	Accent modeling (non-native English by German)	6
	Accent modeling (non-native English by Italian)	6

5.5.1.1 Baseline 1: Acoustic Features Score using HMM

A text dependent accent identification approach based on the approach presented in Figure 1.18 was tested. The accent models were created by adapting the target context independent acoustic models with 16 Gaussian mixtures by using non-native speech with MLLR and MAP. For testing the accent of an unknown speaker, it is carried out by forced aligning the speech of the non-native speaker, and comparing the acoustic scores generated by different accent models.

The results for the experiment are presented in Table 5.21. Using non-native speech to adapt the target acoustic model gave encouraging results for identifying German and Italian accents. However, the approach did not do well for identifying accents from non-native French speakers. For non-native French test, to verify further whether the low correct identification rates are due to the unsuitable decision threshold used, we have added additional score to the Chinese accent model (which is equaled to changing the decision threshold). The results in Table 5.22 show that changing the decision threshold improves the Chinese accent identification, but at the same time

reduces the Vietnamese accent identification. This shows that changing the threshold score does not help. The possible reason for the failure in identifying non-native French accent is due to the small amount of adaptation speech, and the similarity between the two accents.

Table 5.21 Accent identification using accented acoustic models

HMM Model for PR 1	HMM Model for PR 2	Speaker	Correct rate
French + Non-native French (Vietnamese)	French + Non-native French (Chinese)	Vietnamese	53.14%
		Chinese	47.24%
English + Non-native English (German)	English + Non-native English (Italian)	German	87.43%
		Italian	89.38%

Table 5.22 Effect of changing the decision threshold in accent identification (acoustic features)

Threshold score	1000	5000	10000	50000	100000
Vietnamese	53.14%	52.94%	52.94%	51.72%	50.71%
Chinese	47.24%	47.44%	47.65%	48.47%	50.10%

5.5.1.2 Baseline 2: Acoustic Features Score using GMM

A text independent accent identification approach based on Gaussian mixture model (GMM) for speaker recognition was also tested. The Alizé speaker recognition toolkit [Bonastre 2005] was used to create the accent models based on GMM. An accent independent model was initially created. It was then adapted to accented models with few iterations of MAP adaptation, by using the same amount of non-native speech as before. Table 5.23 below shows the accent identification results using 32 Gaussians mixture model. The results again show that recognizing English accent is relatively good, although slightly poorer that previous approach. Non-native French accent is still a problem for recognition.

Table 5.23 Accent identification using Gaussian mixture models

GMM 1	GMM 2	Speaker	Correct rate
Non-native French (Vietnamese)	Non-native French (Chinese)	Vietnamese	58.62%
		Chinese	45.43%
Non-native English (German)	Non-native English (Italian)	German	74.78%
		Italian	73.31%

5.5.1.3 Baseline Method 3: Phonotactic Features Score using LM

In this test, language models are used to capture non-native phonotactic features for accent identification. The approach has earlier been presented in Section 1.5.4.2 (see Figure 1.19). In general, the phoneme trigram models were created using the phoneme strings from the decoding of non-native speech. For a given utterance with an unknown accent, a target language phoneme recognizer decodes the utterance, and the accent model which gives the highest language model score according to equation 1.17 is selected as the hypothesized accent. For our test, the phoneme bigram weight was set at 0.7999, and the unigram at 0.2. The same conditions in term of test and adaptation material were used for testing as mentioned in the previous section, refer to Table 5.20. Table 5.24 shows the accent identification results.

Table 5.24 Accent identification using phoneme language models

LM 1	LM 2	Speaker	Correct rate
Bigram: Vietnamese accent	Bigram: Chinese accent	Vietnamese	78.09%
		Chinese	25.15%
Bigram: German accent	Bigram: Italian accent	German	67.67%
		Italian	83.79%

The results show that it identifies English accent quite well, but like the approach before, it also suffers from classification errors with non-native French accent from Chinese speakers. Again, to verify the accent from non-native French speakers, we have attempted to modify the decision threshold, and Table 5.25 shows the results. The results show that changing the decision threshold improves the identification of the accent of Chinese speakers, and at the same time deteriorate the identification of the accent of Vietnamese speakers.

Table 5.25 Effect of changing the decision threshold in accent identification (language model)

Threshold score	0	1.0	2.0
Vietnamese	78.09%	60.74%	46.04%
Chinese	25.15%	42.80%	57.81%

5.5.1.4 Baseline 4: Phoneme Distribution Score using SVM

In this section, phonotactic features are modeled using SVM, based on the approach described in Section 1.5.4.2 (see Figure 1.20). This is a text independent method. The accent models were built by using SVM classifier through supervised training using instances of multilingual phoneme distributions. Recall that each training utterance is decoded by parallel phoneme recognizers producing phoneme distributions of different languages, which will be merged to become the training instance.

In this test, five phoneme recognizers of French, English, Malay, Chinese and Vietnamese were employed. Their context independent acoustic models were created from BREF120, TIMIT, MSC, CADCC and VN corpora respectively (refer to Table 5.3 for information on the corpora). The phoneme recognizers produced phoneme distributions of five languages with a total of 202 dimensions or phonemes distributions each. The same amount of non-native utterances like before was used for creating the accent models and for testing. Table 5.26 shows the accent identification results. This method is relatively better comparing to the previous phonotactic approach using LM score in term of identifying non-native English accent. However, like the previous two approaches, it also faces the same difficulty in discriminating non-native French accent from Vietnamese and Chinese speakers.

Table 5.26 Accent identification using phoneme distribution features

Speakers	Correct rate	Speaker	Correct rate
Vietnamese	70.00%	German	70.19%
Chinese	48.48%	Italian	89.96%

5.5.2 Proposed Approach: Multilingual Decision Trees

This section will examine our proposed multilingual decision tree approach for accent identification. Note that the experiments were carried out in a text dependent mode which requires the knowledge of the transcription of the utterance spoken by the speaker during testing. No fundamental constraint would prevent us to test this method in a text independent (and unsupervised) mode using a first pass hypothesis of speech recognition system instead of the reference transcription. However, since we are working on non-native speech recognition, such a hypothesis may be seriously degraded compared to the reference. Testing the feasibility of this text independent mode is part of future work. As mentioned in Chapter 4, the accent models consist of multilingual decision trees, which are built by using the hypotheses from the multilingual phoneme recognizers and the references from the force alignment using target language speech recognition system. During accent identification, the speech from the non-native speakers is also decoded by the multilingual phoneme recognizer and the hypothesis is compared to a forced aligned reference. The triphone confusions are created and evaluated using the accent models in the form of multilingual decision trees. The accent model which gives the highest score will be selected as the hypothesized accent.

In this test, the same multilingual phoneme recognizers (French, Mandarin, Vietnamese, English and Malay) employed in the previous test were also used here. The same decision trees utility as in the pronunciation dictionary experiments was used. The floor probability is set at 0.005. The results in Table 5.27 show that most of the decision trees of different languages can be used to identify different accents equally well as the target language. By making use of all the decision trees during evaluation, it produces a better result compared to just using a target language decision trees, except for German speakers. Compared to the baseline, the method seems to work for non-native English and French, although the performance of German speakers

is less than the baseline approaches. The approach seems useful in situation when the training data is limited (case of non-native French).

Table 5.27 Accent identification for non-native French and English speakers using decision trees (DT) of different languages

Language in the DT	Non-native French	Correct rate	Non-native English	Correct rate
French	Vietnamese	75.05%	German	65.30%
	Chinese	55.83%	Italian	90.24%
Chinese	Vietnamese	69.78%	German	54.74%
	Chinese	60.53%	Italian	92.04%
Vietnamese	Vietnamese	67.95%	German	58.41%
	Chinese	70.35%	Italian	92.40%
English	Vietnamese	72.81%	German	75.93%
	Chinese	54.40%	Italian	88.81%
Malay	Vietnamese	65.92%	German	67.39%
	Chinese	66.46%	Italian	89.38%
Combine All	Vietnamese	83.16%	German	68.46%
	Chinese	70.76%	Italian	95.62%

5.5.3 Conclusions from Accent Identification

This section describes our preliminary experiments on accent identification, which is useful for non-native speech recognition. We have examined a multilingual decision trees approach for accent identification and compared it against some baseline approaches. This approach gives consistently good results even when only small amount of non-native speech is available to create the accent models. On the other hand, baseline approaches seems to work well when large amount of speech is available for creating the accent models (non-native English case), but not very satisfactory when small amount of training data is available (non-native French case). However, the baseline approaches presented here except the one using acoustic features have the added advantages of being text independent. The combination of these approaches may be useful to overcome the pros and cons of each other.

Conclusions and Future works

Conclusions

Automatic speech recognition system has been increasingly applied in various fields. However, speech recognition systems still suffer from various difficulties in treating non-native speech. The accuracy of the systems in recognizing non-native speech is at least twice lower than native speakers. The high error rate is due to the difference in native and non-native speech characteristics. In general, non-native speakers commit development and transfer errors when they learn a new language. These two types of errors are obvious and occurr in different levels of language, namely phonology, pronunciation, vocabulary and grammars.

The L2 phonology of non-native speakers may be different from native speaker. Mastering L2 speech sounds is not easy for language learners because of interference from the L1 of the speaker. A Speech Learning Model (SLM) has been proposed by Fledge for describing the process a non-native speaker goes through when he learns a new language. In term of prosody, there are also differences between non-native and native speakers. Even though the difference in prosody does not change the meaning of the utterance, it might influence the speech recognition performance. In term of pronunciation, the L1 pronunciation rules of the speaker may influence how the he pronounces L2 words. They may also simplify the pronunciation for complex syllables through insertion, deletion and substitution of speech sounds. They are also likely to use the wrong vocabulary. One reason is the negative transfer of L1 vocabulary to L2. Another is the wrongly association of meaning to L2 vocabulary. The grammars used by non-native speakers are also more general and common. In grammars, non-native speakers make more mistakes, and part of the mistakes is due to the influence from their L1 as well as unfamiliarity. These differences in phonology, pronunciation, vocabulary and grammars in non-native speech cause mismatch in the acoustic, pronunciation and language model used in speech recognition system for treating the non-native speech.

For studying non-native speech recognition, we have recorded a non-native French read speech corpus in the tourism domain. The corpus was uttered by Vietnamese and Chinese speakers. The highly accented speech has been evaluated using phonetic approach through perception and acoustic analysis, as well as through data-driven approach. The results show that both approaches show comparable results. However, data-driven approach is easier to conduct, and it is able to provide some additional insight into the data. The results show that cross-lingual transfer is evident in the non-native speech. These observations and also from other existing linguistic studies have shown that non-native speakers tend to transfer their native phoneme close to the target phoneme (according to IPA) to substitute the target language phoneme. This

generalized procedure can be useful for predicting the possible cross-lingual transfer by non-native speakers.

Existing non-native acoustic modeling approaches can generally be divided into acoustic model reconstruction, acoustic model interpolation, acoustic model merging and the general speaker adaptation approaches. These approaches either use small amount of non-native speech or the native language of the speaker for adaptation because of the difficulty to acquire large amount of non-native speech. We have proposed to use multilingual resources for adapting target acoustic model for non-native speakers. Three types of resources have been identified. There are the native language of the speaker (L1), any non-native language (L2), and languages close to the native language of the speakers (L3). Besides that, existing approaches do not address the problem of how the linguistic knowledge can be utilized for modeling cross-lingual transfer.

We have proposed to use the generalized knowledge from IPA and existing linguistic study to determine the source phoneme transfer of non-native speakers. It can also be estimated using data-driven phoneme confusion approach. After matching the target and source phonemes, non-native adaptation can be carried out. Depending on the type of multilingual resources available, different acoustic modeling approaches have been proposed. If multilingual acoustic models are available, the hybrid of acoustic model interpolation and merging has proven to be useful for modeling cross-lingual transfer. The idea is to interpolate the target and source Gaussian that are close, and merge them if they are far from each other. This is similar to the hypotheses of the Speech Learning Model (SLM), where language learners may use the L1 speech sounds, the target language speech sounds, or the intermediate speech sounds between L1 and target language depending on the experience of the speaker and the perceived difference between L1 and target language. The results from our proposed hybrid are better that the traditional interpolation or merging approaches.

Moreover, in cases where multilingual corpora are available, we have proposed different interpolation approaches for adaptation. Three types of interpolation have been proposed to be used under different constraints. In situation when only multilingual corpora are available, a manual interpolation can be performed by predicting the a priori weight to be assigned. If the non-native speaker can provide some non-native speech, the interpolation weights can be estimated using weighted least square. We have also shown that the results from traditional eigenvoices approach can be improved by creating a bi-lingual space in the eigenspace for adapting non-native speakers. Interpolation in this case is carried out by using eigenvectors. Overall, among the three kinds of resources that have been proposed, L2 even though it is from a language different from the target language, can perform as good as L1 for adaptation. On the other hand, L3 can also improve the target acoustic model, although not as much as L1 and L2. Our proposed interpolation approaches are equal in performance of the hybrid approach, but the interpolation approaches produce models with less number of Gaussians mixture compared to hybrid approach. Weighted least square is able to estimate the interpolation weights rather accurately. The approach is slightly better than eigenvoices in non-native adaptation: it is easier to carry out; it is able to take advantage of limited resource, and L3 resources for adaptation. Overall, both approaches perform better than traditional MLLR given L1 resources. When these

approaches are combined with MLLR, it further reduces the error rate of speech recognition system for non-native speakers.

Existing studies also do not address deeply the issue of context dependent modeling for non-native speakers. We have suggested that the hybrid of interpolation and merging proposed for cross-lingual transfer to be applied also for modeling context variation. The results show that, by giving appropriate (intermediate) weight, the approach performs better than state tying or increasing the amount of Gaussians mixture per state. At the same time, the improvement is attained with only a slight reduction in error rate of native speech recognition system.

Pronunciation modeling approaches can be divided based on the component where the pronunciation variants are modeled. The possible locations are pronunciation dictionary, acoustic model, language model, and rescoring module. Existing studies found that using only the native speech of the speaker for modeling pronunciation variants do not give a significant improvement. Hence, we have modified two approaches which model the variants in the pronunciation dictionary and the rescoring module, so that they work even with limited non-native speech. For the pronunciation dictionary approach, two passes were employed to find the pronunciation variants using decision trees. The purpose of using two passes is due to the low accuracy of the phoneme recognizer. By using two passes, the objective is to reduce those unlikely observations in the first pass. As for modeling pronunciation variants at the rescoring module, triphone confusion model has been used with phoneme recognizer to rescore n-best lists. Experiments carried out found that modeling pronunciation variants using pronunciation dictionary produce better results than modeling it at rescoring level. In addition, the rescoring approach requires more processing than the traditional dictionary approach. We have also proposed an original pronunciation clustering approach using eigenvectors. It is called latent pronunciation analysis, by analogy with latent semantic analysis. The eigenvectors are derived from supervectors which are created from speaker dependent pronunciation decision trees. The results have shown that the speakers can be separated based on their origin using the first eigenvector. Given some non-native speech, the approach can also be used for pronunciation adaptation. The results show that the pronunciation adaptation approach is able to predict the pronunciation habits quite well. However, when the accent of the speaker is known in advance, the results show that it is sufficient to use only the pronunciation dictionary approach.

Accent identification techniques are generally based on acoustic or phonotactic features. Many accent identification approaches are available, but it is unknown how well these systems do in situation when non-native speech available to create the accent models is limited. Our experiments carried out showed that the baseline approaches using acoustic and phonotactics methods may fail to identify the non-native accents when non-native speech is limited, but when the amount of speech is sufficient and the accent of the speakers differ significantly, the approaches can perform quite well. We have proposed a new multilingual decision tree approach for accent identification. It is however a text dependent approach which requires the existence of the transcription of the utterance spoken. Decision trees are used because they have proven to be useful in generalizing the observation for example in pronunciation modeling and state tying. On the other hand, using multilingual models improve the result of language identification system.

The results show that the proposed approach identifies the accent of the speakers better than the baseline approaches examined, in situation where only limited amount of speech is available for training.

Future works

In this work, we have presented approaches for adapting acoustic and pronunciation model using multilingual resources. Three types of languages have proven to be beneficial for adapting the non-native speakers from our experiments. They are the native language of the speaker (L1), any non-native language spoken by the same native group (L2) and language close to the native language of the speaker (L3). In current work, we do not address the characteristics that determine the closeness of a language. An in dept analysis into this will certainly be useful, so that relevant L3 resources can be identified for non-native adaptation purpose. At the same time, it will also be interesting to know how well the languages from the same family, classified by linguists performed in adaptation. In addition, so far in the experiments, all the target language phonemes are mapped to only one source phoneme. Multiple source language resources are used to adapt the target language acoustic model by simply adapting it with one after another. In fact, this do not necessary have to be the case. The relevant phones from different source multilingual corpora can be used for adapting the target phone. A more intelligent approach which takes into consideration the type, the context and the amount of the speech in each corpus may give a further boost in improving to the target acoustic model. However, if some non-native speech is available, it can probably be used to select the best units from multilingual resources to adapt the target acoustic units, probably using existing distance measures such as HMM distance [Juang 1985], Polyphone Decision Tree Specialization [Schultz 2000] and others proposed for multilingual acoustic modeling. This may produce an even better result if correctly executed.

The hybrid approach of interpolation and merging is a promising method for modeling cross-lingual transfer and context using only acoustic models. Currently, we do not propose a way to estimate the weights given some non-native speech. If there are some non-native speech from the speaker, a simple solution which can be used is to carry out forced alignment with the pre-adapted acoustic models (from hybrid approach), and measure the acoustic scores. The one that gives the highest score will be chosen for that speaker. However, it would be better to be able to estimate the weights instead, which is more flexible and may give a better result.

Because of the difficulty in acquiring non-native speech, we have not addressed the problem of adapting the language model for non-native speakers. For testing language model adaptation, it is prerequisite to acquire spontaneous non-native speech which we do not posses. We learned in Chapter 1 that non-native speakers are likely to transfer their native vocabulary and grammar rules to the target language. It would be interesting to see whether it is possible to use the native language of the speaker to adapt the language model. Factored language models, which are an extension of the n-gram language models, may be useful for improving non-native language model. Part of speech or semantic class can be assigned to the words in the target and source languages. The source language grammar structure which is represented by part of speech with

trigram, and vocabulary which have similar graphemes to the target language can be transferred to the target language, for instance by interpolating the both the target and source language model.

Code switching occurs more and more often among speakers nowadays. It is a phenomenon where speakers use more than one language or dialect in their speech. Normally, it happens when the persons involved know the languages or dialects used. In speech which contains code switching, it has been found that 84% are single word switch, 10% phrase switch and 6% clause (a group of words consisting of a subject and its predicate) switch [Skiba 1997]. Code switching is common when the speaker has difficulty to express himself using the target language. As a result, he has to temporary switch to another language to express the idea. For example, in the areas of science and technology, where the native language of the speaker may not be able to present clearly the idea, it is common to find speakers to switch to English. Code switching is becoming increasing popular in the conversation of the speakers with the introduction of international languages such as English because of its richness and possibly the higher social position associate with it. Code switching can also happen because of social reason, for example to associate to a particular group or identity. For instance, Singlish is a type of English spoken by Singaporean which is a mixture of English, Malay, Hokkien, Teochew and Cantonese. It is associated to the Singaporean identity. Finally, code switching can also happen when the speakers involved want to limit part of the conversation to a particular group of speakers.

Code switching has increased the challenges in the speech recognition area. A typical language identification system will have difficulty because the period the switching occurred is not known and the duration is very short, since it may involve only one word. Hence, combining the speech recognition systems of different languages with a language identification system for treating code switching may not be effective. Another possibility is to treat code switching as a single system. This means that the existing target models have to be adapted to recognizing acoustic units, vocabulary and possibly grammar from other languages. In term of acoustic modeling, the interesting question concerns the modeling of different acoustic units. Our proposed non-native acoustic modeling approaches proposed here may be useful. The advantage is that the same acoustic units can serve for similar phonemes in different languages. For the pronunciation dictionary, this means that the languages that shares the same phoneme set can employ the same acoustic unit from the acoustic model. However, for the new phoneme in other languages, they may have to be adapted and added to the target acoustic model. Code switching also means that foreign words from languages involved have to be added to the pronunciation dictionary and take into consideration in the language model. If all the words in the language involved are added to the target language, this will create an exponential increase in the possible words. This will reduce the accuracy of speech recognition system. However, not all words and word combinations are possible. An in dept study into the habits of the speaker in code switching will be necessary. In term of language model, interesting question is how to adapt the target language model to take into consideration the possibility of code switching, since it is normally observable in conversation but not in writing. Code switching is not a random process and it follows certain ‘rules’ or constraints. According to [Skiba 1997], there are two constraints which restrict the speaker from switching between languages. First, the free morpheme constraint which states that the speaker only switches to words from other language that has a certain similar form

to the target language. Second, the equivalence constraint which indicates that switching is only possible if it does not violate the grammars of either language. It would be interesting to see how these kinds of linguistic rules can be combined with statistical n-gram language models. Code switching therefore is a new territory that is interesting and worth studying on.

Confidence scoring is another interesting area which may be applied to non-native speech recognition. The aim of confidence scoring is to estimate the quality of the decoding of the utterance. Hence, it has a great potential to be used in non-native speech recognition. Firstly, in the field of computer-assisted language learning (CALL), the score can give language learners an idea about their pronunciations. The existing confidence scoring algorithm may need to be modified to allow it to analyze and compare the pronunciation of non-native speaker at different levels for example phoneme, syllable, word and sentences, to help them know the types of error they often commit. Second, confidence scoring can be combined with non-native speech recognition to evaluate the decoding, so that for those decoding under the threshold, some additional processing can be carried out on it, for example by going through another pass of decoding, or the speaker can be asked to repeat what he said. This would improve the efficiency of the speech recognition system.

Appendix

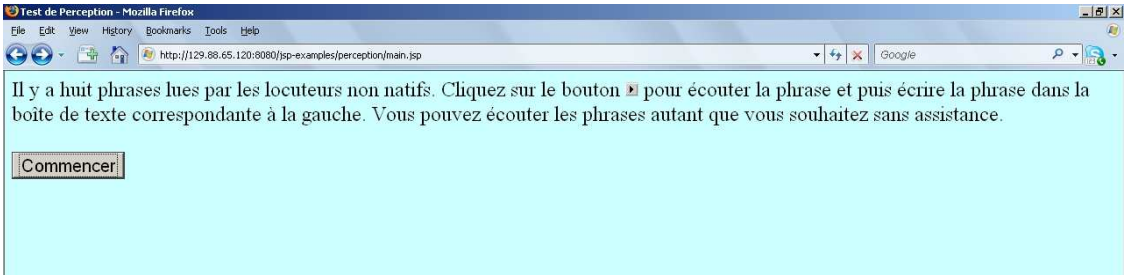


Figure A1 Intelligibility test screen: welcome screen

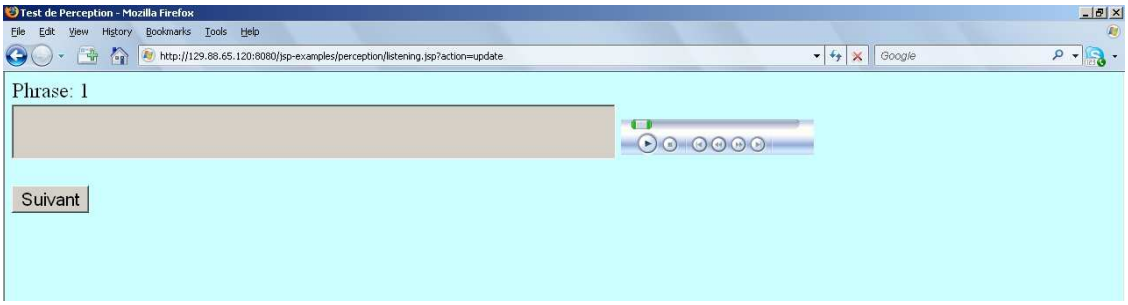


Figure A2 Intelligibility test screen: listening screen

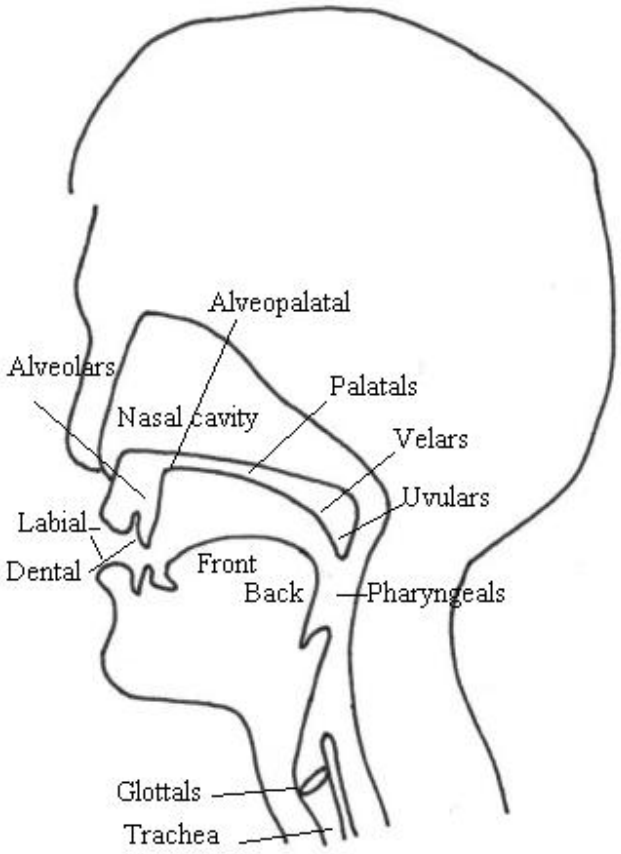


Figure A3 Human vocal tract

Résumé étendu en français

Contexte et résumé des contributions

Dans un monde de plus en plus globalisé, la capacité à communiquer en plusieurs langues donne beaucoup d'avantages aux locuteurs. Par ailleurs, le besoin d'apprentissage d'une nouvelle langue est rendu nécessaire par les migrations qui sont de plus en plus communes en particulier pour des raisons économiques. Aux États-Unis par exemple, 37,5 millions de personnes, soit près d'une personne sur cinq, était d'origine étrangère en 2006 [Ohlemacher 2007]. Parallèlement à cela, l'activité touristique génère aussi d'importants mouvements de personnes : en France par exemple, 78 millions de touristes ont visités le pays en 2006 [LExpansion 2007].

Ces « locuteurs étrangers » (touristes, migrants) sont de plus en plus confrontés à l'utilisation de services vocaux interactifs, certains intégrant la reconnaissance vocale. Alors que la reconnaissance automatique de la parole atteint désormais des performances souvent satisfaisantes pour les locuteurs natifs, la performance de reconnaissance sur les locuteurs non natifs reste encore insuffisante. Ce problème est un frein au développement des services vocaux.

Cette thèse aborde les problèmes qui concernent la reconnaissance automatique de la parole pour des locuteurs non natifs. Des études montrent que la performance des systèmes de reconnaissance vocale est au moins deux fois plus faible pour des locuteurs non natifs. La parole des locuteurs non natifs a des caractéristiques différentes. Pour les apprenants d'une langue, c'est un défi de prononcer les nouveaux phonèmes qui n'existent pas dans leur langue maternelle. Ils ont donc tendance à emprunter les sons de leur langue maternelle. D'autre part, même pour les phonèmes similaires qui existent à la fois dans la langue cible et la langue maternelle du locuteur, les locuteurs non natifs peuvent avoir des difficultés à changer certaines habitudes d'articulation spécifiques à leur langue maternelle.

La reconnaissance automatique de la parole, dans le cadre de l'approche statistique, utilise trois modèles principaux, à savoir le modèle acoustique, le modèle de prononciation et le modèle de langage. Ces modèles sont créés en utilisant les approches fondées sur les données avec les données des locuteurs natifs uniquement dans la plupart des cas. En conséquence, il y a des disparités entre la parole non native et les modèles utilisés ce qui réduit le taux de reconnaissance par rapport à la parole native. L'obtention de parole non native pour une langue cible donnée (afin de créer un modèle acoustique qui modélise la parole non native) peut par ailleurs être longue et parfois irréalisable.

Dans cette thèse, nous nous sommes intéressés aux méthodes de modélisation non native qui peuvent être employées sous différentes contraintes de ressources. Nous proposons d'utiliser des ressources multilingues pour surmonter la difficulté d'obtenir la parole non native. Au cas où des phrases en parole non native sont disponibles, celles-ci peuvent également être exploitées.

La modélisation acoustique et la modélisation de prononciation pour la parole non native sont étudiées dans cette thèse. Sur le thème de la modélisation acoustique, nous examinons l'utilisation de ressources multilingues pour adapter le modèle acoustique de la langue cible. Les locuteurs non natifs réalisent parfois un transfert entre les unités phonétiques de leur langue maternelle et celles de la langue cible quand ils apprennent une nouvelle langue. En utilisant les ressources multilingues et cette information de transfert, nous pouvons représenter l'espace de la parole non native dans un espace acoustique multilingue (voir la Figure Ra). Selon le type de ressources multilingues (modèles acoustiques ou corpus) disponibles, différentes techniques sont proposées.

Concernant la modélisation de prononciation, nous revisitons deux approches conventionnelles de modélisation de prononciation. Ces approches sont modifiées pour être utilisées même en situation où peu de parole non native est disponible. Nous proposons également une technique originale de regroupement des locuteurs suivant leurs habitudes de prononciation. Cette approche peut être aussi utilisée pour l'adaptation du dictionnaire de prononciation. Un ensemble de locuteurs non natifs est représenté dans un « espace de prononciation » de faible dimension. Pour un locuteur inconnu, les variantes de prononciation de ce locuteur peuvent être estimées à partir de sa position dans l'espace de prononciation, estimée à partir de quelques phrases de ce locuteur.

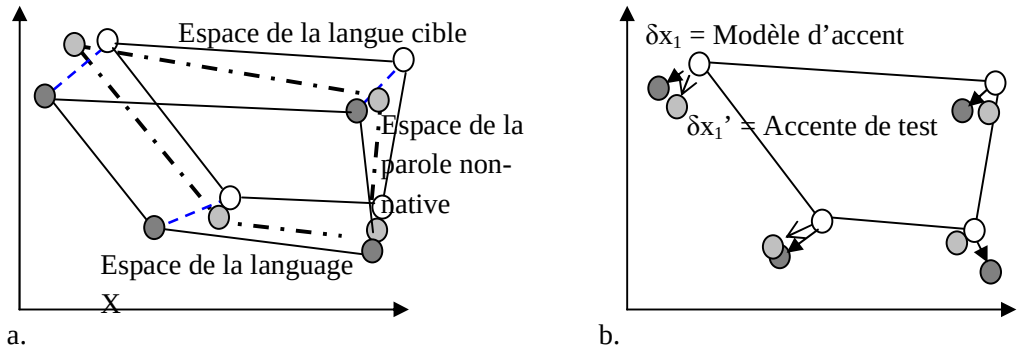


Figure R. Modélisation acoustique non native et modélisation d'accent en utilisant des ressources multilingues. a) Utilisation de l'espace de la langue cible et de la langue X (par exemple langue maternelle du locuteur) pour estimer l'espace de la parole non native en langue cible prononcée par un locuteur d'origine X. b) Utilisation du modèle de la langue X et quelques phrases non natives pour créer le modèle d'accent δx, qui servira de référence pour positionner le locuteur X' inconnu lors du test (=> modèle d'accent δx').

En plus de la modélisation acoustique et de prononciation, une étude préliminaire sur l'identification automatique d'accents est également proposée, à partir de ressources multilingues. Les ressources multilingues se sont révélées utiles pour l'amélioration des systèmes d'identification de la langue. Notre système d'identification d'accent utilise lui aussi des paramètres phonotactiques. Les ressources multilingues sont utilisées pour capturer la structure et le degré de changement avec la parole non native (voir la Figure Rb). Cette approche préliminaire donne des résultats encourageants lorsque peu de parole est disponible.

Organisation du mémoire

Ce mémoire débute par une brève introduction sur l'architecture et les composants d'un système de reconnaissance automatique de la parole. Une présentation sur le sujet de l'acquisition de la langue maternelle et de la langue seconde est ensuite donnée. Ces deux points nous permettent de comprendre les mauvaises performances des systèmes de reconnaissance vocale pour les locuteurs non natifs par rapport aux locuteurs natifs. La fin du chapitre est consacrée à l'état de l'art des approches dans le domaine de la reconnaissance automatique de la parole non native. Les techniques de modélisation acoustique, de prononciation et de langage pour la parole non native sont présentées, ainsi que les rares systèmes d'identification d'accent existant.

Le chapitre 2 présente nos propositions pour la modélisation acoustique non native en utilisant les ressources multilingues. En général, deux approches différentes sont proposées selon le type de ressources multilingues qui sont disponibles. Si les modèles acoustiques multilingues sont disponibles, l'approche hybride d'interpolation et de fusion peut être appliquée pour adapter le modèle cible. Par contre si les corpus multilingues sont à notre disposition, les approches par interpolation peuvent être utilisées. L'une des approches d'interpolation présentée peut être employée sans avoir du tout de parole non native.

Le chapitre 3 concerne nos travaux sur la modélisation de prononciation. Les approches de modélisation de prononciation conventionnelles comme l'approche par modification du dictionnaire de prononciation (en utilisant des arbres de décision) ou re-ordonnément de listes des n-best en utilisant un score de prononciation, sont revisitées. Nous proposons en plus une approche originale pour le regroupement de locuteurs que nous appelons « analyse de prononciation latente », par analogie avec l'analyse sémantique latente. Cette approche est utilisée pour l'adaptation du dictionnaire de prononciation. Après les travaux en modélisation de prononciation, nous présentons nos travaux préliminaires dans le domaine de l'identification d'accent. Nous décrivons une approche phonotactique qui utilise les ressources multilingues pour la création des modèles d'accent sous la forme d'arbres de décision.

Afin d'évaluer les approches proposés dans les chapitres précédents, nous décrivons, dans le chapitre 4, un corpus français non natif que nous avons enregistré. Avant d'utiliser ce corpus pour les tests, nous l'avons évalué en effectuant des tests de perception et une analyse acoustique.

Dans le chapitre 5, nos expériences sont présentées pour la modélisation acoustique, la modélisation de prononciation et le système préliminaire d'identification d'accent. Nous

définissons les conditions d'expérimentation : le moteur de reconnaissance automatique de la parole, les corpus multilingues d'apprentissage, le dictionnaire de prononciation et le modèle de langage utilisé. Les tests sont réalisés sur notre corpus de parole non native en français, mais des tests supplémentaires sont effectués sur de la parole anglaise non native. Nous concluons ce mémoire par une discussion de nos travaux et de leurs résultats et par une présentation des perspectives.

Principaux résultats et conclusions

Corpus de parole non native en français

Un corpus français non natif dans le domaine de tourisme a été acquis. Nous avons enregistré des phrases lues par des locuteurs vietnamiens et chinois. L'analyse du corpus (par des méthodes d'analyses automatiques ou manuelles) montre que le transfert des unités phonétiques de la langue maternelle à la langue seconde est évident pour les locuteurs non natifs. Ces observations et aussi les résultats d'autres études linguistiques montrent que les locuteurs non natifs ont tendance à transférer le phonème de leur langue maternelle (L1) qui est le plus proche du phonème cible (selon le tableau API). Cette information est utile pour la modélisation acoustique non native.

Modélisation acoustique multilingue pour la reconnaissance automatique de la parole non native

Les approches de modélisation acoustique non native peuvent généralement être groupées en quatre catégories : la reconstruction de modèles acoustiques, l'interpolation de modèles acoustiques, la fusion de modèles acoustiques et les techniques d'adaptation au locuteur. Ces approches utilisent soit la langue maternelle du locuteur, soit une petite quantité de parole non native pour adapter le modèle acoustique cible. Nous avons utilisé, dans nos travaux, des ressources multilingues pour adapter le modèle acoustique cible. Trois types de ressources ont été examinées : des données correspondant à la langue maternelle du locuteur (L1), des données de parole non native, mais dans une langue différente de la langue cible (L2), et des données correspondant à une langue proche de la langue maternelle du locuteur (L3).

Nous avons comparé deux différentes approches pour obtenir le transfert de phonèmes entre langues source et cible (méthodes fondées sur les connaissances, et méthodes automatiques utilisant un décodeur acoustico-phonétique). Ces méthodes se sont révélées équivalentes pour obtenir le transfert de phonèmes source / cible. Une fois que le phonème correspondant en langue source pour chaque phonème de la langue cible est déterminé, l'adaptation du modèle acoustique standard en langue cible peut être effectuée. Selon le type de ressources multilingues disponibles, différentes techniques de modélisation acoustique ont été proposées. Si les modèles acoustiques multilingues sont disponibles, une approche hybride d'interpolation et de fusion s'est révélée utile pour la modélisation acoustique non native. L'idée principale consiste à interpoler les distributions Gaussiennes cible et source qui sont proches, et à les fusionner si elles sont

éloignées les unes des autres. Le concept est similaire par rapport aux hypothèses sur la parole non native de Fledge, qui spécifient que les locuteurs non natifs peuvent utiliser les sons de leur langue maternelle, les sons en langue cible, ou les sons intermédiaires entre la langue maternelle et la langue cible, en fonction de leur expérience. Nos résultats montrent que l'approche hybride proposée est plus performante que l'approche conventionnelle d'interpolation ou de fusion.

Nous évaluons également quelques approches d'interpolation pour profiter de données multilingues pour adapter le modèle acoustique cible. Trois types d'approche d'interpolation ont été proposés pour l'adaptation de modèle acoustique sous différentes contraintes. Au cas où seulement les corpus multilingues sont disponibles, une interpolation manuelle peut être effectuée par la prévision du poids a priori à attribuer. Avec quelques phrases non natives du locuteur, les poids d'interpolation peuvent être estimés avec la formule des moindres carrés pondérés. Enfin, nous montrons que les résultats de l'approche par *eigenvoices* pour les locuteurs non natifs peuvent être améliorés par la création d'un espace d'adaptation propre bi-lingue (langue source + langue cible). L'interpolation dans ce cas est réalisée en utilisant des vecteurs propres. Nos résultats montrent que parmi les trois types de ressources que nous avons proposées, la ressource L2 même si elle est d'une langue différente de la langue cible, est utile pour l'adaptation des modèles acoustiques à la parole non native, et comparable à ressource L1. La ressource L3 d'autre part peut améliorer le modèle acoustique cible, mais pas aussi bien que les ressources L1 et L2. Les approches d'interpolation proposées sont comparables en performance à l'approche hybride, mais les approches par interpolation conduisent à des modèles de moindre complexité (moins de distributions gaussiennes). En général, les deux approches sont meilleures que l'approche conventionnelle d'adaptation au locuteur type MLLR (testée en mode supervisé). Lorsque ces approches sont combinées avec MLLR, le taux d'erreur est encore amélioré.

Les études existantes n'abordent pas en profondeur la question qui concerne la modélisation du contexte phonétique pour les locuteurs non natifs. L'approche hybride d'interpolation et de fusion que nous avons proposée précédemment peut être appliquée aussi pour la modélisation de contexte (interpolation/fusion entre le modèle indépendant du contexte et le modèle dépendant du contexte). Les résultats montrent que, en accordant un poids moyen, l'approche donne un meilleur résultat par rapport à la mise en commun des états (state-tying) ou à l'augmentation du nombre de distributions gaussiennes par état. En plus, l'amélioration des taux erreur des locuteurs non natifs est obtenue avec seulement une légère réduction des taux d'erreur pour les locuteurs natifs.

Modélisation de la prononciation et reconnaissance d'accents

Les approches de modélisation de prononciation peuvent être divisées en fonction du composant du système où les variantes de prononciation sont modélisées. Les composants possibles sont le dictionnaire de prononciation, le modèle acoustique, le modèle de langage, et le module de re-ordonnément de liste n-best (*n-best rescoring*). Les études existantes montrent que l'utilisation seulement de parole en langue maternelle pour la modélisation des variantes dans le dictionnaire de prononciation ne produit pas une amélioration significative. Par conséquent, nous avons examiné et modifié deux approches qui modélisent les variantes dans le dictionnaire de

prononciation et le module de réévaluation, pour qu'elles puissent traiter de la parole non native en faible quantité. Nos expériences montrent que la modélisation des variantes dans le dictionnaire de prononciation (l'approche par arbres de décision) produit des meilleurs résultats que la modélisation effectuée dans le module de réévaluation. En outre, l'approche de réévaluation des *n-best* nécessite plus de traitement que l'approche conventionnelle avec le dictionnaire de prononciation. Nous proposons également une approche de regroupement des locuteurs selon leurs habitudes de prononciation. Nous nommons cette approche « analyse de prononciation latente », par analogie avec l'analyse sémantique latente. Les vecteurs propres sont dérivés de supervecteurs créés à partir d'arbres de décision qui modélisent la prononciation de chaque locuteur. Nos résultats montrent que les locuteurs non natifs peuvent être regroupés en fonction de leur origine en utilisant cette analyse latente. La même approche peut également être utilisée pour l'adaptation de prononciation étant données quelques phrases non natives. Les résultats montrent que l'analyse de prononciation latente est capable de prédire des variantes de prononciation lorsque l'accent du locuteur n'est pas connu à l'avance.

Nous proposons aussi dans cette thèse une technique originale qui utilise les arbres multilingues de décision pour identifier l'accent du locuteur. C'est cependant une approche dépendante du texte qui nécessite la transcription du signal de parole servant à identifier l'accent. L'approche proposée utilise des ressources multilingues pour créer les modèles d'accent sous la forme d'arbres de décision qui ont montré leur potentiel pour généraliser les observations par exemple dans la modélisation de prononciation. Les résultats montrent que l'approche proposée est plus efficace par rapport aux approches « état de l'art » dans le cas où la quantité de parole non native disponible pour l'entraînement est limitée.

Quelques perspectives

Nous avons proposé d'utiliser des ressources multilingues pour la modélisation acoustique non native. Nos expériences montrent que trois types de données de parole sont utiles pour adapter le modèle non natif. Ce sont les données correspondant à la langue maternelle du locuteur (L1), des données de parole non native prononcées par des locuteurs de même origine, mais dans des langues différentes de la langue cible (L2), et des données correspondant à une langue proche de la langue maternelle du locuteur (L3). Dans ce manuscrit, nous n'étudions pas les caractéristiques qui déterminent la proximité des langues. Une analyse approfondie dans ce sujet sera certainement utile pour trouver davantage de ressources multilingues utiles pour la modélisation non native. De plus, l'adaptation acoustique multilingue est faite en trouvant pour chaque phonème de la langue cible avec un phonème correspondant de la langue source. En fait, les ressources multilingues peuvent être employées d'une différente manière qui est plus intelligente pour créer des modèles non natifs plus performants. De multiples phones de différents corpus multilingues peuvent être utilisés pour adapter un phone d'une langue cible. Une approche qui prend en considération le type, le contexte et la quantité de la parole dans chaque corpus source pourrait certainement améliorer davantage le modèle acoustique pour les locuteurs non natifs. Si quelques phrases non natives sont disponibles, elles peuvent probablement être utilisées pour sélectionner les meilleures phones des ressources multilingues pour l'adaptation, en utilisant des

mesures de distance par exemple la distance de HMM [Juang 1985], l'approche PDTS (Polyphone Decision Tree Specialization) [Schultz 2000] et d'autres qui sont proposées pour la modélisation acoustique multilingue.

L'approche hybride d'interpolation et de fusion est une méthode prometteuse pour modéliser le transfert translingue et le contexte pour des locuteurs non natifs. La performance de l'approche dépend du poids à priori qui est assigné. Nous n'avons pas proposé une méthode automatique pour estimer le poids de modélisation hybride. Une solution simple qui peut être appliquée si nous avions quelques phrases du locuteur, est de faire des alignements forcés avec les modèles acoustiques pré-adaptés (de l'approche hybride), et de mesurer le score acoustique. Le modèle qui donne le score le plus élevé sera choisi. Toutefois, une méthode plus souple qui peut estimer automatiquement les poids serait préférable.

En raison de la difficulté à acquérir la parole non native spontanée, nous n'abordons pas le sujet de l'adaptation du modèle de langage. Les études montrent que les locuteurs non natifs sont susceptibles de transférer leur vocabulaire et leur grammaire native à la langue cible. Il serait intéressant de voir s'il est possible d'utiliser la langue maternelle du locuteur pour adapter le modèle de langage. Les modèles de langage factorisés (*Factored language models*), qui sont une extension des modèles de langage n-gramme classiques, pourraient être utiles pour cela. La classe lexique ou sémantique peut être attribuée aux mots dans la langue cible et source. Les trigrammes de classe lexique qui représentent les règles de grammaire, et les vocabulaires de langue maternelle du locuteur qui ont les graphèmes similaires par rapport aux vocabulaires de la langue cible pourraient être transférés à la langue cible, par exemple en interpolant les modèles de langage cible et source.

L'alternance codique (code switching) est de plus en plus courante parmi les locuteurs de nos jours. Il s'agit d'une alternance de deux ou plusieurs langues ou dialectes dans une même conversation. Généralement, les personnes impliquées connaissent les langues ou dialectes utilisés. Dans des discours qui concernent l'alternance codique, les études montrent que 84% des cas impliquent l'alternant d'un seul mot, 10% alternant d'une phrase, et 6% alternant d'une proposition (clause) [Skiba 1997]. L'alternance codique est utilisée comme une stratégie par les locuteurs pour surmonter la difficulté à exprimer une idée en langue courante. En conséquence, il doit passer temporairement à une autre langue. Par exemple, dans le domaine de la science et de la technologie, les locuteurs peuvent avoir la difficulté à présenter clairement les termes ou les idées en langue maternelle, donc ils peuvent passer à l'anglais pour l'exprimer. L'alternance codique peut également se produire pour des raisons sociales, par exemple pour s'identifier comme appartenant à un groupe particulier. Par exemple, l'anglais singapourien (Singlish, dialecte anglais parlé à Singapour) est un mélange d'anglais, de malais, de minnan, de teochew et de cantonais. Il est associé à l'identité des Singapouriens. Enfin, l'alternance codique peut également se produire lorsque les locuteurs impliqués veulent limiter une partie de la conversation à un groupe particulier. L'alternance codique est une difficulté supplémentaire pour les systèmes de reconnaissance automatique de la parole. Un système d'identification de la langue typique aura la difficulté parce que la période de changement n'est pas connue et la durée d'alternance est très courte puisqu'elle peut concerner seulement un mot. Par conséquent, la

combinaison de systèmes de reconnaissance vocale en plusieurs langues, avec un système d'identification de la langue pour traiter l'alternance codique n'est pas une bonne solution. Une autre possibilité consiste à traiter l'alternance codique comme un système indépendant. Cela signifie que les modèles cibles contiennent les unités acoustiques, les vocabulaires et la grammaire pour toutes les langues impliquées. En terme de modélisation acoustique, la question intéressante est ce qui concerne la modélisation des différentes unités acoustiques. Nos approches proposées pour la modélisation acoustique peuvent être utiles. Les phonèmes similaires de différentes langues sont modélisés une fois seulement dans un modèle acoustique. Toutefois, les nouveaux phonèmes qui n'existent pas dans la langue cible doivent être adaptés et ajoutés. Les mots des langues impliqués doivent être ajoutés au dictionnaire de prononciation et pris en compte dans le modèle de langage cible. Une étude approfondie des habitudes du locuteur à propos de l'alternance codique peut être également nécessaire. En terme de modèle de langage, il est difficile d'adapter le modèle de langage cible puisque l'alternance codique est observable seulement en conversation, mais pas dans les textes écrits. L'alternance codique n'est pas un processus aléatoire et elle suit certaines règles ou contraintes. Deux contraintes limitent l'alternance chez les locuteurs [Skiba 1997]. Premièrement, les contraintes articulatoires impliquent que le locuteur alterne avec les mots de la langue qui ont une certaine forme similaire à la langue cible. Deuxièmement, une autre contrainte indique que l'alternance codique n'est possible que si elle ne viole pas les grammaires des deux langues. Il serait intéressant de savoir comment ces règles linguistiques peuvent être combinées avec les modèles de langage n-gramme. L'alternance codique est donc un nouveau territoire qui est intéressant et mériterait d'être étudié.

La mesure de confiance pour la parole non native est aussi un sujet intéressant. L'objectif de la mesure de confiance est d'évaluer la qualité du décodage. Dans le domaine de l'apprentissage des langues assistées, la mesure donne aux apprenants une idée de la qualité de leur prononciation. La plupart des techniques existantes sont construites pour analyser la parole native. Des tests doivent être effectués pour savoir si ces approches peuvent être utilisés pour la parole non native. Le système doit être capable d'analyser et de comparer la prononciation des locuteurs non natifs à différents niveaux, par exemple les phonèmes, les syllabes, les mots et les phrases, pour les aider à connaître les types d'erreurs qu'ils font souvent. En plus, la mesure de confiance peut être intégré dans des systèmes de reconnaissance automatique de la parole non native. Pour les décodages en dessous du seuil de mesure, un traitement supplémentaire peut être effectué, ou le locuteur peut répéter ce qu'il a dit. Cela permettrait d'améliorer la performance du système de reconnaissance de parole non native.

Personal Publications

[Tan 2006] Tan, T.-P., and Besacier, L., 2006, A French Non-Native Corpus for Automatic Speech Recognition, LREC'06: Genova, p. 1610-1613.

[Tan 2007a] Tan, T.-P., and Besacier, L., 2007, Acoustic Model Interpolation for Non-Native Speech Recognition, ICASSP'07: Hawaii, p. 1009-1012.

[Tan 2007b] Tan, T.-P., and Besacier, L., 2007, Modeling Context and Language Variation for Non-Native Speech Recognition, Interspeech'07: Antwerp, p. 1429-1432.

[Tan 2008a] Tan, T.-P., and Besacier, L., 2008, Modélisation acoustique multilingue pour la reconnaissance automatique de la parole non native, JEP'08: Avignon.

[Tan 2008b] Tan, T.-P., and Besacier, L., 2008, Improving Pronunciation Modeling for Non-Native Speech Recognition, Interspeech'08: Brisbane.

References

[Altmann 2006] Altmann, H., 2006, The Perception and Production of Second Language Stress: A Cross Linguistic Experimental Study: PhD Dissertation, University of Delaware.

[Arslan 1996] Arslan, L. M., and Hansen, H. H. L., 1996, Language Accent Classification in American English: *Speech Communication*, vol. 18, p. 353-367.

[Barkova 2004] Bartkova, K., and Jouvet, D., 2004, Multiple Models for Improved Speech Recognition for Non-Native Speakers, SPECOM'04: St. Petersburg.

[Béchet 2005] Béchet, F., 2005, LIA_PHON: Avignon, Laboratoire Informatique d'Avignon.

[Berkling 1998] Berkling, K., Zissman, M., Vonwiller, J., and Cleirigh, C., 1998, Improving Accent Identification Through Knowledge of English Syllable Structure, ICSLP'98: Sydney, p. 89-92.

[Besacier 2001] Besacier, L., Blanchon, H., Fouquet, Y., Guilbaud, J. P., Helme, S., Mazenot, S., Moraru, D., and Vaufreydaz, D., 2001, Speech Translation for French in the NESPOLE! European Project, Eurospeech'01: Aalborg, Denmark, p. 1291-1294.

[Blackburn 1993] Blackburn, C. S., Vonwiller, J. P., and King, R. W., 1993, Automatic Accent Classification Using Artificial Neural Networks, Eurospeech'93: Berlin, p. 1241-1244.

[Boersma 2007] Boersma, P., and Weenink, D., 2007, Praat: doing phonetics by computer (Version 5.0.02).

[Bohn 1995] Bohn, O.-S., 1995, Cross-Language Speech Perception in Adult First Language Transfer Doesn't Tell It All, *in* Strange, W., editor, *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, Baltimore: York Press, p. 279-303.

[Bonastre 2005] Bonastre, J.-F., and Meignier, S., 2005, ALIZE, A Free Toolkit for Speaker Recognition, ICASSP'05: Philadelphia, p. 737-740.

[Bouselmi 2005] Bouselmi, G., Fohr, D., and Haton, J.-P., 2005, Fully Automated Non-Native Speech Recognition Using Confusion -Based Acoustic Model Integration, Eurospeech'05: Lisboa, p. 1369-1372.

[Bouselmi 2006] Bouselmi, G., Fohr, D., Illina, I., and Haton, J.-P., 2006, Multilingual Non-Native Speech Recognition using Phonetic Confusion-Based Acoustic Model Modification and Graphemic Constraints, ICSLP'06: Pittsburgh, p. 109-112.

[Chen 2001] Chen, T., Huang, C., Chang, E., and Wang, J., 2001, Automatic Accent Identification Using Gaussian Mixture Models, ASRU'01: Madonna di Campiglio, Trento, p. 343-346.

[Choueiter 2008] Choueiter, G., Zweig, G., and Nguyen, P., 2008, An Empirical Study of Automatic Accent Identification, ICASSP'08: Las Vegas, p. 4265-4268.

[Clarkson 1997] Clarkson, P., and Rosenfeld, R., 1997, Statistical Language Modeling Using the CMU-Cambridge Toolkit, Eurospeech'07: Greece, p. 2707-2710.

[Compernolle 2001] Compernolle, D. V., 2001, Recognizing Speech of Goats, Wolves, Sheep and ... Non-Natives: *Speech Communication*, vol. 35, p. 71-79.

[CCC 2005] Chinese Corpus Consortium, 2005, CCC Resources : Online, Chinese Corpus Consortium, http://www.d-ear.com/CCC/resources.htm.

[CMU 2000] CMU, 2000, SphinxTrain Documentation: Online, CMU, http://www.speech.cs.cmu.edu/sphinxman

[Davis 1980] Davis, S. B., and Mermelstein, P., 1980, Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences: *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, p. 357-366.

[Deng 2006] Deng, Y., X., L., Kwan, C., Xu, R., Raj, B., and Williamson, D., 2006, An Integrated Approach to Improve Speech Recognition Rate for Non-Natives Speakers, ICSLP'06: Pittsburgh, p. 1734-1737.

[Duanmu 2002] Duanmu, S., 2002, *The Phonology of Standard Chinese*: New York, Oxford University Press.

[Duda 2000] Duda, R. O., Hart, P. E., and Stork, D. G., 2000, *Pattern Classification*: New York, Wiley-Interscience.

[Fisher 1986] Fisher, W. M., Doddington, G. R., and Goudie-Marshall, K. M., 1986, The DARPA Speech Recognition Research Database: Specifications and Status, Proceedings of DARPA Workshop on

Speech Recognition, p. 93-99.

[Flege 1987] Flege, J., 1987, The Production of 'New' and 'Similar' Phones in a Foreign Language: Evidence for the Effect of Equivalence Classification*: Journal of Phonetics*, vol. 15, p. 47-65.

[Flege 1995] Flege, J., 1995, Second Language Speech Learning Theory, Findings, and Problems, *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, Baltimore: York Press, p. 233-277.

[Flege 1997] Flege, J., Frieda, E., and Nozawa, T., 1997, Amount of Native-Language (L1) Use Affects the Pronunciation of an L2: *Journal of Phonetics*, vol. 25, p. 169-186.

[Flege 2004] Flege, J., and MacKay, I. R. A., 2004, Perceiving Vowels in a Second Language: *Studies in Second Language Acquisition*, vol. 26, p. 1-34.

[Flury 1988] Flury, B., and Riedwyl, H., 1988, *Multivariate Statistics: A Practical Approach*: London, Chapman and Hall.

[Gauvain 1994] Gauvain, J.-L., and Lee, C.-H., 1994, Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains: *IEEE Transactions on Speech and Audio Processing*, vol. 2, p. 291-298.

[Ghesquiere 2002] Ghesquiere, P.-J., and Compernolle, D. V., 2002, Feature Subset Selection for Flemish Accent Identification, SPS'02: Leuven, Belgium, p. S02-1-S02-4.

[Goronzy 2001a] Goronzy, S., Kompe, R., and Rapp, S., 2001, Generating Non-Native Pronunciation Variants for Lexicon Adaptation, ISCA'01: Sophia Antipolis, France, p. 143-146.

[Goronzy 2001b] Goronzy, S., Sahakyan, M., and Wokurek, W., 2001, Is Non-Native Pronunciation Modeling Necessary?, Eurospeech'01: Aalborgh, Denmark, p. 305-308.

[Goronzy 2002] Goronzy, S., 2002, *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition*: Berlin, Springer Verlag.

[Gravier 2004] Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., and Choukri, K., 2004, ESTER, Une Campagne D'évaluation Des Systèmes D'indexation Automatique D'émissions Radiophoniques En Français, JEP'04: Maroc, p. 253-256.

[Gruhn 2004] Gruhn, R., Markov, K., and Nakamura, S., 2004, A Statistical Lexicon for Non-Native Speech Recognition, ICSLP'04: South Korea.

[Howard 2000] Howard, A., 2000, *Elementary Linear Algebra*: New York, John Wiley & Sons.

[Huang 2001] Huang, X., Acero, A., and Hon, H.-W., 2001, *Spoken Language Processing: A Guide to Theory, Algoirthm, and System Development*: New Jersey, Prentice Hall PTR.

[Humpries 1996] Humpries, J. J., Woodland, P. C., and Pearce, D., 1996, Using Accent-specific Pronunciation Modelling for Robust Speech Recognition, ICSLP'96: Philadelphia, p. 2324-2327.

[Humpries 1997] Humpries, J. J., and Woodland, P. C., 1997, Using Accent-Specific Pronunciation Modelling for Improved Large Vocabulary Continuous Speech Recognition, Eurospeech'97: Rhodes, Greece, p. 2367-2370.

[Itahashi 1992] Itahashi, S., and Yamashita, T., 1992, A Discrimination Method Between Japanese Dialects, ISCA'92: Banff, Alberta, p. 1015-1018.

[Iverson 2003] Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Yoh'ich, o., Kettermann, A., and Siebert, C., 2003, A Perceptual Interference Account of Acquisition Difficulties for Non-Native Phonemes: Cognition, vol. 87, p. B47-B57.

[Jelinek 2001] Jelinek, F., 2001, *Statistical Methods for Speech Recognition*: London, MIT Press.

[Juang 1985] Juang, B.-H., and Rabiner, L. R., 1985, A Probabilistic Distance Measure for Hidden Markov Models: AT&T Technical Journal, vol. 64, p. 391-408.

[Kent 2002] Kent, R. D., and Read, C., 2002, *The Acoustic Analysis of Speech*: Canada, Singular Thomson Learning.

[Killer 2003a] Killer, M., 2003, Grapheme Based Speech Recognition: Master Theses, Pittsburgh, Carnegie Mellon.

[Killer 2003b] Killer, M., Stuker, S., and Schultz, T., 2003, Grapheme Based Speech Recognition, Eurospeech'03: Geneva, p. 3141-3144.

[Kim 1997] Kim, K. H. S., Relkin, N. R., Lee, K.-M., and Hirsch, J., 1997, Distinct Cortical Areas Associated with Native and Second Languages: *Nature*, vol. 388, p. 171-174.

[Kim 2003] Kim, W., and Khudanpur, S., 2003, Language Model Adaptation Using Cross-Lingual Information, Eurospeech'03: Geneva, p. 3129-3132.

[Kuhl 2000] Kuhl, P. K., 2000, A New View of Language Acquisition: *Proceedings of the National Academy of Science*, p. 11850-11857.

[Kuhl 1997] Kuhl, P. K., Andruski, J. E., Christovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, vol. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F., 1997, Cross-Language Analysis of Phonetic Units in Language Addressed to Infants: *Science*, vol. 277, p. 684-

686.

[Kuhn 1998a] Kuhn, R., Nguyen, P., Goldwasser, L., Niedzielski, N., Fincke, S., and Contolini, M., 1998, Eigenvoices for Speaker Adaptation, ICSLP'98: Sydney, Australia, p. 1774-1777.

[Kuhn 1998b] Kuhn, R., Nguyen, P., Junqua, J. C., and Goldwasser, L., 1998, Eigenfaces and Eigenvoices: *Dimensionality Reduction for Specialized Pattern Recognition*, p. 71-76.

[Kuhn 1999] Kuhn, R., Nguyen, P., Junqua, J. C., Boman, R., Niedzielski, N., Fincke, S., and Contolini, M., 1999, Fast Speaker Adaptation Using A Priori Knowledge, ICASSP'99: Phoenix, USA, p. 749-752.

[Kumpf 1996] Kumpf, K., and King, R. W., 1996, Automatic Accent Classification of Foreign Accented Australian English Speech, ICSLP'96: Philadelphia, p. 1740-1743.

[Kumpf 1997] Kumpf, K., and King, R. W., 1997, Foreign Speaker Accent Classification Using Phoneme-Dependent Accent Discrimination Models and Comparisons with Human Perception Benchmarks, Eurospeech'97: Rhodes, Greece, p. 2323-2326.

[Ladefoged 2000] Ladefoged, P., 2000, *Vowels and Consonants*: Los Angelas, Blackwell Publishing.

[Lamel 1991] Lamel, L. F., Gauvain, J. L., and M., E., 1991, BREF, a Large Vocabulary Spoken Corpus for French, Eurospeech'91: Genoa, p. 505-508.

[Le 2004] Le, V.-B., Do-Dat, T., Casteli, E., Besacier, L., and Serignat, J. F., 2004, Spoken and written language resources for Vietnamese LREC'04: Lisbon, p. 599-602.

[Le 2005] Le, V.-B., and Besacier, L., 2005, First Steps in Fast Acoustic Modeling for a New Target Language: Application to Vietnamese, ICASSP'05: Philadelphia, USA, p. 821-824.

[Le 2006] Le, V.-B., 2006, Reconnaissance automatique de la parole pour des langues peu dotées: PhD Dissertation, Groupe d'Etude sur l'Oral et le Dialogue, Université Joseph Fourier, Grenoble.

[Leggetter 1995] Leggetter, C. J., and Woodland, P. C., 1995, Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models: *Computer Speech and Language*, vol. 9, p. 171-185.

[LExpansion 2007] LExpansion, 2007, Le nombre de touristes étrangers en France en 2006, L'Expansion: Paris.

[Li 1998] Li, F., and Yan, Y., 1998, Correlation Generated Targets for Neural Network Speech Recognition, ICSP'98: Beijing, p. 718-721.

[Liu 1999] Liu, W.-K., and Fung, P., 1999, Fast Accent Identification and Accented Speech Recognition, ICASSP'99: Phoenix, Arizona, p. 221-224.

[Liu 2003a] Liu, Y., and Fung, P., 2003, Modeling partial pronunciation variations for spontaneous Mandarin speech recognition: *Computer Speech and Language*, vol. 17, p. 357-379.

[Liu 2003b] Liu, Y., and Fung, P., 2003, Partial Change Accent Change Models for Accented Mandarin Speech Recognition, ASRU'03: St. Thomas, U.S. Virgin Islands, p. 111-113.

[Liu 2006] Liu, Y., and Fung, P., 2006, Multi-Accent Chinese Speech Recognition, ICSLP'06: Pittsburgh, p. 133-136.

[Logan 1991] Logan, J. S., Lively, S. E., and Pisoni, D. B., 1991, Training Japanese Listeners to Identify English /r/ and /l/: A First Report: *Journal of Acoustical Society of America*, vol. 89, p. 874-886.

[Mennen 1998] Mennen, I., 1998, Can Language Learners Ever Acquire the Intonation of a Second Lanugage?, STILL: Marholmen, Sweden, p. 17-20.

[Mennen 2004] Mennen, I., 2004, Bi-directional Interference in the Intonation of Dutch Speakers of Greek: Journal of Phonetics, vol. 32, p. 543-563.

[Mennen 2006] Mennen, I., 2006, Phonetic and Phonological Influences in Non-Native Intonation: An Overview for Language Teachers: Edinburgh, Queen Margaret University College.

[Menzel 2000] Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., and Souter, C., 2000, The ISLE Corpus of Non-Native Spoken English, LREC'00: Athens, p. 957-963.

[Minematsu 2003] Minematsu, N., Osaki, K., and Hirose, K., 2003, Improvement of Non-native Speech Recognition by Effectively Modeling Frequently Observed Pronunciation Habits, Eurospeech'03: Geneva, p. 2597-2600.

[Montgomery 2001] Montgomery, D. C., Peck, E. A., and Geoffrey Vining, G., 2001, *Introduction to Linear Regression Analysis*, 3rd Edition, Wiley.

[Morgan 2004] Morgan, J. J., 2004, Making a Speech Recognizer Tolerate Non-Native Speech through Gaussian Mixture Merging, ICALL'04: Venice, Italy.

[Nguyen 1998] Nguyen, P., 1998, Fast Speaker Adaptation: Technical Report, Institut Eurécom.

[Nock 1998] Nock, H. J., and Young, S. J., 1998, Detecting and Correcting Poor Pronunciations for Multiword Units, Modeling Pronunciation Variation for Automatic Speech Recognition: Rolduc, The Netherlands, p. 85-90.

[O'Grady 2000] O'Grady, W., and Archibald, J., 2000, *Contemporary Linguistic Analysis: An Introduction*: Toronto, Addison Wesley Longman.

[Oh 2006] Oh, Y. R., Yoon, J. S., and Kim, H. K., 2006, Acoustic Model Adaptation based on Pronunciation Variability Analysis for Non-Native Speech Recognition, ICASSP'06: Toulouse, France, p. I-137 - I-140.

[Ohlemacher 2007] Ohlemacher, S., 2007, Number of Immigrants Hits Record 37.5M, Washington Post: Washington.

[Ou 2007] Ou, S.-C., 2007, Linguistic Factors in L2 Word Stress Acquisition: A Comparison of Chinese and Vietnamese EFL Learners' Development, ICPhS'07: Saarbrücken, p. 1681-1684.

[Pierrehumbert 1980] Pierrehumbert, J., 1980, The Phonology and Phonetics of English Intonation, Linguistics and Philosophy: Cambridge, Massachusetts Institute of Technology.

[Quinlan 1993] Quinlan, J. R., 1993, *C4.5: Programs for Machine Learning*: California, Morgan Kaufmann.

[Raab 2007] Raab, M., Gruhn, R., and Noeth, E., 2007, Non-Native Speech Databases, ASRU'07: Kyoto, p. 413-418.

[Rabiner 1993] Rabiner, L. R., and Juang, B.-H., 1993, *Fundamental of Speech Recognition*: New Jersey, Prentice Hall PTR.

[Raux 2004] Raux, A., 2004, Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition, ICSLP'04: Jeju Island, Korea, p. 613-616.

[Ravishankar 2006] Ravishankar, M. K., 2006, Sphinx3 Decoders: Online, http://cmusphinx.sourceforge.net/sphinx3/doc/s3_overview.html.

[Rochet 1995] Rochet, B. L., 1995, Perception and Production of Second-Language Speech Sounds by Adults, *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, Baltimore: York Press, p. 380-410.

[Schultz 1998] Schultz, T., and Waibel, A., 1998, Language Independent and Language Adaptive Large Vocabulary Speech Recognition, ICSLP'98: Sydney, p. 1819-1821.

[Schultz 2000] Schultz, T., and Waibel, A., 2000, Polyphone Decision Tree Specialization for Language Adaptation, ICASSP'00: Istanbul, p. 1707-1710.

[Schultz 2002] Schultz, T., Jin, Q., Laskowski, K., Tribble, A., and Waibel, A., 2002, Speaker, Accent, and Language Identification. using Multilingual Phone Strings, HLT-2002: San Diego.

[ScienceDaily 2008] ScienceDaily, 2008, Linguist Tunes In To Pitch Processing In Brain: Online, ScienceDaily, http://www.sciencedaily.com/releases/2008/02/080216114856.htm.

[Skiba 1997] Skiba, R., 1997, Code Switching as a Countenance of Language Interference: Online, The Internet TESL Journal, http://iteslj.org/Articles/Skiba-CodeSwitching.html

[Steidl 2004] Steidl, S., Stemmer, G., Hacker, C., and Nöth, E., 2004, Adaptation in the Pronunciation Space for Non-Native Speech Recognition, ICSLP'04: South Korea, p. 2901-2904.

[Strik 1998] Strik, H., and Cucchiarini, C., 1998, Modeling pronunciation variation for ASR: overview and comparison of methods, ESCA workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition: Rolduc, p. 137-144.

[Strik 1999] Strik, H., and Cucchiarini, C., 1999, Modeling Pronunciation Variation for ASR: A Survey of the Literature*: Speech Communication*, vol. 29, p. 225-246.

[Taylor 1998] Taylor, P. A., Black, A., and Caley, R., 1998, The Architecture of the Festival Speech Synthesis System, The Third ESCA Workshop in Speech Synthesis: Jenolan Caves, Australia, p. 147-151.

[Tomokiyo 2000] Tomokiyo, L. M., 2000, Linguistic Properties of Non-Native Speech, ICASSP'00: Istanbul, Turkey, p. 1335-1338.

[Tomokiyo 2001] Tomokiyo, L. M., and Waibel, A., 2001, Adaptation Methods for Non-Native Speech, Workshop on Multilinguality in Spoken Language Processing: Aalborg.

[Turk 1991] Turk, M. A., and Pentland, A. P., 1991, Eigenfaces for Recognition: *Journal of Cognitive Neuroscience*.

[Tychtl 1999] Tychtl, Z. k., and Psutka, J., 1999, Speech Production Based on the Mel-Frequency Cepstral Coefficients, Eurospeech'99: Budapest, p. 2335-2338.

[Uebler 1999] Uebler, U., and Boros, M., 1999, Recognition of Non-native German Speech with Multilingual Recognizers, Eurospeech'99: Budapest, p. 911-913.

[Ullakonoja 2007] Ullakonoja, R., 2007, Comparison of Pitch Range in Finnish (L1) and Russian (L2), ICPhS'07: Saarbrücken, p. 1701-1704.

[Ueyama 1996] Ueyama, M., and Jun, S.-A., 1996, Focus Realization in Japanese English and Korean English Intonation, Japanese Korean Linguistics Conference: Los Angeles, p. 629-645.

[Vaufreydaz 2000] Vaufreydaz, D., Bergamini, J., Serignat, J. F., Besacier, L., and Akbar, M., 2000, A New Methodology for Speech Corpora Definition from Internet Documents, LREC'00: Athens, Greece, p. 423-426.

[Wang 2003a] Wang, Z., and Schultz, T., 2003, Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization, Eurospeech'03: Geneva, Switzerland, p. 1449-1452.

[Wang 2003b] Wang, Z., Schultz, T., and Waibel, A., 2003, Comparison of Acoustic Model Adaptation Techniques on Non-native Speech, ICASSP'03: Hong Kong, China, p. 540-543.

[Wayland 2006] Wayland, R., Guion, S. G., and Landfair, D., 2006, Native Thai Speakers' Acquisition of English Word Stress Patterns: *Journal of Psycholinguistic Research*, v. 35, p. 285-304.

[Weikum 2007] Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastian-Gallés, N., and Werker, J. F., 2007, Visual Language Discrimination in Infancy: *Science*, vol. 316, p. 1159-1159.

[Weinberger 2007] Weinberger, S. H., 2007, The Speech Accent Archive: Online, Virginia, George Mason University, http://accent.gmu.edu/.

[Westwood 1999] Westwood, R. J., 1999, Speaker Adaptation Using Eigenvoices, University of Cambridge.

[WelFord 2007] WilFord, J. N., 2007, Languages Die, but Not Their Last Words: Online, The New York Times, New York, http://www.nytimes.com/2007/09/19/science/19language.html

[Witt 1999a] Witt, S., 1999, Use of Speech Recognition in Computer Assisted Language Learning: PhD Dissertation, Department of Engineering, University of Cambridge.

[Witt 1999b] Witt, S., and Young, S., 1999, Off-Line Acoustic Modelling of Non-Native Accents, Eurospeech'99: Budapest, Hungary, p. 1367-1370.

[Woodland 1993] Woodland, P. C., and Young, S. J., 1993, The HTK Tied-State Continuous Speech Recognition, Eurospeech'93: Berlin, p. 2207-2210.

[Woodland 1994] Woodland, P. C., Odell, J. J., Valtchev, V., and Young, S. J., 1994, Large Vocabulary Continuous Speech Recognition Using HTK, ICASSP'94: Adelaide, p. 125-128.

[Young 1993] Young, S. J., and Woodland, P. C., 1993, The Use of State Tying in Continuous Speech Recognition, Eurospeech'93: Berlin, Germany, p. 2203-2206.

[Zissman 1996a] Zissman, M. A., 1996, Comparison of Four Approaches to Automatic Language Identification of Telephone Speech: *IEEE Transactions on Speech and Audio Processing*, vol. 4, p. 31-44.

[Zissman 1996b] Zissman, M. A., Gleason, T. P., Rekart, D. M., and Losiewicz, B. L., 1996, Automatic Dialect Identification of Extemporaneous Conversational, Latin American Spanish Speech, ICASSP'06: Atlanta, p. 777-780.