A FRAMEWORK FOR AUTOMATIC CODE SWITCHING SPEECH RECOGNITION WITH MULTILINGUAL ACOUSTIC AND PRONUNCIATION MODELS ADAPTATION

BASEM H. A. AHMED

UNIVERSITI SAINS MALAYSIA

2014

A FRAMEWORK FOR AUTOMATIC CODE SWITCHING SPEECH RECOGNITION WITH MULTILINGUAL ACOUSTIC AND PRONUNCIATION MODELS ADAPTATION

by

BASEM H. A. AHMED

Thesis submitted in fulfillment of the requirements

for the degree of

Doctor of Philosophy

May 2014

ACKNOWLEDGEMENTS

In the name of Allah SWT, the Most Gracious and the Most Merciful, I offer my humble gratitude for giving me the strength to complete this thesis.

I would like to express my sincere gratitude and appreciation to my supervisor, Dr. Tien-Ping Tan, for his constructive ideas, criticism, guidance, and patience throughout the long duration of preparing this thesis. He successfully guided me through some stressful times and was always willing to sharpen my understanding on this thesis and other academic activities. A special thanks as well to my co-supervisor, Associate Professor Chan Huah Yong, for his effort in this research. I would also like to take this opportunity to thank the University Science Malaysia (USM) and School of Computer Science.

To my parents and sisters, who have given me their prayers, encouragement, and unfailing support throughout this long journey.

Thank you to Assistant Professor Mohd Radi and my other colleagues for their support, guidance, encouragement, and friendship.

Finally, and most importantly, I would like to extend my gratitude and affection to my beloved wife, Heba Abdullah Ahmed, and my children, Hussein, Sara, and Amr. Thank you for your support, patience, love, encouragement, and inspiration that greatly facilitated the completion of this challenging effort.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AM	Acoustic Model
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BIC	Bayesian Information Criteria
CSR	Continuous Speech Recognition
CWR	Connected Word Recognition
DTW	Dynamic Time Warping
DVQ	Distributed Vector Quantization
DCT	Discrete Cosine Transform
Eng	English
GMM	Gaussian Mixture Models
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
IWR	Isolated Word Speech Recognition
LVCSR	Large Vocabulary Continuous Speech Recognition
LBD	Language Boundary Detection
LDA	Linear Discriminative Analysis
LID	Automatic Language Identification
LM	Language Model
LPC	Linear Predictive Coding
Mal	Malay
Man	Mandarin

MAP	Maximum A Posteriori
MFC	Mel-Frequency Cepstral
MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
PM	Pronunciation Model
PPRLM	Parallel Phone Recognizer Followed by Language Modeling
SDC	Shifted Delta Cepstral
SVM	Support Vector Machines
VQ	Vector Quantization
WER	Word Error Rate
WRR	Word Recognition Rate

SUATUKERANGKA UNTUK PENGECAMAN PERTUTURAN PERALIHAN KOD AUTOMATIK DENGAN PEMODELAN AKUSTIK BERBILANG BAHASA DAN PENGADAPTASI MODEL SEBUTAN

ABSTRAK

Pengecaman pertuturan peralihan kod ialah suatu masalah yang mencabar kerana tiga sebab. Peralihan kod bukan penggabungan secara mudah dua bahasa, tetapi ia mempunyai fonologi, leksikal, dan variasi tatabahasa yang tersendiri. Kedua, sumber-sumber bahasa untuk pengecaman kod seperti korpus pertuturan dan teks adalah terhad dan sukar untuk dikumpul. Oleh itu, pembinaan model peralihan kod mungkin memerlukan strategi yang berbeza daripada yang biasanya digunakan untuk pengecaman pertuturan automatik bagi satu bahasa. Ketiga, segmen peralihan bahasa dalam sesuatu ucapan bolehlah sangat pendek dengan hanya melibatkan satu perkataan, atau sepanjang ucapan itu sendiri. Ini membuat pengenalan bahasa automatik sesuatu yang sangat sukar. Dalam tesis ini, kami mencadangkan satu pendekatan baru untuk pengecaman automatik pertuturan peralihan kod. Kaedah yang dicadangkan terdiri daripada dua fasa: pengecaman pertuturan automatik, dan penilaian semula. Kerangka sistem ini menggunakan beberapa pengecaman pertuturan automatik secara selari untuk pengecaman pertuturan. Kami juga mencadangkan penggunaan pendekatan model akustik yang dikenali sebagai pendekatan hibrid interpolasi dan gabungan untuk saling mengadaptasi model akustik bagi bahasa yang berbeza untuk mengecam pertuturan peralihan kod dengan lebih baik. Untuk pemodelan sebutan, kami mencadangkan satu pendekatan untuk memodel pengecaman pertuturan automatik bagi

penutur bukan asli. Kami telah menguji pendekatan kami pada dua korpus peralihan kod: Melayu-Inggeris dan Mandarin-Inggeris. Kadar kesilapan perkataan untuk pengecaman pertuturan peralihan kod bagi bahasa Melayu-Inggeris menurun daripada 33.2% kepada 25.2% apabila pendekatan yang dicadangkan digunakan, dan kadar kesilapan perkataan untuk bahasa Mandarin-Inggeris pula menurun daripada 81.2% kepada 56.3%. Ini menunjukkan bahawa pendekatan yang dicadangkan berpotensi untuk mengecam pertuturan peralihan kod.

A FRAMEWORK FOR AUTOMATIC CODE SWITCHING SPEECH RECOGNITION WITH MULTILINGUAL ACOUSTIC AND PRONUNCIATION MODELS ADAPTATION

ABSTRACT

Recognition of code-switching speech is a challenging problem because of three issues. Code-switching is not a simple mixing of two languages, but each has its own phonological, lexical, and grammatical variations. Second, code-switching resources, such as speech and text corpora, are limited and difficult to collect. Therefore, creating code-switching speech recognition models may require a different strategy from that typically used for monolingual automatic speech recognition (ASR). Third, a segment of language switching in an utterance can be as short as a word or as long as an utterance itself. This variation may make language identification difficult. In this thesis, we propose a novel approach to achieve automatic recognition of code-switching speech. The proposed method consists of two phases, namely, ASR and rescoring. The framework uses parallel automatic speech recognizers for speech recognition. We also put forward the usage of an acoustic model adaptation approach known as hybrid approach of interpolation and merging to cross-adapt acoustic models of different languages to recognize code-switching speech better. In pronunciation modeling, we propose an approach to model the pronunciation of non-native accented speech for an ASR system. Our approach is tested on two code-switching corpora: Malay–English and Mandarin–English. The word error rate for Malay–English code-switching speech recognition reduced from 33.2% to 25.2% while

that for Mandarin–English code-switching speech recognition reduced from 81.2% to 56.3% when our proposed approaches are applied. This result shows that the proposed approaches are promising to treat code-switching speech.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Speech is the most convenient medium for people to get their message across (Paul, 2009). Spoken languages in the world number more than 6000 (Katzner, 2002; Wagner & Venezky, 1999; Wagner, Venezky, & Street, 1999). Nowadays, many people can speak more than one language. Multilingual speakers tend to switch from one language to another, a phenomenon known as code switching.

A multilingual speaker may have various reasons to code switch when speaking. For example, switching to English technical words or phrases is simpler than attempting to recall the equivalent expressions in Malay.

Bilinguals or multilinguals often code switch when the language they use does not have a specific word or when they cannot find a word to express themselves (Choy, 2011; Coulmas, 2005; Scotton, 1988). Moreover, the code-switching form is often used to strengthen a statement (Gal, 1979) and to express more semantically significant information (Auer, 1999; Baoueb, 2009; Gumperz, 1982). According to Scotton (1995), speakers code switch because of many possible language choices. Some studies (Scotton, 1988; Su, 2001) also state that ethnic minority communities can show their cultural

identity through code switching. Often, people attempt to form friendly relationships with others by speaking in their language as much as possible.

Most linguists assert that code switching is not an accidental occurrence, but has its own rules and constraints. Code-switching constraints are linguistic constraints that prohibit switching from one language to another. The constraints can be structure, size-of-constituent, and free morpheme constraints.

- Structure constraint: occurs between two languages L1 and L2 in a discourse, where elements of both languages do not violate a syntactic rule (Berk-Seligson, 1986; Redouane, 2005).
- Size-of-constituent constraint: occurs between two languages L1 and L2 at phrase structure boundaries. It can be categorized into two classes. The first class is the higher-level constituents, such as sentences and clauses, which are major constituents and tend to be switched frequently. The second is the lower-level constituents (Poplack, 1980).
- Free morpheme constraint: occurs between a free and a bound morpheme. The free morpheme indicates that a morpheme can appear alone in a language, and a bound morpheme indicates that the morpheme cannot appear alone in a language (Gumperz, 1982; Redouane, 2005).

Code-switching speech consists of more than one language within a speech utterance. Furthermore, the code-switching speaking style is common in several multilingual societies. English–Spanish in the US, French–German in Switzerland,

Mandarin–Taiwanese in Taiwan, Cantonese–English in Hong Kong (J. Chan, Ching, & Lee, 2005; Shia, Chiu, Hsieh, & Wu, 2004; Zentella, 1997), and Malay–English and Mandarin–English in Malaysia are some examples of language combinations that exist in such societies (T.-P. X. Tan, Xiong; Tang, Enya Kong; Chng, Eng Siong; Li, Haizhou 2009). Therefore, code switching is common in societies where more than one language is spoken. The automatic speech recognition (ASR) task in code-switching speech is more difficult than in a monolingual speech. In the last decade, the research on ASR with monolingual speech has shifted to that with multilingual speech (Joachim Kohler, 2001; Lyu, Lyu, Chiang, & Hsu, 2008; Uebler, 2001). Similar to the requirement of the monolingual ASR task, multilingual speech recognition tasks have to obtain a large speech corpus for each language to achieve a small word error rate (WER). Although most of the major languages, such as English, French, Spanish, and German, have large speech corpora (Joachim Kohler, 2001; Kumar, Mohandas, & Haizhou, 2005; Lyu et al., 2008; Uebler, 2001; Walker, Lackey, Muller, & Schone, 2003), not all languages have a large speech corpus.

The ASR system is also known as a speech-to-text system, which decodes an utterance to text. Figures 1-1 and 1-2 show the two main phases of a speech recognition system: decoding and training. The three models created during the training, namely, acoustic, pronunciation, and language models, are used to decode speech to text. The acoustic model describes the basic units of speech, such as phones, syllables, or words; the pronunciation model contains language units, such as words or syllables; and the language model presents the statistics of language structure and syntax.

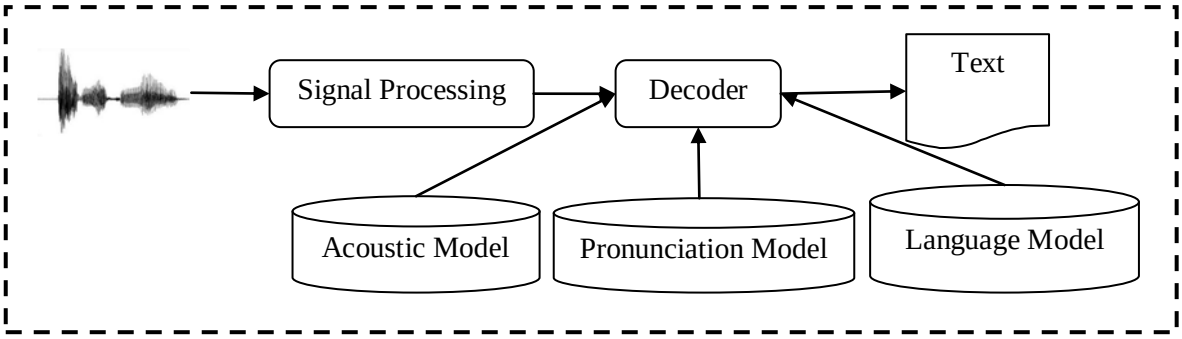

Figure 1-1: Automatic Speech Recognition Decoding Process

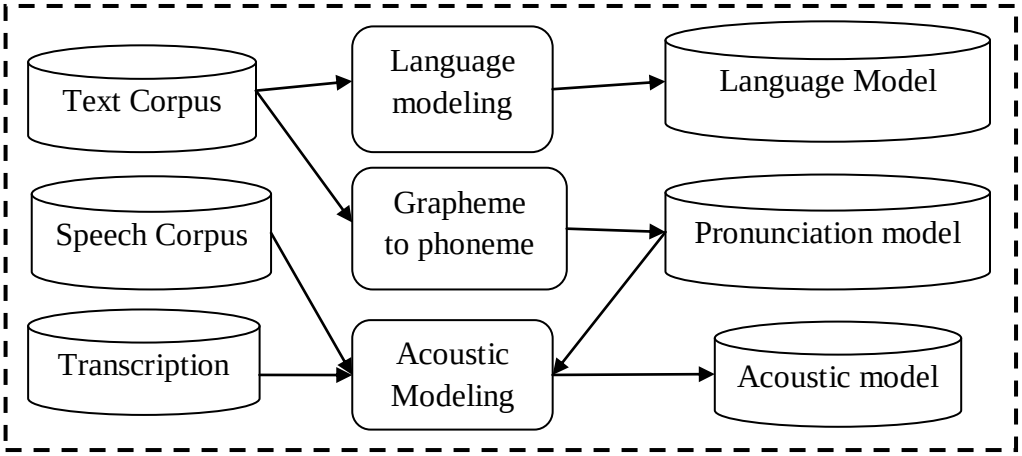

Figure 1-2: Automatic Speech Recognition Training Processes

The ASR system involves two main phases: training and recognition. A rigorous acoustic modeling procedure is followed to model the basic speech units, such as phones, syllables, or others, along with the acoustic observation in the acoustic model. In the training phase, known speech is recorded and preprocessed. It then enters the first stage, feature extraction, where the front-end algorithm is used to extract discriminative features. The next stage is the hidden Markov Model (HMM) training.

Pronunciation modeling uses the acoustic units to create words or syllables. In cases where the acoustic units are phonemes or phones, using a typical dictionary to build the

pronunciation dictionary is possible because most dictionaries use the International Phonetic Alphabet (IPA) to describe word pronunciation. The linguistic rules are used to transform the graphemes to phonemes automatically when a pronunciation description is not available. Manual verification is necessary to correct words that are exceptions to the rule. If the linguistic rules that transform the graphemes to phonemes are not available, then the pronunciation model can be modeled using context-dependent graphemes as the acoustic units (Charoenpornsawat, Hewavitharana, & Schultz, 2006; Killer, Stüker, & Schultz, 2003).

The language model represents the grammar of a language (De Mori, 2007). It presents the syntax and morphology rules of the language. ASR systems use n-gram language models to determine the correct sequence of words by estimating the likelihood of the n^{th} word based on the n-1 preceding words.

The common approaches to represent the grammar of a language are formal language model and stochastic language model. The formal language model is a knowledge-based approach that uses linguistic knowledge to represent the language model, whereas the stochastic language model is a data-driven approach that employs text corpus to extract rules that represent the language model (Huang, Acero, & Hon, 2001).

The recognition phase starts with the conversion of the speech signal into a series of acoustic features. Then, the acoustic score for the features is computed for every state, and the most probable word sequence is determined. The most probable word sequence can be

achieved by maximizing the posterior probability for the given feature vectors using the language model (Jiang, 2005).

This thesis examines automatic code-switching speech recognition for code-switching speech. As described above, the statistical speech recognition system models speech at different levels using three kinds of models, namely, acoustic, pronunciation, and language models. The mismatch in acoustic, phones, pronunciation, and language model during the decoding of code-switching speech deteriorates ASR accuracy. Therefore, an ASR architecture that decodes code-switching speech and models that define code-switching speech better is essential.

1.2 Motivation

The need for multilingual speech applications is growing. In Asia especially, it is quite the norm for people to speak interchangeably in a mixed language even within one sentence. As more and more systems feature the capability to understand speech, the demands for systems that can understand code-switching speech are increasing, but current technology is unable to fulfill these needs because of poor performance. As a result, users may be forced to speak in a certain way for the system to understand them.

1.3 Significance of Work

The main contribution of our work in multilingual ASR is to improve the recognition rate of automatic code-switching speech recognition and to avoid deteriorating the recognition

on monolingual speech. In this work, we use existing monolingual resources, such as text and speech corpora, to create models that recognize code-switching speech.

1.4 Problem Statement

An ASR system requires large amounts of resources in term of speech and text from the target language (monolingual) to create robust acoustic and language models that recognize the target speech. However, not all languages can boast an extensive speech database (Lyu & Lyu, 2008). Obtaining the resources to build acoustic and language models to handle code-switching speech is difficult because these resources are limited and difficult to acquire and scarce. Code switching speech happens most in dialog, and transcribing these resources are expensive and time consuming (Lyu et al., 2008). Only major languages like English, French, Spanish, and German have many speech corpora (Joachim Kohler, 2001; Lyu & Lyu, 2008).

Identifying different language segments in a code-switching utterance is a difficult task because a multilingual speaker can switch to a different language in a word before switching it back. Typical code-switching speech recognition requires a language identification (LID) system to identify different language segments in the utterance before a suitable ASR can be used to decode each segment. Lyu and Lyu (2008), Mehrabani & Hansen (2011) used LID for identifying different language segments in a code-switching utterance and give low accuracy. Low accuracy of LID will subsequently decrease the accuracy of multilingual speech recognition. Suggest a code-switching speech recognition

framework without using LID that does not deteriorating monolingual speech recognition will be a promising direction to improve the accuracy.

Code-switching speech consists of more than one languages, each of which has its own phonological, lexical, and grammatical variation. Hence, the main problem with acoustic modeling is how to build robust large vocabulary continuous speech recognition (LVCSR) system for a new target language (small speech database available) using speech corpus or an acoustic model from various source languages. If the acoustic model or speech corpus is used efficiently, better results will be obtained. By contrast, one of the languages used in code-switching speech is often a non-native language of the speaker. The proposed approaches in Z. Wang et al. (2003), (Tien-Ping Tan & L. Besacier, 2007) adapt the acoustic model using acoustic model interpolation for recognizing non-native speech. They reduce the WER from 49.3% to 36%. S. Witt and Young (1999) and (T.-P. Tan, 2008) used the acoustic model merging approach for recognizing non-native speech. The average baseline WER of 28.3% improved to 20.6%. The acoustic model can be utilized more effectively to enhance the WER.

Non-native pronunciation modeling similar to acoustic modeling is also required. Pronunciation variations by non-native speakers arise from the influence of their mother tongues (Adams & Munro, 2009; Meierkord, 2004; Strevens, 1992). Consequently, the way they pronounce the words in the second language will be different from the pronunciation of a native speaker. The majorities of existing pronunciation models are developed based on native speakers, and does not take in consideration that the non-native

speaker can't pronounce the complex pronunciation rules. Thus, our focus is to determine how the existing model can be adapted to match the pronunciation of non-native speakers.

Building the language model for code-switching speech will require us to acquire a code-switching text. However, a code-switching text may be limited or not available because most code-switching speech only exists in the form of conversation or dialogue. One possibility is to manually transcribe spontaneous speech data, but this requires time and money. Therefore, we will attempt to use existing monolingual text to model the syntactical grammar of code-switching speech.

In this research, we propose a framework to improve the decoding of code-switching speech. We model the pronunciation of non-native accented speech and propose an approach that fully uses the existing acoustic models to model code-switching speech.

1.5 Research Question

In the section 1.4, we describe the problem of building an automatic code switching speech recognition, such as limitation in the current language identification accuracy, the limited resources for building code switching model for the automatic speech recognition system and the non-native characteristics in code switching speech. The problems explored in this research are as follows:

1. How to improve the automatic recognition of code-switching speech and at the same time does not cause the accuracy of monolingual speech recognition system to deteriorate?
2. How to model the pronunciation of code-switching and non-native speech?
3. How to use the existing speech corpora to improve the accuracy of automatic code switching speech recognition?

1.6 Objective

The objectives of our study are as follows:

1. To propose a framework that will improve the decoding of code-switching speech. We present the proposed framework in Section 3.2 and its result in Section 4.3.7.
2. To model the pronunciation of non-native accented speech. We present the proposed approach in Section 3.3 and the result in Section 4.3.3.
3. To propose an approach that fully uses the existing acoustic models to model code-switching speech. We present the proposed approach in Section 3.3 and the result in Section 4.3.5

1.7 Methodology Overview

New acoustic and pronunciation modeling approaches for code-switching speech are proposed in this paper. Moreover, we propose a framework to improve the recognition of code-switching speech without deteriorating the accuracy of native speech.

For acoustic modeling, we extend the hybrid of interpolation and merging approach used in non-native speech adaptation for cross-adapting acoustic models for code-switching speech recognition.

For pronunciation modeling, we propose an approach that predicts the word pronunciation of code-switching speakers by analyzing the pronunciation of the sub-words. The frequently used pronunciation of the sub-word is selected as the pronunciation for the target sub-word.

Finally, we propose an approach to achieve automatic recognition of code-switching speech by using parallel automatic speech recognizers from the corresponding languages involved in code-switching.

1.8 Contributions

The main contributions of the present study are as follows:

1. An ASR framework for code-switching speech recognizing without deteriorating monolingual speech recognition. We present the framework in Section 3.2 and the result in Section 4.3.7.
2. An acoustic model adaptation approach that crosses-adapts the acoustic models for code-switching speech recognition. We present the proposed approach in Section 3.3 and the result in Section 4.3.5.

3. A pronunciation modeling approach that adapts a pronunciation dictionary for non-native speakers. We present the proposed approach in Section 3.3 and the result in Section 4.3.3.

1.9 Scope and Limitation of the Research

This research focuses on code-switching speech recognition. Among the ASR models, only pronunciation modeling and acoustic modeling are the main concerns of this research. We do not work on the language model directly, but instead the limitation in modeling code switching language model is solved indirectly in our framework. Our work was carried out only on Malay-English and Mandarin-English code switching speech. Besides that, in this research we concurred with code-switching speech consist of two different languages. We did not test on code switching speech that consists of more than 2 languages.

1.10 Thesis Outline

Chapter 2 examines recent studies on code-switching in acoustic modeling, pronunciation modeling, and LID. Chapter 3 explains the proposed framework for code-switching speech recognition, code-switching acoustic modeling using multilingual resources, and the proposed pronunciation modeling using multilingual resources. Chapter 4 presents our experiments, the corresponding results, and the discussion of the results. Finally, Chapter 5 presents the conclusions of the proposed work and the suggestions for future research in code-switching speech recognition.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

A review of literature on ASR systems demands attention toward Alexander Graham Bell's discovery of the process of converting sound waves into electrical impulses and the first speech recognition system developed by Davis and Biddulph for recognizing telephone-quality digits spoken (K. Davis & Biddulph, 1952; Ghai & Singh, 2012).

Various approaches and types of speech recognition systems gradually came into existence in the last five decades (Ghai & Singh, 2012). This evolution had a remarkable impact on the development of ASR systems for various languages worldwide. ASR system is a speech-to-text conversion system in which the output is displayed as text corresponding to the recognized speech. Thus far, ASR systems have been developed for just a fraction of the approximately 7,300 existing languages worldwide. Russian, Portuguese, Chinese, Vietnamese, Japan, Spanish, Filipino, Arabic, English, Bengali, Tamil, Malayalam, Sinhala, and Malay are the most prominent among them. Maximum work for recognition has been done for the English language.

This chapter introduces the topic of ASR with an emphasis on multilingual speech recognition. Related literature is discussed to outline the current state of speech recognition technology.

2.2 Approaches to ASR System

There are three approaches to automatic speech recognition (Anusuya & Katti, 2010). They are acoustic–phonetic approach, pattern-recognition approach and artificial intelligence approaches, which include knowledge-based approach, connectionist approach and support vector machine approach.

2.2.1 Acoustic–Phonetic Approach

According to Al-Zabibi (1990); Espy Wilson (1987); Liberman and Whalen (2000), a fixed number of distinctive phones exists in a spoken language. Each phone is described by a set of acoustic–phonetic features. The message-bearing components of speech are extracted explicitly with the determination of relevant continuous features, such as ratio of high-low frequencies and formant locations, and of binary acoustic features, such as friction, nasality, and voiced-unvoiced (T.-P. Tan, 2008). Acoustic properties of phoneme, which changes according to many factors, such as acoustic context, speaker gender, age, and emotional state, hinder the commercial application of the acoustic–phonetic approach (Tran, 2000). This approach is normally implemented in the following sequence: spectral analysis, feature detection, segmentation and labeling, and finally, recognizing valid words.

2.2.2 Pattern-Recognition Approach

Pattern training and pattern comparison are the two main steps in pattern-recognition approach (Devijver & Kittler, 1982; Zhu, De Silva, & Ko, 2002). To recognize a spoken word, pattern-recognition approach compares the spoken words with the pattern learned in the training step.

2.2.2.1 Template-based Approach

In this approach, a speech dictionary is built from a set of speech patterns (Connell & Jain, 2001). To recognize a spoken word, the spoken word is compared with each record in the speech dictionary and the best record that matches the spoken word is selected. One of the main drawbacks in this approach is that each word should be previously included in the speech dictionary.

2.2.2.2 Stochastic approach

Stochastic approach can deal with speaker variability and confusing sounds because it is established on the use of probabilistic models (Sankar & Lee, 1995). This approach is more general compared with the template-based approach.

2.2.3 Knowledge-based approach

The knowledge-based approach is a hybrid of the pattern-recognition and acoustic–phonetic approaches (King et al., 2007). Neither the success of the acoustic–phonetic nor template-based approach has been isolated to fully explore human speech processing (Tripathy, 2008).

In knowledge-based approach, the production rules are generated using linguistic knowledge or speech spectrogram observations. Knowledge is useful in defining speech units and selecting suitable input representations (Das, 2013). Samouelian (1994) proposed a data-driven methodology for continuous speech recognition (CSR), in which the knowledge on the structure and characteristics of the speech signal is acquired explicitly from the database through inductive inference (Samouelian, 1994). This approach has the ability to solve inter- and intra-speaker speech variability problems and to generate decision trees. However, the recognition performance of this approach falls short because of the very small number of speakers. Tripathy (2008) proposed a knowledge-based approach by using a fuzzy inference algorithm to classify spoken English vowels. This technique gives better results over the standard Mel-frequency cepstral coefficient (MFCC) feature analysis.

2.2.4 Connectionist approach

In connectionist models, knowledge or constraints are distributed across numerous simple computing units that connect to form a network (Bourlard & Morgan, 1994). Connectionist learning attempts to organize a network of processing elements (El Ayadi, Kamel, & Karray, 2011; Gupta, Radha Mounima, Manjunath, & Manoj, 2012).

2.2.5 Support vector machines approach

Support vector machines (SVMs) use a discriminative approach to optimize the margin between the samples and the classifier border and to generalize unseen patterns (Schuller, Rigoll, & Lang, 2004). SVMs utilize linear and nonlinear splitting to classify data. They

cannot classify variable-length data vectors, only fixed-length data vectors. Before using SVMs, variable-length data have to be transformed to fixed-length vectors (Padrell-Sendra, Martın-Iglesias, & Dıaz-de-Marıa, 2006). SVMs are generalized linear classifiers with maximum-margin fitting functions that provide regularization to improve the generalization process. Padrell-Sendra and colleagues worked on a pure SVM-based continuous speech recognizer by applying SVM for decision making at the frame level and a Token-Passing algorithm to obtain the chain of recognized words. The Token Passing model is an extension of the Viterbi algorithm intended for CSR to manage the uncertainty on the number of words in a sentence. The results obtained from the experiments indicated that recognition accuracy improves with SVMs with a small database, but is obtained at the expense of huge computational effort with a large database (Padrell-Sendra et al., 2006).

2.3 Stochastic Approach to Automatic Speech Recognition

Stochastic approach use probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness occur due to phone confusion, speaker variability's and contextual effect. Thus, stochastic approach is suitable to speech recognition. The most popular stochastic approach is hidden Markov modeling. General literature on HMM is presented in this section. The HMM is introduced, with an emphasis on the three interesting problems it solves, namely, evaluation, decoding, and learning problems. We then present HMM in speech recognition, highlighting on HMM architecture, Gaussian mixture model (GMM), speech features, decoding, and lattice rescoring.

2.3.1 HMM

HMM is a finite-state machine that generates a sequence of discrete time observations. At each time unit, HMM specifies how likely every observation is to be generated in each state. An N-state HMM is defined by the state transition probability (A matrix), output probability distribution (B matrix), and initial state probability (π). Figure 2-1 shows that an HMM consists of three states. The starting state is S_1, and the arc presents the probability of transitioning from one state to another.

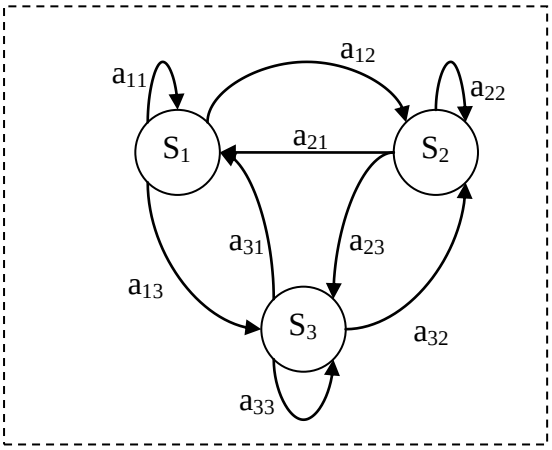

Figure 2-1: HMM example

The three HMM problems are given below:

- Evaluation Problem: Given the HMM model Φ and the observation sequence $O = (o_1, o_2 \dots o_t)$, the evaluation problem calculates the probability that model Φ has generated the sequence O.
- Decoding Problem: Given the HMM Φ and the observation sequence $O = (o_1, o_2 \dots o_t)$, the decoding problem calculates the most likely sequence of hidden states that produce this observation.

- Learning Problem: Given some training observation sequences $O = (o_1, o_2 \dots o_t)$ and a general structure of HMM, determine the HMM parameters that maximize the joint probability (likelihood) and best fit training data. The likelihood can be efficiently calculated using dynamic programming, such as
    - Forward algorithm: Reproduces the observation through the HMM.
    - Backward algorithm: Back traces the observation through the HMM.

2.3.1.1 Evaluation Problem

The evaluation problem computes the likelihood that a given model M produces a given observation sequence $O = (o_1, o_2 \dots o_t)$. The direct computation of this probability is computationally expensive because the total number of state sequences grows exponentially according to the value of T. The problem is solved using forward algorithm, a type of dynamic programming. Given a set of observations $O = (o_1, o_2 \dots o_t)$ derived from HMM with parameters Λ and state sequence $S = (s_1, s_2, s_T)$, we have

$$p(O|\Lambda) = \sum_{\text{all } s} p(S, O|\Lambda) = \sum_{\text{all } s} P(S|\Lambda)p(O, S|\Lambda) \qquad \text{Equation2.1}$$

where the sum is over all the possible values of the state sequence S. Taking into consideration the Markov assumption, the probability of a given state sequence is the product of the corresponding state transition probabilities:

$$P(S|\Lambda) = \pi_{s1} \prod_{t=2}^{T} a_{s_{t-1}\,St} \qquad \text{Equation 2.2}$$

Furthermore, because of the output-independence assumption, we have

$$P(O|s, \Lambda) = \prod_{t=1}^{T} b_{s_t}(O_t) \qquad \text{Equation 2.3}$$

Substituting the two results into Equation 2.3, then the likelihood of the observation O results in

$$p(X|\Lambda) = \sum_{all\ s} \pi_{s1}\, b_{s_{t1}}(O_1) \prod_{t=2}^{T} a_{s_{t-1}\,St}\, b_{s_{t1}}(O_t) \qquad \text{Equation 2.4}$$

2.3.1.2 Decoding Problem

The decoding problem determines the optimal sequence of the states, that is, the state sequence S^* with the highest likelihood given a set of observations and HMM.

A single best state sequence $q^* = (q_1, q_2, \ldots, q_T)$ for a given observation sequence $O = (o_1, o_2, \ldots, o_T)$ can be obtained through the Viterbi algorithm. The Viterbi algorithm generally has three main steps (Jurafsky, 2000):

1 Assign each edge a transition cost.

2 Update all path metrics:

2.1 Calculate P (state)*P (observation)

2.2 For all transitions, calculate P(old state)*P(transition)*P(observation| new state)

2.3 Update all path metrics with the highest probability.

3 Starting from the final state, trace back to the initial state.

Table **2-1** shows the start, transition, and emission probabilities for a sample HMM, which consists of two states and three observations. Given the observations O_1, O_2, and O_3, we want to predict the most probable state sequence for this HMM model.

Table 2-1: Start, transition, and emission probabilities for a sample consisting of two states and three observations

Start probability	
S_1	0.6
S_2	0.4

Transition probability		
	S_1	S_2
S_1	0.7	0.3
S_2	0.4	0.6

Emission probability			
	O_1	O_2	O_3
S_1	0.5	0.4	0.1
S_2	0.1	0.3	0.6

First, we calculate P (state)*P (observation) for O_1.

At O_1, $S_1 = P(S_1)*P(S_1 \text{ at } O_1) = 0.6*0.5 = 0.3$

At O_1, $S_2 = P(S_2)*P(S_2 \text{ at } O_1) = 0.4*0.1 = 0.04$

Table 2-2: First step of predicting the most probable state sequence for HMM model

	O_1	O_2	O_3
S_1	0.3		
S_2	0.04		

Then, we recursively calculate argmax [P (old state)*P (transition)*P (observation| new state)] and store the best path for each state.

At O_2:

S_1 = argmax [$P(S_1)$* P (transition (S_1,S_1)*P (S_1 at O_2),

P (S_1)* P (transition (S_1,S_2)*P (S_1 at O_2)]

= argmax (0.3* 0.7*0.4, 0.3* 0.3*0.4)

=argmax (0.084, 0.036) = 0.84

S_2 = argmax [P(S_2)* P (transition (S_2,S_2)*P (S_2 at O_2),

P (S_2)* P (transition (S_2,S_1)*P (S_2 at O_2)]

= argmax (0.04* 0.6*0.3, 0.04* 0.4*0.3)

= argmax (0.0072, 0.0048) = 0.0072

At O_3:

S_1 = argmax [P(S_1)* P (transition (S_1,S_1)*P (S_1 at O_3),

P (S_1)* P (transition (S_1,S_2)*P (S_1 at O_3)]

= argmax (0.084 * 0.7*0.1, 0.084 * 0.3*0.1)

= argmax (0.00588, 0.00252) = 0.00588

S_2 = argmax [P (S_2)* P (transition (S_2,S_2)*P (S_2 at O_3),

P (S_2)* P (transition (S_2,S_1)*P (S_2 at O_3)]

= argmax (0.0072 * 0.6*0.6, 0.0072 * 0.4*0.6)

= argmax (0.002592, 0.001728) = 0.002592

Table 2-3: Last step of predicting the most probable state sequence for HMM model

	O_1	O_2	O_3
S_1	0.3	0.084 via S_1	0.00588 via S_1
S_2	0.04	0.0072 via S_2	0.002592 via S_2

Thus, the most probable states given the observations O_1, O_2, and O_3 are S_2, S_1, and S_1.

2.3.1.3 Learning Problem

The learning approach is as follows: given HMM and training data (a set of labeled observations), HMM parameters that best describe the data are estimated. The state sequence is unknown and, therefore, ML training cannot be applied directly. The standard solution is to apply a version of the expectation-maximization (EM) algorithm adapted to HMM. This training algorithm is known as the Baum–Welch algorithm or forward-backward algorithm (Huang et al., 2001). Specifically, this algorithm finds Λ, such that $p(O|\Lambda)$ is maximized for a training sequence X. Mathematically, this can be expressed as

$$\Lambda_{\mathrm{ML}} = \max_{\Lambda} \mathrm{p}(\mathrm{O}|\Lambda) \qquad \text{Equation 2.5}$$

2.3.2 HMM in speech recognition

When humans speak, the articulator apparatus modulates air pressure and flow to produce the sounds that constitute the speech signal. Even if speech is a time-varying signal, the signal can be considered as a stationary process in short-time regions. Moreover, a convinced dependency usually exists between sounds in the speech signal that occur after each other, implying that speech is not a memoryless process.

HMMs provide a simple and effective framework to model time-varying spectral vector sequences; therefore, the model is used in the standard ASR.

In HMM-based systems, an input utterance S (t) is converted into word w (or a sequence of words) [i.e., we are looking for the most probable word (W) given time-

varying spectral vector sequences (X)] by evaluating the posteriori probability score P (W|X).

The posteriori probability P (W|X) for acoustical features X from the spoken utterance S can be can be rewritten as follows when applying Bayes' rule (Jurafsky, 2000):

$$\widehat{W} = \underset{W}{\operatorname{argmax}}\, P(W|O)$$

$$= \underset{W}{\operatorname{argmax}} \frac{P(W)P(O|W)}{P(O)} \qquad \text{Equation 2.6}$$

$$= \underset{W}{\operatorname{argmax}}\, P(W)P(O|W)$$

The first term P (W|X) gives the acoustic likelihood for the class and is usually modeled by HMMs (more details on the acoustic model are presented in Section 2.4). The second term P (W) refers to the "a priori probability" of the class W and is usually approximated by a language model (more details on the language model are presented in Section 2.3.8.

The three main ASR system components in a language are as follows (T.-P. Tan, 2008):

- Acoustic model (AM) – phonology of a language
- Pronunciation model (PM) – vocabulary and pronunciations
- Language model (LM) – grammar of a language

The basic units of the acoustic model in an ASR system are phones, phonemes, syllables, and words. Language elements, such as words or syllables, are presented in the pronunciation model using the acoustic units defined in the acoustic model. The language

model expresses the grammar of a language through the pronunciation dictionary vocabulary. Figure 2-2 shows the three models created during the training, namely, the acoustic, pronunciation, and language models, which are used to decode speech to text in the ASR decoding process, as shown in Figure 2-3. The acoustic, pronunciation, and language models are discussed in detail in Sections 2.4.

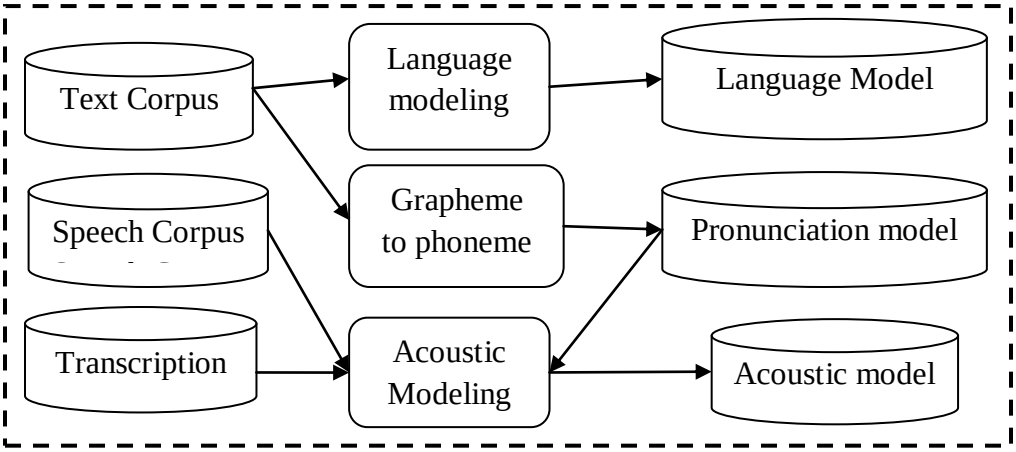

Figure 2-2: ASR Training Processes

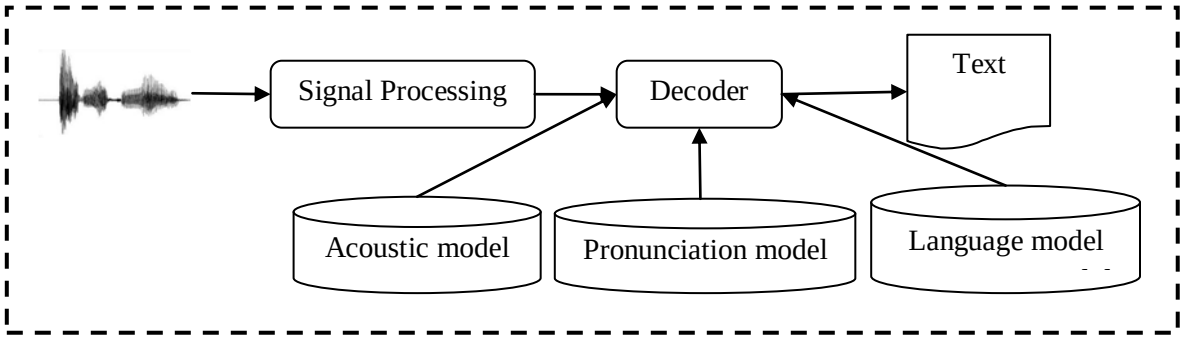

Figure 2-3: ASR Decoding Processes

2.3.2.1 HMM Architecture

Figure 2-4 shows the HMM architecture, which is commonly known as left-to-right HMM. This architecture normally uses three states to model a phone. The observation sequences