Characterizing Phonetic Transformations and Fine-Grained Acoustic Differences Across Dialects

by

Nancy Fang-Yih Chen

B.S., National Taiwan University (2002)
S.M., National Taiwan University (2004)

Submitted to the Harvard-Massachusetts Institute of Technology Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author . . . . . .
Harvard-Massachusetts Institute of Technology Health Sciences and Technology
May 13, 2010

Certified by . . . . . .
Joseph P. Campbell
Assistant Group Leader, MIT Lincoln Laboratory
Thesis Supervisor

Accepted by . . . . . .
Ram Sasisekharan
Director, Harvard-MIT Division of Health Sciences & Technology,
Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering

Characterizing Phonetic Transformations and Fine-Grained Acoustic Differences Across Dialects

by

Nancy Fang-Yih Chen

Submitted to the Harvard-Massachusetts Institute of Technology Health Sciences and Technology
on May 13, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis is motivated by the gaps between speech science and technology in analyzing dialects. In speech science, investigating phonetic rules is usually manually laborious and time consuming, limiting the amount of data analyzed. Without sufficient data, the analysis could potentially overlook or over-specify certain phonetic rules.

On the other hand, in speech technology such as automatic dialect recognition, phonetic rules are rarely modeled explicitly. While many applications do not require such knowledge to obtain good performance, it is beneficial to specifically model pronunciation patterns in certain applications. For example, users of language learning software can benefit from explicit and intuitive feedback from the computer to alter their pronunciation; in forensic phonetics, it is important that results of automated systems are justifiable on phonetic grounds.

In this work, we propose a mathematical framework to analyze dialects in terms of (1) phonetic transformations and (2) acoustic differences. The proposed Phonetic-based Pronunciation Model (PPM) uses a hidden Markov model to characterize when and how often substitutions, insertions, and deletions occur. In particular, clustering methods are compared to better model deletion transformations. In addition, an acoustic counterpart of PPM, Acoustic-based Pronunciation Model (APM), is proposed to characterize and locate fine-grained acoustic differences such as formant transitions and nasalization across dialects.

We used three data sets to empirically compare the proposed models in Arabic and English dialects. Results in automatic dialect recognition demonstrate that the proposed models complement standard baseline systems. Results in pronunciation generation and rule retrieval experiments indicate that the proposed models learn underlying phonetic rules across dialects. Our proposed system postulates pronunciation rules to a phonetician who interprets and refines them to discover new rules or quantify known rules. This can be done on large corpora to develop rules of greater statistical significance than has previously been possible.

Potential applications of this work include speaker characterization and recognition, automatic dialect recognition, automatic speech recognition and synthesis, forensic phonetics, language learning or accent training education, and assistive diagnosis tools for speech and voice disorders.

Thesis Supervisor: Joseph P. Campbell
Title: Assistant Group Leader, MIT Lincoln Laboratory

Acknowledgments

This thesis would not have been possible without my supervisors Wade Shen and Joe Campbell at MIT Lincoln Laboratory. I thank Wade for his sharp insights, vigorous energy, and patient guidance. Despite Wade's busy schedule, he managed to find time to check in with me even when he was traveling. I appreciate Joe for giving me detailed feedback on manuscripts. Joe also always kept his door open in case I needed to discuss issues, be they academic or not.

I would also like to thank my other thesis committee members Adam Albright, Jim Glass, and Tom Quatieri. I thank Tom for always keeping the big picture in mind and making me write a first thesis draft early on, which helped establish the thesis structure and made the writing task less intimidating. I am grateful for Jim's encouragement and useful comments on my papers. I enjoyed Jim's lectures in automatic speech recognition, which are important building blocks of my thesis research. I thank Adam for his helpful literature pointers, linguistic insights, and fresh perspectives on my research.

I have been very fortunate to enjoy the resources at MIT Lincoln Laboratory. I appreciate Cliff Weistein, our group leader, for making it possible for me to work at Lincoln. At the Human Language Technology Group at Lincoln Lab, I have been surrounded by intelligent and enthusiastic researchers. I especially thank Doug Sturim, Pedro Torres-Carrasquillo, Elliot Singer, TJ Hazen, Fred Richardson, Bill Campbell, Bob Dunn, Linda Kukolich, Robert Granville, Ryan Aminzadeh, Carl Quillen, Nick Malyska, and James Van Sciver for helpful and encouraging discussions. I express my gratitude to Reva Schwartz of the United States Secret Service for her support and stimulating discussions. I had endless IT issues with my computers, and fortunately Scott Briere, Brian Pontz and John O'Connor always helped me resolve them. I thank my student labmates, Tian Wang, Zahi Karam, Kim Dietz, Dan Rudoy, and Daryush Mehta. Our interesting yet intellectual conversations in the student office and on the shuttles added a tinge of spice to the long working hours. The group secretaries, Shannon Sullivan and Nancy Levesque, have also been helpful in handling

administrative work and foreign travel affairs.

I would also like to thank the people in the Speech Communication Group at MIT. Ken Stevens, Janet Slifka, Helen Hansen, and Stefanie Shattuck-Hufnagel have inspired me with their passion in science. I thank Caroline Huang for her supportive mentorship and friendship throughout the years. I appreciate the advice I got from senior students: Steven Lulich, Chiyoun Park, Elisabeth Hon Hunt, Youngsook Jung, Tony Okobi, Yoko Saikachi, Xiaomin Mao, Xuemin Chi, Virgilio Villacorta, Julie Yoo and Sherry Zhao. I also thank Arlene Wint for being so considerate.

I thank Lou Braida, Bob Hillman, Bertrand Delgutte and others in the HST administration who made it possible for me to take speech and language pathology classes at MGH Institute of Health Professions. The desire to make speech disorders diagnosis more efficient was a strong motivation behind my thesis work.

I am grateful that Victor Zue encouraged me to apply to graduate school at MIT, so that I could enjoy MIT's open and interdisciplinary research environment. People at Spoken Language Systems Group at MIT CSAIL have also been a source of inspiration throughout grad school. I especially thank Karen Livescu, Paul Hsu, Hung-An Chang, Jackie Lee, Ian McGraw, Tara Sainath, and Mitch Peabody for numerous discussion groups and meetings.

I also benefited from interactions with various researchers in the Linguistics Department at MIT: Donca Steriade, Edward Flemming, Suzanne Flynn, and Michael Kenstowicz, and Feng-Fan Hsieh.

I had a lot of fun dissecting bodies in anatomy class with Erik Larsen, Chris Salthouse, Cara Stepp, Thomas Witzel, Ted Moallem, Tom DiCicco, Manny Simons, and Tamas Bohm. I would also like to thank my friends Wendy Gu, Adrienne Li, Paul Aparicio, Erika Nixon, William Loh, Stephen Hou, Ingrid Bau, Hsu-Yi Lee, Kevin Lee, Vivian Chuang, Shihwei Chang, Shireen Goh, Henry Koh, Lynette Cheah, Kenneth Lim, and Kong-Jie Kah for birthday celebrations, potlucks, and holiday festivals. In particular, I would like to thank Wendy for being such a thoughtful roommate and wonderful baker; there is nothing better than starting your day waking up bathed in the sweet smell of muffins. My rock climbing buddies, Jenny Yuen, Justin Quan, and

Aleem Siddiqui made me look forward to physically challenging Saturdays, so I could de-stress from work.

Finally, I would like to thank Thomas Yeo for his unwavering support and genuine companionship. And thanks to Mom, Dad, and my brother, Eric, who have always believed in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

This thesis is motivated by the gap between the fields of speech science and technology in the study of dialects. (See Figure 1-1.) In speech science or linguistics, discovering and analyzing dialect-specific phonological rules usually requires specialized expert knowledge, and thus requires much time and effort, therefore limiting the amount of data that can be used. Without analyzing sufficient data, there is the potential risk of overlooking or over-specifying certain rules.

On the other hand, in speech technology, dialect-specific pronunciation patterns are usually not explicitly modeled. While many applications do not require such knowledge to obtain good performance, in certain applications it is beneficial to explicitly learn and model such pronunciation patterns. For example, users of language learning software can benefit from explicit and intuitive feedback from the computer to alter their pronunciation; in forensic phonetics, it is important that recognition results of an automated systems are justiable on linguistic grounds [79].

1.2 Proposed Approach

In this work, we generalize and apply the concept of pronunciation modeling [32] from automatic speech recognition (ASR) [72] to the field of dialect recognition. We term

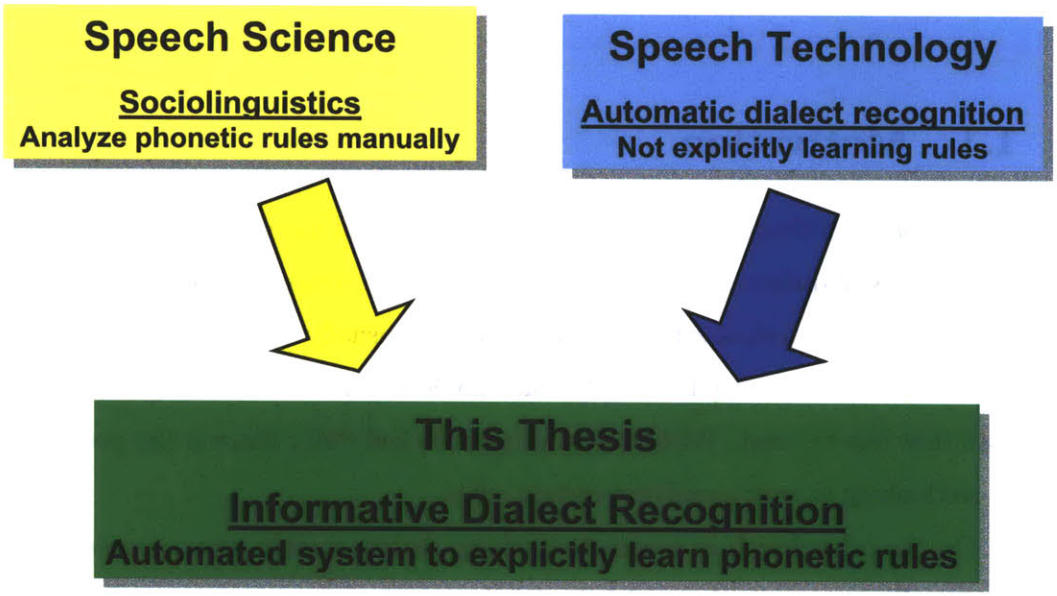

Figure 1-1: This thesis bridges the gap between speech science and technology by combing their strengths together.

this approach as *informative dialect recognition*, as it provides interpretable results that are *informative* to humans [13, 12].

We propose a mathematical framework using hidden Markov Models (HMM) to characterize phonetic and acoustic-based pronunciation variations across dialects. While many dialect recognition methods also take advantage of phonetic and acoustic information, these models are not set up to learn pronunciation rules explicitly for further human interpretation. In contrast, our model design is grounded linguistically to explicitly characterize pronunciation rules. For example, we specify different state transition arcs in our HMM to represent different phonetic transformations.

We employ decision tree clustering to account for data insufficiency and exploit phonetic context. This context clustering approach is similar to binary tree language modeling [64] in spirit, though the probabilistic models are different: in [64], the probability of a current observation is conditioned on a cluster of past observations, whereas in our model the probability of a current observation is conditioned on a reference phone and its context. Using reference phones as a comparison basis, we can explicitly model phonetic and acoustic transformations occurring in different dialects, making the dialect recognizer results interpretable to humans. In addition to standard state clustering used in ASR, we also discuss arc clustering methods to better model deletion transformations.

1.3 Contributions

This thesis proposes *automatic* yet *informative* approaches in analyzing speech variability, which fills in the gaps between speech science and technology research methods. The contributions of this thesis are:

1. Introduce a new interdisciplinary research approach: informative dialect recognition.

2. Propose a mathematical framework to automatically characterize phonetic transformations and acoustic differences across dialects.

3. Demonstrate that the proposed models automatically learn phonetic rules, quantifies occurrence frequencies of the rules, and identify possible regions of interest that could contain phonetic or acoustic information that is dialect-specific.

4. Empirically show that the proposed models complement existing dialect recognition systems in Arabic and English dialects.

5. Survey corpora resources for dialect research, and address challenges in informative dialect recognition.

1.4 Thesis Outline

Figure 1-2 shows the structure of the remaining thesis. In Chapter 2 we review the relevant background of our work in speech science and speech engineering, which includes dialect studies in speech science and linguistics, pronunciation modeling in automatic speech recognition, and automatic language and dialect recognition.

In Chapter 3, we propose a framework using hidden Markov model and decision tree clustering to automatically learn phonetic transformations and acoustic differences across dialects.

In Chapter 4 we summarize the challenges encountered when searching for suitable corpora, analyze corpora related for dialect research, and introduce the 3 databases we chose to empirically evaluate our proposed systems.

Chapters 5 - 7 are the experiments we performed to evaluated the proposed framework. Three different assessment metrics were used. In Chapter 5, we first evaluate if the proposed systems are able to detect dialect differences by conducting dialect recognition experiments. In Chapter 6, we evaluate how well the proposed models generate dialect-specific pronunciations, given a reference dialect's pronunciation. In Chapter 7, we evaluate how well the proposed systems retrieve rules documented in the linguistic literature. Each of these metrics make different assumptions. We attempt to provide a comprehensive analysis of our proposed systems by presenting all three of them.

In Chapter 8, we discuss the characterisitics and implications of top ranking learned rules from the proposed systems.

In Chapter 9 we conclude the contributions of this thesis, discuss future work and potential applications.

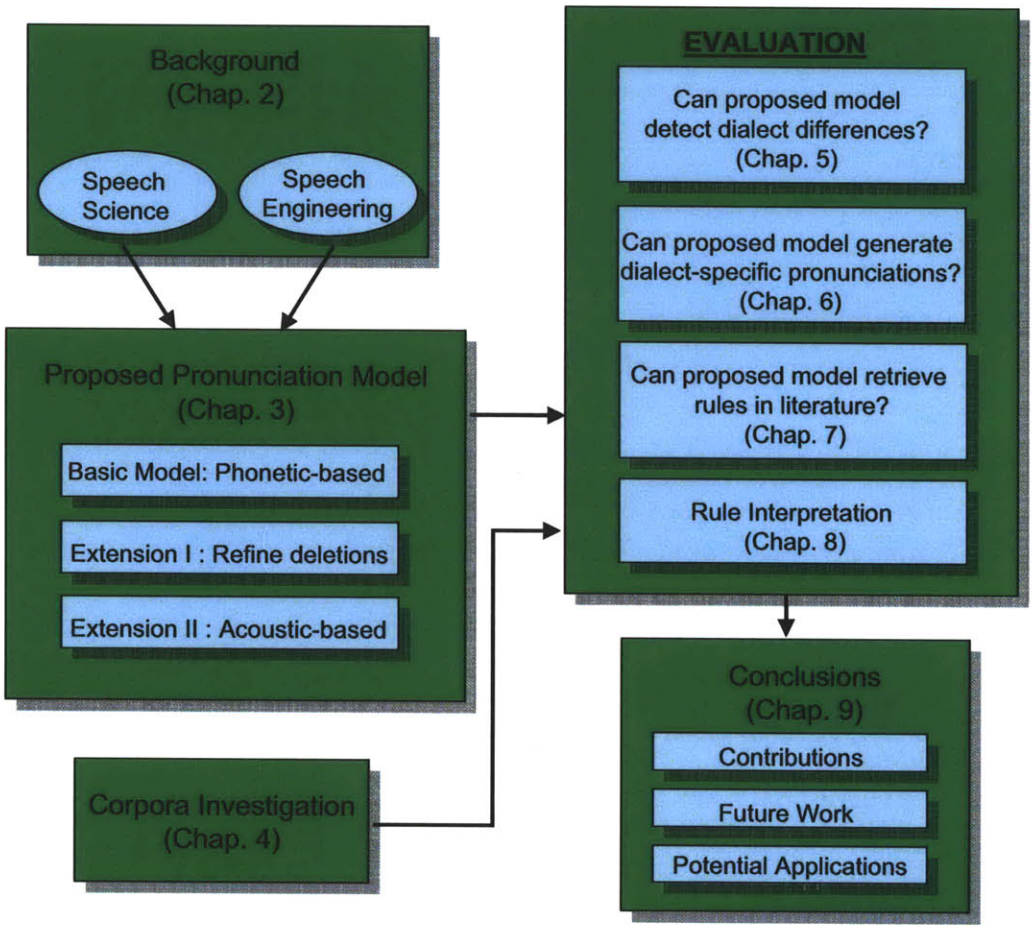

Figure 1-2: Structure of remaining thesis.

Chapter 2

Background

Dialect differences arise from all levels of the linguistic hierarchy, including acoustics, phonetics, phonology, vocabulary, syntax, and prosody [22]. Table 2.1 gives examples of differences between American and British English from various linguistic aspects. Linguistic and speech science studies have shown that many dialect differences exist in the acoustics, phonetic, and phonological levels (e.g., [40, 56, 80, 92]). In this thesis, we will focus on these levels by automatically discovering and analyzing dialect-specific pronunciation patterns.

In speech science or linguistics, discovering and analyzing dialect-specific phonetic rules usually requires specialized expert knowledge, which requires time-consuming manual analysis. In this thesis, we propose automatic approaches that can streamline

Table 2.1: Dialect difference arise from all levels of the linguistic hierarchy. Below are examples for American and British English. For definitions of phonetic symbols, refer to Appendix A.

Acoustics, Phonetics & Phonology (pronunciation)	bath Br: [b aa th] Am: [b ae th]
Lexicon (vocabulary)	Br: lift Am: elevator
Syntax (grammar)	Br: I shall eat. Am: I will eat.
Prosody	speaking rate, pitch, voice quality

traditional methods in studying dialects. Phoneticians, phonologists, and sociolinguists can apply these approaches to locate phonetic rules before detailed analyses. Our research can help speech science and linguistics research be done more efficiently. In addition, such an automatic approach of learning dialect-specific phonological rules is useful in forensic phonetics [79].

In speech technology, dialect-specific pronunciation patterns are usually not explicitly modeled. While many applications do not require such knowledge to obtain good performance, in certain applications it is beneficial to explicitly learn and model such pronunciation patterns. For example, the performance of speech recognition often degrades 20-30% when the dialect of the input speech was not included in the training set [48]. Modeling pronunciation variation caused by dialects can improve speech recognition performance. In addition, there are many other applications that could benefit from explicitly modeling dialect-specific pronunciations. For example, accent training software, dialect identification, and speaker characterization and identification.

In this work, we generalize and apply the concept of pronunciation modeling from automatic speech recognition to the field of dialect recognition. Our proposed model is able to characterize phonetic transformations across dialects explicitly, thus contributing to linguistics and speech science as well. Thus, our interdisciplinary approach of analyzing dialects is related to three bodies of work reviewed below: (1) linguistic and speech science studies that characterize dialects, (2) automatic language and dialect recognition, and (3) pronunciation modeling in automatic speech recognition.

2.1 Terminology and Definitions

- **Phoneme**

  A *phoneme* is a linguistic term referring to the smallest distinctive units in a language. Phonemes are typically encased in "/ /" symbols, as we will show below. If a phoneme of a word is changed, the meaning of the word changes

as well. For example, the phonemes of the word *rice* is /r ai s/. If the first phoneme is changed from /r/ to /l/, then the meaning of the English word changes completely.

What is considered a phoneme varies by language. Although in English /r/ and /l/ are different phonemes, they are the same in Japanese. Thus in the previous example, a native Japanese speaker who only speaks Japanese is unlikely to perceive the difference if /r/ is replaced by /l/ in the word *rice* [39]. Therefore, phonemes are subjective to the native speaker.

- **Phone**

  In engineering, *Phones* are used much more commonly than *phonemes* to refer to units of speech sounds. The categorization of phones depends on acoustic properties and practical modeling considerations. For example, although flaps are only one acoustic realization of the phoneme /t/, it is modeled separately as [dx] since its acoustic properties are distinctive from canonical [t]'s. Note that unlike phonemes, phones are encased in brackets. In this work, we will use reference phones instead of phonemes to categorize speech sounds.

- **Phonetics**

  Phonetics is a branch of linguistics that studies the sounds of human speech. It is concerned with the physical properties of speech sounds (phones): their physiological production, acoustic properties, auditory perception, and neurophysiological status.

- **Phonology**

  Phonology studies how sounds function within a given language or across languages to encode meaning. Phonology studies the systematic patterns of these abstract sound units - the *grammatical* rules of phonemes. For example, *phonotactics* is a branch of phonology that deals with restrictions in a language on the permissible combinations of phonemes. Phonotactic constraints are language specific. For example, in Japanese, consonant clusters like /st/ are not allowed, although they are in English. Similarly, the sounds /kn/ and /n/ are

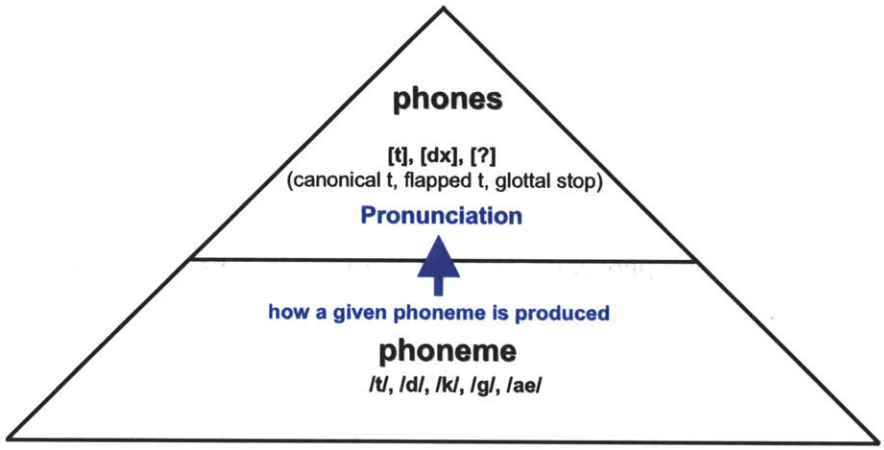

Figure 2-1: Pronunciation is the mapping between an underlying phoneme and its various phonetic implementations. [t] is the canonical t, [dx] is the flapped t, and [?] is glottal stop; all three are different ways a phoneme /t/ can be produced.

not permitted at the beginning of a word in English, but are in German and Dutch.

In this work, since we are building automated systems, we will use phones instead of phonemes. We will borrow the concept of phonology to characterize dialects, while using the basic units as phones instead of phonemes.

- **Pronunciation**

  Pronunciation is defined as how a given phoneme is produced by humans (see Fig. 2-1). Phones are the acoustic implementation of phonemes, and a phoneme can be implemented acoustically in several different phonetic forms. For example, the phoneme */t/* in the word *butter* could be acoustically implemented as a canonical t, a flapped t, or a glottal stop. A native English speaker would still be able to identify all 3 of these phones as the same underlying phoneme /t/ and recognize the word being uttered is *butter*, despite the phonetic differences. This thesis attempts to automatically identify when a phoneme is pronounced differently across dialects, and quantify how the magnitude of these differences.

- **Dialect**

  Dialect is an important, yet complicated, aspect of speaker variability. A dialect

is defined as the language characteristics of a particular population, where the categorization is primarily regional though dialect is also related to education, socioeconomic status, and ethnicity [23, 50, 81]. Dialects are usually mutually intelligible. For example, speakers of British English and American English typically understand each other.

- **Accent**

  The term accent refers to the pronunciation characteristics which identify where a person is from regionally or socially. In our work, we use accent information to distinguish between dialects of native speakers of a language or non-native accents.

- **Language Recognition**

  Language recognition refers to the task of automatically identifying the language being spoken by a person. Language recognition is often referred to language identification (LID) and language detection as well. The three terms will be used interchangeably in this work.

- **Dialect Recognition**

  Dialect recognition refers to the task of automatically identifying the dialect being spoken by a person. Dialect recognition is often referred to dialect identification (DID) as well. The two terms will be used interchangeably in this work.

2.2 Speech Science and Linguistic Studies

Within the same language, there are particular populations who have their own language characteristics. Since our work focuses on speech characteristics at the pronunciation level, we only review studies in acoustics, phonetics, and phonology.

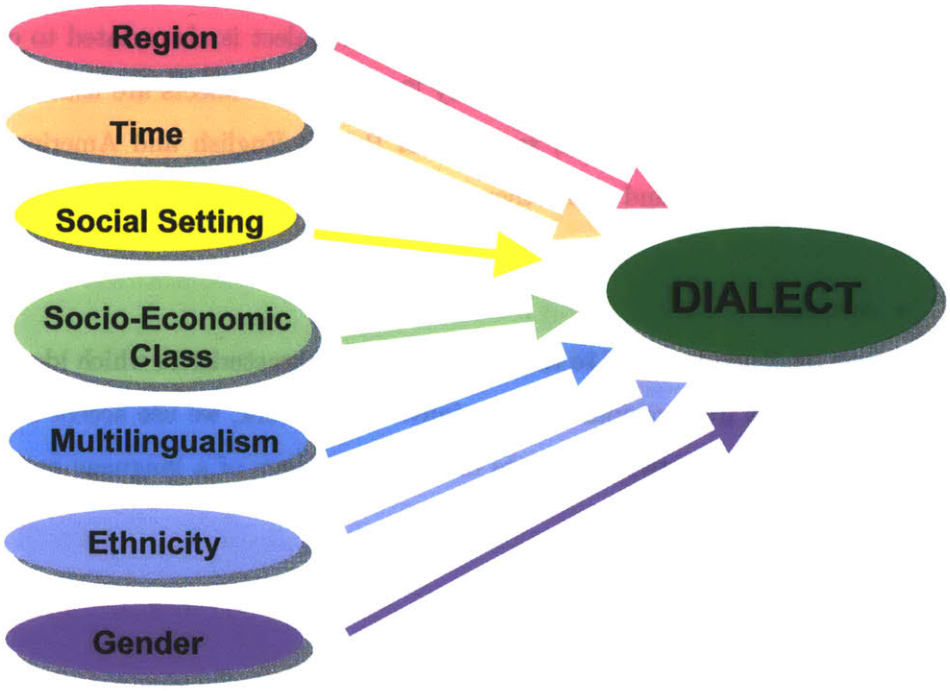

Figure 2-2: *Factors that influence or correlate with dialects.*

2.2.1 Factors that Influence Dialects

Dialect is an important, yet complicated, aspect of speaker variability. A dialect is defined as the language characteristics of a particular population, where the categorization is primarily regional, though dialect is also related to education, socioeconomic status, and ethnicity [23, 50, 81]. Dialects are usually mutually intelligible. For example, speakers of British English and American English typically understand each other.

Figure 2-2 illustrated the numerous factors that influence dialects. Below we discuss these factors in more detail.

- **Region**

  One of the most obvious things a dialect reveals is a person's geographical identity: where he grew up, and perhaps where he lives now. The boundaries of regional dialects could be across nations, such as Levatine Arabic, which include Palestine, Syria, Lebanon, which are located at the eastern-Mediterranean

coastal strip. The boundaries of regional dialects could also be defined by nations, such as American, British, and Australian English, where usually the standard dialect is used for comparison. For example, Received Pronunciation (RP)[1] in the U.K. and General American English (GenAm) in the U.S.A. Dialects could also have much finer boundaries within a country, such as Birmingham, Liverpool, Lancaster, Yorkshire, and Glasgow in U.K., and Boston, New York City, and Texas in the U.S.A [92].

Apart from using towns, counties, state, province, island and country to categorize dialects, there are often common characteristics to all urban dialect against rural ones. In addition, new trends in dialects are proposed to be spread from cities to towns, and from larger towns to smaller towns, skipping over intervening countrysides, which are the last to be affected. For example, *H Dropping* in U.K. has spread from London to Norwich and then from Norwich to East Anglian towns, while the Norfolk countryside still remains /h/-pronouncing [92].

Immigrants of different speech communities have also been proposed to cause sound change [24]. For example, Labov and colleagues' findings show that the Inland North dialect of American English has been undergoing the Northern City Chain Shift (NCCS)[2]. Inland North refers to cities along the Erie Canal and in the Great Lakes region, as well as a corridor extending across central Illinois from Chicago to St. Louis. This switch of dialect characteristics from the North to the Midland is possibly explained by migration of workers from the East Coast to the Great Lakes area (epically Scots-Irish settlers) during the construction of the Erie Canal in the early 19th century [24].

- **Time**

  The fundamental reason why dialects differ is that languages evolve, both spatially and temporally. Any sound change currently in progress can be demon-

[1]Received Pronunciation (RP), also called the Queen's (or King's) English, Oxford English, or BBC English, is the accent of Standard English in England

[2]As illustrated in Fig. 2-3, NCCS is characterized by a clockwise rotation of the low and low-mid vowels: /ae/ is raised and fronted; /eh/, /ah/ and /ih/ are backed; /ao/ is lowered and fronted; /aa/ is fronted [56]

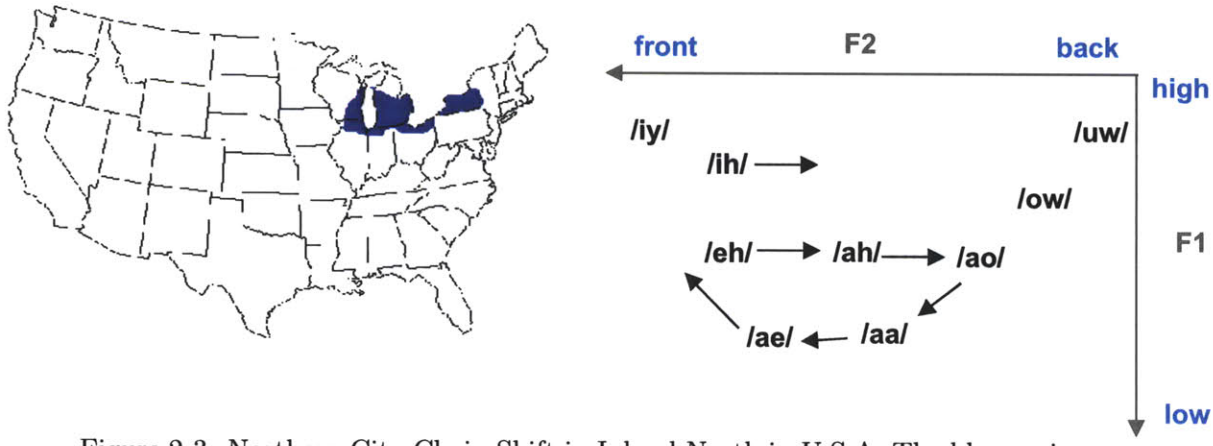

Figure 2-3: Northern City Chain Shift in Inland North in U.S.A. The blue regions on the left indicate the Inland North region, and the vowel plot on the right indicates the vowel shift occurring in this region, proposed by Labov et. al. [56].

strated to be in progress by examining the speech patterns of different age groups. For example, Labov found that the percentage of speakers using non-rhotic vowel in the word *nurse* correlates with age in New York City in 1966. All speakers above age 60 showed no rhotic characteristic in the vowel of *nurse*, while only 4% of speakers between 8 and 19 did so. This phenomenon is explained by the fifty-year olds adopting the General American rhotic vowel, which became the norm for later generations.

- **Social Setting**

  It has long been known that an individual will use different pronunciation patterns in different social circumstances. People tend to use a more standard dialect at formal settings, and switch to a relatively-more native dialect under casual conversations with friends or family. For example, a study showed that the percentage of the alveolar nasal [n] used in words containing *-ing* (as opposed to the velar nasal [ng]) increases as speaking style becomes more casual [92]: reading a list of words, reading passages, formal speech, and casual speech.

  It is also know that speakers of different dialects will accommodate to each others dialect when conversing with each other [35]. Therefore an African American English (AAVE) speaker would show more non-AAVE (e.g., white) dialect characteristics when speaking to a non-AAVE speaker. For example, it has been shown that Oprah Winfrey, an African-American host of popular U.S. daytime

talk show, monophthongizes /ay/ more frequently when the guest is an AAVE speaker [41].

- **Social-Economic Class**

  It has been shown that social-economic class also correlates with dialect differences. For example, Trudgill (1974b:48) found that there is a correlation between /t/ being glottalized in syllable-final positions such as *butter* and *bet* with social-economic class in Norwich, U.K.: around 90% of the working class speakers tend to glottalize syllable-final /t/'s while only half of the middle-class speakers do so.

  Usually the new, fashionable trend originates in the upper or upper-middle class, and spread to other social-economic classes. However, it has been observed that this is not the only direction. For example, H Dropping in U.K. originates from working-class speech, and has spread outwards to other social classes.

- **Multilingualism**

  A person's mother tongue might influence how a person speaks another language acquired later in life. For example, English dialects in India and Singapore are influenced by their native languages (such as Hindi and Hokkien).

- **Ethnicity**

  Many of the accent characteristics often thought of as ethnic are in fact geographical [92]. It is likely that ethnicity correlates with where these speakers live and grow up, which causes dialect differences to be correlated with ethnicity.

- **Gender**

  Holding other factors constant, it has repeatedly been found that women's pronunciation characteristics are closer to the prestige norm than men in studies of English speakers [92]. There are two main explanations for this phenomenon, both related to the sexist characteristic of our society. First, in western societies women are usually more status-conscious than men. Therefore, women make up for this social insecurity through emphasizing and displaying linguistic trends

that are of higher social status. Second, working class accents are connected with masculinity characteristics, which might seem socially inappropriate for women to adopt [92].

2.2.2 How Dialects differ

- **Acoustic Realization**

  The acoustic realization of a phoneme can be different across dialects. For example, in GenAm voiceless stops (/p/, /t/, /k/) are always aspirated unless when proceeded by preceded by /s/. Therefore, when saying the words *spray* and *pray* in GenAm, there is much more air coming out of your mouth in the latter case due to aspiration. However, this aspiration of voiceless stops is not found in Indian English and North England and Scotland.

- **Phonotactic Distribution**

  Phonotactics refers to the constraints of phone sequences in a language or dialect. Rhoticity in English is the most well-known case of different phonotatic distributions across dialects. In non-rhotic accents, /r/ is not allowed at prevocalic positions. Therefore, words such as *farther* sound like *father*, and the word *far* sounds like *fa* with a longer vowel sound. General American English (GAE) is rhotic while RP is not [92].

- **Splits and Mergers**

  Phonemic systems within a dialect change over time. *Splits* occur when new phonemes arise by evolving away from a previous phoneme. For example, the vowels in *trap* and *bath* used to sound the same in RP, but had developed to different phonemes of short /ae/ and long /aa/ in the twentieth century. This so called *trap-bath split*, can be characterized by its phonological environment to some extent, as the split also is word-dependent. The pattern of trap-bath split is that the short vowel /ae/ becomes the long, back vowel /aa/ in RP when it is followed by a voiceless fricative, or nasal that is followed by a consonant: staff, past, path, gasp, ask, castle, fasten, dance, aunt, branch, command, sample.

However, there are words that provide the exact same phonetic context as these words, yet /ae/ remains /ae/ in RP: *gaff, math, mascot, hassle, cancer, ant, mansion, hand, ample.*

Distinct phonemes also *merge* and become indistinguishable. For example, increasing numbers of speakers in the U.S. are merging the vowels in *thought* and *lot.* Traditionally, the northern dialect (as opposed to midland and southern areas) in the U.S. pronounce the following minimal pairs differently: *collar* vs. *caller*, *cot* vs. *caught*, *stock* vs. *stalk*, *don* vs. *dawn*, *knotty* vs *naughty.* The former is /aa/, an open, back vowel, while the latter is /ao/, a lightly-rounded half-open, back vowel. Note that the NCCS phenomenon described above, also included this merger.

- **Lexical Diffusion**

  Differences of lexical diffusion (a.k.a., lexically-specific change) are defined as those differences between accents of a language which are not pervasive throughout all eligible words of the language; i.e., it is impossible to define a structural context in which the alternation takes place. Instead, the alternation refers to limited groups of words. The most common example in English is the alternation between /ay/ and /iy/ in the words *either* and *neither.* The previously mentioned trap-bath split is another example where /ae/ and /aa/ are alternatively used in certain words involving final voiceless fricatives.

2.2.3 Second Language Accents

The origin of non-native accents are different from dialects, but the mechanisms used to characterize non-native accents could be similar. The most common ways to characterize non-native accents are acoustic realizations and phonotactic distributions as mentioned before.

One of the mainstream theories explaining second language (L2) accents is that these speakers carry over phonetic and phonological characteristics of their mother tongue when learning a new language at an age older than 12 years old [28]. These

speakers therefore substitute phonemes from their first language (L1) when they encounter a new phoneme in the second language. For example, in Spanish and Mandarin Chinese there are no short vowels such as [ih] in *bit*. Therefore, the word *bit* is likely to sound like *beat*, substituting the short vowel [ih] to the long vowel [iy] [16].

In addition to substitutions, deletions and insertions can also occur according to the phonetic and phonological system of the native language. The English phoneme /h/ does not exist in French, therefore French speakers systematically delete /h/ when speaking English. For instance, the word *hair* will sound like *air* instead [69]. In Spanish, /s/ must immediately precede or follow a vowel; often a word beginning with /s/ followed by a consonant will be inserted with a schwa before the /s/ (e.g., school, stop, spend) [37].

Besides the age when the speaker starts learning the second language, there are other factors that affect the degree of non-nativeness in second language speakers: exposure time to L2, L1 and L2 similarity, individual differences in acquiring a new language. These factors might also complicate the analysis of L2 accents.

2.3 Automatic Language and Dialect Recognition

Automatic language/dialect recognition refers to the task of automatically identifying the language/dialect being spoken by a person. Their applications fall into two main categories: (1) pre-processing for machine understanding systems, and (2) pre-processing for humans [100]. For example, a multi-lingual voice-controlled travel information retrieval system at a hotel lobby could benefit international travelers by using their native language/dialect to interact with the system. In addition, DID can be used in customer profiling services. For example, Voice-Rate, an experimental dialogue system at Microsoft, uses accent classification to perform consumer profile adaptation and targeted advertisements based on consumer demographics [15]. DID could be applied to data mining and spoken document retrieval [96], and automated speech recognition systems (e.g., pronunciation modeling [60], lexicon adaptation [91], acoustic model training [51].)

In this chapter, we first introduce the basic structure of a LID/DID system in Section 2.3.1, and then we delineate the probabilistic framework of LID/DID systems in Section 2.3.2. In Section 2.3.3 and 2.3.4, we given an historical overview of how research in LID and DID has developed over the past 4 decades. In Section ??, we discuss how our proposed approach connects the field of DID to speech science and pronunciation modeling in automatic speech recognition.

2.3.1 System Architecture

Given the pattern recognition framework, language recognition systems involve two phases: training and recognition. In the training phase, using language-specific information, one or more models are built for each language. In the recognition phase, a spoken utterance is compared to the model(s) of each language and then a decision is made. Thus, the success of a language recognition system relies on the choice of language-specific information used to discriminate among languages, while being robust to dialect speaker, gender, channel, and speaking style variability.

Training

1. **Feature Extraction**

   The goal of LID/DID research to date has generally been to develop methods that do not rely on higher-level knowledge of languages and dialects, but use only the information that is available directly from the waveform. Typical acoustic features used in language and dialect recognition include those used in ASR, such as perceptual linear prediction (PLP) [44] and Mel-frequency cepstrum coefficients (MFCC) [21]. Shifted-delta cepstrum (SDC) [8] has also led to good performance [87]. Features characterizing prosody such as pitch, intensity, and duration, especially in combinations with other features, are sometimes used as well. Phonetic or phonotactic information, captured by using features that have longer time spans such as decoded phones and decoded phone sequences [99].

2. **Training Dialect-Specific Models**

   During the training phase, dialect-specific models are trained. Common model choices include Gaussian mixture model (GMM), HMM, N-grams, support vector machines (SVM), etc. More details on the algorithms of these models are discussed in Section 5.

Recognition

1. **Feature Extraction**

   The same features extracted during the training phase are extracted during the recognition phase.

2. **Pattern Matching**

   During the recognition phase, the likelihood scores of the unknown test utterance O are computed for each language-specific model λ_l: $P(O|\lambda_l)$.

3. **Decision**

   In the decision phase, the log likelihood of each test trial of model λ_l is scored as

$$\log \frac{P(O|\lambda_l)}{\sum_{i \neq d} P(O|\lambda_i)}. \tag{2.1}$$

   As shown below, if the log likelihood of O of model λ_l is greater than a decision threshold θ, the decision output is language l.

$$\log \frac{P(O|\lambda_l)}{\sum_{i \neq l} P(O|\lambda_i)} > \theta, \tag{2.2}$$

   The performance of a recognition system is usually evaluated by the analysis of detection errors. There are two kinds of detection errors: (1) *miss*: failure

to detect a target dialect, and (2) *false alarm*: falsely identifying a non-target dialect as the target. The *detection error trade-off* (DET) curve is a plot of miss vs. false alarm probability for a detection system as its discrimination threshold θ is varied in Eq. (2.2). There is usually a trade-off relationship between the two detection errors [63]. An example of a DET curve is plotted on normal deviate scales is shown in Figure 2-5.

As shown in Figure 2-5, the cross over point between the DET curve and $y = x$ is the *equal error rate (EER)*, indicating the miss and false alarm probabilities are the same. EER is often used to summarize the performance of a detection system.

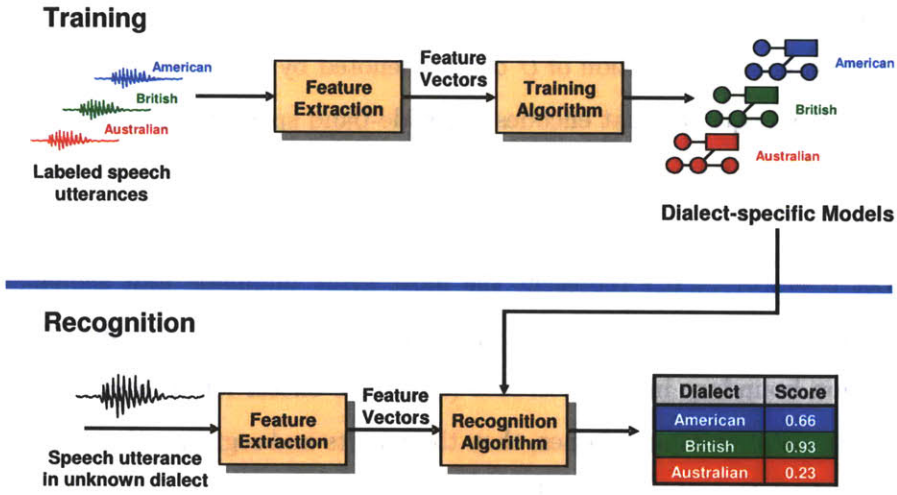

Figure 2-4: *Dialect Recognition System Architecture.*

2.3.2 Probabilistic Framework

Hazen and Zue [42] formulated a formal probabilistic framework to incorporate different linguistic components for the task of language recognition, which we will adopt here to guide us in explaining the different approaches. Let $L = \{L_1, ..., L_n\}$ represent

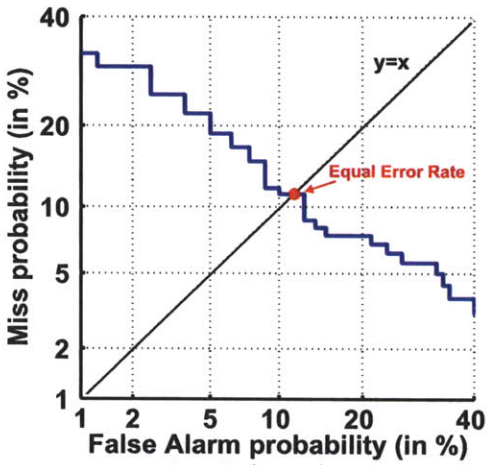

Figure 2-5: *Detection error trade-off (DET) and equal error rate (EER) example.*

the language set of n different languages. When an utterance U is presented to the LID system, the system must use the information from the utterance U to determine which of the n languages in L was spoken.

The acoustic information of U can be denoted by (1) $w = \{w_1, ...w_m\}$, the frame-based vector sequence that encodes the wide-band spectral information, and (2) $f = \{f_1, ..., f_m\}$, the frame-based prosodic feature vector sequence (e.g., the fundamental frequency or the intensity contours).

Let $\upsilon = \{\upsilon_1, ..., \upsilon_p\}$ represent the most likely linguistic unit sequence obtained from some system, and $\epsilon = \{\epsilon_1, ..., \epsilon_{p+1}\}$ represent the corresponding alignment segmentation boundary in the utterance (e.g., time offsets for each unit). For example, if our linguistic units are phones, then these units and segmentations can be obtained from the best hypothesis of a phone recognizer.

Given the wide-band spectral information w, the prosody information f, the most likely linguistic-unit sequence υ, and its segmentations ϵ, the most likely language is found using the following expression:

$$\arg\max_i P(L_i|w, f, \upsilon, \epsilon), \tag{2.3}$$

using standard probability theory this expression can be equivalently written as

$$\arg\max_i P(L_i)P(\upsilon|L_i)P(\epsilon, f|\upsilon, L_i)P(w|f, \upsilon, \epsilon, L_i). \tag{2.4}$$

To simplify the modeling process, we can model each of the four factors in Eq. (2.4) instead of the complicated expression in Eq. (2.3). These four terms in Eq. (2.4) are known as

1. $P(L_i)$: The *a prior* probability of the language.
2. $P(\upsilon|L_i)$: The phonotactic model.
3. $P(\epsilon, f|\upsilon, L_i)$: The prosodic model.
4. $P(w|f, \upsilon, \epsilon, L_i)$: The acoustic model.

While robust methodologies are available for modeling acoustic and phonotactic information, well-developed techniques for automatically characterizing prosody, especially at word- and sentence-levels, are still elusive. Since prosody is beyond the scope or our work, we do not include a background on prosodic modeling approaches. Interested readers can refer to work such as [2, 1, 6] for more details.

Below we go into more detail on some basic approaches that has shaped the development of LID and DID in phonotactic and acoustic modeling.

Phonotactic Modeling

The phonotatic approach is based on the hypothesis that languages/dialects differ in their phone sequence distribution. Assuming the prior distribution of the languages L is uniform, and ignoring the acoustic and prosodic models in Eq. (2.4), the language recognition problem simply becomes:

$$\arg\max_i P(\upsilon|L_i) \tag{2.5}$$

PRLM (Phone Recognition followed by Language Modeling)

A well-known method for modeling phonotactic constraints of languages/dialects is

PRLM (Phone Recognition followed by Language Modeling) [100]. In PRLM, the training data are first decoded through a single phone recognizer. Then an N-gram model is trained on the decoded phones for each language/dialect L_i. typical choices of N are 2 and 3. The interpolated N-gram language model [52] is often used reduce data sparsity issues. For example, a bigram model is

$$\tilde{P}(\upsilon_t|\upsilon_{t-1}) = \kappa_2 P(\upsilon_t|\upsilon_{t-1}) + \kappa_1 P(\upsilon_t) + \kappa_0 P_0, \tag{2.6}$$

where υ_t is the phone observed at time t, P_0 is the reciprocal of the total number of phone symbol types from the phone recognizer, and the κ's can be determined empirically.

Parallel PRLM

Parallel PRLM is an extension to the PRLM approach, using multiple parallel phone recognizers, each trained on a different language. Note that the trained languages need not be any of the languages the LID task is attempting to identify. The intuition behind using multiple phone recognizers as opposed to a single one is to capture more phonotactic differences across languages, since different languages have different phonetic inventories.

Acoustic Modeling

Acoustic modeling has received much attention in the past decade both in language and speaker recognition, due to the simplicity and good performance of GMMs. The recognition problem can be expressed similarly as in phonotactic modeling from Eq. (2.4):

$$\arg\max_i P(w|f, \upsilon, \epsilon, L_i) \tag{2.7}$$

As mentioned in Section 2.3.1, typical acoustic features used in language and dialect recognition include those used in ASR, such as PLP and MFCC.

Gaussian Mixture Model (GMM)

Most acoustic approaches in language and dialect recognition use a GMM, at some

point in their system, to characterize the acoustic space of each language/dialect. GMM assumes that the acoustic vectors w are independent of the linguistic units v, segmentation ϵ, and prosody f, and the acoustic frames are assumed to be i.i.d (independent and identically distributed), simplifying Eq. (2.7) to the following

$$\arg\max_i \Pi_{t=1}^{T} P(w_t|L_i), \tag{2.8}$$

where T is the total number of frames. Assuming that the acoustic distribution is a GMM, the recognition problem can further be expressed as:

$$\arg\max_i \Pi_{t=1}^{T} \sum_{k=1}^{K} \varrho_{ik} N(w_t; \mu_{ik}, \Sigma_{ik}), \tag{2.9}$$

where there are K mixtures, and $\varrho_{ik}, \mu_{ik}, \Sigma_{ik}$ are the weight, mean, and covariance matrix of the k-th Gaussian in dialect i, and N represents the probability density function of the normal distribution.

Universal Background Model (UBM)

In the UBM approach, a dialect-independent GMM is first trained. Then separate GMMs for each dialect is derived by adapting the UBM to the acoustic training data of that dialect using MAP [34]. Advantages of using a UBM approach [75] include the following.

- Performance: The tight coupling between the dialect-specific models and the UBM has shown to outperform decoupled models in speaker recognition [75].

- Insufficient Data: If the training data of a particular dialect is insufficient, MAP can provide an more robust model by weighting the UBM more

- Experiment Speed: Training new dialect models are faster, since it only requires a few adaptation iterations, instead of running EM again.

2.3.3 Historical Development of Language Recognition

Language recognition has been an active research area for nearly 40 years. The pioneering studies date back to Leonard & Doddington (1974) using an acoustic filter-bank approach to identify languages[59]. House & Neuburg (1977) were the first to use language-specific phonotactic constraints in LID [45]. Most other approached proposed in the 1980's were acoustic modeling such as simple frame-based classifiers on formant features [29, 38],

During the past two decades, research in LID developed intensively and rapidly, which is due to at least three reasons: (1) the availability of large and public corpora, (2) the NIST Language Recognition Evaluation (LRE) series, and (3) the influence of the speaker recognition community.

- **Influence of large, public corpora**

  Most of the the basic architectural and statistical algorithmic development in language recognition occurred in the 1990's, which was enabled by large and publicly available corpora[3]. These developments include the popular phonotactics approach of PRLM/PPRLM [42, 99, 94], standard N-grams and bintrees [65], and acoustic approaches using GMM [43]. Although acoustic modeling approaches performed considerably worse in NIST LREs than phonotactics, the two approaches fused well.

- **Influence of NIST Language Recognition Evaluations**

  NIST LRE series (1996 -2009) provided a common basis for comparison on well-defined tasks, enabling researchers to replicate and build on previous approaches that showed good performance. Methods such as GMM-UBM and phonotactic training on phone lattices [83] have consistently shown robust performance across datasets and tasks. New front-end features such as shifted delta cepstra (SDC), also consistently showed better performance than traditional Mel-cepstra features, making the performance of SDC-GMM systems comparable to

[3]e.g., Callfriend, Callhome, and OGI-11L and OGI-22L (including manual phone transcriptions and 100 speakers/language.)

that of PRLM, but required less computational resources [87].

- **Influence of Speaker Recognition**

  Language recognition was also recast as a detection problem in the NIST LREs, which made LID heavily influenced by the speaker recognition community. Studies inspired by speaker recognition include discriminative training of maximum mutual information of GMMs [9] and support vector machines (SVM) [10], subspace-based modeling techniques such as eigenchannel adaptation [49], feature-space latent factor analysis (fLFA) [11], and nuisance attribute projection (NAP) for GMM log likelihood ratio systems [97].

2.3.4 Historical Development of Dialect Recognition

Early studies of DID include work of Arslan and Hansen (1996, 1997), where they used Mel-cepstrum and traditional linguistic features to analyze and recognize non-native accents of American English. Research in DID is motivated by applications in ASR, business, and forensics [4, 15, 13, 7, 26]. The recent addition of dialect tasks in NIST LREs has drawn many researchers in the language and speaker recognition to work on the challenging problem of dialect recognition.

Challenges in DID

Dialect recognition is a much more challenging problem than language recognition.

- **Dialect Differences**: The differences across dialects (of the same language) are often much more subtle than differences across languages.

- **Definition of Dialects**: The definition of dialects is controversial in its linguistic nature. Speakers of the same language evolve their language characteristics over time and space. All these differences in language characteristics can be accounted to dialect differences. Dialects that are very similar, might have little acoustic or phonetic differences (e.g., Central vs. Western Canadian English), while dialects that are very different are sometimes viewed as virtually different

languages (e.g., Arabic dialects). In addition, political factors sometimes come into play, making it difficult to judge whether some classification tasks make linguistic sense in the first place.

- **Dialect Databases**: Unlike language recognition, unifying databases for dialect recognition research are still limited.

Recognition Approaches in DID

The default approach in tackling DID is to view each dialect as a separate language. Therefore, similar to LID, modeling techniques in identifying dialects take advantage of different layers of the linguistic hierarchy [100]. These approaches include (1) acoustic models (e.g. [84, 15, 89, 4, 47, 13, 26]); (2) phonotactic models (e.g. [7, 98, 76]). Research in dialect recognition is also inspired by that in speaker recognition. For example, discriminative training such as hybrid SVM/GMM systems showed great performance [26]; channel compensation techniques such as latent factor analysis also leads to robust performance [88].

Typical acoustic approaches often model spectral features using Gaussian mixture models (GMM). While they can achieve good performance, they do not provide insight into where dialect differences occur. Adapted phonetic models [84] is an extension of GMM, where acoustic information is modeled in phonetic categories, making it easier to pinpoint where the acoustic differences lie in. In our previous work [13], acoustic differences caused by phonetic context were further used to infer underlying phonetic rules, making the dialect recognizer linguistically-informative to humans.

Compared to acoustic models, phonotactic systems usually contain more human-interpretable information. However, most phonotactic systems do not focus on their potentially interpretable results. Exceptions include [76, 7], where discriminative classifiers are trained to recognize dialects [76, 7], and N-grams or context-dependent phones helpful in dialect recognition are discussed. To fill in this research gap, in our work we propose to establish a framework for informative dialect recognition systems.

2.4 Pronunciation Modeling in Automatic Speech Recognition

Automatic speech recognition (ASR) is usually accomplished by classifying the speech signal into small sound units (e.g., phones) and then mapping them to words, and eventually phrases and utterances [32]. This mapping between sound units to words is referred to as pronunciation modeling (see Fig. 2-6). The pronunciation model in an ASR system is often specified as a pronunciation dictionary, which is a list of words and their corresponding pronunciations, shown in terms of the phoneset of the ASR system.

Pronunciation models deal with pronunciation variations caused by factors such as dialect, first language [3, 33, 54, 62, 66, 67], speaking style [32], degree of formality [55, 57], anatomical differences, and emotional status [86].

All data-driven approaches in pronunciation modeling require two steps (1) finding pronunciation rules, and (2) using pronunciation rules (in ASR systems).

2.4.1 Finding Pronunciation Rules

A typical approach of modeling pronunciation variation in ASR is to align canonical phone transcriptions (any phone transcription that can serve as reference to compare with) and alternative phone transcriptions [31, 62, 67]:

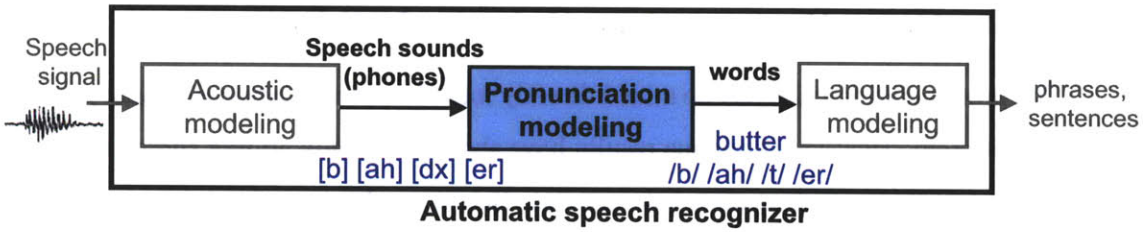

Figure 2-6: Pronunciation modeling in automatic speech recognition. The alternative pronunciation of butter is [b ah dx er]. where /t/ is *flapped*, denoted as [dx]. A flap is caused by a rapid movement of the tongue tip brushing the alveolar ridge. The pronunciation model maps [dx] to /t/.

Step 1: Generate canonical phone transcription using the baseline pronunciation dictionary, word identities of the training data, and the phone recognizer.
Step 2: Generate alternative phone transcriptions by directly decoding the training data using the phone recognizer.
Step 3: *Align* canonical phone transcriptions (from Step 1) and alternative phone transcriptions (from Step 2).
Step 4: New pronunciation learned in Step 3 are selected and added to the baseline pronunciation dictionary to form the new pronunciation dictionary.

This procedure can be iterated by modifying Step 2 to generate alternative phone transcriptions by force-aligning the training data with the new pronunciation dictionary (obtained in Step 4) and word identities. This modification in Step 2 reduces the noise caused by phone recognition errors [62, 85].

This procedure can be used on a word-by-word basis, but it can be further extended to a phonetic basis to learn more generalized pronunciation rules. Decision tree models are useful in learning context-dependent rules, where phones that behave similarly in different phonetic contexts are grouped together [77].

The selection criteria in Step 4 include (1) frequency occurrences of the alternative pronunciations, (2) maximum likelihood, (3) confidence measures, and (4) degree of confusability between variant pronunciations [86].

2.4.2 Using Pronunciation Rules in ASR Systems

The pronunciation rules can be added directly into the pronunciation dictionary in the ASR system. However, previous results have shown that only adding these pronunciation variants to the lexicon is sub-optimal. Better results are achieved by taking the probabilities of the pronunciation variants into consideration (either in the lexicon or language model), and retraining the acoustic models.

Pronunciation modeling is often closely tied with acoustic modeling. Instead of using the updated pronunciation dictionary (learned in Step 4) acoustic models can

Primary Focus	Improve engineering performance	Automatically learn rules
	Automatic Dialect Recognition Richardson & Campbell (2009) Biadsy et al (2010)	*Informative Dialect Recognition* Chen et al (2010, 2011)

Figure 2-7: Comparison of work related to Informative Dialect Recognition in the field of automatic speech recognition.

be retrained according to the learned pronunciation rules. For example, if a pronunciation rule learned is that /er/ is produced as schwa in syllable-final positions, the acoustic model of /er/ is retrained/adapted with speech data that are canonically /er/ but produced as schwa, allowing the acoustic model to tolerate more pronunciation variation. In practice, the pronunciation model and acoustic model are often both adapted to further improve ASR performance.

2.5 Work Related to Informative Dialect Recognition

There is very limited work related to our work. Below we point out some related work from the fields of automatic dialect recognition, sociolinguistics, automatic speech recognition, computer-aided language learning.

In automatic dialect recognition, the primary focus is to improve recognition accuracy. Few studies touch upon the possible linguistic interpretations of their recognition results. In [76, 7], discriminative classifiers are trained to recognize dialects [76, 7], and N-grams or context-dependent phones helpful in dialect recognition are discussed. Our work in informative dialect recognition is very complementary with theirs, as our primary focus is to automatically learn dialect-specific rules, which could be applied to automatic dialect recognition. See Figure 2-7 for comparison.

Figure 2-8 shows comparison of informative dialect recognition research with other work based on the purpose and method used. Some studies in sociolinguistics (e.g.,

purpose / method	Improve speech analysis efficiency	Automatically learn rules
Directly apply ASR tools	*Sociolinguistics* Chen et al (2009) Evanini et al (2009) Yuan & Liberman (2009) *Computer-aided language learning* Kim et al (2004)	*Automatic Speech Recognition* Livescu & Glass (2000), Kim et al (2007)
Generalize ASR concepts	*Informative Dialect Recognition* Chen et al (2010, 2011)	*Informative Dialect Recognition* Chen et al (2010, 2011)

Figure 2-8: Comparison of work related to Informative Dialect Recognition in the fields of sociolinguistics, computer-aided language learning, and automatic speech recognition.

[25, 95]) and computer-aided language learning (e.g., [?]) have taken advantage of tools in automatic speech recognition to reduce manual labor. In particular, forced-alignment of word transcriptions and the pronunciation dictionary has been used to determine the time boundaries of phonetic units. This automated procedure improves speech analysis efficiency, since manual phone transcription is no longer needed (or only fine tuning is required). The rule analysis phase is still primarily manual in these cases.

In informative dialect recognition, concepts in automatic speech recognition are further generalized to make the rule analysis part easier, by postulating phonetic rules to the analyst or pin-pointing regions of potential interest where acoustic characteristics are different across dialects.

In multilingual speech recognition, the pronunciation modeling part does automatically learn rules (e.g., [62, 54]). However, most of these studies do not use models that are linguistically grounded to learn rules explicitly. Our approach in informative dialect recognition is to propose a system that is specifically designed to articulate

the phonetic transformations across dialects.

2.6 Summary and Discussion

In this chapter we surveyed the literature of dialect studies at the acoustics, phonetics, phonology, pronunciation levels in the fields of linguistics, speech science, automatic language and dialect recognition, and pronunciation modeling in automatic speech recognition. We compare our work with the mainstream research themes and methodology in these three fields below.

- **Pronunciation modeling in Automatic Speech Recognition**
  There are numerous differences between pronunciation modeling in ASR and our work.
  - In ASR, all aspects of pronunciation variation is modeled, be it dialect, speaker, speaking style. In contrast, our work focuses on characterizing pronunciation variation due to dialect differences.
  - In ASR, the goal is to decrease word error rate (WER). In contrast, our goal is to analyze and quantify the *generality* of found pronunciation patterns, or how well the found pronunciation patterns characterize a dialect. WER performance on multi-dialect ASR is only one potential approach to indirectly measure how the found pronunciation patterns characterize dialects. In addition, the phone error rate of the proposed pronunciation model (ground-truth phones of a dialect vs. the most likely phone sequence generated by the proposed model) could serve as a metric to quantify rule recovery performance (see Chapter 6 for more details).

- **Automatic Dialect Recognition**
  In this work, explicitly model phonetic rules and acoustic differences across dialects by generalizing and adopting the concept of pronunciation modeling [32]. We employ decision tree clustering to account for data insufficiency and exploit acoustic properties of phonetic context. This context clustering approach is

similar to binary tree language modeling [64] in spirit, though the probabilistic models are different: in [64], the probability of a current observation is conditioned on a cluster of past observations, whereas in our model the probability of a current observation is conditioned on a reference phone and its context. Using reference phones as a comparison basis, we can explicitly model phonetic transformations occurring in different dialects, making the results of our dialect recognition system interpretable to humans.

The primary goal of this work is not to improve the performance of automatic dialect recognition, though the performance of dialect recognition could be an indirect indicator of how well the proposed models learned pronunciation rules. If pronunciations differ across dialects, and the proposed model learns dialect differences perfectly well, it is fair to assume that the proposed model will do reasonably well in dialect recognition experiments.

- **Speech Science and Linguistic Analysis**

  Our work proposes a mathematical framework that can be easily applied to model pronunciation rules automatically despite the language of interest. Due to the automatic nature of the model, our work can process more data efficiently. In addition, the proposed model is free of potential (unconscious) subjective bias in analyzing dialects.

  This framework can serve as a first pass to automatically characterizing unknown/unfamiliar dialects. This procedure helps linguists pinpoint regions of interest more efficiently, and provides an initial hypothesis (e.g. a phonetic transformation rule) for linguists to test. It is possible that the proposed rule is not characterized in the most optimal way, but it provides the linguist with a basis hypothesis to be further refined. In traditional methods, the formulation of this basis hypothesis might take a large amount of perceptual observation and manual analysis. In our work, we simulate this process automatically, so the linguists can save time and effort.

In the following chapter, we discuss approaches to adopt and generalize the concept of pronunciation modeling in ASR (Chapter 3) to automatically characterize dialect variations.

Chapter 3

Pronunciation Model

In this chapter we first introduce the basic structure of Phone-based Pronunciation Model (PPM). We then address the limitations of PPM, and propose refined models.

3.1 Intuition of Phone-based Pronunciation Model (PPM)

The intuition behind the proposed model, Phone-based Pronunciation Model (PPM) is to characterize dialectal pronunciations through phonetic transformations of a reference dialect. The purpose of the reference dialect is to use one dialect as a basis for comparison. Figure 3-1 shows an phonetic transformation of [ae] in American English being substituted by [aa] in British English.

There are three kinds of phonetic transformations: substitution, deletion, and insertion, which we describe below.

3.1.1 Substitution

A substitution occurs when a phone is acoustically realized differently from the reference dialect. For example, in Fig. 3-2 we see the word *bath* is pronounced differently in American and British English: the reference phone /ae/ become [aa] in British English. We know that /ae/ does not always transform to [aa], since the vowel in

trap is [ae] instead of [aa] in British English. Therefore, in the proposed model, we automatically learn when /ae/ is realized as [aa] in British English, and quantify how often this substitution occurs.

3.1.2 Deletion

A deletion occurs when a reference phone is not acoustically realized in the dialect of interest, thus *deleted*. Fig. 3-2 shows an example of non-rhoticity in British English, where /r/ is deleted when preceded by /aa/ and followed by a consonant in words such as *park*.

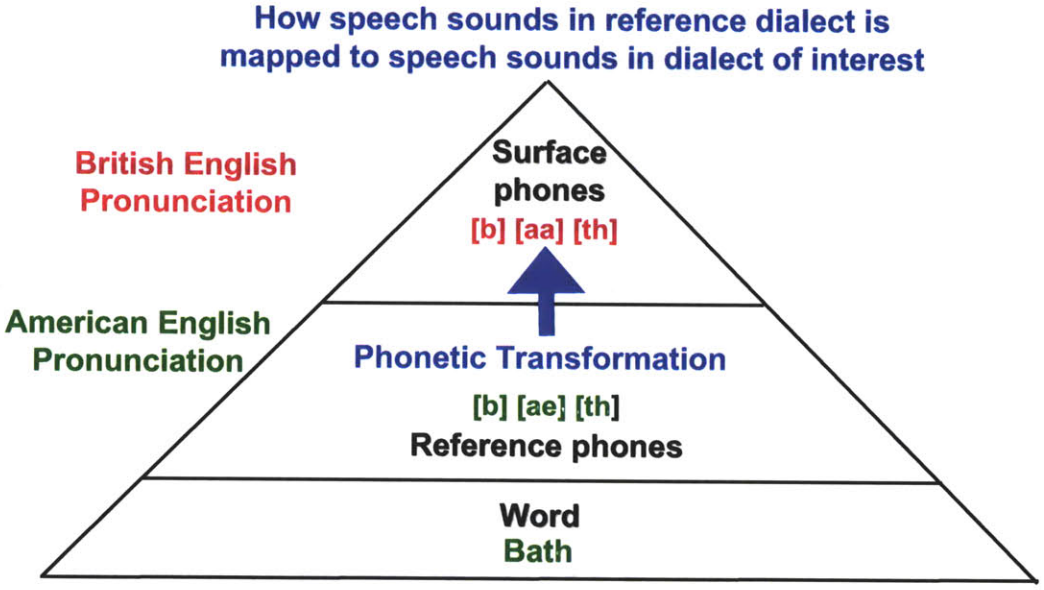

Figure 3-1: Phonetic transformation: an example of [ae] in American English pronunciation (reference phones) transforming to [aa] in British English (surface phones).

3.1.3 Insertion

An insertion occurs when a phone is produced in the dialect of interest, although it has no corresponding phone in the reference dialect. Fig. 3-2 shows two examples of the intrusive r insertion rule: [r] is inserted in between a vowel-ending word and a vowel-initial word in the phrases *the idea(r) is* and *saw(r) a film*. Note that [r] is not inserted after *idea* and *saw* when spoken in isolation.

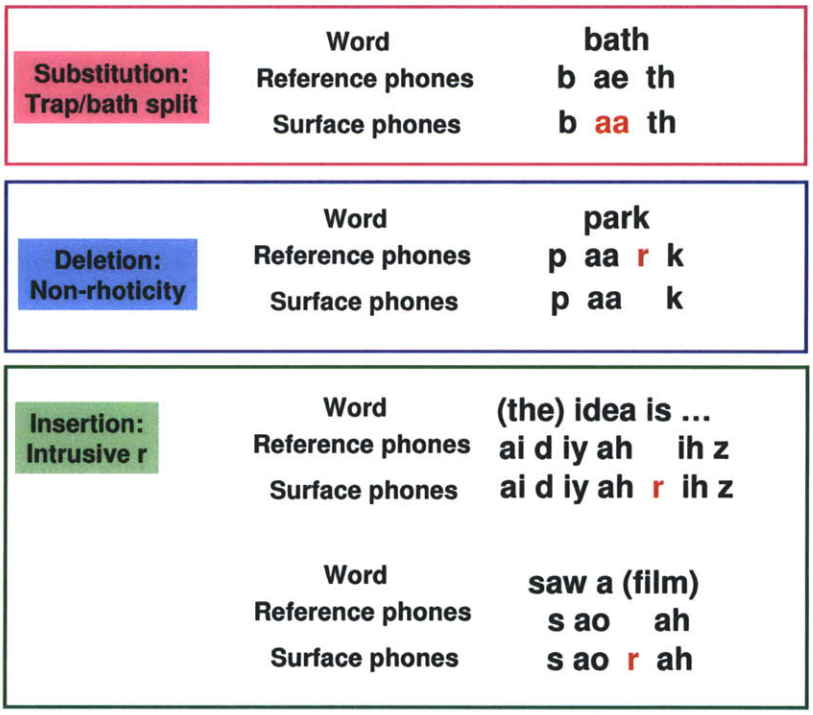

Figure 3-2: Examples of phonetic transformations that characterize dialects. Reference phones are American English pronunciation, and surface phones are British English pronunciation.

3.2 Mathematical Framework

Phonetic-based Pronunciation Model (PPM) is a hidden Markov model (HMM). The reference dialect's pronunciation is modeled by the states, and the pronunciation

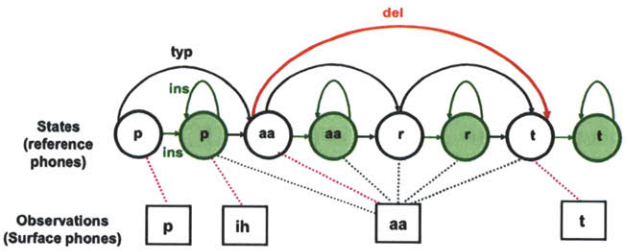

Figure 3-3: *Each reference phone is denoted by a normal state (black empty circle), followed by an insertion state (filled circle); squares denote emissions. Arrows are state transition arcs (black: typical; green: insertion; red: deletion); dash lines are possible alignments.*

of the dialect of interest is modeled by the observations. Phonetic transformations (insertion, deletion, substitution) across dialects are modeled by state transition probabilities.

3.2.1 HMM (Hidden Markov Model) Architecture

The HMM architecture of PPM is illustrated in Figure 3-3. Below we delineate each element of the HMM system.

- **States**

  $\Phi = \{1, 2, ..., N\}$ is a set of states representing the state space. The state at time t is denoted as q_t.

  Suppose the reference phone sequence is $C = c_1, c_2, ..., c_n$. Each reference phone c_i corresponds to two states, a *normal* state s_{2i-1} followed by an *insertion* state s_{2i}. Therefore, the corresponding states of the reference phone sequence C are $S = s_1, s_2, ..., s_{2n}$.

  Figure 3-4 illustrated the motivation of the 1-2 mapping of reference phones and states. Each reference phone is mapped to two states, a normal state and an insertion state, so that insertion phonetic transformations could be handled more gracefully.

  $Q = q_1, q_2, ..., q_T$ represents the possible state transition path taken in S; i.e., Q represents the alignment between the states and observation in Figure 3-3. Q

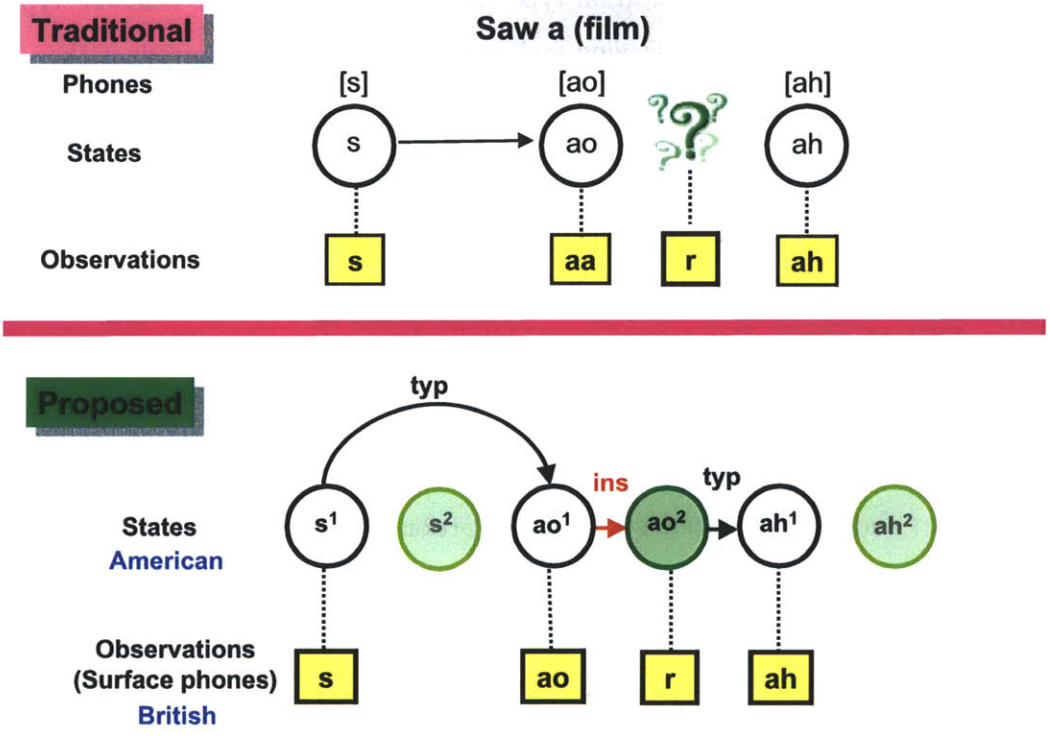

Figure 3-4: A traditional HMM system does not handle insertion transformations, so insertion states are introduced in the proposed HMM architecture.

takes on values of phones in S by a monotonic order:

$$\text{if} \quad q_t = s_i, q_{t+1} = s_j, \quad \text{then} \quad i \leq j. \tag{3.1}$$

Note that although the reference phone sequence C and its corresponding states S are known, we do not know which transition path was taken by Q. This is because the state transition type (insertion, self-insertion, deletion, typical transition) taken when leaving each state is unknown. We introduce the state transition types in the next section.

- **State transition types**

  There are 4 types of state transitions: insertion, self-insertion, deletion, and typical transitions. State transition types are represented by $r \in \{ins, sel, del, typ\}$. Since state transition types are graphically depicted as arcs with arrows, we will use the term *transition arc* interchangeably with *state transition type.*

  1. When transition arc $r = ins$, the origin state is a normal state s_{2i-1}, and the destination state is an insertion state s_{2i}. Only normal states are allowed to have insertion arcs.
  2. When transition arc $r = sel$, the origin and destination states are the same insertion state s_{2i}. Self-insertions are used to characterize consecutive insertions.
  3. When transition arc $r = del$, one or more normal states in S are skipped. This is to say, *if the origin state is* s_l *and the destination is* s_m*, then* $r = del$ *if and only if* $m - l \geq 2$.

  For simplicity, we will only discuss the case where one normal state is skipped here, which can be easily generalized to model consecutive deletions.

  A transition arc $r = typ$, when r is not ins, neither sel, nor del.

- **Observations (Emissions of the States)**

  $V = \{v_1, ..., v_M\}$ is the observation alphabet. The observation at time t is

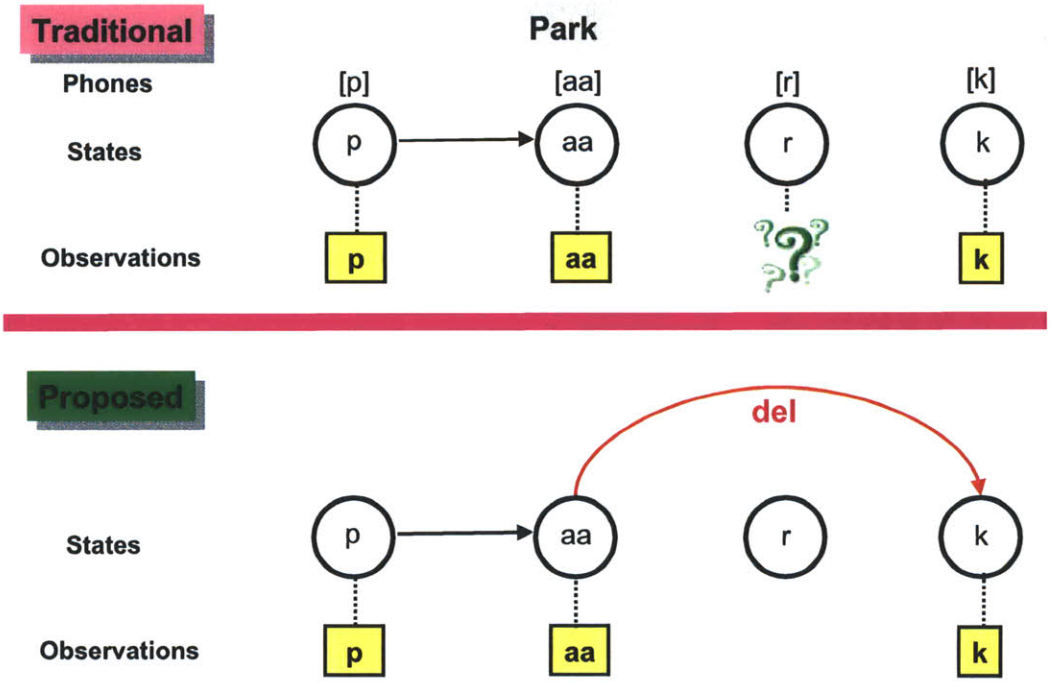

Figure 3-5: Motivation of introducing deletion arcs in proposed HMM network.

denoted as o_t. Let the corresponding observations of the states S be $O = \{o_1, o_2, ..., o_T\}$. In general the length of the states and observations are different; i.e., $n \neq T$.

- **State transition probability**

  The state transition probability from state x to state y through transition arc type r is

$$A_{xry} = P(q_{t+1} = y, r | q_t = x), \tag{3.2}$$

  where $1 \leq x, y \leq N$, transition type $r \in \{ins, sel, del, typ\}$, $\sum_y \sum_r A_{xry} = 1, \forall x$. When traversing over all the possible state transition paths of S, the probability of transitioning from state s_i to state s_j in S through transition type r is

$$a_{irj} = A_{xry}, \tag{3.3}$$

where $1 \leq i, j \leq 2n, s_i = x, s_j = y$. Note that if $r = sel$, then $i = j$.

We are aware that r over-specifies a, since r can be inferred from i and j. We retain r in specifying a for clarity purposes, since the transition arc type r is an distinctive characteristic of our framework to model phonetic transformations.

- **Emission (Observation) distribution in state x**

  The probability emitting observation v_k at any time t given state x is

$$B_x(k) = P(o_t = v_k | q_t = x), \tag{3.4}$$

  where $1 \leq x \leq N, 1 \leq k \leq M$. When traversing over all the possible state transition paths of S, the probability of s_i corresponding to state x and emits v_k is

$$b_i(o_t) = B_x(k), \tag{3.5}$$

  where $s_i = x, 1 \leq i \leq 2n$.

- The entire pronunciation model of dialect d is denoted as $\lambda_d = \{A, B\}$. Figure 3-6 compares the traditional and proposed HMM network. The differences are: insertion states, which emit inserted phones; insertion state transition, which are state transitions whose target state is an insertion state; deletion state transition, which are state transitions who skips normal states.

Phone usually refers to *monophone*, where a phone's surrounding phones do not affect the identity of the phone of interest; e.g., monophone [t] is always referred to as [t], regardless of its surrounding phones. A *biphone* is a monophone in the context of another monophone; e.g., biphone [t+a] is defined as the monophone [t] followed by monophone [a]. Similarly, *triphone* /x-y+z/ is defined as the monophone /y/ whose preceding phone is /x/ and following phone is /z/. Figure 3-7 illustrates some examples.

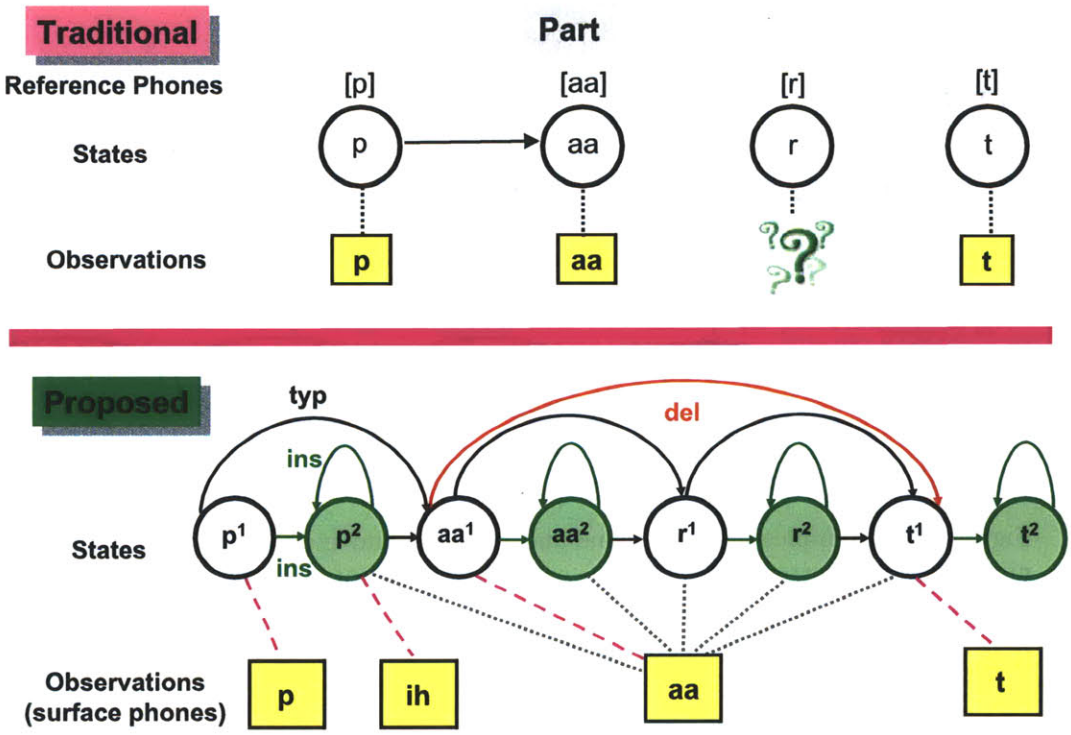

Figure 3-6: Comparison between traditional HMM and proposed HMM network. The underlying word is *part*, which is represented by reference phones of [p aa r t]. In the traditional HMM network, each phone is mapped to one state, but in the proposed HMM network, each phone is mapped to two states, a normal state and an insertions state. The insertion states model atypical yet systematic phonetic transformations of insertion. The insertions emit inserted surface phones that do not have a corresponding normal state. State transitions divided into three different types. (1) Insertion state transitions are state transitions whose target states are insertion states. (2) Deletion state transitions are state transitions that skip normal states. Deletion state transitions are used to model the deletion phonetic transformation, when there is no surface phone mapping to the reference phone. The deletion state transition does so by skipping a normal state, therefore the skipped normal state cannot emit anything. (3) Typical state transitions are state transitions that are neither insertion nor Deleon.

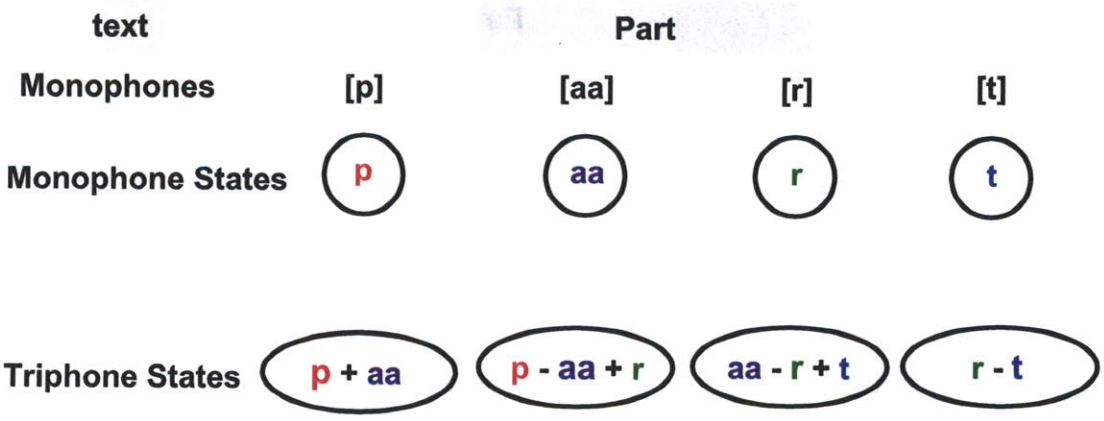

Figure 3-7: Examples of monophone and triphone and their notation. At the beginning and end of utterances biphones are used instead of triphones.

3.2.2 Scoring

Given the observations $O = \{o_1, o_2, ..., o_T\}$, and the states $S = \{s_1, s_2, ..., s_{2n}\}$, and a model $\lambda = \{A, B\}$, we want to compute $P(O|\lambda, S)$, the probability of the observation sequence given the states. The likelihood $P(O|\lambda, S)$ can be obtained by summing over all possible state transition paths Q as shown below.

$$P(O|\lambda, S) = \sum_{Q} P(O, Q|\lambda, S). \tag{3.6}$$

Figure 3-8 shows an example of all the possible alignments given the states and observations.

Forward Algorithm

Consider the forward variable $\alpha_t(i)$ defined as

$$\alpha_t(i) = P(o_1 o_2 ... o_t, q_t = s_i|\lambda, S), \quad 1 \leq t \leq T \tag{3.7}$$

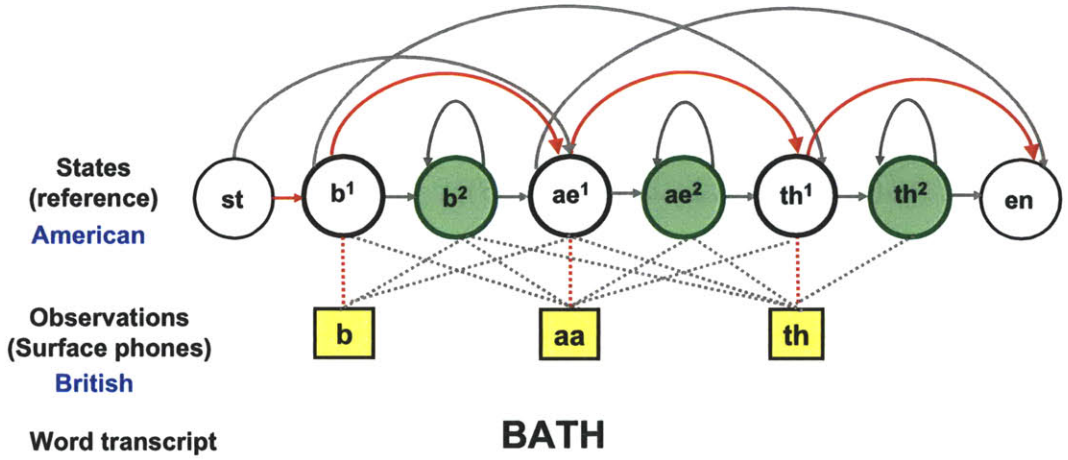

Figure 3-8: Given the states and observations, all possible alignments between the states and the observations are shown. The red alignment path shows the path with the highest likelihood. A likelihood score is computed for each alignment path. During test time, all the likelihood scores of each possible alignment is summed.

that is, the probability of the partial observation sequence until time t, $o_1, o_2, ...o_t$, and the state s_i at time t, given the model λ and the states S.

$\alpha_t(i)$ can be solved inductively, as follows:

1. Initialization

$$t = 1 : \alpha_1(j) = 1 \times b_j(o_1), 1 \leq s_j \leq N \tag{3.8}$$

2. Induction

$$t > 1 : \alpha_t(j) = [\sum_{i=1}^{2n} \sum_{r} \alpha_{t-1}(i) a_{irj}] b_i(o_t) \tag{3.9}$$

3. Termination

$$P(O|\lambda, S) = \sum_{i=1}^{2n} \alpha_T(i) \tag{3.10}$$

The final result is similar to the Forward Algorithm in a typical HMM system.

Backward Algorithm

Similarly, we can consider a backward variable $\beta_t(i)$ defined as

$$\beta_t(i) = P(o_{t+1}o_{t+2}...o_T|q_t = s_i, \lambda, S), 1 \leq t \leq T-1, 1 \leq s_i \leq N \tag{3.11}$$

that is, the probability of the partial observation sequence from time $t+1$ to the end, given the state s_i at time t and the model λ.

1. Initialization

$$\beta_T(i) = 1 \tag{3.12}$$

2. Induction

$$\beta_t(i) = \sum_{j=1}^{2n} \sum_{r} a_{irj} b_j(o_{t+1}) \beta_{t+1}(j), 1 \leq t \leq T-1 \tag{3.13}$$

The forward and backward variables can be used to compute $P(O|\lambda, S)$ efficiently, and is also useful in estimating model parameters $\lambda = (A, B)$ in Section 3.2.3.

3.2.3 Training: Model Parameter Estimation

There is no known way to analytically solve for the model parameters that maximize the probability of the observation sequence in a closed form. Alternatively, we can use an iterative procedure similar to Baum-Welch method [72] (i.e., expectation-maximization method) to locally maximize the likelihood $P(O|\lambda, S)$.

Derivation of re-estimation formulas from the auxiliary Q function

Baum's auxiliary function

$$Q(\lambda', \lambda) = \sum_{\mathbf{q}} P(\mathbf{O}, \mathbf{q}|\lambda', S) log P(\mathbf{O}, \mathbf{q}|\lambda, S) \tag{3.14}$$

P and $logP$ can be further expressed as

$$P(\mathbf{O}, \mathbf{q}|\lambda, S) = \pi_{q_0} \Pi_{t=1}^{T} \Pi_r A_{q_{t-1}, r, q_t} B_{q_t}(\mathbf{o_t}) \tag{3.15}$$

$$log P(\mathbf{O}, \mathbf{q}|\lambda, S) = log \pi_{q_0} + \sum_{t=1}^{T} \sum_{r} log A_{q_{t-1}, r, q_t} + \sum_{t=1}^{T} B_{q_t}(\mathbf{o_t}) \tag{3.16}$$

Without loss of generality, we set $\pi_{q_0} = 1$. Thus,

$$Q(\lambda', \lambda) = \sum_{y=1}^{N} Q_{A_x}(\lambda', \mathbf{A_x}) + \sum_{y=1}^{N} Q_{B_x}(\lambda', \mathbf{B_x}) \tag{3.17}$$

where $\mathbf{A_x} = [A_{x1}, A_{x2}, ..., A_{xN}]$, $\mathbf{B_x} = [B_x(v_1), ..., B_x(v_k)]$, and

$$Q_{A_i}(\lambda', \mathbf{A_x}) = \sum_{y=1}^{N} \sum_{t=1}^{T} \sum_{r} P(\mathbf{O}, q_{t-1} = x, r, q_t = y|\lambda', S) \log A_{irj} \tag{3.18}$$

$$Q_{B_i}(\lambda', \mathbf{B_x}) = \sum_{t=1}^{T} P(\mathbf{O}, q_t = x|\lambda', S) \log B_x(\mathbf{o_t}). \tag{3.19}$$

Since $Q(\lambda', \lambda)$ is separated into three independent terms, we can maximize $Q(\lambda', \lambda)$ over λ by maximizing the individual terms separately, subject to their stochastic constraints

$$\sum_{y=1}^{N} \sum_{r} A_{xry} = 1, \ \forall x \tag{3.20}$$

$$\sum_{k=1}^{K} B_x(k) = 1, \ \forall x \tag{3.21}$$

where the individual auxiliary functions all have the form

$$J(\sigma_j) = \sum_{y=1}^{N} \mu_j \log \sigma_j \tag{3.22}$$

subject to the constraints $\sum_{y=1}^{N} \sigma_j = 1, \sigma_j \geq 0$. We can reformulate the Eq. (3.22) using the Lagrange multiplier ρ, $\rho \geq 0$:

$$J(\sigma_j; \rho) = \sum_{y=1}^{N} \mu_j \log \sigma_j - \rho(\sum_{y=1}^{N} \sigma_j - 1) \tag{3.23}$$

To maximize $J(\sigma_j; \rho)$, we set $\frac{\partial J(\sigma_j; \rho)}{\partial \sigma_j} = 0$:

$$\frac{\partial J(\sigma_j; \rho)}{\partial \sigma_j} = \frac{\mu_j}{\sigma_j} - \rho = 0 \tag{3.24}$$

We obtain $\sigma_j = \frac{\mu_j}{\rho}$. Since $\sum_{y=1}^{N} \sigma_j = 1$, we get $\rho = \sum_{y=1}^{N} \mu_j$. Therefore,

$$\sigma_j = \frac{\mu_j}{\sum_{y=1}^{N} \mu_j}, \; y = 1, 2, ...N \tag{3.25}$$

then the maximization leads to the model re-estimate $\bar{\lambda} = [\bar{\pi}, \bar{A}, \bar{B}]$, where

$$\bar{\pi_x} = \frac{P(\mathbf{O}, q_o = s_i|\lambda)\delta(s_i, x)}{P(\mathbf{O}|\lambda, S)} \tag{3.26}$$

$$\bar{A_{xry}} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} = s_i, r, q_t = s_j|\lambda, S)\delta(s_i, x)\delta(s_j, y)}{\sum_{t=1}^{T} \sum_r P(\mathbf{O}, q_{t-1} = s_I|\lambda, S)\delta(s_i, x)} \tag{3.27}$$

$$\bar{B_x}(k) = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_t = s_i, |\lambda, S)\delta(s_i, x)\delta(o_t, v_k)}{\sum_{t=1}^{T} P(\mathbf{O}, q_t = s_i, |\lambda, S)\delta(s_i, x)} \tag{3.28}$$

where

$$\delta(k,l) = 1 \text{ if } k = l \tag{3.29}$$

$$= 0 \text{ otherwise} \tag{3.30}$$

For simplicity, the above equations can be represented using the variables ξ and γ: $\xi_t(x,r,y)$ is the probability of being in state x at time t and state y at time $t+1$ through transition arc r, given the model and observation sequence and the states S.

$$\xi_t(x,r,y) = \sum_{s_i=x}\sum_{s_j=y} P(q_t = s_i, r, q_{t+1} = s_j | O, \lambda, S) \tag{3.31}$$

$$= \sum_{s_i=x}\sum_{s_j=y} \frac{\alpha_t(i)a_{irj}b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda,S)} \tag{3.32}$$

$$\gamma_t(x) = \sum_{y=1}^{N}\sum_{r} \xi_t(x,r,y) \tag{3.33}$$

Summing $\gamma_t(x)$ and $\xi_t(x,y)$ we get:

$$\sum_{t=1}^{T-1} \gamma_t(x) = \text{expected number of transitions from state } x \tag{3.34}$$

$$\sum_{t=1}^{T-1} \xi_t(x,r,y) = \text{expected number of transitions from state } x \text{ to } y \text{ through arc } r \tag{3.35}$$

$$\bar{A}_{xry} = \frac{\sum_{t=1}^{T-1} \xi_t(x,r,y)}{\sum_{t=1}^{T-1} \gamma_t(x)} \tag{3.36}$$

$$\bar{B}_x(k) = \frac{\sum_{t=1;o_t=v_k}^{T} \gamma_t(x)}{\sum_{t=1}^{T} \gamma_t(x)} \tag{3.37}$$

Re-estimating until local optimal

It has been proven by Baum and his colleagues that

$$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \tag{3.38}$$

$$P(O|\lambda, S) \geq P(O|\lambda', S) \tag{3.39}$$

By maximizing $Q(\lambda', \lambda)$ over λ, we can improve the model parameters by increasing the likelihood $P(O|\lambda, S)$.

Remarks

The condition on the reference phone sequence S in the likelihood computation could be removed without loss of generality. For the *informative* emphasis of our work, we retain the conditioning on S to better characterize dialect-specific phonetic rules/transformations. Conditioning on S helps constrain the possible state transition paths with the most likely paths. With the additional transition paths of deletion and insertion in our HMM system, the possible state transition paths grow exponentially without this constraint.

For future work, we plan to empirically investigate the case where there is no conditioning on S. This investigation has many practical applications, since no transcription is required (at least) in the scoring phase. It will be more challenging to obtain a clean reference phone sequence, if there were no transcription to aid in the training phase. This procedure of *unsupervised* training might not be the most optimal in phonetic rule analysis, but could be helpful in engineering applications, where

the main goal is to achieve good dialect recognition performance.

3.3 Decision Tree Clustering

As discussed in Section 3.1, context independent modeling might not be able to fully capture the phonetic rules characterizing dialects. Therefore, instead of modeling only monophones, triphones (or even quinphones) might more appropriately characterize phonetic rules. The number of phones increases exponentially when considering phonetic context, which often results in poor model parameter estimation due to data insufficiency. Thus in this section, we discuss clustering approaches that could be used to tie model parameters to better characterize dialect differences.

A decision tree is a top-down recursive clustering method commonly used in automated speech recognition to train acoustic models by pooling triphones of similar acoustics together. Decision tree clustering is an automatic procedure that can incorporate linguistically-defined features to better characterize different acoustic implementations of the same phone. Here we adopt decision tree clustering to learn phonetic rules across dialects, and determine which parameters to tie to better estimate the model parameters. It should be noted that other clustering methods are also feasible.

3.3.1 Algorithm

When triphone states are considered, model parameters increases exponentially. To better estimate model parameters, state-tying is often used to pool parameters that share common characteristics.

The log likelihood increase of splitting node k to nodes k_1 and k_2 using the attribute H_f, where f corresponds to feature f, is

$$\Delta \log L = \log \frac{L(O_{k_1}|x \in H_f)L(O_{k_2}|x \notin H_f)}{L(O_k|x)}. \tag{3.40}$$

The attribute chosen to split the data at node k is $\arg\max_{H_f} \Delta \log_L$. This splitting

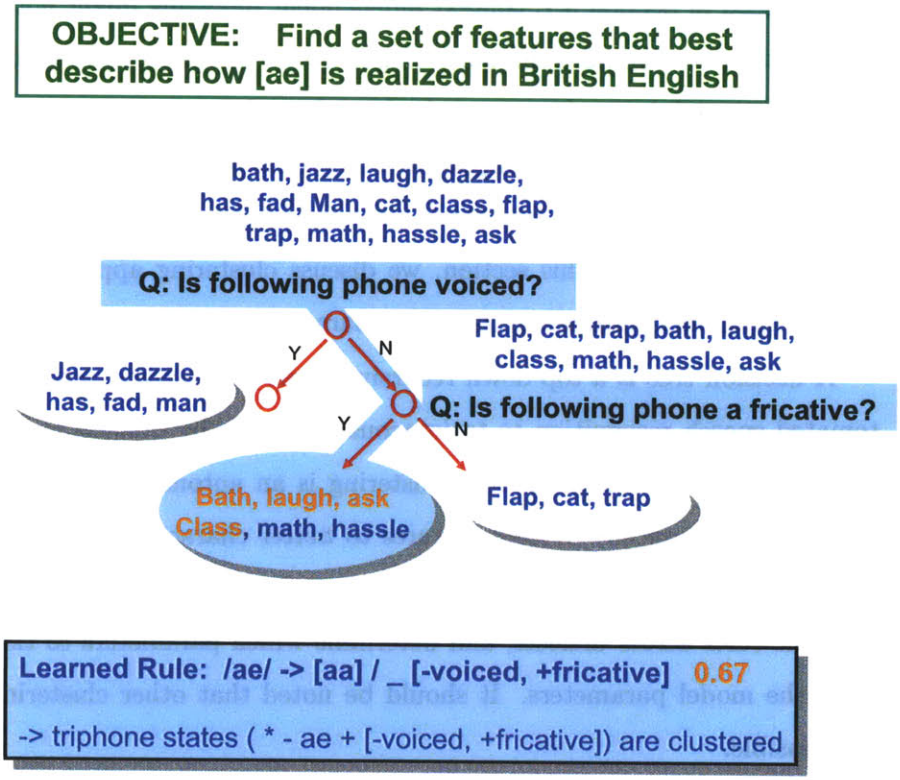

Figure 3-9: An example of decision tree clustering. At each node, a list of yes-no questions are asked, and the questions that provides the best split of data (e.g., the most likelihood increase) is chosen to split the node into two children nodes. The splitting process is repeated recursively until the stop criteria is reached. After clustering, each leaf node represents a rule. Some rules are trivial, mapping [ae] to [ae], but some show interesting phonetic transformations. For example, the light blue leaf node shows that 67% of words containing [ae] followed by voiceless fricatives are transformed into [aa] in British English. The yes-no questions used to split each node are describes the conditioning phonetic context where phonetic transformation occurs.

procedure is done recursively until a stop criterion is reached.

3.3.2 HMM Model Estimation after State Clustering

After state clustering, assume triphone states are clustered into I groups. Group i is specified by $G_i = (\zeta_\ell^i, \zeta_m^i, \zeta_r^i)$, where ζ_ℓ^i specifies the left context state, ζ_m^i specifies the center (middle) state, and ζ_r^i specifies the right context state.

The models estimation equations still have the same form as Eq. (3.27) and Eq. (3.28):

$$\bar{A_{G_i,r}} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, r, q_t^{tri} \in G_i | \lambda, S)}{\sum_{t=1}^{T} \sum_{r \in R} P(\mathbf{O}, r, q_{t-1} \in \Psi | \lambda, S)} (1 - P_D),$$

where triphone $q_t^{tri} = (q_{t-1} - q_t + q_{t+1})$, $R = \{typ, ins\}$ and $\Psi = \zeta_\ell^i \cap (\sigma_1^c \cap \sigma_2^c \cap ... \sigma_I^c)$ and

$$\bar{B_{G_i}}(k) = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_t^{tri} \in G_i, | \lambda, S) \delta(o_t, v_k)}{\sum_{t=1}^{T} P(\mathbf{O}, q_t^{tri} \in G_i, | \lambda, S)} \tag{3.41}$$

3.4 Training Procedure of PPM

Figure 3-10 shows the training procedure for phonetic-based pronunciation model. The reference phones are generated through forced-alignment using the audio, word transcriptions, and pronunciation dictionary for the reference dialect. The surface phones are generated through phone recognition decoding. A monophone HMM is first trained, and its estimated model parameters are used to initialize the triphone HMM. The trained triphone HMM provides the decision tree clustering procedure the triphone likelihoods used during splitting. The decision tree clustering module determines which triphone states are clustered. Triphone states that are in the same clustered group merges their identity to the group identity. Finally, a triphone-

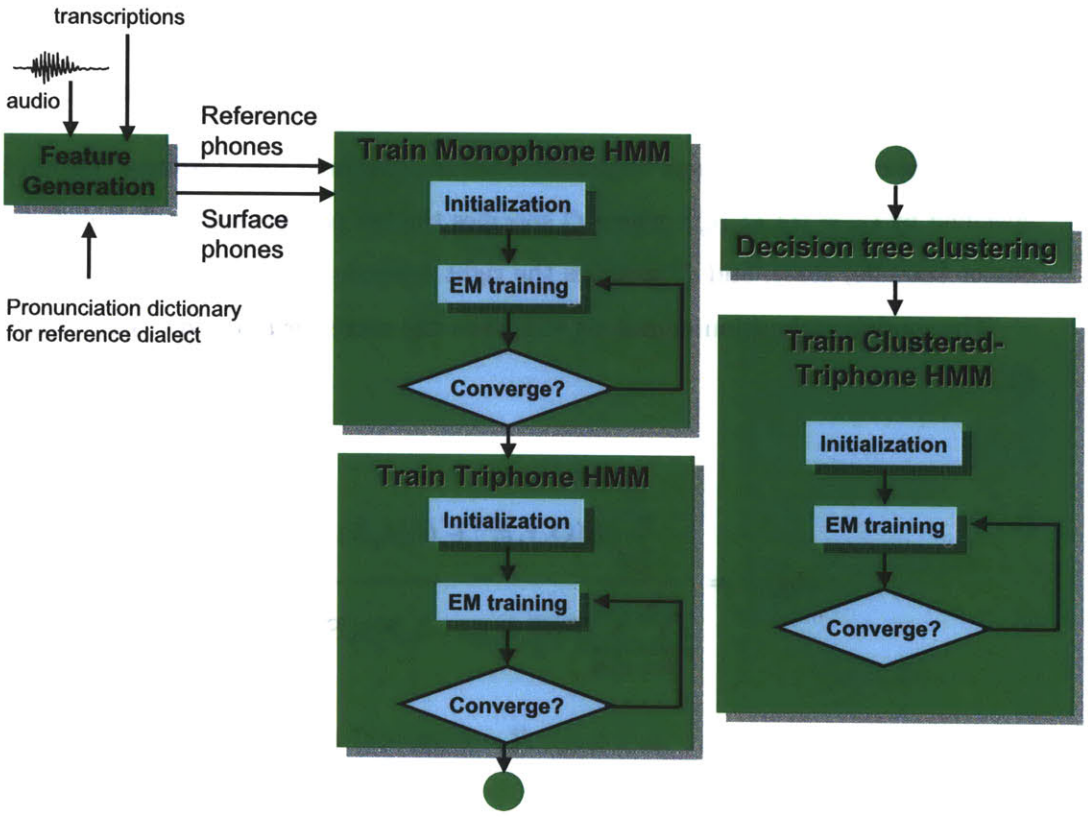

Figure 3-10: An example of decision tree clustering. At each node, a list of yes-no questions are asked, and the questions that provides the each node are describes the conditioning phonetic context where phonetic transformation occurs.

clustered HMM is trained using the clustering results of the decision tree module.

3.5 Limitations of PPM

3.5.1 Constraints in Learning Deletion Rules

Deletion rules are also not characterized comprehensively in PPM. The standard triphone state tying mechanism used in [12] makes two assumptions about deletion rules.

1. The phone preceding a deleted phone is affected and characterized phonetically

through automatic phone recognition or manual phone transcriptions.

2. The phone following a deleted phone does not specify when deletions occur.

These assumptions are over-simplifications and only apply to certain rules.

For example, GAE is rhotic, while Received Pronunciation (RP) in UK is not. Rhotic speakers pronounce /r/ in all positions, while non-rhotic speakers pronounce /r/ only if it is followed by a vowel [92]. Therefore, the word *park* (/p aa r k/) in GAE would sound like *pak* ([p aa: k][1]) in RP, since /r/ is followed by a consonant /k/. Clearly, this non-rhotic rule does not comply with assumption 2. While the vowel /aa/ before /r/ does change its vowel quality by becoming a longer sound [aa:], this phenomenon could be too subtle to be captured practically in automated systems, and might not be true for all deletion transformations across dialects.

In addition, since deletions are modeled by state transition arcs that skip states in PPM, it is expected that arc clustering rather than state clustering is more suitable in determining the tying structure for deletions.

3.5.2 Inability to Capture Fine-Grained Acoustic Differences

PPM is a token-based system, which does not directly exploit fine-grained acoustic differences across dialects. When dialect differences are subtle, such as in the case of GAE (General American English) and AAVE (African American Vernacular English), phonotactic information alone are insufficient to characterize dialects.

In our previous work [13], we expanded the usage of acoustic phonetic models [84] from monophones to biphones to characterize context-dependent phonetic rules. In this work, we take a step further and use clustering to more effectively model acoustic differences across phonetic contexts. This proposed method is similar in spirit to discriminative phonotactics using context-dependent phone classifiers in [7]. However, the focus in [7] is to improve DID accuracy, while our focus is to characterize dialect differences explicitly.

[1] [aa:] represents a long [aa]

3.6 PPM Refinement I: Sophisticated Tying

We refine PPM to include arc clustering, since deletions could be more appropriately modeled through arc clustering instead of state clustering, as mentioned in the previous section.

Consider a triphone state $(s_{k-1} - s_k + s_{k+1})$. First, we use arc clustering to determine which deletion arcs to tie together, and then estimate the tied deletion probabilities accordingly. Next, we estimate the typical and insertion transition probabilities originating from s_k as in the state-tying case with a new Lagrange constraint, since the total sum of deletion probabilities leaving s_k are pre-determined by arc clustering.

3.6.1 Arc Clustering for Deletions

Unlike other transition arcs, deletion transition arcs are not only specified by its origin and target state, but also the normal states that are skipped. Expected counts of the state x being deleted when q_t corresponds to attribute H_f is

$$E_{d=x} = \sum_{t=1}^{T} P(q_{t+1}, r = del, d = x | q_t \in H_f), \tag{3.42}$$

where d represents the deleted/skipped state.

Expected counts of the state x *not* being deleted when q_t corresponds to attribute H_{fk} is

$$E_{d \neq x} = \sum_{t=1}^{T} \sum_{r} P(q_{t+1}, r | q_t = x \in H_f), \tag{3.43}$$

since state x could not have been skipped if there were transition arcs leaving it.

The likelihood of q_t corresponding to attribute H_f is

$$L(x | q_t \in H_f) = \langle \frac{E_{d=x}}{E_{d=x} + E_{d \neq x}} \rangle^{E_{d=x}} \langle \frac{E_{d \neq x}}{E_{d=x} + E_{d \neq x}} \rangle^{E_{d \neq x}} \tag{3.44}$$

Similarly, the likelihood of q_t not belonging to attribute H_f, $L(x | q_t \notin H_f)$ can also be

obtained. Following the split criteria in Eq. (3.40), decision tree clustering is used to determine the arc tying structure.

We assume deletion arcs are clustered into J groups. Group j is specified by $D_j = (\sigma_j, \varsigma_j, \tau_j)$, where σ_j specifies the origin of the transition arc, ς_j specifies the skipped state, and τ_j specifies the target of the arc. The model estimation equation for deletion transitions belonging to clustered group D_j is

$$\bar{A}_{D_j} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} \in \sigma_j, r=del, d \in \varsigma_j, q_t \in \tau_j | \lambda, S)}{\sum_{t=1}^{T} \sum_r P(\mathbf{O}, r, q_{t-1} \in \sigma_j | \lambda, S)} \tag{3.45}$$

3.6.2 State Clustering for Substitutions and Insertions

The sum of all deletion probability leaving triphone state $(s_{k-1} - s_k + s_{k+1})$ is

$$\begin{aligned} P_D =& P(q_{t+1} = s_{k+2}, r = del | q_t = s_k) \\ =& \sum_j P(q_{t+1} = s_{k+2} \in \tau_j, r = del | q_t = s_k, s_{k+1} \in \varsigma_j) \\ & P(s_{k+1} \in \varsigma_j, q_{t+1} = s_{k+2} \in \tau_j, r = del | q_t = s_k) \end{aligned} \tag{3.46}$$

After state clustering, we assume triphone states are clustered into I groups. Group i is specified by $G_i = (\zeta_\ell^i, \zeta_m^i, \zeta_r^i)$, where ζ_ℓ^i specifies the left context state, ζ_m^i specifies the center state, and ζ_r^i specifies the right context state. Similar to using Baum's auxiliary function in typical HMM systems, it can be shown that the tied typical and insertion transition probabilities obtained in state tying are redistributed proportionally as

$$\bar{A}_{G_i,r} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, r, q_{t-1}^{tri} \in G_i | \lambda, S)}{\sum_{t=1}^{T} \sum_{r \in R} P(\mathbf{O}, r, q_{t-1}^{tri} \in G_i^* | \lambda, S)} (1 - P_D),$$

where $q_{t-1}^{tri} = (q_{t-2} - q_{t-1} + q_t)$, $R = \{typ, ins\}$ and $G_i^* = (\zeta_\ell^i \cap (\sigma_1^c \cap \sigma_2^c \cap ... \sigma_I^c), \zeta_m^i, \zeta_r^i)$.

Note that insertion and typical arcs are destination independent, thus it doesn't matter what q_t is in Eq. (3.47).

3.7 PPM Refinement II: Acoustic-based Pronunciation Model

Instead of decoded phone sequences, acoustic observations such as perceptual linear prediction can be used to characterize dialect differences as well. Acoustic observations can characterize pronunciation differences that are not large enough to warrant a phonetic change. Acoustic observations are typically used to describe spectral information of speech, such as the formant frequencies, voicing, frication noise.

The acoustic counterpart of PPM can be obtained by replacing the discrete observation probability $B_x(k)$ in Eq. (3.28) to a continuous *pdf* $B_x(\mathbf{z})$, which can be modeled as a mixture of Gaussians:

$$B_x(\mathbf{z}) = \sum_{l=1}^{M} w_{xl}\mathcal{N}(\mathbf{z}; \mu_{xl}, \Sigma_{xl}), \tag{3.47}$$

where $\mathcal{N}$ is the normal density, $w_{xl}, \mu_{xl}, \Sigma_{xl}$ are the mixture weight, mean vector, and covariance matrix of state x and mixture l, $1 \leq x \leq N$, $1 \leq l \leq M$.

In this work we only implement a simplified version of APM. Only normal states and typical state transition arcs are considered, and triphone states are clustered by standard tying. We leave the complete implementation of APM for future work.

3.8 Remarks

Below are assumptions of the proposed model.

- We assume that underlying phonetic rules governing dialect differences exist, and account for a noticeable degree of dialect differences that can be measured.
- We assume the given pronunciation dictionary uses phone sets that are able to capture phonetic transformations across dialects. For example, if the given phone set does not represent flaps and canonical /t/'s differently, we will not be able to learn that flaps only occur in American English under certain conditions, and not in other dialects of English.

Given these assumptions, our model design is language and dialect independent. Given the word transcript of the different dialects and a pronunciation dictionary of at least one dialect, any dialects can be characterized by our model without additional linguistic knowledge.

3.9 Summary

In this chapter, we delineate our proposed pronunciation model by deriving the mathematical framework of an HMM system with three different state transition types, which are used to model phonetic transformations across dialects. We also introduced sophisticated tying mechanisms that incorporate transition arc clustering to characterize deletion phonetic rules. The proposed pronunciation model characterizes and predicts pronunciation variation across dialects, and can also be used in automatically recognizing dialects. In the following chapters, we empirically evaluate our pronunciation model on various datasets of different corpus sizes, speaking styles, and languages.

Chapter 4

Corpora Investigation

Due to the interdisciplinary nature of this work, one of the greatest challenges was to find corpora suitable for both speech analysis and dialect recognition experiments. Corpora for these two purposes usually have complementary characteristics, and existing corpora that have both are rare. Usually corpora for analysis and interpretation of pronunciation rules require word transcription and ideally phone transcriptions, but corpora with these properties are usually smaller in size, and therefore not suitable for dialect recognition purposes; dialect recognition results are often inconclusive due to too few trials. On the other hand, large-scale corpora used for dialect recognition purposes usually do not have word transcriptions, which makes the interpreting dialect-specific rules challenging.

In Section 4.1, we analyze ideal properties of corpora for informative dialect recognition, the challenges of choosing such corpora due to the complex nature of dialects and the practical constraints of existing resources, and evaluate potential corpora that might be useful in our work. In Section 4.2, we introduce three sets of corpora chosen for this work, and in Section 4.3 we conclude our discussion on corpora.

4.1 Corpora Analysis

In this section we discuss the requirements of the corpora we need in informative dialect recognition, the practical constraints of existing linguistic resources, and eval-

uate corpora that could potentially be used for our work.

4.1.1 Ideal Corpora Properties

The ideal corpora needed for analyzing and characterizing dialects satisfy the following properties.

- **Linguistic dialect labels**: Ideal dialect labels correspond to *linguistic* categorization. However, dialect labels in many available databases do not always correspond with the *true* dialect label.

- **Word transcriptions**: Word transcriptions provide canonical references for comparing pronunciation patterns across dialects. In addition, phone recognition accuracy typically increases substantially if word transcriptions and a reasonable phone recognizer are provided.

- **Pronunciation dictionary**: A pronunciation dictionary of the reference dialect, along with the word transcriptions and a phone recognizer, are used to generate the reference phones through forced-alignment. Ideally, the phoneset used in the pronunciation dictionary should be able to distinguish dialect differences. For example, if the pronunciation dictionary does not include [dx] (flaps), making both flaps and canonical [t] are represented the same way, and if the dialects being compared are American and British English, we would not be able to learn that canonical [t] transforms into flaps when in intervocalic positions in American English in the phonetic-based pronunciation model (PPM). This is not an issue for acoustic-based pronunciation model (APM), since APM is not limited to detecting phonetic transformations, which are defined by the phonesets of the pronunciation dictionary.

- **Sufficient speakers**: If the number and sampling of speakers is limited, the proposed algorithms will learn speaker-specific characteristics instead of general dialect-specific characteristics.

- **Sufficient training data**: Machine learning methods require sufficient training data so the algorithms have enough data to generalize well.

These properties are challenging to fulfil. Below we discuss these challenges in two aspects: (1) the ambiguous nature of dialects, and (2) practical constraints of existing resources.

4.1.2 The Ambiguous Nature of Dialects

The *true* dialect label of a speaker and whether this true dialect label even exists are not always clear cut:

- Some people are multi-dialectal and they often code switch according to the other speaker or to the degree of formality of the scenario.

- There are no distinct boundaries between dialects. Some people might have characteristics of more than one dialect. For example, with the increased amount of traveling and moving, it is more difficult to determine a speaker's dialect just by simple terms such as birthplace or hometown.

- Though traditional categorization of dialects is primarily based on region, it has been shown that other factors such as education, socio-economic status, race are related to dialects as well.

4.1.3 Practical Constraints of Existing Resources

There are practical constraints of existing resources in corpora and pronunciation lexicons for speech technology purposes.

- There are not many large corpora suitable for dialect characterization. If the original data collection was not for dialect analyses purposes, the dialect labels might not be appropriate. For example, self-reported dialect labels might not necessarily correspond well with linguistic definitions.

- Some well-studied or suitable corpora are often small in size (e.g., TIMIT) or not publicly available.

- Pronunciation lexicons and letter-to-sound tools for automatic speech recognition are limited, especially for languages such as English where spelling is irregular. The setup of the pronunciation lexicon can influence the reference canonical pronunciations, which would also influence the pronunciation rules learned.

4.1.4 Evaluation of Corpora Candidates for Informative Dialect Recognition

Due to the intrinsic and practical challenges discussed above, it is important to evaluate whether the chosen corpora are suitable for the informative dialect recognition experiments. This evaluation process can be done by literature review of the dialects of interest, corpora candidate, and pronunciation lexicon. The dialect-specific pronunciation rules in the literature could be used to guide the design and setup of the automatic system. It is important to determine whether the pronunciation lexicon is suitable in capturing dialect variation in pronunciation. For example, if the phoneset in the pronunciation lexicon does not distinguish between the vowels in *marry*, *merry*, and *Mary*, then we would not be able to detect that the dialect spoken in certain northern regions in the U.S. (e.g., Philadelphia, New York City, and New England) which make distinctions between these vowels.

In Table 4.1, we list corpora candidates that were investigated in this thesis work, their properties, and practical issues we encountered in our initial assessment.

In setting up the corpora for the experiments, it is important to make sure dialect labels do not correlate with other factors (e.g., gender, channel) that are much stronger than dialect characteristics [14]. In our investigation, we unfortunately found out that WSJ0, WSJ1, and WSJ-CAM0 are unsuitable for performing dialect recognition experiments between British and American English. A more detailed analysis is documented in Appendix B. Fortunately, we were still able to use WSJ-

CAM0 and the pronunciation dictionary from WSJ0 (American English) to conduct pronunciation generation experiments to assess how well the proposed models are able to convert American pronunciation to British pronunciation.

4.2 Adopted Datasets

Despite the many difficulties and challenges in surveying and selecting suitable corpora, we managed to work around constraints and adopted three databases for our work: WSJ-CAM0, 5-Dialect Arabic Corpus, and StoryCorps. We introduce each of them in the following sections.

4.2.1 WSJ-CAM0

Corpus Description

WSJ-CAM0 stands for *the Wall Street Journal recorded at the University of CAMbridge (phase 0)*. WSJ-CAM0 is the UK English equivalent of a subset of the US American English WSJ0 database [78]. The training data contains speech of 15.3 hr (92 speakers). The test and dev set are each 4 hr (48 speakers). It consists of speaker-independent (SI) read material, split into training, development test and evaluation test sets. There are 9 utterances from each of 92 speakers that are designated as training material for speech recognition algorithms. A further 48 speakers each read 40 sentences utterances containing only words from a fixed 5,000 word vocabulary of 40 sentences from the 64,000 word vocabulary, which will be used as testing material. Each of the total of 140 speakers also recorded a common set of 18 adaptation sentences. The data partition of WSJ-CAM0 is listed in Table 4.2.

Pronunciation Dictionary from Corpus WSJ0

Only having a corpus containing speakers of one dialect is insufficient in characterizing dialect differences. Pooling together different corpora with different dialects is unsuitable for evaluating our proposed models, because the channel differences across

Table 4.1: Analysis of word-transcribed corpora candidates for informative dialect recognition. CTS: conversational telephone speech; Am: American; Br: British; Conv: conversation.

Corpus	Language	Dialects	Type, Channel	Speakers per dialect	Duration per dialect	Issues
AMI [19]	English	Am, UK, Scottish	meeting	14-26	< 100hr	limited speakers
ANAE [56]	English	south & north Am	telephone	N/A	N/A	small
Buckeye [71]	English	Ohio	Conv, studio	40	20 - 40hr	only 1 dialect
BU Radio [68]	English	General Am.	Read, studio	7	12hr	limited speakers
FAE [30]	English	Foreign accents	telephone	> 200	71min	lack transcripts
Fisher [17]	English	south & north Am	CTS	> 100	30hr	noisy dialect labels
NSP [18]	English	6 regions in ANAE	studio	10	5-10min	small & small perceptual difference
StoryCorps [46]	English	AAVE, non-AAVE	studio	> 67hr	> 200	transcriptions incomplete
S. Renals [73]	English	Br	Broadcast news	N/A	50hr	unattainable
TIMIT [27]	English	8 dialects in Am.	read speech	33–102	15–45min	limited test trials
Ivie [20]	English	Br	Read /studio	108	1hr	limited transcripts
WSJ0, WSJ-CAM0 [70, 78]	English	Am & Br	Read, studio	> 83	14hr	see Appendix B
Fisher	Spanish	Caribbean & non-Caribbean	CTS	> 44	> 40hr	not public
5-Dialect Arabic corpus [58]	Arabic	AE, EG, IQ, PS, SY	read & CTS	500	13 hr	read portion is MSA

Table 4.2: WSJ-CAM0 data partition

Set	Speaker number	Duration
Train	92	15.3 hr
Dev	48	4 hr
Test	48	4 hr

corpora are likely to strongly correlate with the dialect labels, even when the two corpora design intended to collect data the same way. This was the case for WSJ0, WSj1, and WSJ-CAM0 as mentioned above.

However, we designed a pronunciation generation experiment to overcome this challenge. By using the pronunciation dictionary from WSJ0, we can obtain the American pronunciation of a given word in WSJ-CAM0, which can serve as the reference phones for the phonetic and acoustic-based pronunciation models. By evaluating how well the proposed models are able to convert the American pronunciation to the British pronunciation, we can assess how well the proposed algorithms are characterizing dialect differences. For experimental details and results, please see Chapter 6.

4.2.2 5-Dialect Arabic Corpus

The corpus includes Arabic dialects of United Arab Emirates (AE), Egypt (EG), Iraq (IQ), Palestine (PS), and Syria (SY) [58]. Each dialect has its own regional pronunciation dictionary. There are 250 telephone conversations with 100 speakers per dialect. A set of 13 pre-selected topics were chosen with the aim of achieving as much as possible an equal distribution across all topics for the final database. The details of the data partition are listed in Table 4.3 Models were trained on 46.3 hours of speech. The development set was 8.4 hours (1,011 30-second trials) and the test set was 8.8 hours (1,061 30-second trials). The gender ratio of male vs. female is 1:1 across all 5 dialects for all three partitions (see Table 4.4 for details). Only the conversational speech data were chosen.

Table 4.3: Data partition and description

Set	**Speaker Number**	**Duration**	**Number of 30-sec trials**
Train	276	46.25 h	N/A
Dev	83	13.9	1,011
Test	88	14.75	1,061

Table 4.4: Number of speakers in each data partition

	Train			Test			Dev		
Dialect	M	F	total	M	F	total	M	F	total
AE	26	26	52	8	9	17	7	8	15
EG	30	30	60	9	10	19	9	9	18
PS	30	30	60	10	10	20	10	10	20
IQ	26	26	52	8	8	16	7	7	14
SY	26	26	52	8	8	16	8	8	16

4.2.3 StoryCorps: AAVE vs. non-AAVE

Two sets of American English dialects were chosen from StoryCorps [46]: (1) African American Vernacular English (AAVE), (2) Non-AAVE American English. AAVE speakers were self-reported as African American. (2) Non-AAVE speakers were self reported as white or of European decent. The conversations are between speakers of the same dialect to minimize accommodation issues [35]. The train set is 22.6 hr (69 speakers), the dev set is 7.1 hr (38 speakers), and the test set is 7 hr (28 speakers). The data partition of StoryCorps is listed in Table 4.5.

Table 4.5: StoryCorps data partition

Set	Speaker number	Duration
Train	82	27.4 hr
Dev	31	8.2 hr
Test	38	9.2 hr

4.3 Summary

In this chapter we investigated corpora issues in informative dialect recognition research, which is the main challenge in this line of work. We analyzed the properties and limitations of existing corpora with dialect labels, and introduced the three corpora we selected to evaluate our proposed systems.

Chapter 5

Dialect Recognition Experiments

Dialect recognition error rates are one measure of how well the phonetic rules are learned in the proposed pronunciation models. Assuming dialect-specific rules occur frequently enough in the training and test set, a pronunciation model that learned these rules should lead to good DID performance.

Note that it is challenging to interpret how well the rules are learned through dialect recognition experiments, partly due to the limited resources of available corpora and characteristics of different dialects. Results could easily be confounded with channel issues, amount of data, numbers of speakers, completeness of corpora documentation, linguistic aspects other than phonetics and acoustics (e.g.,lexicon, grammar, prosody, pragmatics). Despite these potential complications, we still present our experiment results, and analyze them with caveats to watch out for. The key innovation is, again, the introduction of a framework for characterizing dialects quantitatively at the phonetic and acoustic level.

Below we conduct dialect recognition experiments on two datasets[1]: (1) 5-Dialect Arabic Corpus in Section 5.1, and (2) StoryCorps in Section 5.2. We summarize the findings of these two experiments in Section 5.3.

[1]WSJ0 (USA) and WSJ-CAM0 (UK) were inappropriate for DID experiments due to strong channel differences correlating with dialect labels. For more detailed analysis, see Appendix B

5.1 Experiment I: 5-dialect Arabic Corpus

5.1.1 Experimental Setup

Data

The corpus includes Arabic dialects of United Arab Emirates (AE), Egypt (EG), Iraq (IQ), Palestine (PS), and Syria (SY) [58]. Each dialect has its own regional pronunciation dictionary. There are 250 telephone conversations with 100 speakers per dialect. Channel was strictly controlled so there are no channel differences correlating with dialect labels. Gender and conversation topic were controlled across dialects and sets (train, dev, test.) Models were trained on 46.3 hours of speech. The development set was 8.4 hours (1,011 30-second trials) and the test set was 8.8 hours (1,061 30-second trials). For more details about the 5-Dialect Arabic Corpus, please refer to Section 4.2.2.

Pronunciation Dictionaries

We used 2 pronunciation dictionaries: (1) pan-Arabic Dict, combining all regional dictionaries, which is used to generate surface phones; (2) IQ Dict, used to generate reference phones. The reference pronunciations of out of vocabulary words were obtained through an Iraqi letter-to-sound tool trained at MIT Lincoln Laboratory.

Decision of Reference Dialect

There are linguistic and engineering considerations that lead to the determination of the reference dialect being Iraqi Arabic. From the engineering standpoint, we have a lot more additional resources of Iraqi Arabic to train phone recognizers and letter-to-sound tools[2].

From the linguistic standpoint, Egyptian Arabic was avoided as a reference dialect due to the popularity of Egyptian media. Many native Arabic speakers, though not from Egypt, have more knowledge of how Egyptians speaker differently from them.

[2]Letter to sound tools are trained to map text to phones. This is especially important for languages where spelling is irregular.

Native Arabic speakers were asked to perform informal perceptual tests in the initial phase of the experiment design to determine if the experiments are going in the intended direction. Palestinian and Syrian Arabic both belonged to the same Levatine family, so they would be more similar to each other than the other Arabic dialects. If the differences between Palestinian and Syrian Arabic are subtle, it would be difficult to determine if the proposed models are unable to detect very fine-grained differences that exist or the dialect differences are just too subtle to be measured.

Tools: Phone Recognizers and Backend Classifier

The reference dialect is IQ, and so was excluded in dialect recognition experiments. Two IQ phone recognizers were used: (1) the triphone recognizer (38 monophones, 3 states/triphone, 128 Gaussians/state); (2) the monophone recognizer (38 monophones, 3 states/monophone, 2048 Gaussians/state). The time stamps of the conversations in this corpora are not fully specified, therefore we used the triphone IQ recognizer and the word transcriptions to force-align to obtain time stamps for each spoken utterance.

To examine how the proposed PPM systems complement other systems, fusion experiments were performed using a backend classifier [74].

5.1.2 Implementation Details

Mathematical details of how to train and score dialect identification systems are in Section 2.3.1. Below we document the implementation details of our DID experiment.

Systems S_1-S_3: PPM Surface phones obtained through phone decoding

Reference phones were obtained by force-aligning word transcripts with IQ Dict using the triphone IQ recognizer. Surface phones were obtained by direct decoding using the IQ triphone recognizer. The following systems were trained. (1) System S_1: Mono PPM, which models context-independent phonetic rules; (2) System S_2: Tri PPM (standard tying), which models context-dependent phonetic rules by clustering

triphone states, as mentioned in Section 3.3; (3) System S_3: Tri PPM (sophisticated tying), which models context-dependent phonetic rules by clustering deletion arcs and triphone states, as mentioned in Section 3.6.

To simplify the tying mechanism, deletion and typical transitions are constrained to be destination independent; i.e.,

$$a_{irj} = a_{ir}, \qquad r \in \{typ, del\} \tag{5.1}$$

Insertion transitions were destination specific since insertions can only be emitted by insertion states. Consecutive insertions were allowed through the self insertion arcs, while consecutive deletions were not allowed.

Systems F_1-F_3: PPM Surface phones obtained by forced-alignment

The system setup of System F_1-F_3 are exactly the same as System S_1-S_3 except for the surface phone generation, where in System F_1-F_3 surface phones were obtained through force-alignment using pan-Arabic Dict with the triphone IQ recognizer.

Systems A_1-A_2: APM

System A_1 and A_2 are acoustic counterparts of System S_1 and S_2: (1) System A_1: Mono APM, where acoustic characteristics are modeled in monophone categories; (2) System A_2: Tri APM (standard tying), where acoustic characteristics are modeled in clustered triphone categories. System A_{2a} is trained exactly the same way as System A_2, but during test time, scoring is done without transcription aid.

APM were implemented by first training a universal background model using data from both dialects. Dialect-specific data were then used for state clustering and adapting the universal background model using 32 Gaussians/state.

Three systems often used in DID are chosen as baseline systems to be compared with the proposed phone-based and acoustic-based pronunciation models.

System B_1: SDC-GMM

Each GMM has 2048 mixture components, and shifted delta cepstra (SDC) are used as features. The experiment setup is the same as [89]. A universal background model was first trained on all dialects, and then each dialect-specific GMM was adapted.

System B_2: Adapted Phonetic Models (APM0)

Adapted phonetic models can be viewed as an extension of GMM systems, where acoustic information is modeled in phonetic categories, making it easier to pinpoint where the acoustic differences lie in. An adapted phonetic model was trained according to [84]. The IQ monophone recognizer was used to segment the speech signal to monophone units, where the acoustic observations of each phone were modeled as a GMM.

System B_3: Phone recognition followed by language modeling (PRLM)

PRLM [98] is one of the most classical phonotactic approaches in dialect recognition. Adapted tokenizers trained in Section 5.1.2 were used to generate phone sequences to train language models [84] for our PRLM baseline.

5.1.3 Results

EER results of each system are listed in Figure 5-1. The following are some general observations.

- System A_2 (Triphone APM; standard tying) obtains the lowest EER.
- PPM systems ($S_1 - S_3$, $F_1 - F_3$) outperform baseline systems ($B_1 - B_3$).
- Triphone PPM and APM outperform their monophone counterparts.
- Performance of sophisticated tying is comparable to that of standard tying.

Phone-based Pronunciation Model	EER (%)	Confidence Interval (%)
S_1: Mono PPM	17.80	15.91 - 19.97
S_2: Tri PPM (standard tying)	16.51	14.80 - 18.50
S_3: Tri PPM (sophisticated tying)	16.66	14.45 - 18.91

Acoustic-based Pronunciation Model	EER (%)	Confidence Interval (%)
A_1: Mono APM	18.58	16.53 - 20.73
A_2: Tri APM (standard tying)	13.77	12.04 - 15.32
A_{2a}: Tri APM (standard tying; unsupervised at test)	23.89	21.78 - 26.15

Forced Aligned PPM Systems	EER (%)	Confidence Interval (%)
F_1: Mono PPM	17.01	15.10 - 19.07
F_2: Tri PPM (standard tying)	16.68	14.89 - 18.58
F_3: Tri APM (sophisticated tying)	16.71	14.96 - 18.80

Baseline Systems	EER (%)	Confidence Interval (%)
B_1: SDC-GMM	26.80	25.02 - 28.45
B_2: Acoustic phonetic model	24.33	22.80 - 26.20
B_3: PRLM	21.20	19.51 - 23.00

Figure 5-1: *DID performance comparison for 5-Dialect Arabic Corpus.*

Baseline System	EER (%)	Fusion Results (%)					
		+ F_2: Tri-PPM			+ S_2: Tri-PPM		
		EER	Confidence Interval	Relative gain	EER	Confidence Interval	Relative gain
B_1:SDC-GMM	26.80	19.63	17.49 - 21.91	26.75	17.31	15.45 - 19.44	35.41
B_2:APM0	24.33	18.71	16.71 - 20.95	23.10	16.95	15.09 - 19.08	30.33
B_3:PRLM	21.2	14.36	12.53 - 16.07	32.26	14.02	12.31 - 15.97	35.87

Figure 5-2: *Baseline performance and fusion results. Units in %.*

5.1.4 Discussion

Below we do further analysis and discuss the implications of our DID results on the 5-dialect Arabic Corpus.

PPM Systems

The performance of Systems S_1, S_2, S_3 and F_1, F_2, F_3 are similar. System S_1, S_2, S_3 are more practical since they do not require pronunciation dictionaries, which are time- and labor-intensive to construct. In addition, Systems S_1, S_2, S_3 can learn more rules than Systems F_1, F_2, F_3 since they are not constrained by the pronunciation dictionary, which might be why the relative gains of System S_1, S_2, S_3 in fusion experiments are on average 17.2% relative greater than those of System F_1, F_2, F_3 (see Fig. 5-2.) Systems S_1, S_2, S_3 achieve greater fusion gains despite the additional noise from phone recognition errors when generating surface phones.

PPM Systems exploiting phonetic context perform better than their monophone counterparts: System S_2 outperforms S_1 by 7.25% relative, and System F_2 outperforms F_1 by 1.94% relative. Performance of sophisticated tying comparable to that of standard tying.

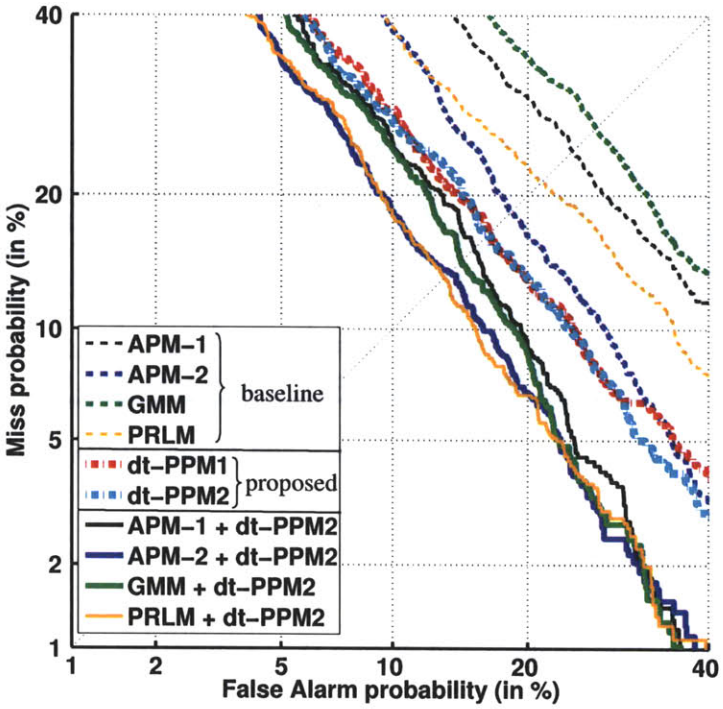

Figure 5-3: *Detection error trade-off (DET) curves of 5-Dialect Arabic Corpus.*

Baselines vs. PPM Systems

All PPM systems perform better than the acoustic and phonotactic baselines (see Figure 5-1) and Figure 5-2. Moreover, the Triphone PPMs fuse well with them: relative gains are 25-36% and 21-32% after fusion with System S_2 and F_2 (see Fig. 5-2), suggesting that PPMs exploit phonetic information not used in these baselines.

Fig. 5-3 shows the detection error trade-off (DET) curves of the baselines systems $B_1 - B_3$, the Triphone PPMs (System S_2 and F_2), and their fusion results between baseline systems and PPM systems. We see that in general DET performance corresponds with the EER performance: Triphone PPM systems perform better than baselines, and error rates of fused systems are even lower.[3]

[3]The trend that EER corresponds with DET performance also holds for other systems, thus their DET curves are excluded for clarity purposes.

APM Systems

We list the different versions of APM in Figure 5-4, and compare their differences according to (1) whether the model exploits phonetic context, (2) word transcriptions are used during training, and (3) word transcriptions are used during test time.

The simplest model is System B_2 (APM0), a monophone acoustic model that doesn't use word transcriptions at training nor test time. It is an extension of GMM, where acoustic information is categorized into phonetic units, which are determined through a phone recognizer. Among the APM systems, it requires the least resources to train. It is expected that its performance to be at least as good as GMM, and possibly better, but worse among the other APM systems. The results in Figure 5-1 and 5-2 match our expectations.

All other versions of APM systems require transcriptions. System A_1 is a monophone acoustic model that requires transcriptions during train and test. It is expected that System A_1 performs better than System B_2, since there are no phone recognition errors in determining phonetic units in System A_1. System A_1 outperforms System B_2 by 23.6% relative, indicating the importance of having a *clean* reference to compare acoustic characteristics across dialects.

It is well-known in linguistics that acoustic or phonetic differences across dialects are often phonetic-context dependent. This phenomenon is empirically demonstrated: System A_2 outperforms System A_1, by 25.9% relative. System A_2 (Tri APM) is by far the best system among APMs and all the other systems.

For mere dialect recognition purposes, it is desirable that transcriptions are not needed during test time. System A_{2a} is the *unsupervised* version of System A_2, where transcriptions are not used during test time. The EER result of System A_{2a} is comparable to System B_2 (performance difference: 1.8% relative), suggesting that it is more practical to use System B_1 in pure DID applications, since no transcription is required and performance is similar to System A_{2a}.

APM Systems	Phonetic context	Transcription at training	Transcription at test	EER (%)	Description
B_2: Mono APM	N	N	N	24.33	Extension of B_2: GMM
A_1: Mono APM	N	Y	Y	18.58	Acoustic counterpart of S_1: mono PPM
A_2: Tri APM	Y	Y	Y	13.77	Acoustic counterpart of S_2: Tri PPM
A_{2a}: Tri APM	Y	Y	N	23.89	*Unsupervised* version of A_2: Tri APM

Figure 5-4: *Different versions of APM System.*

Baselines vs. APM Systems

We fused the best performing pronunciation model, System A_2 (Triphone APM) with the baseline systems, as shown in Figure 5-5. Relative gains are largest when System A_2 is fused with System B_1 (SDC-GMM), reaching 56%. Relative gains are smallest when System A_2 is fused with System B_2 (APM0), which is still pretty high, reaching 42.5%.

Figure 5-3 shows DET curves of the baselines systems $B_1 - B_3$, System S_2 (Triphone PPM), System A_2 (Triphone APM), and fusion results. We see that in general DET performance corresponds with the EER performance: System S_2 and A_2 (Triphone PPM and Triphone APM) perform better than the baselines, and error rates of fused systems are even lower.

Word-Usage Differences

Further analysis shows that word-usage differences are large among Arabic dialects. A 1-gram language model of word occurrences achieves EER of 5.43%, outperforming all proposed models and standard baseline systems. This word-usage difference is a confounding factor in our analysis, since it is challenging to tease out how much performance gain of the pronunciation models (and other baseline systems) are from lexical or phonotactic differences across dialects. Fortunately, the StoryCorps corpus does have lexical differences that confound our analysis. We will discuss the dialect

Baseline System	EER (%)	Fusion Results (%) Baseline + System A_2 (Tri-APM)		
		EER	Confidence Interval	Relative Gain
B_1:SDC-GMM	26.80	13.51	11.78 - 15.39	49.59
B_2:APM0	24.33	14.10	12.37 - 16.13	42.47
B_3:PRLM	21.2	11.29	9.64 - 13.07	46.75

Figure 5-5: *Fusion results with System A_2: tri-APM.*

recognition experiments performed on StoryCorps in the next Section, which still show the effectiveness of the proposed systems.

5.1.5 Summary

Assuming that word-usage difference does not compromise our analysis, the empirical DID results in Section 5.1 show that on the 5-Dialect Arabic Corpus:

(1) Triphone APM using standard tying is the best performing system.

(2) The proposed PPM and APM systems all perform better than baseline systems and fuse well with them, achieving relative gains beyond 26-56%.

(3) PPM performance is similar when using standard and sophisticated tying.

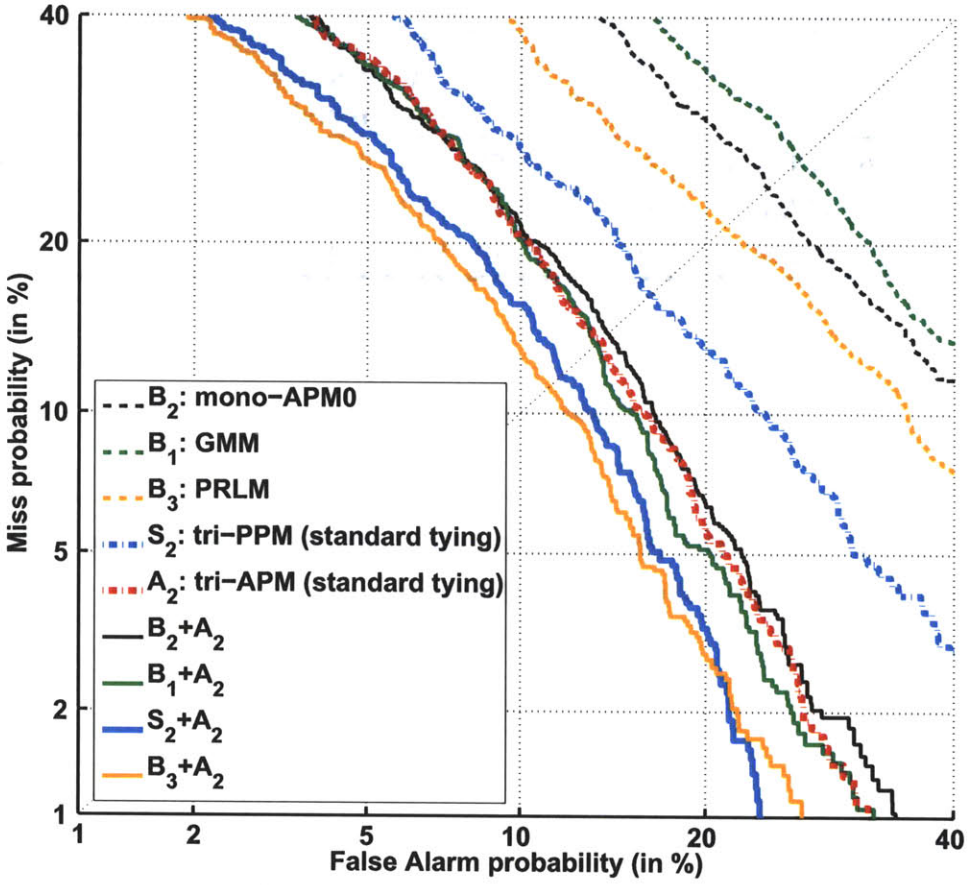

Figure 5-6: *Detection error trade-off: Fusion Results with System A_2: tri-APM (standard tying).*

5.2 Experiment II: StoryCorps

5.2.1 Experimental Setup

Data

Two sets of American English dialects were chosen from StoryCorps [46]. (1) African American Vernacular English (AAVE): speakers self-reported as African Americans. (2) Non-AAVE: speakers self reported as white. The conversations are between speakers of the same dialect to minimize accommodation issues [35]. Gender and age were balanced across all sets and dialects. The train set is 22.6 hr (69 speakers), the dev set is 7.1 hr (38 speakers), and the test set is 7 hr (28 speakers). There is virtually no word-usage difference across AAVE and non-AAVE dialects in StorpyCorps, since 1-gram word language models lead to EER performance close to chance. For the PPM systems, surface phones were obtained through a phone recognizer trained on WSJ0 [70]. The dev and test sets were divided into 30-sec trials. For more information about StoryCorps, please refer to Section 4.2.3 and [46].

5.2.2 Implementation Details

All implementation details are the same as Section 5.1.2, except for Systems F_1-F_3. The StoryCorps data did not come with dialect-specific pronunciation dictionaries as the 5-Dialect Arabic Corpus. Instead, we used phonetic rules converted from the linguistic literature [93], and transformed the WSJ0 American English dictionary into an AAVE-version pronunciation dictionary.

5.2.3 Results

EER results are listed in Figure 5-7. The following are some general observations.

- System A_2 (Triphone APM; standard tying) obtains the lowest EER: 9.97%.
- Performance of APM systems ($A_1 - A_2$) are comparable to Baselines (System $B_1 - B_2$), and better than PRLM (System B_3).

- Triphone PPM and APM outperform their monophone counterparts.
- Sophisticated tying outperforms standard tying in PPM by a relative gain of 6.1%.
- Acoustic systems outperform phonetic systems in distinguishing AAVE and non-AAVE dialects.

5.2.4 Discussion

PPM Systems

The performance of forced-aligned PPM systems are virtually the same despite phonetic context or tying structure, all around 23%. This result implies that while the *ground-truth* phonetic rules encoded in the AAVE pronunciation dictionary are useful in DID to some extent, these phonetic rules are by no means not comprehensive. System S_1, S_2, and S_3, where surface phones are determined by direct-decoding of phone recognition, perform better than the forced-aligned PPMs (System F_1, F_2, F_3) by at least 26% relative. This result suggests that Systems S_1, S_2, S_3 are learning phonetic rules beyond the ground-truth rules.

The triphone PPMs (System S_2 and S_3) outperform monophone PPM (System S_1) by 10.9% and 16.4%, respectively. System S_3 outperform System S_2 by 6%, suggesting that arc clustering is more appropriate for modeling deletion rules than state clustering.

APM Systems

The general trend of DID on StoryCorps is that acoustic systems are superior to phonetic systems, possibly because the main differences between AAVE and non-AAVE are acoustic rather than phonetic. In Figure 5-7, we see that without using phonetic context information, monophone APM (System A_1) already outperforms monophone PPM (System S_1) by 37.9% relative. The relative gain is even higher

Phone-based Pronunciation Model	EER (%)	Confidence Interval (%)
S_1: Mono PPM	17.35	13.32 - 22.07
S_2: Tri PPM (standard tying)	15.46	11.68 - 9.86
S_3: Tri PPM (sophisticated tying)	14.51	10.79 - 18.94

Acoustic-based Pronunciation Model	EER (%)	Confidence Interval (%)
A_1: Mono APM	10.78	6.28 - 13.33
A_2: Tri APM (standard tying)	9.97	7.04 - 13.97
A_{2a}: Tri APM (standard tying; unsupervised at test)	10.73	7.46 - 14.96

Forced Aligned PPM Systems	EER (%)	Confidence Interval (%)
F_1: Mono PPM	23.65	19.11 - 28.72
F_2: Tri PPM (standard tying)	23.29	18.49 - 28.72
F_3: Tri APM (sophisticated tying)	23.30	18.47 - 28.36

Baseline Systems	EER (%)	Confidence Interval (%)
B_1: SDC-GMM	10.86	7.70 - 15.22
B_2: Acoustic phonetic model	10.76	7.06 - 15.53
B_3: PRLM	13.45	9.73 - 18.16

Figure 5-7: *DID performance comparison for StoryCorps (AAVE vs. non-AAVE).*

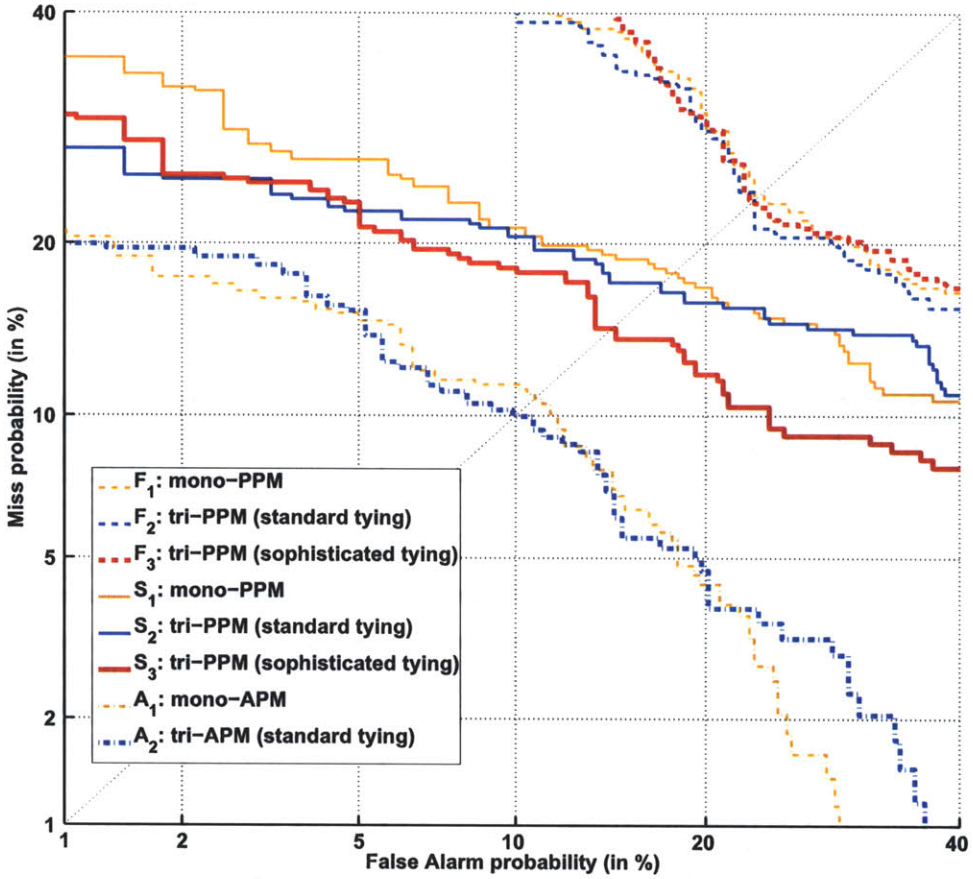

Figure 5-8: *Detection Error Trade-off Curves comparing pronunciation models (StoryCorps).*

when phonetic context is used: triphone APM (System A_4) outperforms monophone PPM (System S_1) by 42.5%.

Note that these gains are obtained only by a simplified APM system. Inferring from the results of the PPM systems, the complete triphone APM system, which uses arc clustering to model deletions, could achieve potentially even better DID performance. We plan to investigate this hypothesis in future work.

In Figure 5-8 we plot the DET curves all the PPM and APM systems. We see that the DET curve results are similar to those of EER: System A_1, A_2 performs the best, followed by System $S_1 - S_3$ (surface phones obtained through phone recognition decoding), and Systems $F_1 - F_3$ (surface phones obtained through forced-alignment with pronunciation dictionary) are the worst.

System A_{2a} is the *unsupervised* version of System A_2, where transcriptions are

not used during test time. Similar to results on the 5-Dialect Arabic Corpus, performance of System A_{2a} is comparable to that of System B_2 (performance difference: 1.2% relative), suggesting that it is more practical to use System B_1 in pure DID applications, since no transcription is required and performance is similar to System A_{2a}.

PPM vs. Baseline Systems

Both PRLM (System B_3) and triphone PPM use phonotactic information to recognize dialects. Their performances are close to each other (in the range of 13%-15% EER), with PRLM performing better than triphone PPM using sophisticated tying (System S_3). SDC-GMM and acoustic phonetic model (System B_1, B_2) both achieve around 10% EER, outperforming triphone PPM with sophisticated tying by at least 25% relative. Fusion results between PPM and baseline systems did not improve performance.

APM vs. Baseline Systems

We fused the best performing system triphone APM (System A_2) with the baseline systems (System $B_1 - B_3$), which resulted in relative gains beyond 25%. PRLM achieved the most relative gain (46.6%) when fused with triphone APM, driving down the EER to 7.18%. This fusion gain is most likely because PRLM is a phonetic system, which complements triphone-APM the most.

In Figure 5-10, we illustrate the performance comparison and fusion results in a DET plot. Again, the DET plot performance is similar to that of EER: Triphone APM performs better than baseline results, and detection errors are even lower when triphone APM fuses with the baselines.

5.2.5 Summary

The StoryCorps DID experiment suggests that (1) acoustic-based systems such as the proposed triphone APM is important in recognizing dialects with subtle differences

Baseline System	EER (%)	Fusion Results (%): Baseline + System A_2 (Tri-APM)		
		EER	Confidence Interval	Relative gain
B_1:SDC-GMM	10.86	8.10	4.83 - 12.20	25.4
B_2:APM0	10.76	7.91	4.35 - 12.85	26.6
B_3:PRLM	13.45	7.18	3.94 - 11.94	46.6

Figure 5-9: *Fusion results with System A_2: tri-APM on StoryCorps.*

such as AAVE and non-AAVE, (2) APM has useful applications in DID, and (3) the complete implementation of triphone APM, which uses sophisticated tying instead of standard tying can achieve even better DID performance.

5.3 Summary

We performed DID experiments on dialect from two different languages (Arabic: 5-Dialect Arabic Corpus; American English: StoryCorps.) Although the word-usage difference across Arabic dialects poses uncertainty on our analysis in Section 5.1, our results on StoryCorps are free from such criticism. In addition, many conclusions drawn from results of both experiments are similar. We summarize the main findings below.

(1) Proposed triphone APM is the best system, and fuses well with baseline systems on both corpora.

(2) Proposed PPM systems are able to learn rules beyond pronunciation dictionary or linguistic literature, even though surface phones obtained through phone

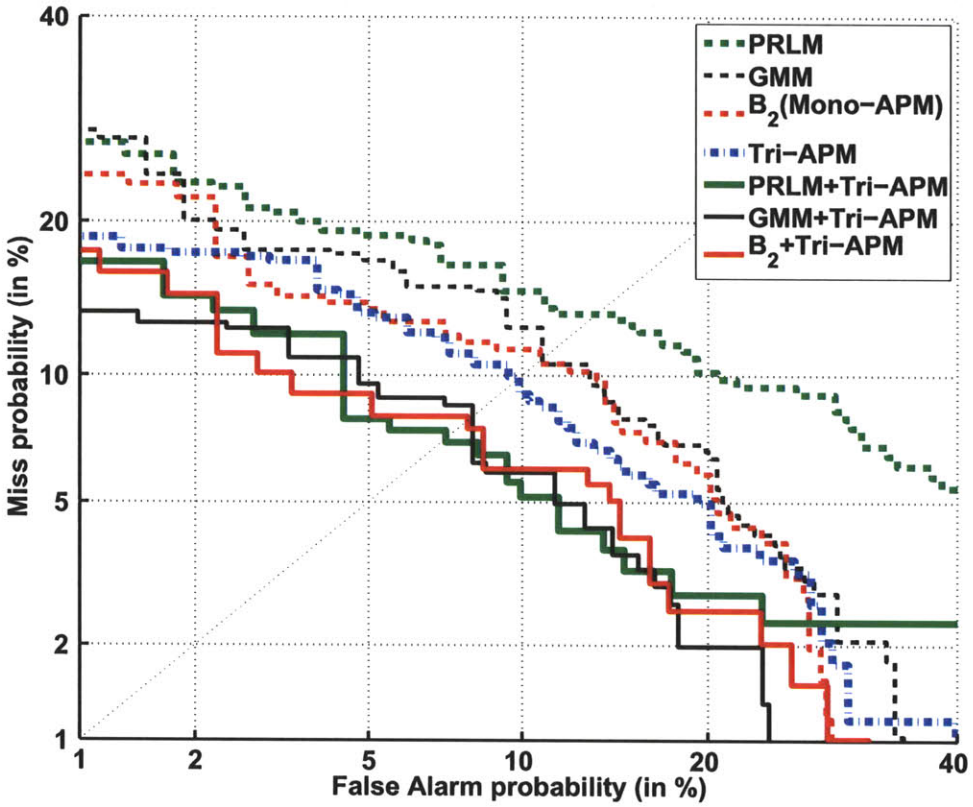

Figure 5-10: *Detection error trade-off (DET) curves of StoryCorps.*

recognition decoding are prone to errors.

(3) Phonetic context improves performance in PPM and APM systems.

(4) Performance of sophisticated tying in PPM is comparable to standard tying on the Arabic corpus; performance of sophisticated tying in PPM is slightly better than standard tying on the American English corpus.

(5) Performance of *Unsupervised* APM systems (without use of transcriptions during test time) are comparable to standard baseline systems in DID on both corpora.

Chapter 6

Pronunciation Generation Experiments

In the last chapter, we ran dialect recognition experiments to assess how well the proposed models learns phonetic rules. In this chapter, we conduct pronunciation generation experiments, where we evaluate how well a model has learned phonetic rules by generating dialect-specific pronunciations given a reference dialect's pronunciation. We ran experiments on two datasets[1]: (1) WSJ-CAM0 (see Section 6.1), and (2) 5-Dialect Arabic Corpus (see Section 6.2, and summarize the findings in Section 6.3.

6.1 Experiment I: WSJ-CAM0

The objective of this experiment is to assess how well a pronunciation model generates British pronunciations given American pronunciations.

6.1.1 Assumptions

1. All pronunciation variations across dialects are governed by underlying phonetic rules.

[1]We did not run this experiment on StoryCorps because we do not have pronunciation dictionaries for AAVE to use as ground-truth surface phones.

2. The phonetic transcriptions provided by WSJ-CAM0 are *ground-truth* surface phones O^*.

3. The ability to predict ground-truth surface phones from the trained pronunciation model indicates how well the phonetic rules are learned from the pronunciation model algorithms.

We are aware that some of these assumptions might be oversimplifications in reality, but they are useful for analysis purposes.

6.1.2 Experimental Setup

Data: WSJ-CAM0

WSJ-CAM0 [78] is the UK version of the American English WSJ0 database [70]. The training set is 15.3 hr (92 speakers); the test and dev set are each 4 hr (48 speakers). For more details of WSJ-CAM0 please refer to Section 4.2.1 and [78].

PPM and APM systems were trained using WSJ-CAM0's train set. The reference phones C are determined by the WSJ0 American English pronunciation dictionary and the ground-truth surface phones O^* are the phonetic transcriptions provided by WSJ-CAM0, as mentioned in Section 6.2.1.

PPM Systems

Given a trained pronunciation model, we generate the most likely observations $\hat{O}$ given the reference phones C in the test set[2]. Dynamic programming is used to align $\hat{O}$ with the ground-truth O^*. (See Figure 6-1.)

The phone error rate (PER) between O^* and $\hat{O}$ is listed in Figure 6-3. As shown in Figure 6-2, the baseline system S_0 is the case where no pronunciation model is used; i.e., the PER between the ground-truth surface phones O^* and the reference phones C (obtained from the American WSJ dictionary.)

[2]Total number of phones in the test set: 299,853.

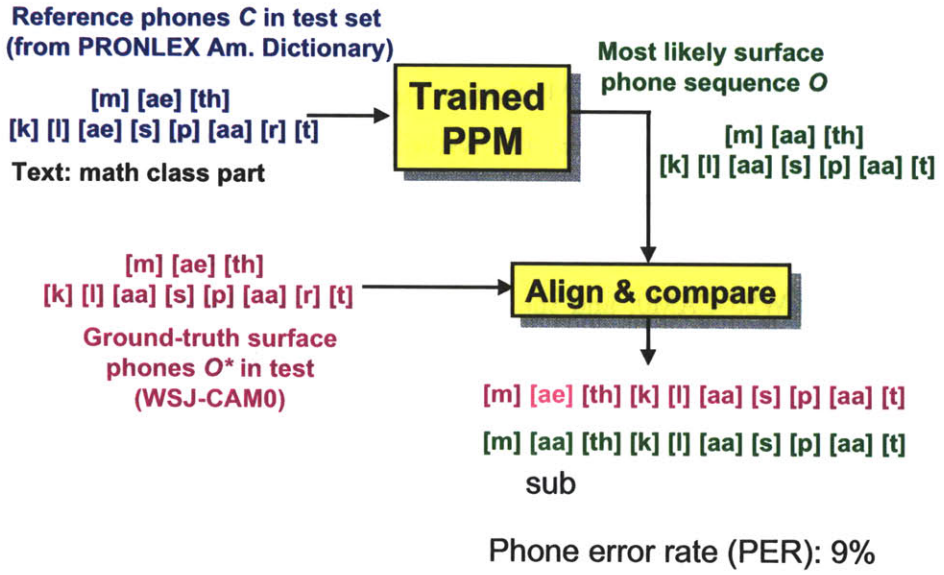

Figure 6-1: *Experimental setup for pronunciation generation experiment.*

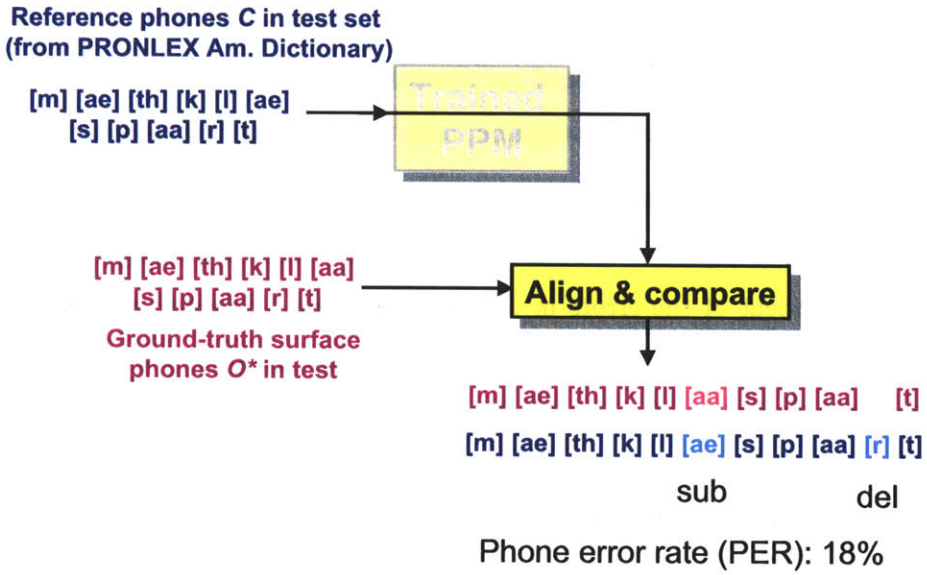

Figure 6-2: *Baseline for pronunciation generation experiment.*

APM Systems

We assess how well triphone APM (System A_2) is modeling British dialect acoustics by comparing it with two systems: (1) System S_u, an oracle APM system trained on WSJ-CAM0's audio and pronunciation dictionary, and (2) System S_b, a baseline APM system trained on WSJ0's audio and pronunciation dictionary. All 3 systems decode the test set of WSJ-CAM0, and the decoded phones are aligned with ground-truth O^* to compute PER (see Figure 6-4.) All systems used standard tying for triphone state clustering.

Note that PPM systems are given the reference phones at test time (i.e., supervised), while the APM systems decode phones without transcription aid at test time (i.e., unsupervised). Thus, it is expected that PPM systems obtain lower PER.

Statistical Test

We used the matched pairs test in [36] to evaluate whether the performance difference of the two systems being compared are statistically significant. Errors were divided into deletion, insertion, and substitution; each type of error was analyzed separately.

Method. Let us suppose that we can divide the output stream from a pronunciation model system into segments in such a way that the errors in one segment are statistically independent of the errors in any other segment. Suppose we are comparing the performance difference of Y_1 and Y_2 Let N_1^i be the number of errors made in the i-th segment by System Y_1 , and N_2^i the number of errors made by System Y_2. Note that the type of error is unimportant, as long as the method of counting errors is consistent for each segment and for both systems.

Let $Z_i = N_1^i - N_2^i$, $i = 1, ..., n$, where n is the number of segments. Let μ_z be the unknown average difference in the number of errors in a segment made by the two Systems. We would like to ascertain whether $\mu_z = 0$. The maximum likelihood

estimate of μ_z and the variance of Z_i are

$$\hat{\mu_z} = \sum_{i=1}^{n} \frac{Z_i}{n} \tag{6.1}$$

$$\hat{\sigma_z^2} = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \mu_z)^2 \tag{6.2}$$

If W is defined as

$$W = \frac{\hat{\mu_z}}{\hat{\sigma_z}/\sqrt{n}}, \tag{6.3}$$

then assuming n is sufficiently large, W will approximate a standard normal distribution $\mathcal{N}(0,1)$. We can test the null hypothesis H_0: $\mu_z = 0$, by computing $P = 2Pr(Z > |w|)$, where Z is a random variable with distribution $\mathcal{N}(0,1)$ and w is the realized value of W.

Implementation Details. We divided the generated surface phone outputs into segments where no errors have occurred for some minimal time period T (*good* segments) and segments where errors occur (*bad* segments), according to [36]. T is required to be sufficiently long to ensure that after a good segment, the first error in a bad segment is independent of any previous errors. T was swept on the development set (ranging from values of 9 to 402 phones), and all resulted in similar p-values ($p < 0.001$) on the test set. The number of segments n ranged from 756 to 32491, which is assumed to be sufficiently large enough for W to be normally distributed, where a reasonable estimate of the variance of Z_i can be obtained.

6.1.3 Results

Results of the PPM systems are shown in Figure 6-3, and those of APM are shown in Figure 6-4. All improvements are shown to be statistically significant ($p < 0.001$).

System	Phone Error Rate (PER) (%)			
	Overall	Deletion	Insertion	Substitution
S_0: Baseline	21.7	4.0	3.6	14.2
S_1: Mono PPM	15.1	2.0	3.3	9.8
S_2: Tri PPM (standard tying)	9.0	2.1	1.9	5.0
S_3: Tri PPM (sophisticated tying)	9.0	1.4	2.6	5.0

System	Relative improvement to baseline (%)			
	Overall	Deletion	Insertion	Substitution
S_1: Mono PPM	30.4	50	8.3	31
S_2: Tri PPM (standard tying)	59	48	47.2	64.8
S_3: Tri PPM (sophisticated tying)	59	65	27.8	64.8

Figure 6-3: *PPM system performance in generating British pronunciation from American pronunciation.*

System	Training data		Phone Error Rate (PER) (%)			
	Audio	Dictionary	Overall	Del.	Ins.	Sub.
A_u: Oracle (upper bound)	WSJ-CAM0	WSJ-CAM0	25.3	4.2	6.6	14.5
A_b: Baseline	WSJ0	WSJ0	42.5	9.0	6.0	27.5
A_2: Tri APM	WSJ-CAM0	WSJ0	27.9	5.3	5.0	17.5

System	Relative Improvement to Baseline System A_b (%)			
	Overall	Deletion	Insertion	Substitution
A_2: Tri APM	34.4	41	16.7	36.4

System	Relative Difference from Oracle System A_u (%)			
	Overall	Deletion	Insertion	Substitution
A_2: Tri APM	9.3	20.8	-32	17.1

Figure 6-4: *APM system performance in generating British pronunciation from American pronunciation.*

6.1.4 Discussion

PPM Systems: Sophisticated Tying is Suitable in Modeling Deletions

Without using phonetic context, monophone PPM (System S_1) already beats the baseline (System S_0) by 30.4%. Relative gains from triphone PPMs (System S_2 and S_3) are even greater, both reducing the baseline PER by 59% relative. However, the matched pairs test [36] shows that the two systems are making statistically different errors ($p < 0.001$).

Compared to Monophone PPM (System S_1), System S_2 shows *negative* improvement in deletion errors (-5%), implying that standard tying over-generalizes deletion rules. When considering deletion errors, sophisticated tying (System S_3) beats standard tying (System S_2) by 33% relative, supporting our hypothesis that arc clustering is suitable in modeling deletions. It also corresponds with linguistic knowledge that the phone of interest is generally affected more by its right-context than left-context; e.g., R-dropping in RP [92]. Among the /r/'s that were incorrectly deleted in standard tying (System S_2), sophisticated tying (System S_3) correctly generated 24% of these /r/'s. Though sophisticated tying (System S_3) reduces deletion errors, its insertion errors increases when compared to standard tying (System S_2). This phenomenon might be caused by data sparsity, since sophisticated tying requires more model parameters than standard tying.

APM System

From Figure 6-4, we see that triphone APM (System A_2) beats the baseline System A_b by 34.4% relative. The sub-category relative error reductions are 41%, 16.7%, and 36.4% for deletions, insertions, and substitutions, respectively. We empirically discovered large channel differences between WSJ0 and WSJ-CAM0 (see Appendix B). Therefore, these gains also include channel differences, making them appear overly optimistic. Nonetheless, the performance of triphone APM (System A_2) is not far from that of the oracle system A_u: System A_u beats triphone APM (System A_2) only by 2.6% absolute and 9.3% relative, suggesting that triphone APM is fairly

capable of transforming American pronunciations to British pronunciations.

6.2 Experiment II: 5-Arabic Dialect Corpus

6.2.1 Assumptions

We adapt the assumptions in Section 6.1 to the following.

1. All pronunciation variation across dialects are governed by underlying phonetic rules.

2. All pronunciation variation across dialects are captured in pan-Arabic Dict.

3. The ground-truth surface phones O^* of each dialect can be obtained by force-alignment using pan-Arabic Dict, word transcripts, and the IQ phone recognizer.

4. The ability to predict ground-truth surface phones O^* using the trained PPM system indicates how well the underlying phonetic rules are retrieved from the PPM algorithm.

6.2.2 Experimental Setup

Data: 5-Dialect Arabic Corpus

All experimental setup are the same as Section 6.1.2, except that the ground-truth surface phones are obtained through force-aligning the test set audio with pan-Arabic Dict.

6.2.3 Results

The phone error rate between the ground-truth surface phones and the estimated surface phones (generated from the trained PPMs system) of the test set are averaged across the four dialects (AE, EG, PS, SY). Results of the PPM systems are shown in

System	Phone Error Rate (PER) (%)			
	Overall	Deletion	Insertion	Substitution
S_0: Baseline	14.73	4.83	2.35	7.55
S_1: Mono PPM	14.61	4.83	2.35	7.43
S_2: Tri PPM (standard tying)	14.56	4.88	2.35	7.33
S_3: Tri PPM (sophisticated tying)	14.48	4.83	2.35	7.30

System	Relative improvement to baseline (%)			
	Overall	Deletion	Insertion	Substitution
S_1: Mono PPM	0.81	0	0	1.59
S_2: Tri PPM (standard tying)	1.15	-1	0	2.91
S_3: Tri PPM (sophisticated tying)	1.70	0	0	3.31

Figure 6-5: *PPM system performance in generating dialect-specific (AE, EG, PS, SY) pronunciation from reference (IQ) pronunciation.*

Figure 6-5. All improvements are shown to be statistically significant ($p < 0.01$). The baseline system S_0 is the case where no pronunciation model is used; i.e., the PER between the ground-truth surface phones O^* and the reference phones C (obtained from forced-alignment using IQ-Dict.)

6.2.4 Discussion

Relative gains of the PPM systems are small, but statistically significant. The main performance gain is from substation errors. Insertion errors showed no improvement, but was low (2.35%) to begin with. Monophone PPM (System S_1) improves the baseline (System S_0) by 0.08% relative. The relative gain from System S_2 (triphone PPM; standard tying) is slightly larger (1.15%). Similar to the WSJ-CAM0 case, System S_2 shows *negative* improvement in deletion errors (-1%) when compared to the baseline (System S_0) and monophone PPM (System S_1). This result implies that standard tying is not suitable for modeling deletions. System S_3 (triphone PPM; sophisticated tying) is the best performing system, reducing the substitution error by 3.31% relative, and overall error by 1.71%.

6.3 Summary

In this chapter, we evaluate the ability of a pronunciation model to learn phonetic rules by how well it is at generating dialect-specific pronunciation from a reference dialect. We ran experiments on two datasets: (1) WSJ-CAM0, where we generate British pronunciation given the American pronunciation, and (2) 5-Dialect Arabic Corpus, where we generate pronunciations of non-Iraqi Arabic dialects (AE, EG, PS, SY) from Iraqi Arabic pronunciation. Our results suggest that

1. Phonetic context improves performance of generating dialect-specific pronunciations from a reference dialect.

2. Standard tying increases deletion errors, while sophisticated tying does not.

3. Triphone APM could potentially perform even better if sophisticated tying is used instead of standard tying.

Chapter 7

Rule Retrieval Experiment

In the last two chapters, we ran dialect recognition and pronunciation generation experiments to assess how well the proposed models learns phonetic rules. We were not able to run pronunciation experiments on StoryCorps due to the lack of an AVVE pronunciation dictionary. Therefore, in this section we run an information retrieval experiment on StoryCorps to assess how well our proposed pronunciation models are able to retrieve rules documented in the literature.

7.1 Experimental Setup

We compare automatically learned rules with the linguistic literature, and use information retrieval to quantify our results. Only systems exploiting phonetic context are considered since these rules often depend on context.

7.1.1 Data: StoryCorps

Two sets of American English dialects were chosen from StoryCorps [46]. (1) African American Vernacular English (AAVE): speakers self-reported as African Americans. (2) Non-AAVE: speakers self-reported as white. The conversations are between speakers of the same dialect to minimize accommodation issues [35]. Gender and age were balanced across all sets and dialects. The train set is 22.6 hr (69 speakers), the dev

set is 7.1 hr (38 speakers), and the test set is 7 hr (28 speakers). For more details of StoryCorps please refer to Section 4.2.3 and [46].

7.1.2 Ground-Truth Rules

We adopted descriptions of the AAVE dialect from the literature (e.g., [93]), converted them to 31 phonetic rules with the help of linguists. These rules serve as the *ground-truth* rules in this experiment. We list these rules in Figure 7-1 and Figure 7-1.

7.2 Implementation Details

For a given ground-truth rule, a rule retrieval experiment was done. True trials are triphone states in the test set that match the ground truth rule's center phone and phonetic context. False trials are triphone occurrences in the test set that match the center phone of the ground truth rule but not the phonetic context. The duration-normalized log likelihood ratio of each trial was used to compute recall[1] and precision[2] for each rule retrieval experiment. To compare pronunciation models, we used two metrics: (1) the precision rate (when recall is fixed at *0.1*), and (2) the optimal *F measure*[3], determined by tuning the decision threshold on the dev set. These measurements were averaged across the 30 rule retrieval experiments and listed in Figure 7-3.

7.3 Results and Discussion

Results from the rule retrieval experiment are listed in Figure 7-3 and Figure 7-4.

From Figure 7-3 and Figure 7-4 we see that triphone APM (System A_2) outperforms the triphone PPM using standard tying (System S_2) by 42% in F measure, and 48% in precision. Triphone APM (System A_2) outperforms the triphone PPM using

[1] $recall = P(\text{retrieved rules}|\text{ground truth})$
[2] $precision = P(\text{ground truth}|\text{retrieved rules})$
[3] $F = \frac{2 \times precision \times recall}{precision + recall}$

CONSONANTS
Interdental fricatives become labial fricatives
[dh] -> [v] / [+vowel] _ [+vowel] [th] -> [f] / [+vowel] _ [+vowel]
Coronal fricatives become stops
[dh] -> [d] / [+vowel] _ [+vowel] [th] -> [t] / [+vowel] _ [+vowel] [th] -> [d]/ _ [n] [s] -> [d]/ _ [n]
velar nasal become alveolar
[ng] -> [n]
Other
[t] -> [k] / [s] _ [r]
VOWELS
/eh/ and /ih/ merger
[eh] -> [ih] [eh] -> [ih] / _ [+nasal]
[eh] diphthongization
[eh] -> [ey] / _ [l] [eh] -> [ey] / _ [r]
[ih] elongation
[ih] -> [iy]
[ih]-> [ae] / [th] _ [ng]
[ao] vowel shift
[ao] -> [aa]
[ay] monophthongization
[ay] -> [ae]
[ay] -> [aa] / _ [+cons]
[ay] -> [ae] / _ [+cons]
[ay] vowel shift
[ay] -> [oy]
[aw] vowel shift
[aw] -> [ow]
[aw] -> [ay] / _ [l]
[aw] -> [aa] / _ [l]
[aw] -> [uw] / _ [t]
[aw] -> [ow] / _ [t]
[uw] vowel shift
[uh] -> [uw]/ _ [l]

Figure 7-1: *Ground-truth substitution rules.*

Consonant cluster deletion
[-voiced +cons] -> ø / [+nasal +cons] _ [+voiced +cons] -> ø / [+nasal +cons] _
R vocalization
r -> ø / [+vowel] _ r-> ø / [+vowel] _ [+vowel] r -> ø / _ [+cons +stop +front]
L vocalization
l -> ø / [+vowel] _ [+cons]

Figure 7-2: *Ground-truth deletion rules.*

System	Average F measure	Average Precision (Recall=0.1)
S_2: Tri PPM (standard tying)	0.24	0.14
S_3: Tri PPM (sophisticated tying)	0.13	0.08
A_2: Tri APM (standard tying)	0.42	0.27

Figure 7-3: *Comparison of rule retrieval results.*

sophisticated tying (System S_3) by 70% in both F measure and precision. These results imply that the triphone APM system retrieves dialect-specific rules much better than the PPM systems, as is also suggested in the DID task.

The lack of deletion rules in the ground-truth list might be one reason why System S_3 (sophisticated tying) performed poorly. A caveat to this experiment is that the ground truth rules selected here are not comprehensive. In addition, some of these rules might be anecdotally used to describe the AAVE dialect, but lack empirical and statistical verification. In future work, rule candidates from the false alarms could be further analyzed and potentially complement existing linguistic knowledge.

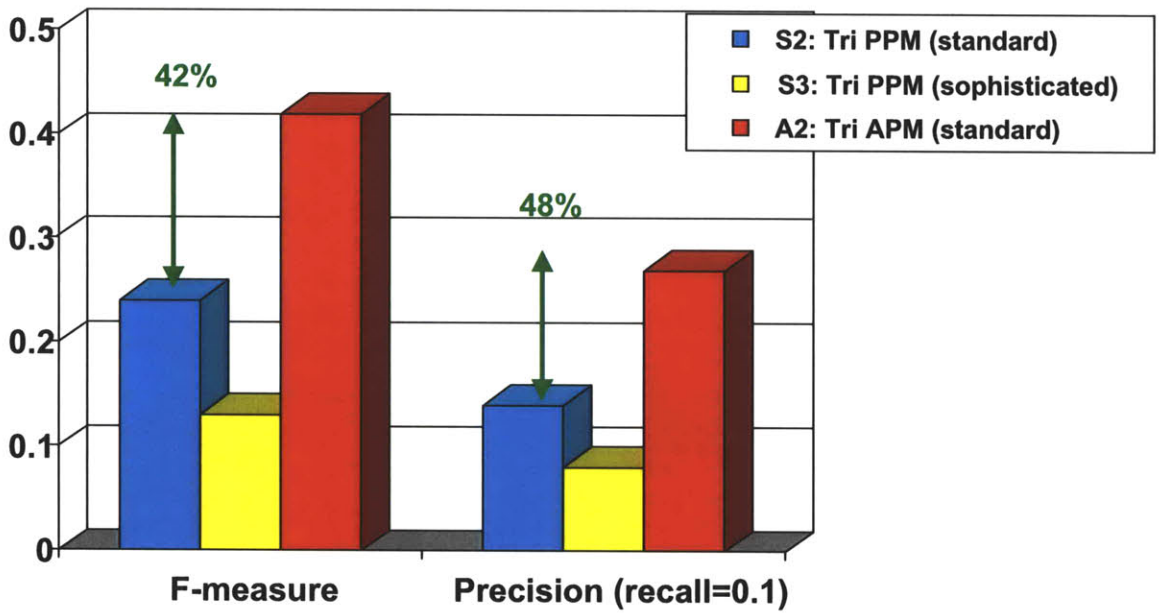

Figure 7-4: *System A_2 Tri-APM improves retrieval rate by at least 42% relative.*

7.4 Summary

In this chapter, we ran a rule retrieval experiment to compare the learned rules with linguistic descriptions of the AAVE dialect. We found that the acoustic-based pronunciation model (APM) outperforms the other pronunciation models. In the next chapter, we will examine the top ranking rules in the proposed pronunciation models and discuss their implications.

Chapter 8

Discussion of Automatically Learned Rules

Automatically learned rules from our proposed systems could be challenging to interpret at first sight, as its appearance seems different from linguistic rules. However, they provide rich implications that can be further explored and examined. In this chapter, we compare the top ranking rules from the proposed systems with linguistic literature descriptions, and discuss the potential implications of these rules, advantages and limitations of the model, and how these results could help linguists further examine dialect-specific characteristics. We also illustrate examples of where the learned rules correspond to the linguistic literature, suggesting that the proposed models are moving in the right direction of learning dialect-specific rules.

8.1 Determination of Top Ranking Rules

We list top ranking rules and compare them with linguistic descriptions in Figure 8-3, Figure 8-10, Figure 8-11, Figure 8-12, Figure 8-5, Figure 8-1, Figure 8-2, Figure 8-13, and Figure 8-14. The occurrence frequency of the phonetic transformation given the phonetic context of the learned rule (denoted as *Prob.*) is also listed. Linguistic descriptions were extracted from [93] for AAVE, [92] for RP, and [90, 53] for Arabic

dialects[1].

The ranking is determined by log likelihood ratio computed just as in in Section 7.1.[2] For substitution and insertion rules, only the top 2% were selected. For deletion rules, the top 25% were selected.[3]

8.2 Rule Analysis and Interpretation

8.2.1 Refined-Rules with Quantification of Occurrence Frequency

We see in Figure 8-1 that System S_2 is able to learn the *trap-bath split* rule: [ae] transforms to [aa] when [ae] is followed by a voiceless fricative. However, this linguistic rule corresponds to two automatically learned rules. The frequency of this transformation is also context dependent. If the following voiceless fricative is [+front] (i.e. [f]) the transformation is more likely to occur as opposed to the case where [ae] is preceded by a voiceless phone and followed by a *un-fronted* voiceless fricative (i.e., [s], [sh], [th], [ch]); the probabilities are 0.843 vs. 0.52. According to these automatically learned rules, the transformation is more likely to occur in a word like *laugh* than a word like *pass*

These results give us insights into how to further analyze linguistic descriptions, refine the conditioning of the rules, and quantify how frequently these rules take place. We discuss another example in the AAVE dialect below. It is documented in the literature that the vowel [ao] transforms to [aa] in the AAVE dialect [93]. However, our results in Figure 8-3 shows that [ao] is is transformed to [aa] when it is preceded by vowels with mid height (i.e., [ah], [eh], [er], [ey], [ow], [uh]). Again, we see that the probability of this phonetic transformation taking place depends on the

[1]We only found linguistic descriptions of substation rules in Arabic dialects [90, 53], so we only compare learned substitution rules with them in Figure 8-13 and Figure 8-14.

[2]For RP (WSJ-CAM0), the log likelihood was computed instead because there is no non-target model.

[3]The definition of top ranking is different for substitution/insertion rules and deletion rules because there are far more substitution and insertion rules (42 times more).

Literature	Proposed System	
Rule	**Learned Rule**	**Prob.**
[ae] -> [aa] /_ [+fric, -voiced] (trap-bath split)	[ae] -> [aa] /_ [+fric, -voiced, +front]	0.843
	[ae] -> [aa] / [-voiced]_ [+fric, -voiced, -front]	0.520
[r] -> ø / _ [+cons, #] (R Dropping)	$[er]_{ins}$ -> [ah] / [+vowel, -nasal] _ [-vowel, -sil, -glide, +fric, -dipth, -liq, -syl, -stop, +affric]	1.00
	[er] -> [ah] / [+syl, +liq] _ [-short, +fric, -vowel, -dipth, -glide, -liq, -syl, +voiced, -stop, +affric]	1.00

Figure 8-1: *RP substitution rule comparison. Learned rules from System S_2 (Triphone PPM; standard tying.)*

phone following [ao]: the probability is higher (0.254) if the following phone is a stop as opposed to a non-stop (0.196).

These frequency differences and narrowing of phonetic context conditioning could reflect the reality that some phonetic contexts more easily lead to transformations. It should be noted that there are other possible reasons for this phenomenon to occur. For example, since the training data is not infinite, it is possible that this phenomenon is a result of under-sampling of certain words, causing rules to be over-specified. More training data and further investigation is required to verify if under-sampling is an issue.

Literature	Proposed System	
Rule	**Learned Rule**	**Prob.**
[r] -> ø / _ [+cons] [r] -> ø / _ [#] (R Dropping)	[r] -> ø / [+low, +long] _ [-vowel]	0.926
	[r] -> ø / [-low] _ [-syl]	0.02
	[er] -> ø	0.006
??	[ah] -> ø / [-glide, -voiced, +fric] _ [-syl, +nasal, +cent]	0.772
??	[ah] -> ø / [+cent, -liq, -stop] _ [+syl]	0.623

Figure 8-2: *RP deletion rule comparison. Learned rules from System S_3 (Triphone PPM; sophisticated tying.)*

Literature	Proposed System	
Rule	**Learned Rule**	**Prob.**
[th] -> [d] /_ [n]	$[th]_{ins}$ -> [d]	0.079
[eh] -> [ih] [eh] -> [ih] / _ [+nasal]	[eh] -> [ih] / _ [+nasal]	0.624
[eh] -> [ey] /_[l,r]	$[eh]_{ins}$ -> [d]	0.078
[ao] -> [aa]	[ao] -> [aa] / [+mid] _ [+stop]	0.254
	[ao] -> [aa] / [+mid] _ [-stop]	0.196
	[ao] -> [ah] / _ [-mid, +voiced]	0.247
	$[ao]_{ins}$ -> [l] _ [-syl, -liq, -nasal, +voiced]	0.402
[ay] -> [ae]	[ay] -> [ae] / [-liq, +sil] _	0.124
[ay] -> [ae] / _ [+cons]	[ay] -> [ae] / [-liq, -sil, +nasal] _	0.112
[ay] -> [aa] / _ [+cons]	[ay] -> [ae] / [-liq, -sil, -nasal, -front] _	0.209

Figure 8-3: *AAVE substitution rule comparison. Learned rules from System S_2 (Triphone PPM; standard tying; surface phones obtained through phone recognition.*

8.2.2 Redundant phonetic context descriptions

At first sight, automatically learned rules might seem very different from rules from the linguistic literature. However, this is not necessarily always true. For example, the first learned rule corresponding to R-dropping in Figure 8-1 has a lot of redundancies in its phonetic context. It can more tersely be expressed as $[er]_{ins} \rightarrow$ [ah] / [+vowel] _ [+affric][4].

To illustrate that the R-dropping phenomenon is occurring in this phonetic context, we show the speech spectrogram of speaker c21 producing the utterance c21c0224 from WSJ-CAM0 in Figure 8-4. The yellow highlighted region illustrates where the reference phones and surfaces differ. The reference phone [er] becomes non-rhotic, [ah]. The non-rhoticity of [er] is illustrated by the rising F3 near 1.25 second, since rhoticity causes a low F3 near 2K Hz.

8.2.3 Triphone APM Pinpoints Regions with Potential Acoustic Differences

For the APM systems, we list triphone examples that have high log likelihood ratio ranking. In this case, these triphones indicate regions of interest, since the acoustic characteristics in these regions are different across dialects. These differences might not be large enough to warrant a phonetic transformation, but further examination might reveal certain acoustic characteristics that are dialect-specific.

Figure 8-6 shows an example of a top ranking triphone [uw-l] in AAVE. In the surface phone sequence, and the speech spectrogram, [l] is deleted in the word *cool*. This corresponds with the l vocalization rule in AAVE descriptions. Figure 8-8 shows multiple examples of triphones of [ay]. The light blue regions indicate [ay] triphones that correspond to top ranking triphones in APM. The orange regions indicate [ay] triphones that do not correspond with top ranking triphones in APM. We see that APM is able to predict when [ay] deglides and becomes like a monophone, and when it does not deglide. More examples of top-ranking triphones corresponding to the

[4][+fric] was defined as fricative or affricate consonants.

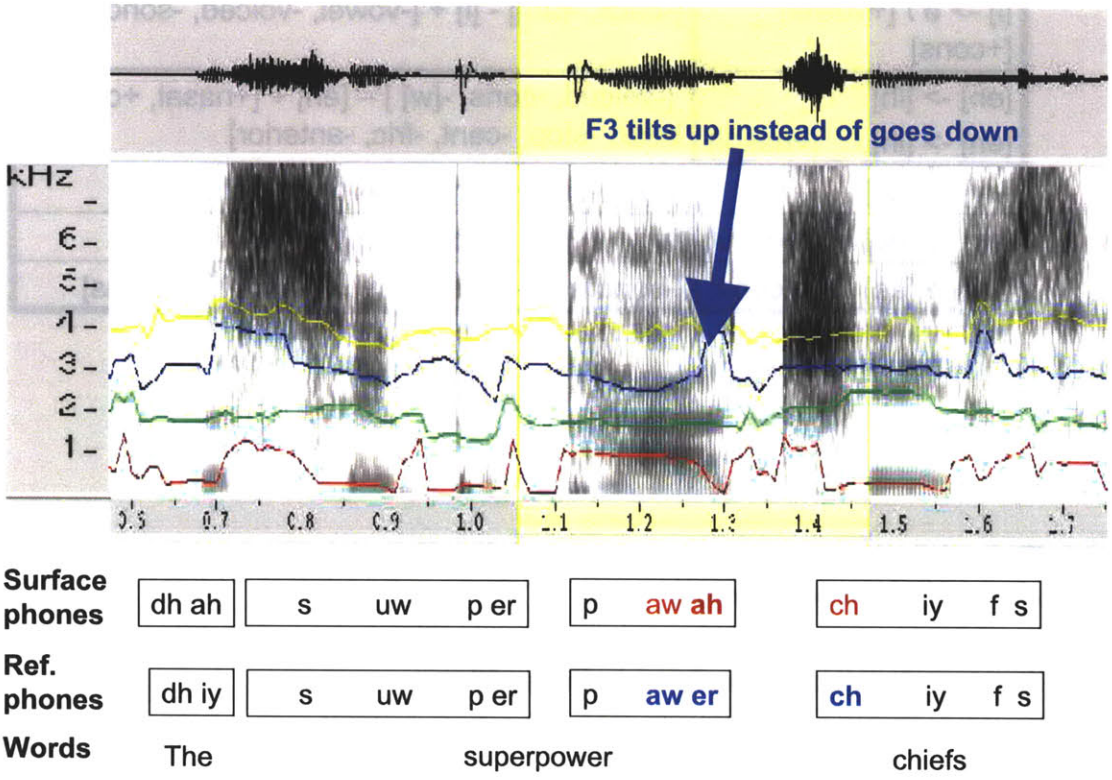

Figure 8-4: *Example of learned rule* $[er]_{ins} \rightarrow$ *[ah] / [+vowel] _ [+affric]. Speech spectrogram of a British speaker saying the utterance, "The superpower chiefs". The yellow highlighted region illustrates where the reference phones and surfaces differ. The reference phone [er] becomes non-rhotic, [ah]. The non-rhoticity of [er] is illustrated by the rising F3 near 1.25 second, since rhoticity causes a low F3 near 2K Hz.*

Literature	System S_3 (Triphone APM – standard tying)
Rule	**Dialect-Specifc Triphone Examples**
[l] -> ø / [+vowel] _ [+cons]	[-back, -[ah]] - [l] + [-vowel, -voiced, -sonorant]
[eh] -> [ih] [eh] -> [ih] / _ [+nasal]	[-voiced, -cons, -[w]] – [eh] + [+nasal, +cons, -back, -stop, -cent, -fric, -anterior]
[ay] -> [ae]	[- [r]] - [ay] + [+glide, +voiced, -stop, -fric]
[ay] -> [aa] / _ [+cons]	[ay] + [+[w], -back, -stop]
[ay] -> [ae] / _ [+cons]	[-sonorant, -cons] – [ay] + [-fric, -hh, +cons]

Figure 8-5: *AAVE rule comparison. Examples of learned rules from System A_2 (Triphone APM; standard tying.)*

literature descriptions are shown in Figure ??.

8.2.4 Sophisticated Tying for Deletion Rules

In Figure 8-2 we see that the R-dropping rules are more concisely expressed, the closest to the linguistic equivalent is [r] $\rightarrow \emptyset /$ [+low, +long] _ [-vowel], with probability of 0.926 of deleting /r/ in such a context. Phones that belong to [+low, +long] are [aa], [ao]. [aw], and [-vowel] is almost the same as [+cons]. When /r/ is preceded by [+low, +long] vowels, it usually forms a syllable: [aa r]. [ao r], [aw r], which are often word boundaries (denoted as # in the linguistic rule.)

We see that compared to the R-dropping rules learned in System S_2 (standard tying), the R-dropping rules learned from System S_3 is more general and doesn't just cover special cases where [r] is followed by an affricate (see Figure 8-4 and Figure 8-9.)

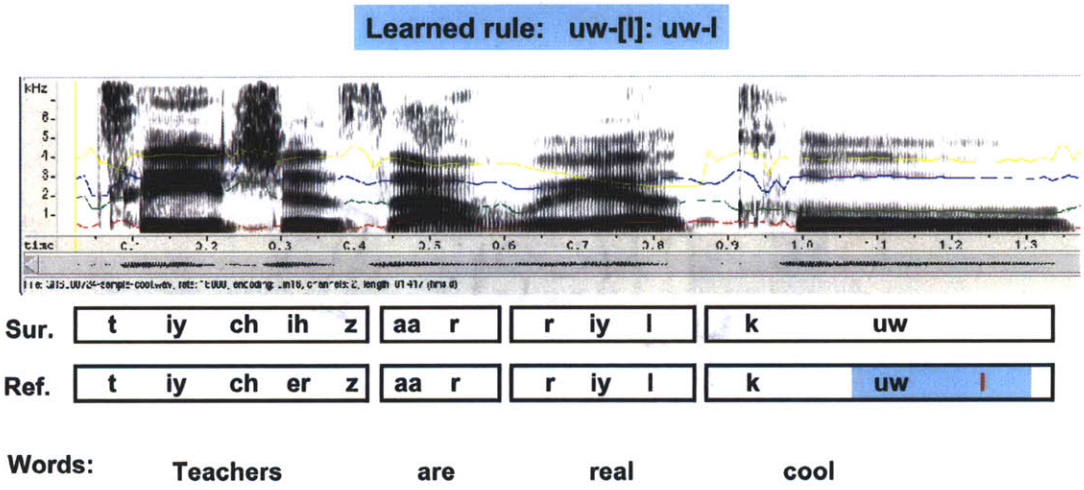

Figure 8-6: *An example of a top scoring triphone of APM corresponding to the /l/ vocalization rule in AAVE.*

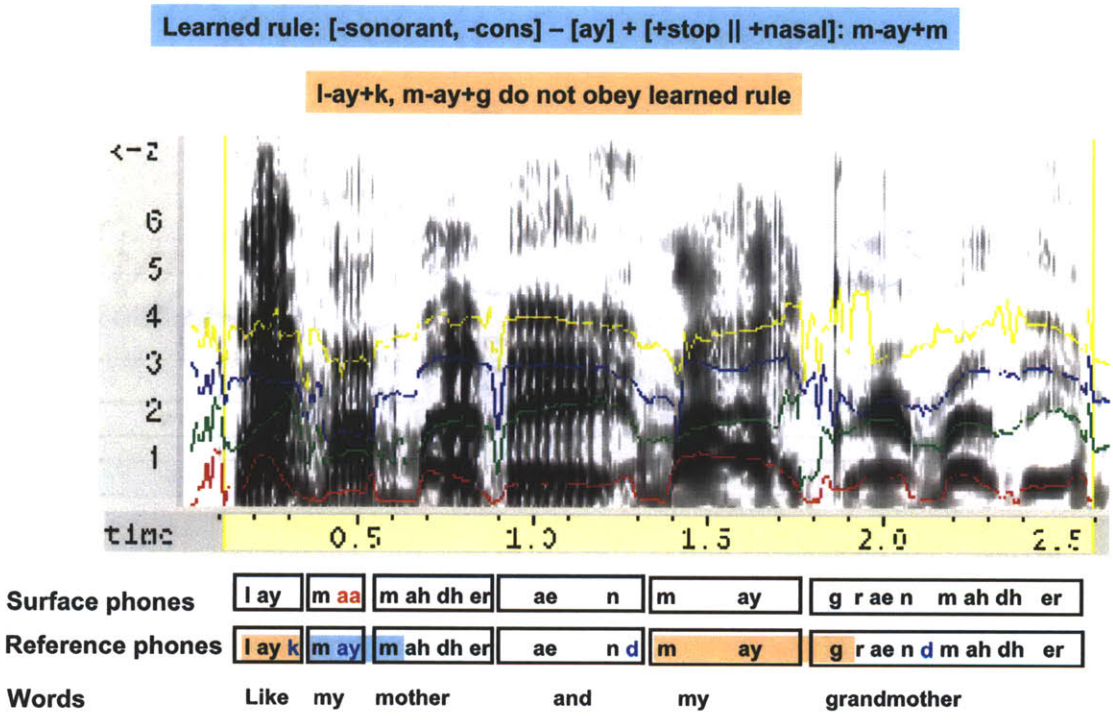

Figure 8-7: *Example of a top scoring triphone of APM corresponding to the /ay/ monophthongization rule in AAVE.*

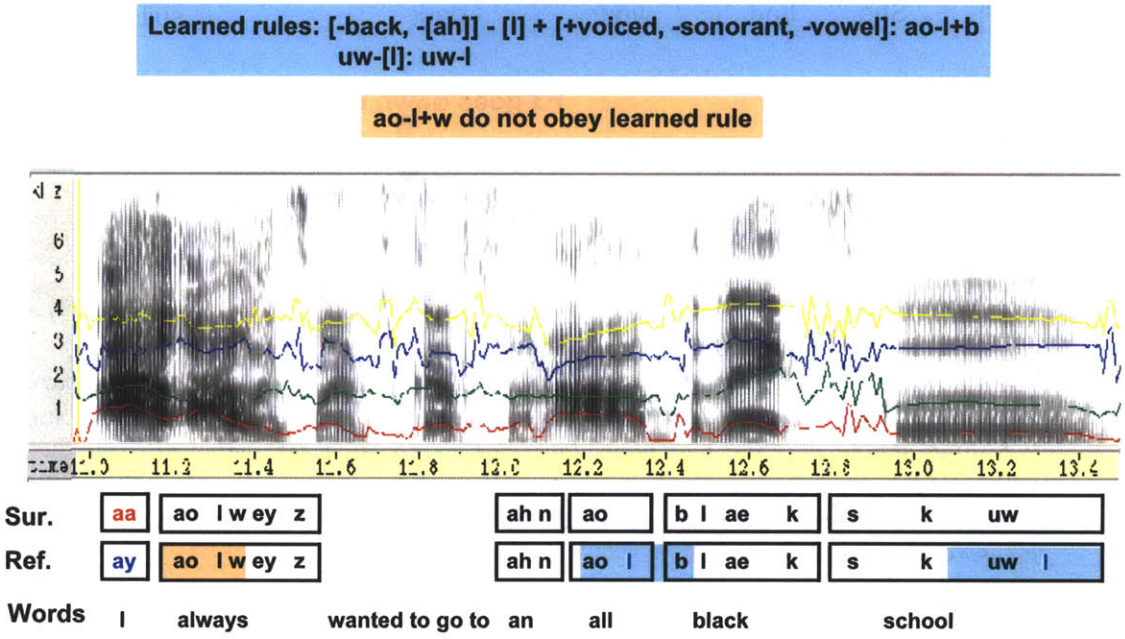

Figure 8-8: *Example of a top scoring triphone of APM corresponding to the /l/ vocalization and /ay/ monophthongization rules in AAVE.*

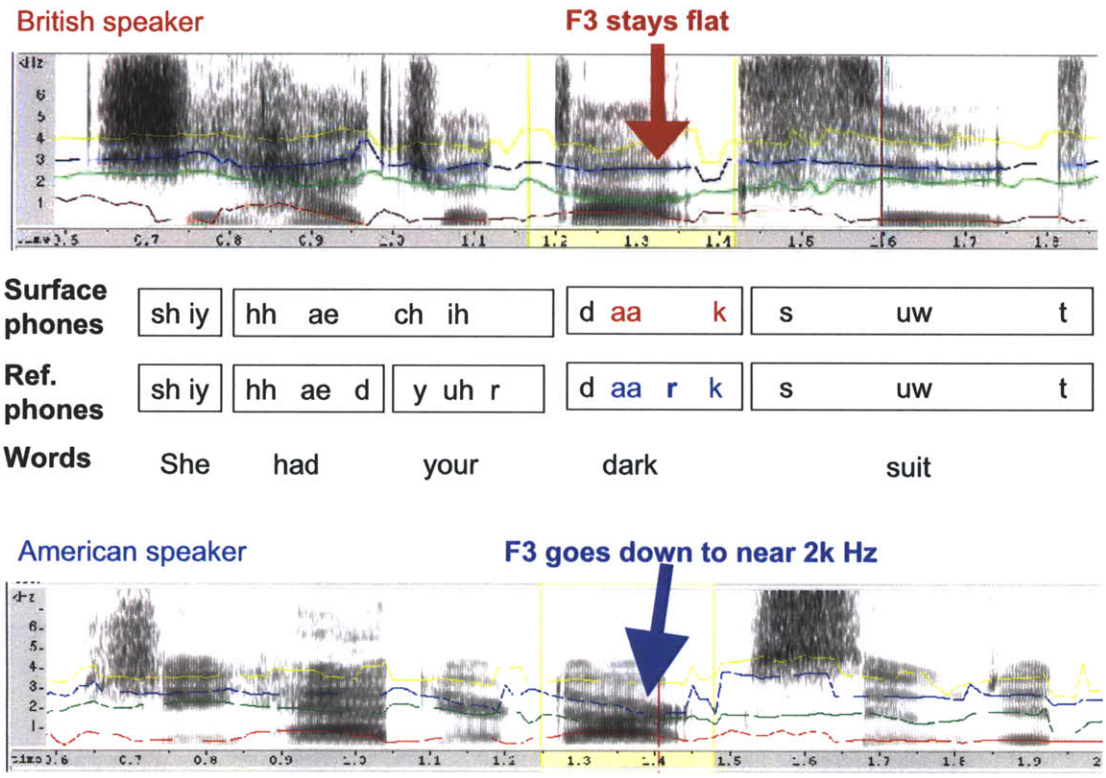

Figure 8-9: *Example of learned rule [r] → ∕ [+low, +long] _ [-vowel]. Comparison between of British speaker (top panel) and an American speaker (lower panel) saying the same sentence, "She had your dark suit in greasy wash water all year". The yellow highlighted region illustrates where the reference phones and surfaces differ. The British speaker's F3 stays flat in the vowel of dark, while the American Speaker's F3 goes down near 2k Hz.*

This is another indication that sophisticated tying is more suitable in modeling deletion phonetic transformations, in addition to the DID and pronunciation generation results in the previous chapters.

To further illustrate the rhoticity difference between RP and GAE, we contrast the speech spectrograms of a British speaker and an American speaker (from TIMIT) saying the same sentence, "She had your dark suit in greasy wash water all year" in Figure 8-9. The yellow highlighted region illustrates where the reference phones and surfaces differ. The British speaker's F3 stays flat in the vowel of *dark*, while the American Speaker's F3 goes down near 2k Hz.

Literature	Proposed System	
Rule	**Learned Rule**	**Prob.**
[l] -> ø / [+vowel] _ [+cons]	[l] -> ø / [+short] _ [-vowel, -dipth, -glide, -liq, -syl, -voiced]	0.224
	[l] -> ø / [-short] _ [-short, -dipth, +front]	0.132
[r] -> ø / [+vowel] _ [r] -> ø / [+vowel] _ [+vowel] [r] -> ø / _ [+const,+stop, +front]	[r] -> ø / [+cent] _ [+vowel, -low, -front, -dipth]	0.146
	[r] -> ø / [-round] _ [-vowel, -dipth, +glide]	0.11
	[r] -> ø / [-roundl] _ [-vowel,-dipth, -glide, -liq, -syl, +voiced]	0.261
	[r] -> ø / [-roundl] _ [-vowel,-dipth, -glide, +liq]	0.113
	[r] -> ø / [-roundl] _ [-vowel,-dipth, -glide, -liq, -syl, +voiced]	0.114
	[r] -> ø / [+roundl] _ [-vowel,-dipth, -glide, -liq]	0.114
[+voiced +cons] -> ø / [+nasal +cons] _	[jh] -> ø / [+nasal] _ [-vowel]	0.117
	[z] -> ø / [+nasal, -short] _ [-vowel]	0.069
[-voiced +cons] -> ø / [+nasal +cons] _	[th] -> ø / _ [-vowel, +voiced, -dipth, +glide]	0.051
	[s] -> ø / [-stop] _ [-stop, -cent, -front, -vowel]	0.091

Figure 8-10: *AAVE deletion rule comparison. Learned rules are from System S_3 (Triphone PPM; sophisticated tying; surface phones obtained through phone recognition.*

Though Sophisticated Tying performed poorly in the information retrieval experiment (StoryCorps) in Section 7.1, we see that in Figure Figure 8-10 and Figure 8-12, most of the ground-truth deletion rules have corresponding automatically learned rules. These results imply that the information retrieval result could be overly pessimistic for sophisticated tying.

8.2.5 False Alarms: Potential New Rules

In Figure 8-2, we see the last two learned rules correspond to question marks in the literature column, indicating that we did not find corresponding rules in the literature for deleting schwa [ah]. One possible explanation is that the schwas in *fun*, *nation*, *sun* are heavily nasalized, causing the schwa vowel to appear as if it has disappeared/been

deleted. It might be a common reduction in spoken English, which may or may not be dialect-specific (we do not have suitable ground-truth surface phones for American English to determine if this phonetic reduction also occurs in GAE and how often it occurs.) Further investigation is required to examine if these rules with high rankings are indeed dialect-specific rules or mere experimental artifacts.

8.3 Future Model Refinement

An interesting observation is that while the insertion states were originally meant to model insertions, it appears that it is sometimes used to learn uncommon substation rules. For example, in Figure 8-3 we see that the insertion state of [th], denoted as $[th]_{ins}$, models the transformation [th] $\rightarrow$ [d]. Another example is [eh] in Figure 8-3. The transformation of [eh] $\rightarrow$ [ih] is more common than [eh] $\rightarrow$ [ey]. The former is modeled by a normal [eh] state, while the latter is modeled by the insertion state of [eh].

In our current model setup, it is more challenging to interpret insertion rules, given that it is possible for substation rules modeled by insertion states as previously mentioned. In addition, for the dialects we are investigating in this thesis, there appears to be fewer insertion rules documented in the literature, making it more challenging to interpret them. In future work, we plan to investigate how to model and interpret insertion rules with more precision and detail.

8.4 Summary

We compared the automatically learned rules for RP and Arabic Dialects with those in the literature and discussed the properties of learned rules. Our results suggest that our system could be used to further analyze phonetic rules (in terms of their occurrence frequency and context conditions) and acoustic characteristics across dialects.

Literature	Proposed System	
Rule	Learned Rule	Prob.
[eh] -> [ih] [eh] -> [ih] / _ [+nasal]	[eh] -> [ih]	0.23
[dh] -> [d] / [+vowel] _ [vowel]	$[dh]_{ins}$ -> [th]	0.064
[uh] -> [uw] /_[l]	[uh] -> [uw]	0.004
[ao] -> [aa]	[ao] -> [aa]	0.009
[aw] -> [uw]	$[aw]_{ins}$ -> [uw]	0.012

Figure 8-11: *AAVE substitution rule comparison. Learned rules are from System F_2 (Triphone PPM; standard tying; surface phones obtained through forced-alignment.)*

Literature	Proposed System	
Rule	Learned Rule	Prob.
[r] -> ø / [+vowel] _ [r] -> ø / [+vowel] _ [+vowel] [r] -> ø / _ [+cons, +stop, +front]	[r] -> ø / [+vowel] _ [-cent, -vowel]	0.09
	[r] -> ø / [+vowel] _ [+cent, -vowel]	0.097
	[r] -> ø / [+vowel] _ [+cent, +vowel]	0.101
[-voiced +cons] -> ø / [+nasal +cons] _	[t] -> ø / _ [+voiced]	0.112
	[hh] -> ø / _ [-high, -short, +vowel, +dipth]	0.028
[+voiced +cons] -> ø / [+nasal +cons] _	[y] -> ø / _ [-round, +vowel]	0.163
[l] -> ø / [+vowel] _ [+cons]	[l] -> ø / [+vowel, +short] _ [-vowel, -dipth, +glide]	0.034
	[l] -> ø / [+vowel, -short] _ [-vowel]	0.069
	[l] -> ø / [+vowel, +short] _ [-vowel, -dipth, -glide]	0.094
	[l] -> ø / [-vowel] _ [+vowel, -dipth]	0.033

Figure 8-12: *AAVE deletion rule comparison. Learned rules from System F_3 (Triphone PPM; sophisticated tying; surface phones obtained through forced-alignment.)*

Literature		Proposed System		
Description	**Dialect**	**Learned Rule**	**Prob**	**Dialect**
Interdental fricatives become stops	EG PS SY	[th] -> [t] / _ [+long]	0.79 0.70 0.87	EG PS SY
		[dh] -> [d] / [-back] _	0.57	EG
		[th] -> [t] / [-short] _ [-long]	0.62	EG
vowel [o] exists (usually only [a], [i], [u] exist)	IQ	[o:] -> [u:] / _ [+fricative, -voiced]	0.68	EG
		[o:] -> [a] / _ [+fricative, +voiced]	0.51	EG

Figure 8-13: *Examples of learned rules from System F_2 (Triphone PPM; standard tying) trained on 5-Dialect Arabic Corpus.*

Literature		Proposed System		
Description	**Dialect**	**Learned Rule**	**Prob**	**Dialect**
Palatal voiced affricate becomes palatal approximant	AE	[dZ] -> [j] /_ [+syl]	0.32	AE
Palatal voiced affricate becomes voiced stop	EG	[dZ] -> [d]	0.25	EG
vowel [o] exists	IQ	[o:] -> [a]	0.28; 0.27; 0.32; 0.27	AE, EG, PS, SY
Interdental fricatives become stops	EG PS SY	[th] -> [t]/_ [-short]	0.60	EG
		[th] -> [t] / [-low] _ [+short]	0.59	
		[th] -> [t]	0.42; 0.43	PS, SY
		[dh] -> [d]	0.24; 0.29	PS, SY
		[dh] -> [d] / [-front] _	0.33	EG

Figure 8-14: *Examples of learned rules from System S_2 (Triphone PPM; standard tying) trained on 5-Dialect Arabic Corpus.*

Chapter 9

Conclusion

We conclude the work of this thesis by summarizing our contributions and discussing research directions for future work and potential applications.

9.1 Contributions

The contributions of this thesis are:

1. Proposed ***automatic*** yet ***informative*** approach in analyzing speech variability. This interdisciplinary research direction in dialect studies, combining the strengths of speech science and engineering, is termed *Informative Dialect Recognition.*

2. Proposed mathematical framework to characterize phonetic and acoustic transformations across dialects in a rich and explicit manner.

3. Empirical results in rule retrieval, pronunciation generation, and dialect recognition indicate that proposed systems exploit underlying rules across dialects.

4. Proposed models complement existing dialect recognition systems, suggesting the proposed models exploit information not used in traditional dialect recognition systems.

5. Proposed system postulates rules from large corpora to a phonetician to discover, refine, and quantify rules.

6. Surveyed corpora resources for dialect research, and address challenges in informative dialect recognition.

9.2 Discussion and Future Work

9.2.1 Characterizing Rules

Learning Right-Context Driven Rules More Comprehensively

One limitation of the current implementation setup of the proposed pronunciation model is that *right-context driven* substitution /insertion rules might not be learned *comprehensively*. For illustration purposes, assume a dialect difference between American and British English is vowel nasalization: all vowels followed by a nasal consonant will be fully nasalized in American English; i.e., [+vowel] $\rightarrow$ [+vowel, +nasal] / _ [+nasal]. In the current setup, these right-context rules might be learned in a fragmented way separately for each vowel, as illustrated in Figure 9-1. (The rule is right-context driven because the driving force of the phonetic change is caused by the nasal consonant on the right context of the vowel.) These vowel nasalization rules are still being learned, but perhaps not in a manner that fully exploits the generality of the rule.

This limitation can be gracefully handled by reversing all directions of the state transition arcs as shown in Figure 9-2. Then both left and right-context driven rules can be learned appropriately. It is expected that the fusion of these two systems would yield more gains in dialect recognition experiments.

Rules Beyond Triphone Contexts

Some phonetic rules are influenced by phones beyond their directly neighboring phones. For example, words such as *dance*, *chance*, *can't*, and *stamp* also transform the vowel [ae] to [aa] in British English. In these phonetic transformations, [ae]

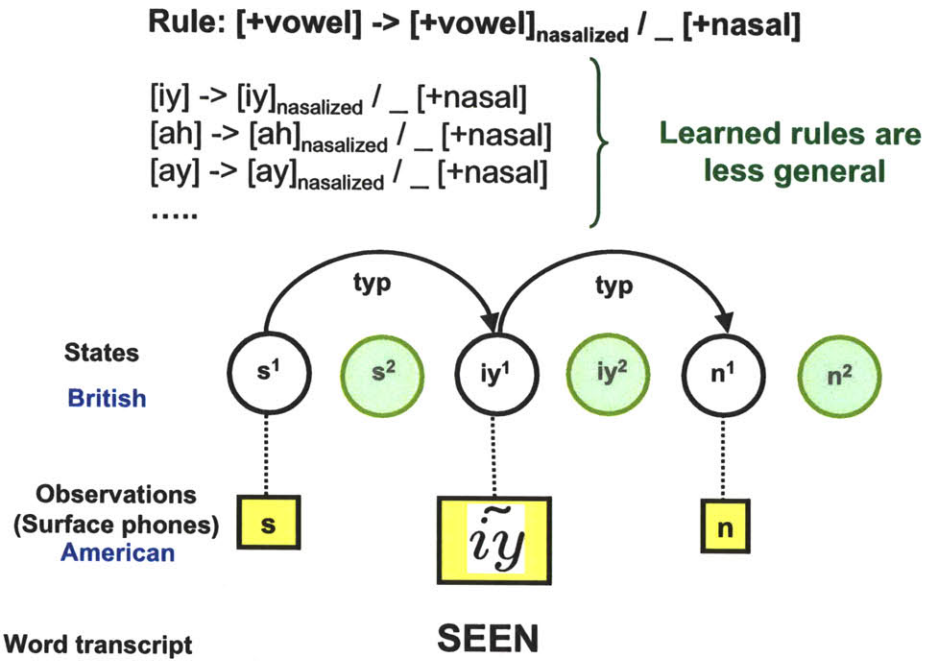

Figure 9-1: Example of rule learning limitation in current system setup.

is followed by a nasal that is followed by a consonant. These rules could be modeled by expanding the current implementation setup from triphones to quinphones (or phones conditioned on even more neighboring phones). Since quinphones face more data sparsity problems than triphones, more smoothing procedures might be required to estimate model parameters better.

Articulating Specific Discriminating Acoustic Properties

Characterizing dialect differences (or any kind of speech variability) is a challenging task, because dialect differences are often not discrete and binary. There are different degrees of acoustic implementation, which might not fit into any phonetic category. APM (Acoustic-based Pronunciation Model) is able to handle these fine-grained changes. It pinpoints locations that are acoustically different across dialects. In the current system, human examination is required to further understand which acoustic aspects are different. It could be more efficient for the speech scientist if

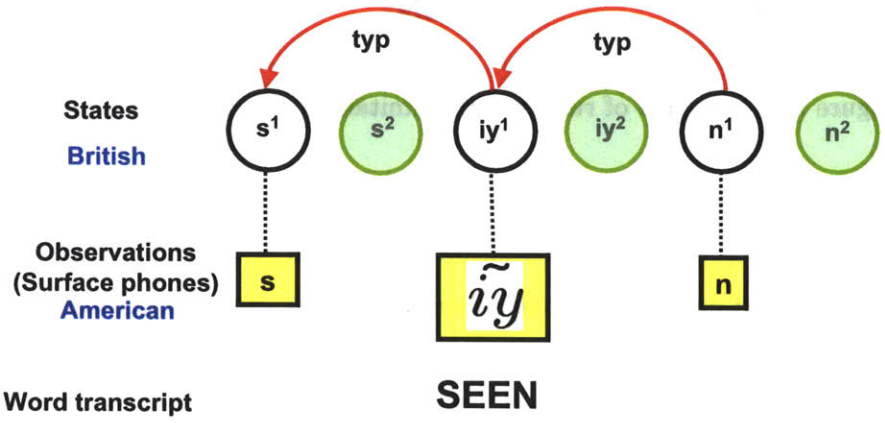

Figure 9-2: Limitation shown in Figure 9-1 can be elegantly dealt with simply by reversing the direction of all state transition arcs.

APM (Acoustic-based Pronunciation Model) could further pinpoint which acoustic properties might be different (e.g., voicing, formant transitions, nasalization).

Discriminative Clustering

The splitting criteria for state and arc clustering could be changed to log likelihood ratio of a target dialect and a non-target dialect. This discriminative clustering approach could potentially learn more rules and improve DID performance.

Sophisticated Tying in Acoustic-based Pronunciation Model (APM)

Inferring from the StoryCorps PPM results, sophisticated tying could potentially further improve the performance of triphone APM. This hypothesis could be verified empirically.

Integrating Higher Linguistic Component

Other linguistic components that contribute to pronunciation differences (such as prosody, vocabulary, syntax) could also be considered to make the pronunciation model more comprehensive.

Some of these higher level characterization could be easily integrated at the decision tree level by adding yes-no questions such as "is the phone at a syllable/word/utterance final position".

Characterizing Dialects using Distinctive Features

Instead of using phonetic transformations (substitution, insertion, deletion) we can characterize dialect differences at a finer level, such as distinctive features [61]. The main challenge is existing resources would be more even more limited, since it requires corpora of different dialects with detailed, fine-grained/feature manual transcriptions.

9.2.2 Redefining Dialects through Unsupervised Clustering

In this work, we used region of residence or ethnicity as a proxy for dialect. As discussed before, many factors influence and correlate with dialect, so it is challenging to to determine the single *ground-truth* dialect of a speaker.

Dialect groups are often defined by convenient cultural indicators rather than on the basis of similarity [50]. One of the challenges of dialect research stems from this ill-defined nature of dialects. However, if we categorize speakers based on their pronunciation characteristics, we could examine dialect characteristics from a different perspective. For example, we can apply unsupervised clustering to find speakers with similar pronunciation characteristics. Each speaker could belong to different clustered groups at the same time. For example, a speaker might be 90% rhotic and 10% non-rhtoic. These probabilistic and finer-grained results could better characterize dialects. These unsupervised clustered groups can be compared with traditional dialect labels, and further our understanding and analysis of dialects.

9.2.3 Further Verification on Model Robustness

We ran experiments on a diverse set of corpora to test out our proposed algorithms, obtaining empirical results from different languages (Arabic and English) and different speaking styles (conversation vs. reading). It would be beneficial to empirically verify that our proposed models work on other dialects, languages, and speaking styles.

9.3 Potential Applications

There are numerous potential applications for this thesis work (see Figure 9-3). We describe some of them below.

9.3.1 Speech Technology: Dialect and Speaker Recognition

Our dialect recognition results suggest that the proposed systems exploit information that existing systems do not. Speaker differences could also potentially be char-

acterized with our proposed systems, and complement existing speaker recognition systems.

Unsupervised Scoring in Phonetic-based Pronunciation Model (PPM)

Unsupervised scoring (i.e., scoring without transcriptions) can be done under the current PPM implementation, but the DID accuracy is low when scoring is unsupervised without reference phones. The DID accuracy is at least partially due to the constraints on insertion and deletion state transitions, which could be relaxed to improve performance.

9.3.2 Speech Analysis: Verify, Refine, and Propose Rules

As mentioned before, our model can serve as a first pass for linguists to analyze new rules or re-examine existing rules in a more efficient manner than traditional approaches. Measures to verify validity of rules include (1) perceptual experiments done by native bilingual subjects, and (2) further acoustic analysis by speech scientists.

9.3.3 Healthcare: Characterizing Speech and Voice Disorders

Instead of characterizing dialects, our proposed framework can also be applied to speech disorders. For example, some phonological disordered children do not pronounce word-final consonants, so *cat* sounds like /k ae/ instead of /k ae t/ [5].

9.3.4 Forensic Phonetics

The proposed system can also help forensic phoneticians tease out which speech characteristics are dialect-specific and which are speaker-specific [82], which is often not well-documented in the literature.

9.3.5 Education: Language Learning or Accent Training Software

Other potential applications include accent training/language learning education, where the automated system can provide explicit feedback on which pronunciation patterns that require the most improvement.

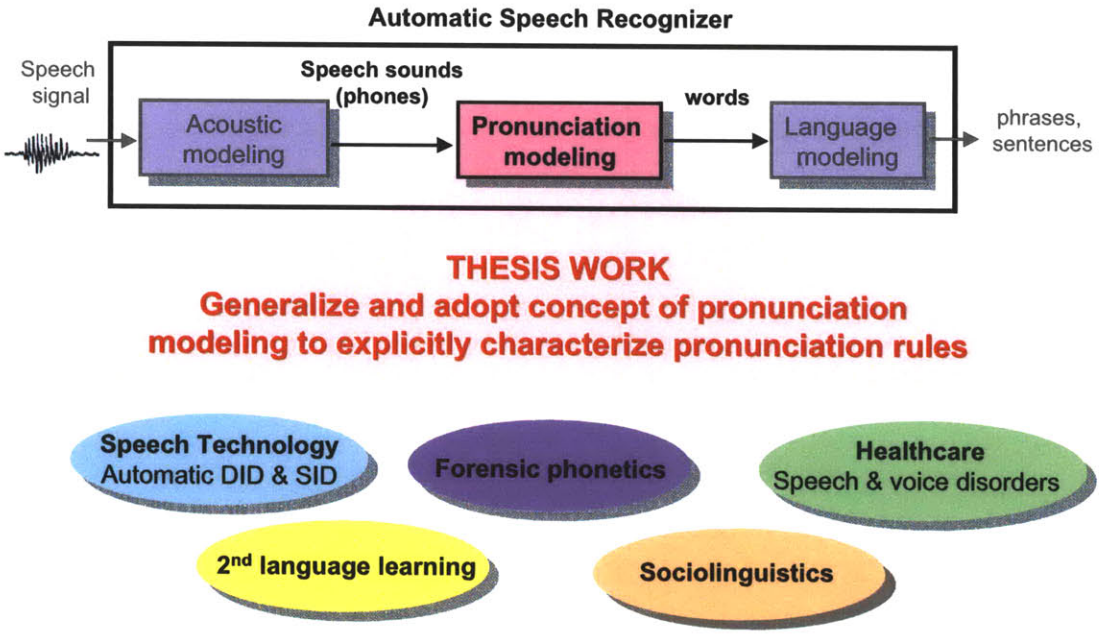

Figure 9-3: Potential applications of this thesis.

Appendix A

The Phonetic Alphabet

A.1 English

Figure A-1 lists the English phonetic Alphabet used in this work (the datasets WSJ-CAM0 and StoryCorps). The bold font in the word examples highlight which part of the word is represented by the phone symbol acoustically. A phone symbol can be mapped to more than one type of sound.

The third column shows the features that belong to the phone; *affric* is short for affricate, *cent* is short for center, *cons* is short for consonant, *dipth* is short for diphthong, *fric* is short for fricative, *syl* is short for syllable. Since affricates do not occur often and have fricative properties as well, affricatives were also lumped into the fricative feature for practical reasons in the experiments; i.e., fricative includes affricate. The feature [syl] means that the phone itself could be a syllable. These features are used in decision tree clustering.

A.2 Arabic

Figure A-2 lists the Arabic phonetic alphabet used in this work (the dataset 5-Dialect Arabic Corpus). The second column shows the features that belong to the phone; *affric* is short for affricate, *cent* is short for center, *cons* is short for consonant, *retro* is short for retroflex *fric* is short for fricative, *syl* is short for syllable. Unlike English,

Phone symbol	Word Examples	Features
[aa]	**b**o**b**	[vowel], [cent], [low], [long]
[ae]	**b**a**t**	[vowel] , [cent], [low], [short]
[ah]	b**u**t, **a**bout	[vowel] , [cent], [mid], [short]
[ao]	**b**ou**ght**	[vowel] , [cent], [round], [low], [long]
[aw]	b**ou**t	[vowel] , [cent], [low], [dipth], [long]
[ay]	**b**i**te**	[vowel], [cent], [low], [dipth], [long]
[b]	**b**ee	[cons], [voiced], [stop], [front],
[ch]	**ch**oke	[cons], [fric], [affric], [back],
[d]	**d**og	[cons], [voiced], [stop], [cent]
[dh]	**th**at	[cons], [fric], [voiced], [cent]
[eh]	b**e**t	[vowel], [cent], [mid], [short]
[er]	b**ir**d, but**t**er	[vowel], [cent], [mid], [short], [syl]
[ey]	b**ai**t	[vowel], [cent], [mid], [dipth], [long]
[f]	**f**ox	[cons], [fric], [front]
[g]	**g**od	[cons], [voiced], [stop], [back]
[hh]	**h**at, **h**ead	[cons], [back], [glide]
[ih]	b**i**t	[vowel], [front], [high]
[iy]	b**ea**t	[vowel], [front], [high]
[jh]	**j**oke	[vowel], [fric], [affric], [voiced], [back]
[k]	**k**ey	[cons], [stop], [back]
[l]	**L**ake, bo**tt**le	[cons], [cent], [liq], [syl]
[m]	**m**oon	[cons], [nasal], [voiced], [front]
[n]	**n**oon	[cons], [nasal], [voiced], [cent]
[ng]	si**ng**, washi**ng**	[cons], [nasal], [voiced], [back]
[ow]	**b**oa**t**	[vowel], [back], [round], [mid], [dipth], [long]
[oy]	**b**oy	[vowel], [back], [round], [high], [dipth], [long]
[p]	**p**lay	[cons], [stop], [front]
[r]	**r**ock	[cons], [voiced], [cent], [liq]
[s]	**s**ea	[cons], [fric], [cent]
[sh]	**sh**e	[cons], [fric], [cent]
[t]	**t**oy, bu**tt**er	[cons], [stop], [cent]
[th]	**th**eeth	[cons], [fric], [cent]
[uh]	b**oo**k	[vowel], [back], [round], [mid], [short]
[uw]	b**oo**ts	[vowel], [back], [round], [high], [long]
[v]	**v**est	[cons], [voiced], [front]
[w]	**w**ash	[cons], [voiced], [front], [round], [glide]
[y]	**y**acht	[cons], [voiced], [back], [glide]
[z]	**z**oo	[cons], [voiced], [cent]
[zh]	a**z**ure	[cons], [fric], [affric], [voiced], [back]

Figure A-1: *English phone symbols used in this thesis. The third column shows the features that belong to the phone; affric is short for affricate, cent is short for center, cons is short for consonant, dipth is short for diphthong, fric is short for fricative, syl is short for syllable. Since affricates do not occur often and have fricative properties as well, affricatives were also lumped into the fricative feature for practical reasons in the experiments; i.e., fricative includes affricate. The feature [syl] means that the phone itself could be a syllable.*

there are many different variants of affricates and fricatives, therefore they represent distinct features in the Arabic phonetic alphabet.

Phone symbol	Features
[a]	[vowel], [voiced], [front], [low], [short], [syl]
[a^long]	[vowel] , [voiced], [front], [low], [long], [syl]
[b]	[cons] , [voiced], [stop], [front], [lab]
[d]	[cons] , [voiced], [stop], [front]
[dZ]	[cons] , [affric], [front]
[d^retro]	[cons], [voiced], [stop], [front], [retro]
[D]	[cons], [fric], [voiced], [front]
[D^retro]	[cons], [fric], [voiced], [front], [retro]
[e^long]	[vowel]. [voiced], [front], [mid], [long], [syl]
[f]	[cons], [fric], [front], [lab]
[g]	[cons], [voiced], [stop], [back]
[gs]	[cons], [stop], [back]
[gs^retro]	[cons], [stop], [back], [retro]
[G]	[cons], [fric], [voiced], [back]
[h]	[cons], [fric], [back]
[i]	[vowel], [voiced], [front], [high], [short], [syl]
[i^long]	[vowel], [voiced], [front], [high], [long], [syl]
[j]	[cons], [voiced], [cent], [glide], [syl]
[k]	[cons], [stop], [back],
[l]	[cons], [voiced], [cent], [liq], [syl]
[l^retro]	[cons], [voiced], [cent], [liq], [syl], [retro]
[m]	[cons], [nasal], [voiced], [front], [syl], [lab]
[n]	[cons], [nasal], [voiced], [cent], [syl]
[o^long]	[vowel], [voiced], [back], [round], [mid], [long]
[p]	[cons], [stop], [front], [lab]
[q]	[cons], [voiced], [stop], [back]
[r]	[cons], [voiced], [cent], [rhotic], [syl]
[S]	[cons], [fric], [cent]
[s]	[cons], [fric], [front]
[s^retro]	[cons], [fric], [front], [retro]
[t]	[cons], [stop], [front]
[tS]	[cons], [afric], [front]
[t^retro]	[cons], [stop], [front]
[T]	[cons], [fric], [front]
[u]	[vowel], [voiced], [back], [round], [high], [short]
[u^long]	[vowel], [voiced], [back]. [round], [high], [long]
[v]	[cons], [fric], [voiced], [front], [lab]
[w]	[cons], [voiced], [front], [round], [glide], [syl], [lab]
[x]	[cons], [fric], [back]
[X^pala]	[cons], [fric], [cent]
[z]	[cons], [fric], [voiced], [front]
[Z]	[cons], [fric], [voiced], [cent]

Figure A-2: *Arabic phone symbols used in this thesis. The second column shows the features that belong to the phone; affric is short for affricate, cent is short for center, cons is short for consonant, retro is short for retroflex fric is short for fricative, syl is short for syllable. Unlike English, there are many different variants of affricates and fricatives, therefore they represent distinct features in the Arabic phonetic alphabet.*

Appendix B

Channel Issues in WSJ0, WSJ1, and WSJ-CAM0

Many investigated corpora were not suitable, and the brief reasons are listed in Table. 4.1. Below we document the analysis of the channel variations of the WSJ0, WSJ1, WSJ-CAM0 corpora. From our analysis, we determined that these corpora are ineligible for meaningful dialect recognition experiments.

B.1 DID Experiment Setup

We list the data partition of the training and test set in WSJ0 and WSJ-CAM0 in Table 4.2 and Table ??. For American English, WSJ1 was used supplemented in the test set.

Table B.1: WSJ training data

Dialect	Speaker number	Duration
American (WSJ0)	84	16.5 hr
British (WSJ-CAM0)	92	15.3 hr

Table B.2: WSJ test data

Dialect	Speaker number	Duration	Number of 30sec trial
American (WSJ0, WSJ1)	53	4.2 hr	507
British (WSJ-CAM0)	48	4 hr	484

B.2 DID Baseline Experiments

Baseline experiments using SDC-GMM [89], adapted phonetic models [84], and PRLM [100] all reach 0% EER. For the SDC-GMM system, number of mixture components did not influence the EER performance. Mixture components of 2048 or 1, all lead to 0% EER. This superior performance might be due to (1) British and American dialects are very different, or (2) there are other factors other than dialect that strongly correlates with the dialect labels. Since gender was balanced across all sets and both dialects, it is unlikely these superior performances are from gender identification. There is also no speaker overlap between the datasets, so speaker identification is unlikely either. Since these two corpora are recorded at different locations, it is possible there are channel differences. In the next section, we investigate the hypothesis that channel difference is predominant across the two dialects, making the DID baseline performance 0%.

B.3 Channel Difference Investigation

Good performance of baseline experiments predominantly due to channel:

B.3.1 Long-Term Average Power Spectra

As shown in Figure B-1, long-term average power spectra is different across recording sites. Note that the recording locations of WSJ1 involves 3 sites: MIT (Massachusetts Institute of Technology), SRI International(founded as Stanford Research Institute), and TI (Texas Instruments). These 3 sites already have different long-term average power spectra. These results imply that recording location could lead to noticeable

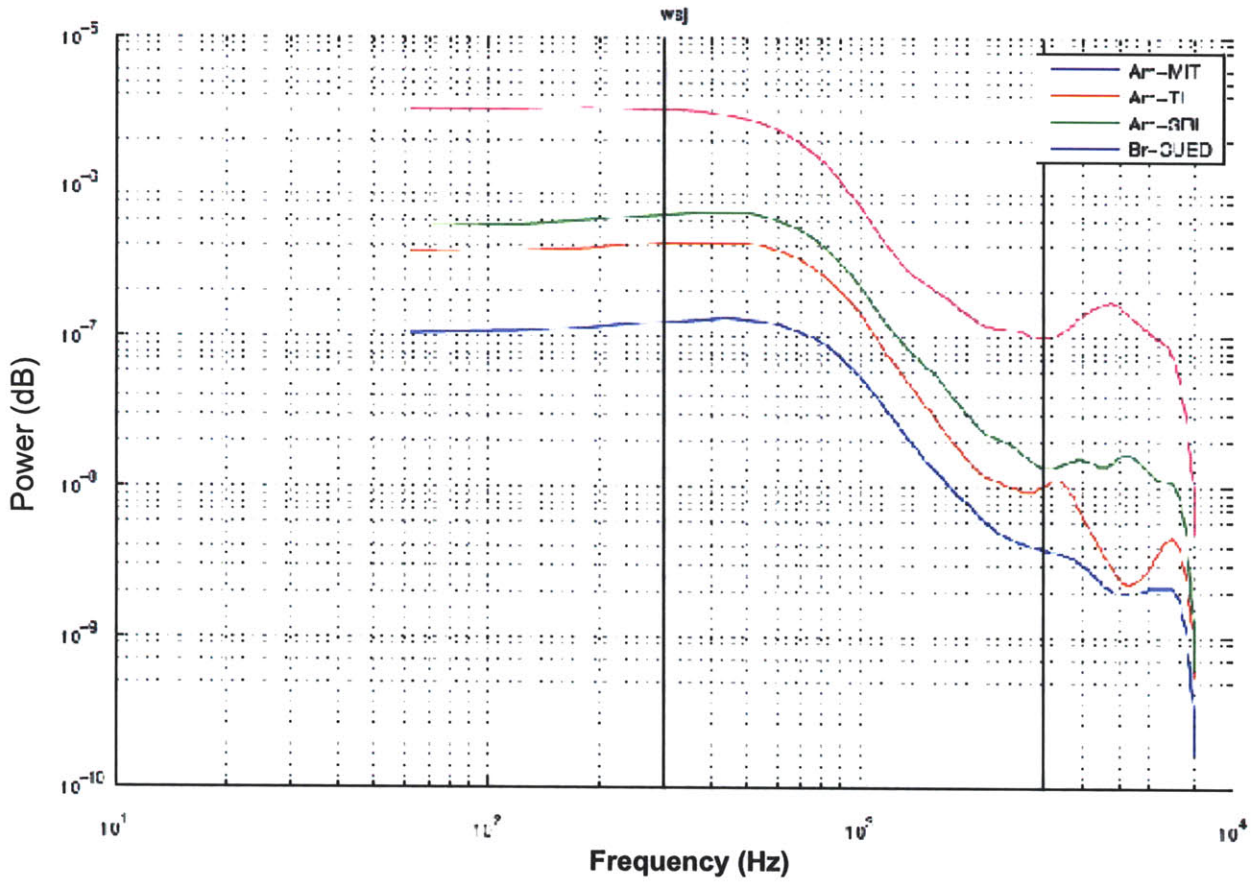

Figure B-1: Long-Term Average Power Spectra of 4 recording sites. American English: MIT (Massachusetts Institute of Technology), SRI (Stanford Research Institute), and TI (Texas Instruments).. British English: CUED (Cambridge University Engineering Department.)

channel differences.

B.3.2 WSJ1 Recording Site Identification

To further investigate whether recording site differences are correlated with acoustic difference. We perform a site identification experiment on WSJ1. WSJ-CAM0 was taken out so that no dialect difference could be the confounding factor.

Our results show that SDC-GMM system (1024 mixture components) with channel compensation (RASTA, feature normalization) is able to achieve decent identification

Table B.3: WSJ1 Recording site detection rate

Recording site	EER (%)
SRI	14.8
MIT	8.2
TI	12.3

of recording sites. Detecting the MIT site is the easiest, reaching 8.2% EER. The most challenging site to detect is SRI, which still obtains EER or 14.8% (see Table B.3.)

It is reasonable to assume that the researchers in WSJ1 tried their best efforts to make recording conditions match across the 3 sites. Even under such condition, site identification error rates are still lower than 15%. It is not unreasonable to assume that site identification error rate between CUED and any of the American sites (MIT, SRI, TI) would be at least 15%. These results suggest that acoustic differences correlating with the British and American dialect labels are probably noticeable.

B.3.3 Monophone APM on Non-Dialect-Specific Phones

In the last section, we used SDC-GMM to model the acoustics of WSJ1. In this experiment, we only model acoustic characteristics of phones that are not known to be dialect-specific across British and American English dialects. We use System A_1 (Monophone APM) to perform this experiment, but only using selective phones. We chose three sets of phones (1)fricatives and silence: [s], [sh], [f], [v], [z], [zh], [sil], (2) [f], [v], and (3) [sil]. Figure B-2 shows the detection error trade-off curves of these three experiments. We see that just by using phone set (1), the EER already is below 10%. Only using [f], [v] (phone set (2)) only makes the EER go up a little beyond 10%. Finally, if only silence is used (phone set (3)), EER is around mid-20%; by only using non-dialect acoustic characteristics, detection error is still decently below chance. This last result strongly suggest that non-dialect acoustic characteristics are correlated with the dialect labels, making the dialect recognition results appear overly optimistic that no meaningful conclusion can be drawn regarding comparing different models.

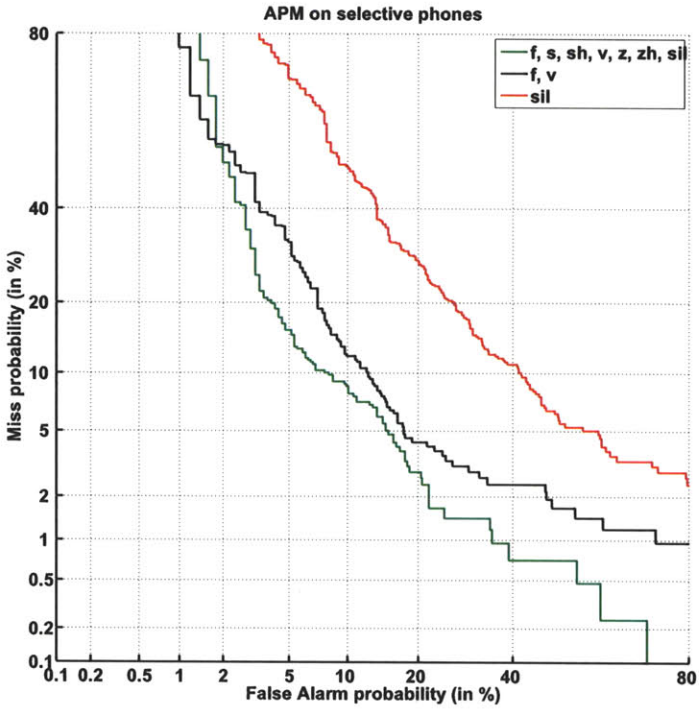

Figure B-2: Monophone APM scoring only selective phones that are not dialect-specific.

B.3.4 Conclusion

Our investigations and analyses suggest that non-dialect acoustic characteristics are correlated with the dialect labels among the WSJ corpora (WSJ0, WSJ1, WSJ-CAM0). Baseline dialect recognition results appear overly optimistic, making it impossible to draw meaningful conclusions from empirical comparisons of the baseline systems and our proposed systems. Therefore, it is unsuitable to use WSJ0 and WSJ-CAM0 to perform dialect recognition experiments.

Bibliography

[1] A. Adami. *Modeling Prosodic Differences for Speaker and Language Recognition.* PhD thesis, OGI School of Science & Engineering at OHSU, 2004.

[2] A. Adami and H. Hermansky. Segmentation of speech for speaker and language recognition. In *Proc. of Eurospeech*, pages 841–844, 2003.

[3] Ingunn Amdal, Filipp Korkmazskiy, and Arun C. Surendran. Data-driven pronunciation modelling for non-native speakers using association strength between phones. In *Automatic Speech Recognition*, 2000.

[4] P. Angkititrakul and J. Hansen. Advances in phone-based modeling for automatic accent classification. *IEEE TASLP*, 2006.

[5] J. Bernthal and N. Bankson. *Articulation and Phonological Disorders.* Allyn & Bacon, 2003.

[6] F. Biadsy and J. Hirschberg. Using prosody and phonotactics in arabic dialect identification. In *Proc. Interspeech*, 2009.

[7] F. Biadsy, H. Soltau, L. Mangu, J. Navartil, and J. Hirschberg. Discriminative phonotactics for dialect recognition using context-dependent phone classifiers. In *Odyssey*, 2010.

[8] B. Bielefeld. Language identification using shifted delta cepstrum. In *Fourteenth Annual Speech Research Symposium*, 1994.

[9] L. Burget, P. Matejka, and J. Cernocky. Discriminative training techniques for acoustic language identification. In *Proc. of ICASSP*, pages 209–212, 2006.

[10] W. Campbell. A covariance kernel for SVM language recognition. In *Proc. of ICASSP*, 2008.

[11] W. Campbell, P. Torres-Carrasquillo, and D. Reynolds. A comparison of subspace feature domain methods for language recognition. In *Proc. of Interspeech*, 2008.

[12] N. Chen, W. Shen, J. Campbell, and P. Torres-Carrasquillo. Informative dialect recognition using context-dependent pronunciation modeling. In *Proc. ICASSP*, 2011.

[13] Nancy Chen, Wade Shen, and Joseph Campbell. A linguistically-informative approach to dialect recognition using dialect-specific context-dependent phonetic models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5014–5017, 2010.

[14] Nancy Chen, Wade Shen, Joseph Campbell, and Reva Schwartz. Large-scale analysis of formant frequency estimation variability in conversational telephone speech. In *Interspeech*, pages 2203–2206, 2009.

[15] G. Choueiter, G. Zweig, and P. Nguyen. An empirical study of automatic accent classification. In *ICASSP*.

[16] F.M. Chreist. *Foreign Accent.* Prentice-Hall, Englewood Cliffs, NJ, USA, 1964.

[17] C. Cieri, D. Miller, and K. Walker. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, 2004.

[18] Cynthia G. Clopper and D. B. Pisoni. The nationwide speech project: A new corpus of american english dialects. *Speech Communication*, 48:633–644, 2006.

[19] AMI Corpus. http://corpus.amiproject.org, last accessed Nov, 2010.

[20] The Ivie Corpus. http://www.phon.ox.ac.uk/ivie/index.php, Last accessed, June 2009.

[21] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[22] A dialect map of American English. http://www.uta.fi/fast/us1/ref/dialmap.html, Last accessed 2009.

[23] Merriam Webster Online Dictionary. http://www.merriam-webster.com/dictionary/dialect, Last accessed 2009.

[24] K. Evanini. *The permeability of dialect boundaries: a case study of the region surrounding Erie.* PhD thesis, University of Pennsylvania, 2009.

[25] Keelan Evanini, Stephen Isard, and Mark Liberman. Automatic formant extraction for sociolinguistic analysis of large corpora. In *Interspeech*, 2009.

[26] M. Collins F. Biadsy, J. Hirschberg. Dialect recognition using a phone-gmm-supervector-based svm kernel. In *Interspeech*, 2010.

[27] William M. Fisher, George R. Doddington, and Kathleen M. Goudie-Marshall. The darpa speech recognition research database: Specifications and status. In *Proceedings of DARPA Workshop on Speech Recognition*, pages 93–99, 1986.

[28] J.E. Flege. Factors affecting degree of perceived foreign accent in english sentences. *J. Acoust. Sot. Amer.*, 84(6):70–79, 1988.

[29] T.J. Foil. Language identication using noisy speech. In *ICASSP*, 1996.

[30] CSLU foreign-accented English corpus. http://www.cslu.ogi.edu/corpora/fae, last accessed Nov, 2010.

[31] Eric Fosler-Lussier. Multi-level decision trees for static and dynamic pronunciation models. In *Eurospeech*, 1999.

[32] Eric Fosler-Lussier. A tutorial on pronunciation modeling for large vocabulary speech recognition. In S Renals and G Grefenstette, editors, *Text- and Speech-Triggered Info.*, pages 38–77. Springer-Verlag Berlin Heidelberg, 2003.

[33] P. Fung and Y. Liu. Effects and modeling of phonetic and acoustic confusions in accented speech. *J. Accoust. Soc. Am.*, 2005.

[34] J.L. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains. *IEEE Trans. Speech Audio Process*, 2, 1994.

[35] H. Giles, D. Taylor, and R. Bouhis. Towards a theory of interpersonal accommodation through language. *Language in Society*, 2:177–192, 1973.

[36] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. of ICASSP*, 1989.

[37] B. Goldstein, L. Fabiano, and P.S. Washington. Phonological skills in predominantly english-speaking, predominantly spanish-speaking, and spanishenglish bilingual children. *Language, Speech, and Hearing Services in Schools*, 36:201–218, 2005.

[38] F. Goodman, A. Martin, and R. Wohlford. Improved automatic language identification in noisy speech. In *Proc. of ICASSP*, volume 1, pages 528–531, 1989.

[39] A. Gopnik. How babies think. Phoneix Paperbacks, 2001.

[40] L. Green. *African American English.* Cambridge University Press, 2002.

[41] J. Hay, S. Jannedy, and N. Mendoza-Denton. Oprah and /ay/: Lexical frequency, referee design and style. In *the 14th International Congress of Phonetic Sciences*, 1999.

[42] T. Hazen and V. Zue. Automatic language identication using a segment-based approach. In *Proc. of Eurospeech*, 1993.

[43] T. Hazen and V. Zue. Segment-based automatic language identication. *Journal of the Acoustical Society of America*, 101(4):2323–2331, 1997.

[44] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1739–1752, April 1990.

[45] A. S. House and N. Neuberg. Toward automatic identification of the language of an utterance. 1. preliminary methodological considerations. *J. Acoust. Sot. Amer.*, 62(3):708–713, 1977.

[46] StoryCorps: http://www.npr.org/series/4516989/storycorps, last accessed, Apr. 6 2011.

[47] R. Huang and J. Hansen. Unsupervised discriminative training with application to dialect classification. *IEEE TASLP*, 15(8):2444–2453, 2007.

[48] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Prentice Hall, 2001.

[49] V. Hubeika, L. Burget, and P. Matejka. Discriminative training on channel compensation for acoustic language recognition. In *Proc. of Interspeech*, 2008.

[50] Mark Huckvale. ACCDIST: a metric for comparing speakers' accents. In *Interspeech*, 2004.

[51] J.J. Humphries and P.C. Woodland. The use of accent-specific pronunciation dictionaries in acoustic model training. In *Proc. of ICASSP*, pages 317–320, 1998.

[52] F. Jelinek. *Readings in Speech Recognition*, chapter 8: Self-organized language modeling for speech recognition, pages 450–506. Morgan Kaufmann, San Mateo, CA, USA, 1990.

[53] A. S. Kaye and J. Rosenhouse. *Arabic Dialects and Maltese in The Semitic Languages.* Robert Hetzron, Routledge, London, 1998.

[54] Mina Kim, Yoo Rhee Oh, and Hong Kook Kim. Non-native pronunciation variation modelling using an indirect data driven method. In *ASRU*, pages 231–236, 2007.

[55] W. Labov. *Sociolinguistic patterns.* University of Pennsylvania Press, 1972.

[56] W. Labov, S. Ash, and C. Boberg. *The Atlas of North American English: Phonetics, Phonology, and Sound Change.* Mouton de Gruyter, 2006.

[57] J. Laver. *Principles of Phonetics.* Cambridge University Press, 1994.

[58] Y. Lei and J. H.L. Hansen. Factor analysis-based information integration for arabic dialect identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4337–4340.

[59] R. Leonard and G. Doddington. Automatic language identication. Technical report, Air Force Rome Air Development Center, 1974.

[60] M.K. Liu, B. Xu, T.Y. Huang, Y.G. Deng, and C.R. Li. Mandarin accent adaptation based on context-independent/ context-dependent pronunciation modeling. In *Proc. of ICASSP*, volume 2, pages 1025–1028, 2000.

[61] K. Livescu. *Feature-Based Pronunciation Modeling for Automatic Speech Recognition.* PhD thesis, MIT, 2005.

[62] K. Livescu and J. Glass. Modeling of non-native speech for automatic speech recognition. In *ICASSP*, 2000.

[63] A. Martin, G. Doddington, and T. Kamm. The det curve in assessment of detection task performance. In *Eurospeech*, volume 4, pages 1899–1903, 1997.

[64] J. Navratil. Recent advances in phonotactic language recognition using binary-decision trees. In *Interspeech*, 2006.

[65] J. Navratil and W. Zuhlke. Phonetic-context mapping in language identification. In *Proc. of Interspeech*, volume 1, pages 71–74, 1997.

[66] Yoo Rhee Oh, Mina Kim, and Hong Kook Kim. Acoustic and pronunciation model adaptation for context-independent and context-dependent pronunciation variability of non-native speech. In *ICASSP*, pages 4281–4284, 2008.

[67] Yoo Rhee Oh, Jae Sam Yoon, and Hong Kook Kim. Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication*, pages 59–70, 2007.

[68] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. The boston university radio news corpus. Technical report, Boston University Technical Report No. ECS-95-001, 1995.

[69] C. Paradis and D. LaCharit. Guttural deletion in loanwords. *Phonology*, 18(2), 2001.

[70] D. Paul and J. Baker. The design for the wall street journal-based csr corpus. In *DARPA Speech and Natural Language Workshop*, pages 357–360. Morgan Kaufmann, 1992.

[71] M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier. Buckeye corpus of conversational speech, 2007. Ohio State University.

[72] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall Singal Processing Series, Prentice Hall, Englewood Cliffs, NJ, 1993.

[73] Steve Renals, Dave Abberley, David Kirby, and Tony Robinson. Indexing and retrieval of broadcast news. *Speech Communication*, 2000.

[74] D. Reynolds, W. Campbell, W. Shen, and E. Singer. Automatic language recognition via spectral and token based approaches. In J. Benesty et al., editor, *Springer Handbook of Speech Processing*. Springer, 2007.

[75] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verication using adapted gaussian mixture models. *Digital Signal Processing*, (19):19–41, 2000.

[76] F. S. Richardson, W. M. Campbell, and P. A. Torres-Carrasquillo. Discriminative n-gram selection for dialect recognition. In *Interspeech*, 2009.

[77] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 1999.

[78] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. Wsj-cam0: A british english corpus for large vocabulary continuous speech recognition. In *ICASSP*, 1994.

[79] Phillip Rose. *Forensic Speaker Identification*. Taylor and Francis, New York, NY, 2002.

[80] P. Sailaja. *Indian English*. Ediburgh University Press, 2009.

[81] K.R. Scherer and H. Giles. *Social Markers in Speech*. Cambridge University Press, 1979.

[82] R. Schwartz, W. Shen, J. Campbell, S. Paget, J. Vonwiller, D. Estival, and C. Cieri. Construction of a phonotactic dialect corpus using semiautomatic annotation. In *Interspeech*, 2007.

[83] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer. Experiments with lattice-based pprlm language identication. In *Proc. of Odyssey-06*, 2006.

[84] Wade Shen, Nancy Chen, and Douglas Reynolds. Dialect recognition using adapted phonetic models. In *Interspeech*, pages 763–766, 2008.

[85] Helmer Strik and Catia Cucchiarini. Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, 29:225–246, 1999.

[86] Helmer Strk. Pronunciation adaptation at the lexical level. In *Adaptation Methods for Speech Recognition*, 2001.

[87] P. Torres-Carrasquillo, E. Singer, M. Kohler, R. Greene, D. Reynolds, and J. Deller. Approaches to language identication using gaussian mixture models and shifted delta cepstral features. In *Proc. of ICSLP*, 2002.

[88] P.A. Torres-Carrasquillo, D. Sturim, D. Reynolds, and A. McCree. Eigenchannel compensation and discriminatively trained gaussian mixture models for dialect and accent recognition. In *Proc. of Interspeech*, 2008.

[89] Pedro Torres-Carrasquillo, Terry Gleason, and Doug Reynolds. Dialect identification using gaussian mixture models. In *Odyssey - The Speaker and Language Recognition Workshop.*

[90] K. Versteegh. *The Arabic Language.* New York: Columbia University Press, 1997.

[91] W. Ward, H. Krech, X. Yu, K. Herold, G. Figgs, A. Ikeno, D. Jurafsky, and W. Byrne. Lexicon adaptation for lvcsr: Speaker idiosyncrasies, non-native speakers, and pronunciation choice. In *ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation*, 2002.

[92] John Wells. *Accents of English: Beyond the British Isles.* Cambridge University Press, 1982.

[93] W. Wolfram and N. Schilling-Este. *American English: Dialects and Variation, 2e. Appendix: An Inventory of Socially Diagnostic Structures.* Blackwell Publishing Professional, 2005.

[94] Y. Yan and E. Barnard. An approach to automatic language identication based on language-dependent phoneme recognition. In *Proc. ICASSP-95*, pages 3511–3514, 1995.

[95] J. Yuan and M. Liberman. Investigating /l/ variation in english through forced alignment. In *Proc. of Interpseech*, 2009.

[96] B. Zhou and J. H. L. Hansen. Speechfind: An experimental on-line spoken document retrieval system for historical audio archives. In *Proc. of ICSLP*, pages 1969–1972, 2002.

[97] X. Zhou, J. Navaratil, J. Pelecanos, G. Ramaswamy, and T. Huang. Intersession variability compensation for language recognition. In *Proc. of ICASSP*, 2008.

[98] M. Zissman, T. Gleason, D. Rekart, and B. Losiewicz. Automatic dialect identification of extemporaneous conversational latin american spanish speech. In *ICASSP.*

[99] M. Zissman and E. Singer. Automatic language identication of telephone speech messages using phoneme recognition and n-gram modeling. In *Proc. ICASSP-94*, volume 1, pages 305–308, 1994.

[100] Marc Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE TSAP*, 1996.