RECOGNIZING NON-NATIVE SPEECH: CHARACTERIZING AND ADAPTING TO NON-NATIVE USAGE IN LVCSR

Laura Mayfield Tomokiyo

CMU-LTI-01-168

School of Computer Science
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee

Alex Waibel, Chair
Jared Bernstein
Maxine Eskenazi
Tanja Schultz
Wayne Ward

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

©Laura Mayfield Tomokiyo, 2001

Abstract

Low-proficiency non-native speakers represent a significant challenge for large-vocabulary continuous speech recognition (LVCSR). Acoustic models are confused by a heavy accent; language models are confused by poor grammar and unconventional word choice. Lack of comfort with the spoken language affects the fundamental properties of connected speech that have been a focus of LVCSR research; cross-word and interword coarticulation, disfluency, and prosody are among the features that differ in native and non-native speech.

In this dissertation, I first address the problem of *characterizing* low-proficiency non-native speech. One population is examined in great detail: learners of English whose native language is Japanese. Properties such as fluency, vocabulary, and pace in read and spontaneous speech are measured for both general and proficiency-controlled data sets. I further show that native and non-native speech can be distinguished using a variety of statistical metrics, including perplexity and Kullback-Leibler divergence. Patterns in reading errors and grammaticality of spontaneous speech are quantitatively described. This analysis, while focusing on one speaker population, provides a model for characterizing non-native speech that the broader LVCSR community may find useful. The generalizability of this model is demonstrated by contrasting the speech of native speakers of Mandarin with that of our primary speaker set.

Second, I explore methods of *adapting* to non-native speech. The test set is controlled for language exposure and proficiency, and the task is a simplified read news task tailored toward the lower-proficiency speakers, who experienced limited success in more difficult reading tasks like the widely-used Wall Street Journal readings. I find that the largest gains in recognition performance come through acoustic adaptation, and present evaluations of adaptation and training techniques incorporating native-language and accented data. From a speaker-adapted baseline of 63.1% WER (the same models perform at 8% for Broadcast News F0 speech), a 29% relative improvement is achieved through a combination of adaptation and training. In contrast, gains from lexical modeling were found to be extremely small, even when investigated in conjunction with retraining. I describe data-driven and linguistically-motivated algorithms for lexical modeling, presenting experimental results and discussing possible reasons why the improvement was not larger.

Finally, I present a novel method for detecting non-native speech. Without using any acoustic features, I show how bilateral and multilateral discrimination can be accomplished on the basis of features present in text. Both recognizer output and transcripts of non-native speech are identified with high accuracy through naive Bayes classification. The word and part-of-speech sequences that are found to be indicative of non-native speech provide an additional resource for characterizing non-native speech, which leads to further insights about the properties of non-native spoken language.

Contents

List of Figures

List of Tables

List of Terms

CND Children's news database. A collection of news articles written for children from the magazine *Time for Kids*. CND was established specifically for this dissertation with the goal of providing content similar to the widely-used Wall Street Journal readings at a reading level appropriate for lower-proficiency speakers.

GA General American (Giegerich, 1992, p.47). The variety of English modeled in the baseline. lexicon.

L1 One's native language

L2 A second language (Ellis, 1997, p.3)

JL1 speaker In this dissertation, native speakers of Japanese speaking English.

L1 speech data Recordings of speakers for whom English is not a native language speaking their native language.

L2 speech data Recordings of speakers for whom English is not a native language speaking English.

Non-native speaker In this dissertation, the term *non-native speaker* refers specifically to speakers of low-to-mid English proficiency.

NIST The National Institute of Standards and Technology

LDC The Linguistic Data Consortium

F0 speech Standard broadcast speech as defined by the NIST evaluation specifications (NIST, 1999). Called "F0" speech because it excludes the following conditions: speech over telephone channels; speech in the presence of background music; speech under degraded conditions; speech from non-native speakers.

LVCSR Large-vocabulary continuous speech recognition.

WER Word error rate. The standard evaluation measure for recognizer performance.

$\mathrm{WER} = 100 - \left(100 * \frac{refN - subN - delN - insN}{refN}\right)$ (Lee, 1990)[1]

[1] Lee actually defines word *accuracy* rate; WER is found by subtracting WAR from 100.

Chapter 1

Introduction

In the early 1990's, dialogues like the following were the target of intensive basic research in speech recognition and natural language understanding (Ward, 1990) .

System:	Welcome to Pittsburgh Airlines' automated flight information service. Do you know the flight number?
User:	Five-twelve.
System:	Would you like arrival or departure information?
User:	Arrival.
System:	Flight 524 arriving on Tuesday, Feb 27th. Is that correct?
User:	No.
System:	What would you like to change, the flight number or the date?
User:	The flight number.
System:	Please say the flight number.
User:	Five one two.
System:	Flight 512 arriving on Wednesday, Feb 28th. Is that correct?
User:	Yes.
System:	Flight 512 arrives in more than one city. Would you like Boston Logan or Charleston, SC?
	...

Today, at least one major airline provides an automated interaction exactly like this as the front-line response to caller flight information queries. Clearly, speech recognition has begun to be embraced by industry, which sees it as an efficient replacement for human operators, and tolerated, however grudgingly, by users. We can expect to see more and more applications of speech recognition in the real world as the technology improves and public acceptance spreads.

With real-world deployment, however, user diversity demands serious consideration. Although the claim that all users should be offered equal access to speech recognition is not strong (people with poor reading skills do not have the same access to newspapers as the highly literate, for example), it does appear that speech recognition applications are on a trajectory to become a default interface for information dissemination systems. Accommodation of users whose language skills are somehow impaired is not only an interesting research problem but also a significant functional concern.

In the dialogue transcribed above, I, a native speaker, called up to make the query. Although the flight number was misrecognized at first, overall, the recognition quality was good, and I was able to get the information I required. I also asked a friend, a native speaker of Japanese who had lived in the United States for two years, to call for the same information.

System:	Welcome to Pittsburgh Airlines' automated flight information service. Do you know the flight number?
User:	Yes.
System:	Please say the flight number.
User:	えーと five hundred and twent- five hundred and twelve.
System:	Would you like arrival or departure information?
User:	Arrival.
System:	Flight 283 arriving on Tuesday, Feb 27th. Is that correct?
User:	あれ? No flight five hundred and twelve please
System:	What would you like to change, the flight number or the date?
User:	は? So... flight number is five hundred and twenty no five hundred and twelve
System:	Please say the flight number.
User:	Flight number is five hundred and twelve.
System:	Flight 94 arriving on Wednesday, Feb 28th. Is that correct?
User:	No.
System:	What would you like to change, the flight number or the date?
User:	Flight number, please.
System:	Please say the flight number.
User:	Five hundred and twenty.
System:	Flight 520 arriving on Wednesday, Feb 28th. Is that correct?
User:	Yes.
System:	Flight 520 arrives in more than one city. Would you like Chicago O'Hare or Kansas City, MO?
User:	あっ, no, no. Not twenty, twelve.
System:	Please repeat or rephrase that.
User:	ん... ん?
System:	Please repeat or rephrase that.
User:	<*click*>

In this case, the speaker is prevented from finding out what she wants to know by a combination of recognition errors and her difficulty responding appropriately to the system's prompts. One can envision other situations in which reduced English proficiency would diminish the effectiveness of speech-driven applications. Dictation systems, for example, are used by people ranging from physicians recording patient information to graduate students with typing injuries. Both of these groups have significant non-native populations. Conversational transcription systems such as meeting record and surveillance systems cannot assume that all subjects will be fluent speakers of the language. If a speech translation system is available to facilitate English-Japanese communication, it may be used not only by native English speakers but also by the many others who speak English better than they speak Japanese. Language learning systems are limited in their ability to offer recognition-based lessons to the degree that speech recognition of new learners is not reliable.

There are many levels at which non-native speech differs from native speech. The term non-native speech, as widely used in the speech recognition community (Byrne et al., 1998; Langlais et al., 1998; Livescu and Glass, 2000; van Leeuwen and Orr, 1999; Witt and Young, 1997), covers an enormous range of proficiencies and speech types. For a language like English, this range is in fact much greater than the range of native speech, even when regional variation is considered. There are a few parameters, however, that seem particularly useful for encoding non-native speech. Accent, mode, lexical choice, syntactic soundness, and fluency are aspects of spoken language that can both describe variation in native speech and be used to distinguish it from non-native speech.

Accent

The word *accent* is the subject of some controversy. The confusion (and genuine lack of an absolute distinction) between *accent* and *dialect*, coupled with increasing awareness of negative associations with marked accents and dialects, has prompted many to abandon both terms in favor of the more neutral and more vague *variety*. One of the reasons that it is so difficult to assign a scholarly definition to the word *accent* is that in the lay sense, accent is by definition not absolute; a listener perceives an accent when the speaker's speech is different from his own. Although academic publications emphasize time and time again that there is no such thing as "unaccented" English (Lippi-Green, 1997; Wardhaugh, 1998), the sense of the word *accent* that is shared by native speakers will always be relative to one's own speech, and it is this understanding that is the foundation for recovery strategies.

While we may lack a clear set of features that characterize accent (Lippi-Green defines accents as "loose bundles of prosodic and segmental features distributed over geographic and/or social space"), lay listeners seldom have difficulty identifying presence or absence of accent; although the boundaries of accent may differ from speaker to speaker, I submit that there are speakers whom any educated native speaker would identify as having a foreign accent. If we adopt Wardhaugh's definition of accent as "how [people] pronounce what they say" and accents as often having "clear regional and social associations" (Wardhaugh, 1998), we can define *foreign accent* as "pronunciation that is associated with a country or region in which English[1] is not the primary language spoken."

Mode

The amount of planning and attention required to generate an utterance can be quite different for native and non-native speakers; attention used for utterance generation can also impact the actual production to the degree that the number of cognitive cycles available for sentence generation and articulation is reduced (Pawley and Syder, 1983, p.208). Variables describing the speech task, level of formality, and spoken language performance have sometimes been borrowed to describe degree of attention as well, but since we cannot assume that the correlation between these variables and attention is the same for native and non-native speech I will modify the definition of the term *mode* as used in e.g. (Finke and Waibel, 1997) to describe

[1] English is used as the default "native language" in many of the examples and definitions in this thesis. This is for convenience only; all definitions, theories, and applications are meant to be extensible to any human language.

the degree of attention paid to utterance generation.

Careful speech and casual speech are often offered in speech recognition literature as examples of speaking *styles* (Eskenazi, 1997), e.g. Although Labov (1972) supported the idea that "styles can be ranged along a single dimension, measured by the amount of attention paid to speech", more recent definitions incorporate formality level (Wardhaugh, 1998) and relationship between speaker and listener (Bell, 1984). Rampton (1987) argues that particularly in the case of the non-native speaker, for whom attention to speech may be distributed very differently from native speakers, Labov's definition is not appropriate. Generally speaking, the term *style* is currently used to describe systematic linguistic choices associated with particular situations (Finegan, 1994). One can separate situational appropriateness from degree of planning, and I will therefore restrict the definition of *style* to formality and difficulty level (audience-directed lexical and structural choices) and use the variable *mode* to encode the degree of planning that goes into formulating an utterance. The variable *register* will be used to describe task- and context-directed lexical and structural choices.

Mode, then, as I have defined it, varies along a continuum and is closely related to proficiency among non-natives. It also directly affects performance. A native speaker and a non-native speaker of low proficiency could be speaking with the same style and in the same register (asking a stranger on the street for directions, for example), but with modes representing very different levels of attention. The greater cognitive load consumed by attention for the non-native speaker may affect his ability to articulate difficult phone sequences, resulting in a stronger accent than he would normally exhibit for isolated words. I assume that mode is different from the other parameters discussed here in that it is not directly evident in the speech that is produced; rather, it exerts an influence on how speech is produced that is different for native and non-native speakers.

Syntactic Soundness

Learners of a language are generally exposed to L2 grammar in the early days of their study, yet incomplete mastery of syntax is one of the features that can mark even highly proficient speech as non-native. One theoretical view of second language acquisition takes the Chomskian position that acquistion of L1 grammar occurs as children instantiate the biologically endowed Universal Grammar, it does not agree on whether L2 learners have access to this resource (Ellis, 1997, p.66). It is clear that adult learners struggle with principles, for example, co-reference through a reflexive, that are instantiated differently (or uniquely) in L1 and/or L2. It has also been observed that attention and learning stage can interfere with production of even those syntactic concepts that L1 and L2 share, as with acquisition of definiteness for Polish learners of English (Van Dyke, 1997).

Native speakers certainly do not always demonstrate prescriptively correct syntax. Soundness in instantiation of basic principles like definiteness marking, however, is common to native speakers. For the most poorly educated native speaker, the sentence "Flight number is five hundred and twelve" just sounds wrong, for reasons he would not know how to explain other than to say "you have to say *the*."

Incorrect instantiation of syntactic principles does not necessarily result in a syntactically incorrect sentence. Native speakers of German frequently confuse past and past perfect in English. Imagine that a party was thrown on Saturday night. On Monday morning, to be asked "did you go to the party?" would not seem unusual; the perfectly grammatical "have you been to the party," on the other hand, would perplex, causing one to wonder if the party were still going on. This type of syntactic misinstantiation is a subtle yet sometimes jarring sign of non-nativeness.

Lexical Choice

The words chosen by a speaker to express a thought can also reveal whether he is native. A sentence can be semantically meaningful and syntactically correct yet noticeably non-native. Let us consider the following sentence pairs.

(1.1) a. What is the cost of a ticket for the concert
b. How much does a ticket for the concert cost

(1.2) a. I'm going to have a jelly and peanut butter sandwich
b. I'm going to have a peanut butter and jelly sandwich

(1.3) a. Let's disassemble the puzzle
b. Let's take apart the puzzle

In each of these examples, the first is technically correct but less likely to be spoken by a native speaker than the second. There are many regional differences in the way native speakers choose words (British "lift" and General American (GA) "elevator" being a familar example). A lack of awareness of familiar lexical patterns, however, results in noticeable idiosyncrasies, as contrasted with regionalisms, in non-native speech. This variable can cause a particular problem for speech recognition as the language model encodes the distribution of words in native speech.

Fluency

The fluency variable describes the pace and smoothness of speech. Native speech is often disfluent; native speakers backtrack, stutter, pause in the middle of a sentence, and speak in fragments in conversational speech. These effects show similarities even across languages (Eklund and Shriberg, 1998). Speech disfluencies are not limited to conversational "modes;" they are found in read speech as well, when readers stumble over the text. Pace, too, varies greatly in native speech. Some natives speak quickly; others speak slowly. Some speak in bursts, others with an even rhythm. However, it appears that measurements of fluency can be used to distinguish native and non-native speech. Cucchiarini et al. (2000), among others, show that pace correlates closely with perception of proficiency. Some non-native reading errors in speech are distinctive and quantifiable (Mayfield Tomokiyo and Jones, 2001). While some disfluencies seem to follow universal patterns, others, including the native-language interjections seen in the dialogue transcribed above, strongly indicate that the speaker is non-native.

It seems clear that native speakers are able to recognize non-native speakers based on features like accent, syntax, and fluency. Children can identify and imitate specific characteristics of speech that mark it as typical of a non-native group. When a listener is first exposed to a variety of non-native speech, he may initially struggle to understand it, but if he is a cooperative listener, he can often adapt very quickly. Humans are incredibly well equipped to understand speech, and tolerate deviation relatively well.

Unfortunately, neither of these skills have come as naturally to the machine. Computer understanding of speech is based on statistical models of patterns found in training corpora. When the accent, syntax, and lexical choice of the speaker are not well-represented in a training corpus, the models must somehow be adapted if good recognition is to be achieved. We might imagine several angles for attacking such adaptation.

The *acoustic model* specifies the expected mapping of acoustic events to phonetic units. In a fully-continuous context-dependent system such as the one that will be described in later chapters, this is an extremely fine-grained representation. Acoustic events are modeled on a sub-phonetic level, and many more variations are recognized as would be in a traditional phonetic analysis; in the recognizer used in this dissertation, 118 distinct realizations of /t/ in GA are modeled. The acoustic model would be the natural place to represent phonetic differences in realization for a given speaker's accent.

The *lexicon*, which describes the phonemic makeup of words, would lend itself to modeling of phonemic differences and phonological adaptation in production. By altering the lexeme specifications, phonemic substitutions, epenthesis, elision, and in some cases phonetic realizational differences can be easily represented. The problem that arises is that the altered lexicon may not interact with the acoustic model as expected. However, lexical modeling is a straightforward approach that has been used with modest success for varieties of native speech (Humphries and Woodland, 1997; Huang et al., 2000) and non-native speech for non-LVCSR tasks (Fung and Liu, 1999).

The recognizer's understanding of how words occur in sequence is encoded in the *language model*. Absent a natural language understanding component, the recognizer has no understanding of the meaningfulness of a hypothesized utterance, and must rely on a statistical model to determine the likelihood of a sequence of words having been uttered. By adapting the language model, the restrictions on probable word sequences could be relaxed for increased tolerance of deviation from native patterns of speech. Alternatively, one could envision training a statistical model of non-native speech, explicitly representing patterns that are common in the speech of non-natives.

Finally, the *system* itself could be adapted for greater flexibility in processing non-native speech. Just as human listeners are able to ask the speaker to repeat himself, delay processing while building context, and silently induce lexical, syntactic, and phonetic mappings from both positive and negative examples, a system that endeavors to understand non-native speech could incorporate learning strategies with the aid of dialogue and natural language understanding components.

This investigation will be restricted to the recognizer components that model pronunciation, namely the acoustic model and the lexicon.

In this dissertation, I concentrate principally on native speakers of Japanese. This speaker population offers great potential for experimental control; English education is standardized in Japan, and the Japanese population in Pittsburgh is large enough that finding speakers with similar educational backgrounds and exposure to English was not difficult. The nature of Japanese-influenced English is well known, if not well studied, from both lexical and phonotactic points of view. The many English words that have worked their way into everyday Japanese speech have undergone semantic and phonological transformations that can help us to predict how Japanese natives will approach production of English. Because nativized foreign words are represented in the Japanese script, an array of orthographic mappings is accessible that may provide further aid in developing a model of Japanese-influenced English.

Applications of this work are also likely to be of interest in Japan. Language tutoring systems that model a particular native language (L1) well can present feedback in the context of linguistic elements that are known to be problematic for speakers that share the user's L1. The Japanese government is currently so concerned about the English language ability of its citizens that it is considering the dramatic step of making English an official language (Kawai, 2000). Such a requirement would increase the demand for English training, and possibly for English versions of natural language systems currently available in Japanese. In such an eventuality, tolerance of non-native English would be critical.

Problem statement

Speech recognition systems consistently perform poorly on all but the most fluent non-native speakers. As speech recognition technology moves into general use, accommodation of non-native speakers is both an interesting research problem and an important functional concern.

Thesis statement

Speech recognition performance for lower-proficiency non-native speakers of English, specifically native speakers of Japanese, can be significantly improved through phonological modeling of the non-native condition.

Organization

This document is organized into seven chapters and three appendices. **Chapter 2: Background and Related Work** surveys the rich history of the study of language acquisition as well as relatively recent research in speech recognition for non-native speakers; **Chapter 3: Non-native Speech Database: Composition and Characterization** provides a description of elicitation and transcription methods and a thorough analysis of the JL1 and ML1 English read and spontaneous speech corpora; **Chapter 4: Acoustic Modeling** describes detailed experiments in acoustic modeling for JL1 English; **Chapter 5: Lexical Modeling** describes linguistically-motivated and data-driven modeling of phonological interference at the lexical level; **Chapter 6: Hypothesis-driven Accent Classification** presents a novel and extremely effective method for detecting non-native speech that can be used to invoke the non-native modeling methods described in previous chapters; and finally, **Chapter 7: Conclusion** summarizes the main contributions of this work and discusses directions for future research. **Appendix A: Data Collection and Speaker Proficiency Evaluation** lists database statistics and demographic information for the speakers; **Appendix B: Phonological Transformation Rules** gives the rules used for linguistically-motivated lexical modeling of non-native speech; and **Appendix C: IPA-Arpabet Mappings** provides a chart relating the International Phonetic Association (IPA) symbols used for linguistic discussions to the ASCII symbols commonly used in the context of speech recognition.

Chapter 2

Background and Related Work

The idea of specialized recognition of non-native speech has developed from two separate directions. In language learning research we have seen increased efforts to use output from speech recognition applications to provide feedback and guidance to the student. The relationship between acoustic scores and human perception has been the focus of much interest in this area, as have methods for measuring distance between the student's speech and a model of "good" native speech. Research in speech recognition, on the other hand, has turned toward non-native speech as the systems become accurate enough and realistic enough for non-native speakers to want to use them. Progress in recognition of non-native speech is measured primarily by reduction in word error, which is not a metric that can be directly linked to such features as intelligibility. The goal of an LVCSR system is to model speech so that the word the speaker intended to say is recognized; this may be accomplished by building a model that is incorrect from a prescriptive standpoint and undesirable from a pedagogical standpoint but represents the speaker's intent.

This chapter begins with a discussion of second language acquisition (SLA) research, which has influenced the way computational modeling of non-native speech is approached. I then give an overview of how non-native speech has been approached in the disciplines of computer-aided language learning and LVCSR, and conclude by discussing issues in elicitation and recording.

2.1 Second Language Acquisition

Do non-native speakers carry over pronunciation habits from their first language to their second? This is the question that research in second language acquisition may help to answer. The assumption that learners systematically substitute L1 phones for L2 phones is widespread in speech recognition.

> "Accent usually comes from the articulation habits of the speaker in her/his own native language." (Fung and Liu, 1999, p.1)

> "An alternative approach to [modeling] non-native speech is to assume that non-native speakers will

> dominantly use their native phones, presumably by mapping the phones of the language they are speaking (L2) to their native language (L1)." (van Leeuwen and Orr, 1999, p.1)

> "[The] techniques introduced here are based on the underlying idea that a non-native speaker... will substitute sounds of his or her mother tongue for those foreign sounds he or she cannot produce." (Witt and Young, 1999, p.1)

Studies in SLA do not agree on this point, however. While the fact that native speakers of a language can often guess a non-native speaker's L1 based on their articulation of specific phones is not disputed, whether any sort of trajectory in phonetic space between specific L1 and L2 phones is common to speakers of the same L1 is the subject of many years of debate.

2.1.1 Contrastive Analysis

Contrastive Analysis (CA) is a branch of applied Linguistics introduced in the 1930's which is concerned with "producing inverted (*i.e.* contrastive, not comparative) two-valued typologies (a CA is always concerned with a *pair* of languages), and founded on the assumption that languages can be compared" (James, 1980, p.3). CA theory claimed that "speakers tend to hear another language and attempt to produce utterances in it in terms of the structure of their own language, thus accounting for their 'accent' in L2," where accent refers not only to phonological accent, but to all elements in the presentation of speech that mark the speaker as foreign (Ferguson, 1989, p.82). In SLA-oriented CA, comparable features of L1 and L2 are identified and described, and mismatches are identified that are likely to lead to error on the part of the learner; CA is said to be able to *predict* and *diagnose* errors. This application is based on the concept of linguistic transfer, which is said to happen when knowledge about one language is applied (correctly or not) to another and intuitively would seem to explain why language learners make the mistakes they do.

The most serious arguments against CA were that its foundations were in structuralism and behavioralism, which had begun to lose favor, and that in practice, it was not an effective method for predicting errors that learners actually make. Brière (1966) reported on an experiment in which American students were played non-English sounds from Arabic, Vietnamese, and French and asked to reproduce them. While there were some cases of clear L1 transfer, Brière found that in other cases the students approximated one non-English sound with another (/ʀ/ for /γ/), which would not be predicted by CA. Furthermore, it was observed that some of the non-English sounds were easier than others for the American students to learn (Brière gives the example of a voiceless non-aspirated fortis dental stop as being easier than the dentalized version), a phenomenon for which CA does not provide an explanation.

2.1.2 Error Analysis

Dissatisfaction with CA led to the development of a paradigm known as Error Analysis (EA). James (1998, p.2) identifies two ways in which language learners "stop short of native-like success in a number of areas of the L2 grammar" (Towell and Hawkins, 1994): "when their L2 knowledge becomes fixed or *fossilized*, and when they produce errors in their attempts at it." This distinguishes them from native speakers, who are defined as knowing their language perfectly. While much is made in speech recognition research of the *im*perfection with which native speakers use their language, in the Chomskian tradition this is a performance issue and should be distinguished from language *competence*.

EA looks for systematic behavior in groups of learners, asking what types of errors out of all of the language errors produced by learners can be clustered together and be classified as "errors that native speakers of language X are likely to make," or "errors that speakers who do not control a system of case marking are likely to make." In EA, only L2 and the intermediate language IL, which represents the learner's understanding of L2 at a given time, are compared for mismatches (recall that in CA, *L1* and L2 were contrasted).

One major argument against EA is said to be that it does not account for the fact that speakers often avoid elements of L2 that they find difficult (further discussion to follow) and therefore do not make errors that EA would predict; another is that it incorrectly ignores the effects of transfer from L1.

2.1.3 Transfer Analysis

Recognition of the theoretical shortcomings of EA led to a return to favor of CA. Wardhaugh (1970) suggested that the problem with early CA was that it claimed to be able to *predict* errors by comparing only L1 and L2. EA was not quite a complete solution to this problem; although it could predict errors more accurately using its model of the learner's current understanding of L2, it did not take into account influences of L1, which cannot be ignored. An alternative, weaker version of CA was proposed, which claims only to be able to "explain (or diagnose) a subset of actually attested errors – those resulting from [L1] interference" (James, 1998, p.5). This incarnation of CA is referred to as *language transfer*, *transfer analysis*, or *weak CA*, and is different from EA in that the intermediate language IL is compared to L1 and not L2; it is used primarily as one tool within an EA-based analysis framework.

2.1.4 Towards a model of non-native speech

The idea of the intermediate language IL, often known as *interlanguage*, as a legitimate, working language has been developed to the point where it can really be taken as the basis of a computational model of a learner's speech. The problem for speech recognition, of course, is that each speaker has an individual model, representing the level of L2 understanding he has reached and the influences of L1 and other languages to which he has been exposed; one would need a way to generalize in order to apply ideas from interlanguage theory to a speech system. Nevertheless, it provides a theoretical background for thinking about implementing an error model for LVCSR.

Corder (1967) introduced the term *transitional competence* to reflect the independent system of the language that learners (both native and non-native) develop. Children acquiring their native language do not control the full adult version of the language, but rather an intermediate language, just as L2 learners do. This concept was then revised, and the idea of the *idiosyncratic dialect* developed to better describe the language spoken by the learner: it is a dialect in that it shares important parts with other varieties of the language, and can be considered one version of that language (as opposed to a separate language), but is idiosyncratic in that there are not enough speakers of that version to claim that they form any sort of language community, a characteristic that speakers of social dialects share. This definition emphasizes the transitional and unstable nature of the intermediate language.

Tarone et al. (1983) discuss strategies that language learners use to overcome difficulties in four major areas: phonological, morphological, syntactic, and lexical. They identify the strategy classes of *transfer*, *overgeneralization*, *prefabrication*, *overelaboration*, *epenthesis* and *avoidance*, most having an application in all four domains. *Avoidance* is further broken down into topic avoidance, semantic avoidance, appealing to authority (asking, using a dictionary), paraphrase, message abandonment, and language switch. It is interesting to consider these strategies in two of the contexts that concern us in speech recognition: system development and data collection. Clearly, many of the strategies outlined can be directly applied to error modeling in the speech system; phonological epenthesis and transfer (e.g. phoneme substitution), morphological overelaboration (choosing uncontracted forms), and lexical overgeneralization can be explicitly represented. The discussion of communication strategies, particularly avoidance strategies, has implications for training data collection as well, however, perhaps even more for task-oriented systems than freely conversational systems, which are traditionally considered more difficult. We often speak of the need to elicit during data collection words and expressions that will appear in real-world use of the system. How important is it to elicit the same *strategies* that will be triggered when non-native speakers try to use a speech system? Or, conversely, to avoid during data collection the triggering of strategies that would not be invoked in real-world use? It may be the case that in conversation, speakers have more flexibility to appeal to strategies such as avoidance in order to hide an inability to pronounce certain words or ask certain questions; they can choose another word or another topic, or choose silence as their avoidance strategy. When they need to find out specific information, however, they may resort to different strategies to express themselves than they would in conversation.

Tarone (1978a) investigates interlanguage *phonology*. For the specific case of Japanese learners of English (and building on L. Dickerson's 1974 dissertation (Dickerson, 1974)), she notes that "certain phonological environments are more favorable to the production of [s] and [z] than others." This effect has important implications for acoustic modeling, as we will see in Chapter 4. Tarone looks with particular interest at the role of the syllable in L2 phonological acquisition, asking why American speakers, for example, struggle with the articulation of /ʒ/ in any syllabic context other than that in which it appears in English. She extends this discussion to the various strategies speakers of many languages invoke to help with the articulation of non-CV (consonant-vowel) syllables. Disagreeing with Oller (1974), who emphasized the difference between the ways consonant clusters are simplified in L1 acquisition (deletion, reduction) and in L2 learning (epenthesis), Tarone supports the idea of the CV syllable as a "universal articulatory and perceptual unit such that the articulators tend to operate in basic CV programs in all languages" (Tarone, 1978b). She found a tendency in learners to simplify even consonant clusters which appeared in L1 using both epenthesis and deletion – she found that the preference for a CV syllable was independent from the strategy used to form it *and* L1. Incorporating this idea of an L1-independent preference for the CV syllable, Tarone identifies five processes and five constraints associated with L2 phonological acquisition.

Processes:

1. negative transfer from L1
2. first language acquisition processes
3. overgeneralization
4. approximation
5. avoidance

Constraints:

1. the inherent difficulty of certain L2 sounds and phonological contexts
2. the tendency of the articulators to rest position
3. the tendency of the articulators to a CV pattern
4. the tendency to avoid extremes of pitch variation
5. emotional and social constraints

These processes and constraints interact to define the learner's interlanguage phonological system and can be the basis for phonological error analysis.

The consistent observation that few L2 pronunciation errors can be traced to direct L1 transfer is not easy to reconcile with the clear consensus that there *are* identifiable foreign accents, a dilemma that Beebe (1987) attacks in a study of myths about interlanguage phonology. If non-native pronunciation errors do not have their roots in differences between L1 and L2 phonology, why can a non-linguist classify foreign accents by country when they cannot so easily classify grammatical errors?

Beebe presents a study which supports findings from earlier studies (e.g. Dickerson, 1974) but which presents data from five language groups, making it more comprehensive than previous studies. Beebe suggests that while native American listeners may classify a non-native phoneme that they hear as a particular native one using recovery strategies based on the English phonological system, the phoneme may not be the one that the speaker intended, and acoustically, may actually be quite distant from the phoneme that the listener thinks he heard. This intuition could shed some light on the agreement among native speakers on characteristics of particular foreign accents while at the same time explaining the lack of success of CA in explaining L2 pronunciation errors. Beebe makes the further observation that most substitution errors are *phonetic*, and not phonemic as it it may appear to native listeners. Looking at the distribution of English /l/ attempts in native speakers of four Asian languages (Japanese, Chinese, Korean, and Indonesian), Beebe found that although the pronunciation error rate was 46% (the calculation of this error rate was not discussed), the rate of substitution of an *r*-variant was only 3%. Three-quarters of the /r/ errors were phonetic deviations from /l/, and not phonemic substitutions of /r/.

Beebe's distinction between phonemic and phonetic errors is important when trying to teach pronunciation, as her findings indicate that while her students may appear to be confusing r and l, theirs are not the sort of errors that minimal-pair training would correct. What they need to understand is why their approximations of /l/ do not sound to a native speaker like /l/, not how /r/ is different from /l/. It is difficult to know whether this distinction would be meaningful to the speech recognizer. On the one hand, one might conclude that if it sounds like an /r/ to a native speaker, it will sound like an /r/ to the recognizer, and since the recognizer can accept pronunciation variants very easily, it would be simple to add /raǐk/ for "like" to the internal lexicon. On the other hand, Beebe's research suggests that while human listeners hear it as an /r/, they do so not because it is spectrally like an /r/, but rather because of the complex interaction between human auditory recovery strategies and phonological and semantic expectations.

2.2 Computer-Aided Language Learning

As technologies for processing human language have matured, it has become possible to view them as pedagogically valuable tools. Advances in speech recognition and parsing have been enthusiastically received in the field of computer-aided language learning (CALL), although the application of "technology" in language learning systems ranges from the very simplistic to the overly optimistic.

Noting that this work focuses specifically in the application of *speech* technology to language learning, let us first consider some common roles of speech in CALL systems.

Interactive: record and playback functions, adding variety to otherwise tedious drills

Quantitative: providing feedback regarding acoustic features like duration and F1/F2

Probabilistic: estimating the likelihood of an acoustic model having produced the acoustic event provided by the speaker

Communicative: incorporating speech with natural language understanding to act as a conversation partner

In an *interactive* context, speech is used to give the learner instant and repeated access to his own pronunciations, and to those of native speakers that he wishes to emulate. Critical issues include monitoring (if the learner has full control over the interaction, will he proceed in the way that is most beneficial to him?) and feedback (without evaluation from a teacher, will the learner know what he is doing wrong?) as well as authenticity, individual learning styles, and limitations in the hard-coded processing domain (Garrett, 1995).

At least one of these concerns can be addressed by providing *quantitative* feedback to the user so that deficiencies and improvements in his speech are clearly visible. Speaking rate and pause frequency are known to have significantly different distributions in native and non-native speech (Mayfield Tomokiyo, 2000) and correlate well with fluency ratings given by speech therapists and phoneticians (Cucchiarini et al., 1998). Eskenazi and Hansma (1998) have found that prosodic features that can be extracted directly from the speech signal are also good indicators of fluency and pronunciation quality.

While systems that offer this kind of quantitative feedback without requiring the user to utter isolated phones do need an acoustic model to generate a time-phone alignment, they are not making a statement about the relationship between the learner's speech and the information about native speech contained in the model. Many CALL systems use *probabilistic* output from the acoustic model to derive a pronunciation score. The scores themselves are then evaluated by comparing them to scores given by human listeners; a scoring algorithm is considered effective if it produces scores that correlate well with scores that experienced humans, such as language teachers, give the speakers. Pronunciation scores can also be given at different levels; a sentence-level score would give a speaker an idea of how good his overall pronunciation is, whereas a phone-level score would be useful for training articulation of specific phonemes.

Bernstein et al. (1990) presented the first HMM-based pronunciation evaluation system. They compared performance of sentence-level models and monophone models for a read speech task, finding that grading results correlate best with decisions by human graders when the sentences were first aligned using the sentence models and scores calculated using the phoneme models. They reported high reliability and agreement among human graders for ratings of pronunciation quality.

Franco et al. (1997) describe HMM-based phone log-likelihood scores and phone log-posterior probability scores that were used to evaluate American learners of French. They found that the posterior-based scores correlate better with human raters than the log-likelihood-based scores. They theorize that this is because their calculation of the posterior score includes a normalizing term in the denominator that would balance out effects of individual speaker characteristics or acoustic channel conditions. The authors also looked at duration and found that duration-based pronunciation scores performed similarly to the posterior-based scores at the speaker level and somewhat worse, but better than the log-likelihood scores, at the sentence level. A combination of posterior and duration scores at the sentence level improved correlation with human raters somewhat over posterior scores used alone. The maximum correlation they were able to achieve was 62%, compared to 65% inter-grader and 76% intra-grader correlation. Extending the approach to scoring of individual phonemes, Kim et al. (1997) report that at the phone level, while posterior-based scores still correlate best with human graders, correlation of duration-based scores is very poor. They hypothesize that this is because of the high variability of phone duration.

In related work at SRI, Ronen et al. (1997) assigned weights to phones based on how damaging mispronunciation of the phone is to expert ratings of overall intelligibility (as perceived by professional teachers). Ronen *et al.* found that incorporating the weights in calculation of the overall sentence score improved correlation with human graders.

Neumeyer et al. (1996) move towards *text independence* by introducing a class of algorithms which do not require a reference sentence or network to align the recognized speech to. They are able to structure the exercises in such a way that the responses expected from the user are highly constrained, yet the user is provided with the illusion of flexibility. Variations on this theme have also been successful (Eskenazi and Hansma, 1998; Ehsani et al., 1997).

Eskenazi (1996) showed that acoustic scores from the recognizer can be used to detect speaker pronunciation errors, and that prosodic features that distinguish non-native from native speech are present in the speech signal. Comparing acoustic phone scores across speakers (10 native and 20 non-native speakers were studied) for individual segments, Eskenazi found significant differences between native and non-native pronunciation for several phonemes, indicating that pronunciation error detection based on acoustic score would be successful. Working with expert tutors, Eskenazi examined possible measures of prosodic errors contributing to accent, finding that segment duration ratios, number of pitch peaks in a segment, and amplitude are features that correspond well with information human experts use to characterize accent. Eskenazi's *Fluency* pronunciation tutor incorporates this information to provide speakers with feedback on their pronunciation (Eskenazi and Hansma, 1998).

Kawai and Hirose (1997) report similar results, using Japanese monophone models for training native speakers of Chinese in pronunciation of the Japanese long vowels and geminate consonants known as tokushuhaku. Duration is phonemic in Japanese, and while short vowels are similar to Chinese vowels, Chinese speakers often have difficulty producing the long vowels. Using average phone durations of 20 native speakers as a guide, their system was able to tell speakers whether their pronunciations were too long, too short, or acceptable.

It has been pointed out that for speech systems designed specifically for pronunciation training of a predetermined phoneme set, a template-based recognizer may provide more useful scores than an HMM-based system (Dalby et al., 1998). In their experience, while an HMM-based recognizer showed better overall recognition accuracy, a template-based system was more accurate at recognizing vowel and nasal contrasts.

It is not clear that speech recognition technology has reached the point at which it can make judgements as to correctness of pronunciation that correspond to human judgements at a satisfactory level (Langlais et al., 1998), although Kawai (1999) claims to have done so for some specific sound types.

Some systems combine native models of the target L2 with native L1 models and non-native L2 models to build a system that can tell learners when their pronunciation is closer to an L1 phone than the target L2 phone. In his doctoral thesis, Kawai (1999) develops systems for English-speaking learners of Japanese and Japanese-speaking learners of English. He uses a bilingual monophone phoneme set and allows free transitions during alignment between English and Japanese phonemes. In this way, he is able to model L1 interference in L2 articulation, providing valuable feedback to the user.

Ronen et al. (1997) use a framework in which native and non-native models are trained and free transitions are allowed between the native and non-native phoneme sets in decoding. Non-native models were trained on speakers that were given low pronunciation scores by human graders. They used monophone models, having determined that system performance did not improve significantly with the introduction of context-dependent models (their experience is shared by Witt and Young, (1997), who found that context-independent models allow better acceptance/rejection accuracy). It is not mentioned whether the path through the mispronunciation network corresponds with human listeners' judgements of nativeness of pronunciation of the individual phonemes, but they did report a lower correlation between machine and human judgements of goodness of pronunciation calculated with this approach than with an approach in which each each utterance is decoded twice, once using native models and once using non-native models, and the HMM log-likelihood scores are combined to calculate a pronunciation score.

Auberg et al. (1998) present an accent coach that teaches English pronunciation to Japanese speakers. They use the IBM ViaVoice system for the recognition component of their system, which tries to teach users to discriminate, identify, and produce sounds that are recognized as being problematic for Japanese learners of English. They describe the extensions that they made to the pronunciation dictionary to account for expected mispronunciations, notably the inclusion of variants to reflect insertion of epenthetic vowels in consonant clusters. Although they were successful in modifying the available tools to a degree that suited their purposes, using off-the-shelf recognition software not designed to recognize non-native speech can undermine the effectiveness of CALL systems (Price, 1998).

Communicative systems address relevance and authenticity concerns about CALL by not only evaluating but also understanding and responding to what the user says. The SUBARASHII Japanese tutoring system (Ehsani et al., 1997) allows beginning learners of Japanese to interact with a ficticious Japanese person to perform simple problem-solving tasks. As the goal of SUBARASHII is not to correct speakers' mistakes but rather to give speakers experience using the language, significant flexibility is allowed at the syntactic and lexical level. Within the context of four constrained situations (as an example, one of the situations involves asking whether the ficticious character would like to go see a movie on Friday), the model of acceptable responses from the user is augmented with probable errors and variations in phrasing. This allows the user flexibility in what he is allowed to say (correct sentences are not rejected just because they are not exactly what the model predicted), and even with some errors, the user is able to interact with the system, as he would in real life with a human listener.

During recognition, monophone acoustic models are used, and the search is constrained by the response model. It would not be possible to take advantage of these restrictions in a full conversational system, but in a system in which the topic and direction of the conversation can be highly *constrained* (as is often the case in language classrooms!), Ehsani *et al.* found that "meaningful conversational practice can be authored and implemented and that high school students do find these encounters useful." The recognition accuracy for grammatically correct and incorrect utterances that were in the response model were 80.8% and 55.6% respectively. Recognition accuracy for utterances that were not in the response model was not reported.

2.3 LVCSR

The CALL research described above focused not on improving recognition quality but rather on using speech recognition, in some form or another, to aid language learning. Accurately recognizing heavily accented and poorly formed non-native conversational speech has not been a priority in CALL, perhaps because even with high-quality recognition, analyzing and providing feedback on conversation is very difficult.

In large-vocabulary continuous speech recognition (LVCSR), the objective is to improve the system's understanding of the speaker, not the speaker's language skills. There are acoustic, lexical, and language models in an LVCSR system, all of which can be adapted to more accurately represent non-native speech. The better the representation, the better the recognition (or so one would hope).

An early study of non-native speakers in LVCSR focused on Hispanic-accented English (Byrne et al., 1998). Initial word error rates were extremely high, averaging 73% in an unrestricted-vocabulary task-based test. It is interesting to note how Byrne *et al.* evaluated the skill levels of their speakers. An *intermediate* skill level implied only some reading knowledge of English, yet the speakers were expected to answer questions such as "What is going on here" and "What will happen next," requiring non-trivial conversational skills. *Advanced* speakers required solid reading comprehension, and were assumed to be able to participate in an argumentative dialogue. It is doubtful that the same correspondence between reading and speaking skills would apply to Japanese speakers. Most Japanese learners of English study the language in Japan before coming to the United States, and can have a high level of competency in reading but extremely limited ability to carry on a conversation. The sociological circumstances surrounding Byrne's speakers' acquisition of English doubtless made his classification the correct one for his target population, but it should be noted that the correspondence between reading and speaking competencies will be different for different target populations, and the data collection protocol and ultimate system design should reflect this.

Studies using more constrained tasks or higher-proficiency speakers have had more success in bringing word error rate to a reasonable level. Witt and Young (1999) have shown that for a simple task, fully-trained source and target language model parameters can be interpolated to form a new set of accent-dependent models that perform well on speakers of different native languages. For high-proficiency speakers and speakers of regional dialects, adaptation using source-language data is effective to the point of being sufficient (Schwartz et al., 1997; Beaugendre et al., 2000), and target-language data may also contribute to WER reductions in some cases (Liu and Fung, 2000a).

The lexical model, or specification of the phones that make up a word can be modified to more accurately represent the phone sequences a speaker is likely to utter. It has been shown that data-driven induction of pronunciation variants can be successful for both foreign-accented speakers and regionally-accented native speakers. Humphries and Woodland (1998) derive a pronunciation dictionary for American-accented English by aligning canonical phonetic transcriptions of words to the result of phoneme recognition using American speech and British acoustic models, and training a decision tree on those alignments. The decision tree is then used to generate an American pronunciation dictionary from a British pronunciation dictionary. Amdall et al. (2000) also collected possible transformations by aligning reference to automatically-generated pronunciations, and show how small gains in accuracy for the WSJ non-native speakers can be achieved by pruning the list of word variants based on the probability of the rules invoked for the individual phone transformations. Livescu and Glass (2000) use a similar alignment-to-phone-hypothesis method to derive pronunciation variants for speakers in the JUPITER weather query system. Their objective, like Amdall's, is to model non-native speech in general, as opposed to focusing on a particular L1 group. Fung and Liu (1999), on the other hand, concentrate on English spoken by Cantonese native speakers. Although their approach to pronunciation variant derivation is not described in detail, it appears that they successfully use predictions from a linguist as to what phone substitutions are likely to develop a lexical model that results in improved recognition on the HKTIMIT isolated phone database.

2.4 Multilinguality

Multilinguality in speech recognition systems has received significant attention as real-world systems begin to be deployed. When the primary focus of research was on developing a reasonable model of speech, the actual language used for development was less important than the learning and modeling techniques that were being refined. Certainly, language-specific issues, including tones in Chinese, liaison in French, and vocabulary specification in German, needed to be resolved, but researchers generally concentrated their efforts in modeling their own language and sometimes another widely used language such as English.

As people came to actually want to use these systems, however, the serious overhead involved in developing a recognizer in a new language, and computational costs involved in running multiple recognizers, made systems that could recognize any of a number of languages attractive. A multilingual system typically uses a common phone set to represent all languages it covers, sharing training data across languages when the phones show similar properties, and making the task of adding a new language easier as the language and phoneme inventory of the overall system grows.

While multilingual systems seem at first to be very close to non-native systems, there are several crucial distinctions. In a multilingual system, users are presumed to be native, or at least near-native, speakers of the recognition target language. While they exhibit the variation that always is seen in native speech, they are expert speakers that fully control the syntax, semantics, and phonology of their language. All that we know about pattern modeling for native speakers is valid for the different languages in a multilingual system because the languages are spoken by natives. There are no issues of L1 interference between speaker groups. Multilingual systems do not face the challenge of modeling inconsistent phonological simplifications beyond what is commonly seen in fluid native speech. In a multilingual system, the objective is to sufficiently represent the phoneme inventory of each language, which has been well-studied for all but the rarest languages. The problem of deciding how models should be shared across languages is a significant one, but should be distinguished from the problem of modeling speakers who have a common target, the L2 phone set, but are achieving it with varying degrees of success.

Schultz and Waibel (1999) describe a method for incorporating new phonemic contexts in the allophonic decision tree. Because the phoneme sequences that occur in each new language can be enumerated based on either existing linguistic analysis or expansion of a text corpus to its phonemic representation, those sequences that do not occur in any of the languages already modeled in the system can be specified. The authors adapt the existing decision tree to the new phonemic environments by pruning back the branches affected by the new polyphones and re-growing those parts incorporating the new acoustic data and re-training the associated distributions. Schultz and Waibel report that his method results in a large decrease in WER with only a fraction of the acoustic data that would have been needed to fully train the new polyphones.

Imperl (1999) describes an algorithm for clustering polyphones across languages. He groups together polyphones with a triphone distance under a certain threshold to share training data and greatly reduces the number of polyphones represented in the system with only a small degradation in WER. Köhler (1999) compares three methods for specifying a phoneme inventory for a context-independent multilingual system, finding that density clustering bootstrapped from the IPA representation of phones in different languages outperforms both depending solely on the IPA symbol and using a purely data-driven clustering approach. Köhler discusses the representational difference of these approaches, noting that the best-performing method operates at a sub-phone level, while using the IPA specification alone does not take advantage of this more specialized modeling.

2.5 Data Collection

Several projects have included the collection and recognition of accented speakers. In addition to the Byrne corpus, the Australian National Database of Spoken Language contains data from non-native speakers, both those who were born in Australia but claim a language other than English as their first and those who arrived in Australia after puberty (Millar et al., 1994). Non-native speakers were mostly of South Vietnamese and Lebanese Arabic backgrounds, although representatives of other migrant populations were also included.

Bratt et al. (1998) describes in detail the methodology used by SRI for collection of read data from American learners of Latin American Spanish. The non-native collection was part of a larger project in which many varieties of Latin American Spanish were recorded. Sentences were primarily taken from Spanish newspaper texts and were balanced for length and phonetic coverage. A subset of the 43,460 utterances from the non-native speakers was phonetically transcribed so that systematic pronunciation errors by the non-native speakers, all native speakers of American English, could be analyzed. In their phonetic transcriptions, transcribers were allowed to choose from the union of the Spanish and English phone sets, and were also provided with diacritics to mark ways in which a Spanish phone sounded non-native if the error was more subtle than substitution of an English phoneme. Inter-coder agreement was measured at the phone level, and it was found that [β], [ð], [ɣ], and [ɾ] were the most consistently transcribed as well as good predictors of native pronunciation.

One of the important questions to ask when developing a speech database is how well disfluencies need to be represented. For language model training, we know that examples of common expressions and constructions are needed and must be elicited during data collection. Does the same care need to be taken with disfluencies? It has been observed that although disfluencies are a significant source of error in Switchboard and hesitation words can be used to better predict other words (Shriberg and Stolcke, 1996), better disfluency modeling does not significantly improve recognition accuracy (Stolcke and Shriberg, 1996). Will this also be the case for non-native speakers? How will non-native speakers differ from native speakers in their disfluency patterns? These questions can only be answered by collecting and analyzing data containing disfluencies. Disfluency behavior appears to be similar across English and Swedish (Eklund and Shriberg, 1998), but we cannot be sure whether a similar relationship exists between other language pairs, and if not whether non-native speakers observe L1 disfluency patterns, L2 disfluency patterns, or a combination, and how disfluencies are distributed when the speaker is not fluent in the language being spoken.

Many of the assumptions ordinarily made when collecting speech data are challenged when working with previously unsampled populations. Eskenazi (1997) points out that speaker competence in linguistic skills and reading ability are among the variables that must be recognized when recording data from children and speakers of languages for which there is not a high standard of literacy. I have observed that when recording non-native speakers who are highly literate in their native language, similar variables must be considered, presenting a special challenge for data collection protocol design and execution (Mayfield Tomokiyo and Burger, 1999). One does not wish to frustrate the speaker, as doing so would tend to both compromise the integrity of the data and leave the speaker with negative feelings.

In disciplines in which recording of speakers for the purpose of analyzing patterns in speech has long been common practice, ethical standards have evolved which we might be encouraged to respect, especially when it could be perceived that our interest in the speaker is because his speech is somehow substandard. In his description of the field methodology in the project on linguistic change and variation, Labov (1984) describes a number of issues in spoken data collection, mentioning among other things how important it is that speakers do not come out of the data collection experience feeling that they have been objectified or misunderstood.

The interview is a technique that is frequently used to gather data for the purposes of sociolinguistic research, and it closely parallels scenario-based data collection in the sense that both are contrived situations designed to elicit natural speech that will be transcribed and analyzed. Both suffer from conflicting definitions of what "natural speech" is and whether it can be elicited in the contrived setting (and whether that matters). The primary difference is the amount and breadth of speech sought; for speech system training we need many hours of speech from a variety of speakers, whereas much sociolinguistic research focuses on the speech of just a few speakers. As we expand our data collection endeavors to cover new speaker populations, we would benefit from the insights of researchers in Sociolinguistics, where speakers of non-standard varieties of languages such as English are often targeted.

In an extensive discussion of the interview, Briggs (1986) makes many points that seem relevant to data collection for LVCSR. He emphasizes the importance of understanding the meaning of the *speech event* (an interview, or an interaction with a speech translation system, e.g.) for the speaker. Recording for a research project may be a familiar event for the researcher, but not for the speaker. Reading aloud is commonplace in American schools, but participants of different backgrounds may be intimidated or even offended when asked to read aloud. While native speakers of English certainly vary in their comfort reading and speaking, when the researchers are also native speakers of English, there are far fewer cultural variables that can lead to misunderstanding.

> [The] hiatus between the communicative norms of the interviewer and interviewee can greatly hinder research, and the problems it engenders have sometimes abruptly terminated the interview . . . if the field worker does not take this gap into account, he or she will fail to see how native communicative patterns have shaped responses; this will lead the researcher to misconstrue their meaning. (Briggs, 1986)

The issue of elicitation of natural speech has been given much attention in areas of Linguistics, especially Sociolinguistics, where entire studies can revolve around the speech of just a few speakers, making it crucial that the speech collected truly represents the natural speech patterns of the speaker being studied. Wolfson (1976) defines the notion of natural speech "as properly equivalent to that of appropriate speech; as not equivalent to unselfconscious speech." She suggests that in some situations, it is *natural* to speak carefully, and that careful speech in such contexts should not be considered unnatural. By the same token, for semi-fluent non-native speakers, whether they are at a real information desk or recording a contrived scenario, their speech will most likely be planned. This means that we can probably allow speakers to make notes of what they plan to say (if that makes them more comfortable). It may also mean that we don't need to make as much of a distinction between read and spontaneous speech; it may be the case that for the purposes of training a non-native recognizer, read dialogues and even read texts may be much more useful than they are for training a native system.

Chapter 3

Non-Native Speech Database: Composition And Characterization

The differences between native and non-native speech can be quantified in a variety of ways, all relevant to the problem of improving recognition for non-native speakers. Differences in articulation, speaking rate, and pause distribution can affect acoustic modeling, which looks for patterns in phone pronunciation and duration and cross-word behavior. Differences in disfluency distribution, word choice, syntax, and discourse style can affect language modeling. And, of course, as these components are not independent of one another, all affect overall recognizer performance.

Understanding how native and non-native speech differ at all levels is clearly an important first step in attacking the problem of non-native recognition. In this chapter, I present an analysis of rhythmic and lexical, and to a certain extent syntactic, differences between the native and non-native speech samples I have collected. This analysis is important for speech recognition, but has implications for other areas of natural language processing such as parsing and discourse processing.

This chapter is structured as follows. Sections 3.1, 3.2, and 3.3 describe the protocol used to build a database of clean wide-band non-native speech. Recording, transcription, and annotation conventions will be presented, as well as evaluation of speaker proficiency. In Section 3.5 I present my analysis of the data, describing lexical distribution, speaking rate and pause distribution, disfluencies, reading errors, and grammaticality in the native and non-native speech.

3.1 Data collection

At the time this thesis work began there were some small databases of non-native speech available. In particular, the LDC Wall Street Journal (LDC, 1994a) and Broadcast News (LDC, 1997) databases have non-native components, and are linked to widely-used native databases so results on non-native speech could be compared to those for native speech. However, because both of these databases were limited to read speech, I would not have been able to compare characteristics of read and spontaneous speech for the same speakers. My goal was also to examine patterns in speech of speakers who were of a lower proficiency level than those in the two read news databases. It was therefore my decision to collect my own database of non-native speech, following as closely as possible the data collection conventions used in developing these well-known databases (LDC, 1996a).

3.1.1 Pilot data collection experiments

In order to determine what type of data I would be able to collect, several pilot experiments were run with five local speakers who were acquaintances and known to be of the target English proficiency level. All were native speakers of Japanese. They returned a number of times to complete different tasks, under different recording conditions, and were asked for their reactions to each situation. I defined my target data based on their feedback. This approach is different from the more conventional method of first characterizing the target data and then designing the task and recording protocal to enable collection of that data. My decision to control for the native language and English proficiency of the speakers limited the potential speaker set, however, and I wanted to maximize the extent to which local speakers would be able to participate.

Spontaneous speech

My decisions in spontaneous speech collection were heavily influenced by the experience of colleagues at the Interactive Systems Labs (ISL) at Carnegie Mellon University who have had many years of experience in collecting spontaneous speech from native speakers in a variety of domains. Data that has been collected at ISL is used as a standard database in multi-site speech system development and evaluation (Burger et al., 2000; Ahlen et al., 1997). In most cases, a scenario is designed and speakers are asked to converse freely in the context of that scenario. The scenario can be goal oriented, asking the speakers to schedule a meeting or make a hotel reservation; it can also be free in form, asking speakers to converse about a certain topic. Completely unconstrained speech can be collected by simply recording conversations and discussions with the permission of the speakers.

The non-native speakers who were invited for pilot experiments resisted both scenario-based and unconstrained spontaneous data collection. They cited their lack of confidence in producing English sentences as the primary reason for their discomfort. One of the speakers refused to do either task. The other four agreed to record, but two of the four complained that the tasks were "hard" and "embarassing." Three of the five speakers said independently that they believed word would get out within the close-knit Pittsburgh Japanese community that the task was unpleasant and I would have difficulty recruiting speakers.

The speakers were asked to do a third task in which they were given prompts designed to elicit natural questions on specific topics. They strongly preferred this task. Although a prompted task can unnaturally constrain the types of words and expressions that are used (speakers tend to repeat the phrases in the

Speaker Group	Native	Non-native	
Prompt type	English prompts	English prompts	Japanese prompts
Perplexity	102.6	59.49	32.73

Table 3.1: Perplexity of spoken English as elicited from native speakers, non-native speakers given English prompts, and non-native speakers given native-language prompts. Perplexity is measured with respect to a native language model.

prompt), I hypothesized that because the speakers were already depending on learned templates in their speech, the prompted speech might not be as distant from spontaneous speech as it would for native speakers. I also thought that giving speakers native-language prompts and asking them to formulate English queries might approximate the pseudo-translation process they were already going through in speaking English, and that such prompts might actually elicit fairly natural utterances, whereas giving them English prompts would encourage them to use phrases that they were not familiar with and introduce speech errors that would not ordinarily occur.

As an informal evaluation of this hypothesis, I examined how well a topic-matched language model predicted the English-prompted utterances, the native-language-prompted utterances, and a set of utterances by native English speakers given the same English prompts. A standard measure of language model performance over a corpus is perplexity, which is defined in e.g. (Manning and Schütze, 1999, p.510) to be

$$\text{ppl} = 2^{\frac{-1}{N} log P(w_1 \ldots N)}$$

where N is the corpus size and the probability of the word sequence $w_1 \ldots N$ is approximated, in the case of a trigram language model, as

$$P(w_1 \ldots N) = \prod_i^N P(w_i | w_{i-2}, w_{i-1})$$

.

The results of the evaluation are shown in Table 3.1. The perplexity is dramatically lower for the Japanese speakers, particularly when the prompts are given in Japanese. This could mean that the Japanese speech is much simpler than the native speech; it could also mean that the Japanese speakers are relying on some fixed phrases that are common in native speech. The difference between the English-prompted and native-language-prompted speech suggests, as does an examination of the transcripts, that Japanese speakers are more predictable in their speech patterns when not influenced by the possibly unfamiliar English phrasing in the prompts.[1]

Read speech

In addition to the spontaneous utterances, I wished to collect samples of read speech from each user as well. As it turned out, this was also not straightforward, as speakers resisted reading text that was too difficult.

[1] It should be noted that the subjecs reported no difficulty in *understanding* the English prompts, only that the phrasings were not the ones they would have chosen.

Wall Street Journal articles, for example, were unanimously ruled impossible by my panel. I chose to have speakers read a modified-vocabulary version of the story of Snow White and several articles from an archive of news articles written for children. This was not a completely altruistic decision; certainly, I did not want to antagonize or embarass the speakers, but I also did not want recordings full of repeated attempts to pronounce unfamiliar words, long pauses in the middle of words, and unintelligible segments. There were plenty of these effects in the simpler readings, and asking the speakers to read texts of higher difficulty would only serve to drive up the word error rate and lower my chances of recruiting volunteers.

Conclusions from pilot experiments

The issue of difficulty was not one I had previously encountered in collection work with native speakers. Talking is something most native speakers do comfortably every day, and getting a gift certificate for ice cream or pizza in exchange for a chat appeals to many people, particularly those of the hungry undergraduate persuasion. In working with the non-native speakers, I became aware of a number of assumptions commonly made about data collection that do not necessarily hold when the speakers are not proficient in the language.

There is a limited supply of speakers.
Although recruiting speakers is never easy, the creative recruiter can always find new venues: recruiting speakers in shopping malls, for example, or festivals. There are simply fewer non-native speakers to be found, particularly when the data collection must be controlled for factors such as speaker proficiency or native language.

There is a risk of alienating the community.
Because non-native speaker communities are often quite tight, opinions of the recording project may spread quickly. It is possible that members may react negatively to the project, feeling perhaps that the researcher is asking them to do something unpleasant when the researcher had no such intention. If word spreads that the project should be avoided, it may become impossible to recruit speakers from the target group. The researcher must be sensitive to cultural norms and possible misinterpretations of the purpose of the project.

All speakers are not equally able to perform the task.
While native speakers vary in their abilities to read aloud or extemporize, this variation is limited to the realm of fluent native speech. Non-native speakers range much more in their abilities to perform different tasks, and care must be taken in planning the data collection to account for variation in speaker ability, something which is not ordinarily a factor in native data collection.

The act of speaking, whether careful or not, represents a major cognitive load for the speaker.
There is some cognitive load associated with speaking for native speakers (Lamble et al., 1999). The more thought that must go into completing the task, the higher the cognitive load, and the more likely speech errors are to occur (Grant, 1999). However, this load is far lighter than the one experienced by non-native

speakers. Non-native speakers must often struggle to remember what the rule is for moving the verb to the end of the sentence, or if the word started with an /r/ or an /l/, not just how best to express their thoughts.

The data collection administrator may not understand intuitively how difficult or easy the task is for the speaker.
Native speakers have a general understanding of what is easy and what is hard for other native speakers. Even if they do not expect that reading a certain Wall Street Journal article will be difficult, they will quickly realize it when they see the speaker struggling. The researcher's intuitions may not be correct with non-native speakers, however.

Some speech tasks (read/careful/spontaneous) may be significantly more difficult than others for the speaker
Again, while I recognize that normal native speakers vary in their abilities to complete certain tasks, this variation just does not compare to the variation among non-native speakers. If a non-native speaker has only been educated using read texts, spontaneous speech will be very difficult for him, whereas a speaker who learned primarily by speaking may find reading aloud far more difficult.

The speakers may perceive the task as a test on which they will be evaluated
This was an issue that I faced many times. Because many speakers have learned the non-native language in school, they often feel that they are being tested when they are asked to use it in artificial situations such as data collection. This worry makes the task more unpleasant for them and may affect their speaking performance. Speakers may or may not be more comfortable if this concern is addressed right away, but it is important for the researcher to understand that speakers may be feeling judged.

Speakers may not have a good idea of what they would say in a given situation, and may not have said and heard something similar before
For example, I had assumed that one speaker's hesitations were because he did not know the right words to complete a hotel reservation. It turned out that he had never had to make a hotel reservation even in his native language, and didn't know what sorts of things he could ask for. He had not told me this when we first explained the task; it took some probing afterwards to understand what had happened.

While it is not strictly necessary to ensure that the data collection experience is a pleasant one for the speakers, it is valuable to consider doing so for two important reasons: concern for the feelings of the speaker, and concern for the integrity of the data. With regard to the latter, in collecting speech data for LVCSR, the goal is to obtain samples of speech that are representative of those that would occur when a speaker is using the speech system for its intended purpose. If speakers are feeling embarassed or tense, if they are frustrated, if they are pressured to use words and expressions that they normally would not, the utterances they produce may be quite atypical of their usual speech. As for the former, respecting and being prepared for differences in expectations of the speech elicitation process have long been a consideration in disciplines such as Sociolinguistics that rely on data collected in field interviews (see Section 2.5 for a discussion).

3.1.2 Data collection protocol

Based on the information gathered during the pilot data collection, the decision was made to have one group do the spontaneous recordings and read the fairy tale, and a second group read a number of articles of children's news from the magazine Time for Kids. The collection of articles read by this second group form the Children's News Database (CND) that was designed for this dissertation. Some speakers were willing to do both tasks, so for those speakers there is an element of overlap in the database. Because it was much easier to recruit volunteers, more speakers were recorded for the news reading task, which in turn influenced the decision to focus primarily on read speech in the acoustic modeling portion of this thesis work.

Speaker recruitment

Most speakers were recruited locally in Pittsburgh, although some were recorded in Japan. The local speakers responded to bulletin board postings around the Carnegie Mellon and University of Pittsburgh campuses and electronic mailing list announcements. The speakers in Japan were members of an English conversation club at Kyushu University. The only requirements were that speakers had studied English for at least six years, continued to experience some difficulty in speaking and understanding it, and had not spent more than one month immersed in an English-speaking environment until after graduating from college.

Potential speakers were given a description of the tasks they would be asked to perform and told that it would take between thirty minutes and one hour, and that they would be given a gift certificate to a local merchant of their choice.

Demographic information

Speakers were asked to fill out a form to record their gender, hometown, dialect, exposure to English, and other characteristics. This information is provided in Appendix A.10.

Anonymization and consent

Each speaker was assigned an identification number that was used to store the recorded data and demographic information. These assignments were known only to one researcher, and the anonymization process was explained to the speakers. Each speaker signed a document stating that he or she agreed to be recorded, with an optional release of their recorded data for the purposes of playing excerpts at research presentations. The document also stated that speakers could terminate their participation at any time, and was provided in both English and Japanese.

Scenario 1. Scenario: Going to a restaurant

You read about the Lemongrass Grill in your guidebook and you would like to try it. Find out the following about the Lemongrass Grill:

- Type of food served
- Prices
- Hours
- Reservation needed?
- Distance from the Plaza Hotel
- Transportation back to the Plaza Hotel

Figure 3.1: Excerpt from elicitation scenario given to JL1 speakers

Recording environment

Based on feedback from the pilot speakers, it was decided to have speakers record onto a tape, alone in a quiet room. A digital audio tape (DAT) was used with a Sennheiser headset. Speakers were given the material with an explanation of the task and shown how to operate the DAT recorder. When they felt comfortable with the device and task and had filled out the paperwork, the test administrator left the room. The speakers always knew how to find the administrator, and often came to ask questions. In only one case did the speaker fail to operate the DAT recorder correctly.

Tasks

Speakers participated in the spontaneous task, the read news task, or both. For the spontaneous task, the speakers were given a set of scenarios consisting of an explanation of the setting and a number of prompts for queries. An English example is given in Figure 3.1. The actual Japanese version is provided in Appendix A.6.

Some speakers that participated in the spontaneous task were also recorded reading the story of Snow White, which is provided in Appendix A.7.

For the read task, speakers were asked to read two or three articles from CND . They were told that they should make their best attempt to pronounce any unfamiliar words, and that if they made an error they could either continue or return to the beginning of the sentence. All speakers read one common article, the text of which is provided in Appendix A.4. The remaining two articles were unique to each speaker.

Amount

Fifty-six native speakers of Japanese were recorded. Of these, twenty-five recorded the spontaneous task only, twenty-three recorded the read news task only, and eight recorded both.

Twelve native speakers were recorded under the same conditions. Of these speakers, one recorded the spontaneous task only, five recorded the read news task only, and six recorded both. The final composition of the database is given in Table 3.2.

3.2 Evaluation of speaker proficiency

So that recognition performance on individual speakers could be put into the context of their level of English proficiency, all speakers were evaluated following the guidelines of the *Speaking Proficiency English Assessment Kit* (SPEAK), a standardized evaluation procedure developed by the Educational Testing Service as part of the *Test of English as a Foreign Language* (TOEFL) program (SPE, 1987; Clark and Swinton, 1979). SPEAK provides guidelines for rating non-native speakers of English in four categories: overall comprehensibility, pronunciation, grammar, and fluency. In a full SPEAK test, proficiency in two or more of these categories is assessed for each of six tasks: reading from text, sentence completion, telling a story depicted by a series of drawings, answering questions about what is happening in a single drawing, answering spoken questions, and describing a printed schedule aloud.

The ratings are on a four-point scale, from 0 to 3. The system assumes that speakers are non-native, so a score of 3 allows for some non-native patterns in pronunciation, prosody, or usage as long as the speech is fully comprehensible and closely approximates native speech. In effect, then, this four-point scale is comparable to a five-point one in which the top score is reserved for native speech.

The SPEAK guidelines provide simple but very specific criteria for assigning proficiency scores. These criteria are listed fully in Appendix A.1. The following is an excerpt, listing the criteria for assigning scores in the comprehensibility category; these criteria cover features found not only in read speech but also in spontaneous speech.

Native language	Prompted		Story		News	
	# speakers	# utterances	# speakers	# utterances	# speakers	# utterances
Japanese	33	2257	13	795	31	3802
English	6	436	6	548	10	1419
Chinese	6	375	6	507	—	—

Table 3.2: General information about the non-native speech database

0 Overall comprehensibility too low in even the simplest type of speech.

1 Generally not comprehensible because of frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control.

2 Comprehensible with errors in pronunciation, grammar, choice of vocabulary items or infrequent pauses or rephrasing.

3 Completely comprehensible in normal speech with occasional grammatical or pronunciation errors.

(SPE, 1987, p.16)

For this thesis, speakers were only rated for proficiency in the first task, reading aloud from text. Each speaker was assessed by two qualified SPEAK raters, whose scores were averaged. In cases where the two raters' diagnostic scores differed by more than 0.95, a third rater assessed the speaker and his score was averaged with the score closest to his to obtain the final rating for the speaker (the outlying score was thrown out). Each speaker was rated on three separate passages, and these three scores were averaged to give a final diagnostic score for that speaker in each of three categories: pronunciation, fluency, and overall comprehensibility. All speakers read the same text, which is given in Appendix A.5.

3.3 Transcription and annotation

Recordings were transcribed by one transcriber and validated by at least one second transcriber. The transcription and annotation conventions were based on those used in the LDC transcriptions of spontaneous speech (LDC, 1996b), with some extensions for transcription of read speech errors. In order to make the extended-format transcriptions compatible with the checking program that was used, the surface form is slightly different from those used in the LDC conventions; the types of annotations made are the same, however, and the transcripts can easily be converted to the LDC format.

The transcribers used the TransEdit transcription tool. TransEdit was written by Susanne Burger and Uwe Meier, graduate students affiliated with the Interactive Systems Labs at CMU. TransEdit allows the transcriber to view and segment the speech waveform and either transcribe and annotate the speech from scratch or annotate a prepared text in an embedded editing window. It runs in a Windows environment.

Examples of transcribed read and spontaneous passages are given in Appendices A.8 and A.9.

3.3.1 Read speech transcription

In transcribing read speech, transcribers worked from the same text that the speakers read, transcribing any departures from the original text. They brought the text up in the text editor and as they listened to each recording they annotated the text to reflect what the speaker actually said.

Allowing the transcribers to work from the original text sped up the transcription progress significantly and also increased the accuracy of the transcriptions. A pilot transcription experiment had suggested that while native transcribers tended to miss some types of reading error made by native speakers, this problem occurred only rarely when transcribing the non-native speech. This is probably because many native reading errors are still examples of natural English and are therefore not as noticeable as non-native reading errors. For example, in the following sentence, both native and non-native speakers exchange singular and plural nouns, but the reading error made by the native speaker results in a smooth and gramatically correct sentence and the reading error was not caught until verification.

(3.1)
a. Then Clinton's lawyers will be given twenty-four hours to present the President's side (text)
b. Then Clinton's *lawyer* will be given twenty-four hours to present the President's side (native reading)
c. Then Clinton's lawyers will be given twenty-four *hour* to present the President's *sides* (non-native reading)

Word-level annotations

To produce a word-level transcription, the original text was preserved and any departures were inserted and marked as reading errors, with the scope of any repeated segments indicated. The following forms of error were annotated:

Insertions the speaker inserts a word that was not written in the text.

(a) Will <;ins the> Fox's film sell as many action figures and fast food meals as The Little Mermaid

Deletions the speaker omits a word that was written in the text.

(a) only a hundred <;del years> ago the rivers of Washington State and Oregon were jumping with salmon

(b) in most places fishermen today catch one third fewer Chinook salmon than they did in the early nineteen <;del hundreds>

Substitutions the speaker misreads a word (or words) as another English word (or words).

(a) will all this effort <;1 &effect> be worth it

(b) settlers arrived in the early eighteen hundreds <;2 &eighteens>

(c) the united states has strongly opposed japan's <;1 &united states> whaling practices

Repairs the speaker "rewinds" one or more words to correct something that he said.

(a) restoring salmon populations to healthy levels will be an {-/upstair=/- upstream} struggle for everyone in the area

(b) {-/Colonists may have used <;ins a> copper/- -/used copp=/- colonists may have used copper}

Repeats/retraces the speaker rewinds one or more words repeating exactly what he said, usually to recover his train of thought or to stall while thinking of what to say next.

(a) {+/the/+ +/the/+ the} N M F S must approve these plans but some groups are already taking steps

(b) {-/some travel hundreds <;del of> miles/- <;meta oh> some travel hundreds of miles}

Neologisms the speaker invents a word[2].

(a) Since nineteen ninety five roaming wolves have killed eighty four sheep <;1 ~sheeps> and seven cattle

(b) The ruins of what appears to be Cleopatra's palace lay buried in layers of mud seaweed <;1 ~seawood> and garbage

Mispronunciations as most speakers are strongly accented, words are only marked as mispronounced if they are articulated in a way that cannot be attributed to native language interference. The majority of mispronounced words are words that were unfamilar to the speaker.

(a) ...near the Columbia and *Willamette [w ih l y ax m eh t] river systems

Unintelligible words words that are only recognizable because the transcriber is looking at the original text, or articulated segments that cannot be marked as insertions or deletions because they are not recognizable as words

(a) The rarer the species the higher the price the animal ((fetches)) abroad

(b) Parents learn the truth about (()) how their children were murdered

Word fragments the speaker either stops or starts in the middle of a word

[2]This differs from native neologisms in that the speaker is not inventing a word in order to better convey meaning; the speaker *thinks* that he is using an established word. Words that are marked as neologisms are made up of recognizable morphemes which, while not combining to form an established word, show the speaker's understanding of English morphology and an attempt to find familiar parts in an unfamiliar word.

(a) In {-/nineteen ni=/- nineteen ninety} wildlife inspectors in Bangkok Thailand found six baby o= orangutans wedged into cra= crates

(b) {-/Environment/- -/environmen=/- environment= *pause* {-/=alist=/- =alists }} the government and ordinary folks <;1 &folk> team up to save the salmon

Noise and meta-utterance transcription

Non-human noises such as microphone noise and distortion and environmental noise were marked in the transcripts. Human noises such as breath sounds, coughing, and laughter were also marked.

In addition to the inserted words marked in word-level annotations, speakers often inserted filler words such as "um" and "uh" both in English and in their native languages, and also meta-level expressions such as "oh" and "I'm sorry." These extra-text words were not annotated as insertions, but rather were given a *filler* or *meta* annotation. This distinction was made so that insertions due to misreading could be isolated. Native-language words were marked as such; a transcriber with a familiarity with the native language of the speaker did these annotations.

Phonetic transcription

For a selection of the recordings, phoneme-level transcriptions were produced by transcribers experienced in phonetic transcription. While the transcribers were restricted to the English phoneme set used by the recognizer, they were permitted to add diacritics indicating such effects as r-coloring, devoicing, nasalization, lengthening, release deletion, and aspiration. There were many times that the transcriber could not identify a phone in the legal phone set that resembled the speaker's articulatory production. In these cases, the expected phone given the canonical pronunciation was used and marked as unrecognizable. The phone set used by the transcribers is provided in Appendix C.

3.3.2 Spontaneous speech transcription

Noise- and phone-level transcriptions for spontaneous speech followed the same conventions as were used for read speech. Word-level transcription conventions followed the LDC's transcription manual for CALLHOME[3] (LDC, 1996b) with the disfluency transcription extentions described in Section 3.3.1. Specifically, the following events were annotated: human noises, non-human noises, filler words (hesitation sounds), unintelligible segments (with or without best guess), foreign-language segments, partial words, idiosyncratic words and neologisms, mispronunciations, and asides and meta-level speech.

Transcribing the spontaneous speech was very challenging because in many cases it was difficult to determine what the speaker was trying to say. For example, for one poorly pronounced utterance, the first and second transcribers disagreed on what was said:

[3] CALLHOME is a two-channel telephone speech task, so not all of the annotations allowed in CALLHOME were needed for the non-native transcriptions.

Transcriber 1 {+/should I/+ should I} go {-/four mo=/- with four months} time

Transcriber 2 {+/should I/+ should I} go {-/four mo=/- with four months} start

Neither of these transcriptions made sense in the context of the prompt, which was roughly "ask what to wear" in the scenario "Going to a play." Only after a third transcriber who was very familiar with Japanese-influenced English listened to the utterance many times did it become apparent that the speaker meant to say the following:

Actual utterance {+/should I/+ should I} go {-/formal/- with formal} style

Transcribers were instructed to transcribe what they thought the speaker said. My reasoning was that a speech recognition system's goal is to match the perceptive skill of a cooperative native listener, and that the reference transcript should reflect what a native speaker hears.

3.4 Training/Test set definitions

In this section, the native and non-native speaker sets that will be used in further experiments are specified. These data sets are used for training, evaluation, cross-validation, and analysis.

Training data
Training data sets are used in the training of acoustic models. Chapter 4 will refer frequently to the training data. For this dissertation, training data was only collected for non-native speakers and read speech.

Evaluation data
Evaluation data is also referred to as *test data*. Recognition experiments always report results on only evaluation data unless otherwise specified.

Cross-validation data
Cross-validation data is used when parameters such as word probabilities or language model weights must be estimated on a data set that is disjoint from the training and evaluation sets.

Analysis data
Analysis data is only used in this chapter, in discussions of data characterization. Analysis data sets are not necessarily disjoint from the corresponding training, evaluation, and cross-validation sets.

Partitioning of the non-native read data into training, evaluation, and cross-validation sets was done based on proficiency; 10 speakers who received a SPEAK score of between 1.83 and 2.17 were selected for the test set, and of the remaining speakers, three were arbitrarily selected for the cross-validation set.

Table 3.3 lists the number of speakers and number of utterances for each of these data set types. An ID tag is also given to each data set for ease of reference throughout the dissertation.

Data set ID	Used for	Type of speech	Domain	Native language	# speakers	# utterances
N-E-R	evaluation	read	CND	English	6	339
N-A-R	analysis	read	CND	English	N-E-R used for analysis	
N-A-story	analysis	read	Snow White	English	6	545
N-A-S	analysis	spontaneous	tourist	English	6	334
NN-E-R	evaluation	read	CND	Japanese	10	419
NN-X-R	cross-validation	read	CND	Japanese	3	125
NN-T-R	training	read	CND	Japanese	15	1343
NN-A-R	analysis	read	CND	Japanese	NN-E-R used for analysis	
NN-A-story	analysis	read	Snow White	Japanese	12	717
NN-A-S	analysis	spontaneous	tourist	Japanese	32	2190
C-A-story	analysis	read	Snow White	Mandarin	6	507
C-A-S	analysis	spontaneous	tourist	Mandarin	6	375

Table 3.3: Specifications for training, evaluation, cross-validation, and analysis sets to be used throughout the thesis. Data set NN-E-R is controlled for proficiency

3.4.1 Common article for read speech evaluation

As noted in Section 3.1.2, each speaker completing the CND task read one article in common with other speakers and one or two articles, depending on length, that was unique to that speaker. This test article will be known as CND1; the text is provided in Appendix A.4.

3.5 Transcript analysis

Because speech recognition has only recently reached the point where we can begin to consider recognition of lower-proficiency speech in LVCSR tasks, the distinctive characteristics of non-native speech, the properties that make it different from native speech, have not been well studied.

3.5.1 Lexical distribution

Although non-native speakers of the proficiency level I am examining do not have the range of vocabulary and expression available to them that native speakers do, it is not clear that their speech, either individually or in the aggregate, could be described as more *restricted* than that of native speakers. In the context of a certain task, native speakers often rely on standard words and phrases, whereas non-native speakers, perhaps performing the task for the first time, may each come up with a unique way to ask the same question. For example, when prompted to ask about dress, most native speakers responded with "what should I wear," while non-native speakers were more creative with their queries:

(3.2) Do we need to wear the formal dress or we can wear the casual one?

(3.3) What kind of clothes do I have to wear for there?

(3.4) In what kind of dresses should I go there?

(3.5) What should I wear to go there?

If we consider this tendency in the context of Jackson's (1932) discussion of "old, well-organized" and "new, now organizing" speech as described in Goldman-Eisler's (1958) observations that in utterance segments of the former type, words are far more predictable than those in segments of the latter type, the hypothesis that the proportion of now-organizing speech is much greater in non-native speech is further motivated, if not explicitly supported.

Pawley and Syder (1983), too, examine "the puzzle of nativelike selection." Although they do not present a statistical analysis, they argue convincingly that "by far the largest part of the English speakers' lexicon consists of complex lexical items including several hundred thousand lexicalized sentence stems" (p.215), showing how such an interpretation of the mystery of nativeness explains how native speakers select "natural and idiomatic" sentences from among those provided by a generative grammar without requiring changes to existing models of English grammar.

In this section, a number of perspectives on the question of how lexical items are distributed in spontaneous non-native speech are presented. It should be noted that the corpora I am examining are very small and not strictly suited to statistical analysis. Nevertheless, by looking at properties like word frequency and corpus entropy it is possible to gain some intuition about the character of non-native speech. One may also make predictions about the behavior of non-native speech by comparing early trends to documented observations about native speech.

Word frequencies

Table 3.4 shows the frequency rankings and occurrence rates of the top 25 words in both the JL1 and native prompted corpora, along with the frequency rankings in the other corpus. For example, the word "could" was the 8th most frequent word in the native corpus, but ranked 105th in the JL1 corpus. "The," on the other hand, ranked first in the JL1 corpus and second in the native corpus.

These frequencies tell us that there are some expected similarities and some striking differences in the way individual words are used by the two speaker groups. Function words such as "the" and " to", and pronouns like "I" and "you" are among the most frequent words in both the corpora. A closer look, however, reveals differences even in the distributions of words with equal ranks. For example, occurences of "the" account for nearly twice as large a percentage of the JL1 corpus as they do in the native corpus. It appears that there are two reasons for this: hypercorrection and structural choices. Many of the instances of "the" in the JL1 corpus are incorrectly used - either no article is necessary, or another word like "a" or "my" would have been more appropriate. The JL1 speakers also tend to use noun phrases where a native speaker would have chosen something else. For example, many of the JL1 speakers asked "What is the cost?" where a native speaker would have said "How much is it?"

"Go" and "get" have rankings that are almost the exact opposite of each other in the two corpora. The contexts in which these words are used are almost identical, but native speakers show a preference for the construction "How do I get to the hotel" and JL1 speakers for the construction "How do I go to the hotel."

Both of these are correct in the grammatical sense; however, a language model trained on native speech is not going to assign as high a probability to the latter as one trained on JL1 speech might.

The word “which” ranks 25th in the JL1 corpus, but only 306th in the native corpus. This is evidence of a strong tendency on the part of JL1 speakers to use non-restrictive relative clauses where native speakers would omit the relative pronoun or use a modifier.

(3.6) a. Please give me the name of the restaurant which is near my hotel.
(Non-native)

b. Are there any good restaurants near the hotel?
(Native)

(3.7) a. What is the leaving time of the return train which is the final one?
(Non-native)

b. What time is the last train back?
(Native)

The words “tell” and “could” are both approximately ten times as frequent in the native corpus as in the non-native corpus. This is partly because native speakers make heavy use of the expression “could you tell me…” in their queries. Although the sentence “Where is the Empire State Building?” is perfectly grammatical, it would probably sound abrupt coming from a native speaker unless he and the person at the (imagined) information desk were already looking at a map and discussing directions. “Tell” is used only rarely by the JL1 speakers, who show a preference for words like “show” and “teach.” This may be an avoidance strategy stemming from confusion about usage of the words “say,” “speak,” “talk,” and “tell,” which English learners of many different language backgrounds report. It also may be evidence of direct translation from Japanese.

The examples that have been given in this section are very specific. The purpose of raising them was not to prove that non-native speakers always use “go” more than “get,” or avoid complex modal forms, although that may be the case. Rather, the objective was to show that there are consistent and significant differences in the distribution of words in the native and non-native speech samples that have been collected, and that there are possible linguistic bases for the divergence. The question of whether these observations hold for other types of non-native data and how they can be exploited in modeling non-native speech is left to future exploration.

Frequent words in JL1 speech			Frequent words in native speech		
Word	JL1 corpus	Native corpus	Word	Native corpus	JL1 corpus
THE	1 (8.37%)	2 (4.87%)	I	1 (5.36%)	6 (2.78%)
TO	2 (4.32%)	3 (4.74%)	THE	2 (4.87%)	1 (8.37%)
IS	3 (3.84%)	7 (2.07%)	TO	3 (4.74%)	2 (4.32%)
HOW	4 (3.59%)	8 (1.99%)	YOU	4 (2.59%)	7 (2.57%)
AND	5 (2.82%)	10 (1.89%)	ME	5 (2.07%)	49 (0.45%)
I	6 (2.78%)	1 (5.36%)	IS	6 (2.07%)	3 (3.84%)
YOU	7 (2.57%)	5 (2.59%)	HOW	7 (1.99%)	4 (3.59%)
WHAT	8 (2.24%)	22 (0.94%)	COULD	8 (1.97%)	105 (0.19%)
CAN	9 (2.08%)	18 (1.19%)	AND	9 (1.89%)	5 (2.82%)
GO	10 (1.60%)	29 (0.73%)	GET	10 (1.75%)	22 (1.02%)
IT	11 (1.56%)	13 (1.51%)	A	11 (1.56%)	20 (1.09%)
DO	12 (1.53%)	22 (0.59%)	IT	12 (1.51%)	11 (1.56%)
STREET	13 (1.48%)	15 (1.35%)	TELL	13 (1.37%)	120 (0.15%)
DOES	14 (1.42%)	91 (0.27%)	STREET	14 (1.35%)	13 (1.48%)
OF	15 (1.36%)	16 (1.21%)	OF	15 (1.21%)	16 (1.36%)
WHERE	16 (1.31%)	47 (0.46%)	SO	16 (1.19%)	123(0.15%)
FROM	17 (1.28%)	61 (0.40%)	CAN	17 (1.19%)	9 (2.08%)
THERE	18 (1.16%)	19 (1.16%)	THERE	18 (1.16%)	19 (1.16%)
A	19 (1.09%)	12 (1.56%)	THAT	19 (1.08%)	51 (0.43%)
RESTAURANT	20 (1.03%)	34 (0.59%)	WELL	20 (0.97%)	107 (0.17%)
GET	21 (1.02%)	11 (1.75%)	WHAT	21 (0.94%)	8 (2.24%)
MUCH	22 (0.99%)	111 (0.22%)	LIKE	22 (0.94%)	98 (0.20%)
TICKET	23 (0.93%)	– (0.00%)	HOTEL	23 (0.89%)	47 (0.48%)
TIME	24 (0.91%)	128 (0.16%)	IN	24 (0.86%)	27 (0.85%)
WHICH	25 (0.86%)	306 (0.03%)	AT	25 (0.75%)	75 (0.29%)

Table 3.4: Word frequencies in prompted speech: frequency rankings and occurrence rates

Common n-grams

The idea of individual word frequencies as an indicator of distance between corpora can be extended to word sequences, which give us more information about how the words are used in context. While the occurrence frequencies are much lower, and the number of unique types much higher than for individual words in the non-native sample, one can still see patterns that suggest ideas for future modeling of non-native word usage.

The most frequent trigram in the non-native data, "where is the," never appeared in the native data at all. This is further evidence that the JL1 speakers favor simple questions where native speakers prefer embedded forms. "Is there any" is another trigram that shows much lower frequency in the native data, although it is part of a generic question (unlike "of fine arts," which is clearly well-represented only because the native speakers were all speaking in the context of a scenario that takes place at the Museum of Fine Arts[4]). It turns out that fully 65% of the instances of this question use "any" improperly with a singular countable (non-mass) or plural noun, an event which occurred only 4% of the time "[be] there any" appeared in the native sample. For example, the Japanese speakers often formed questions like "is there any restaurant

[4]The native speakers in this data set were all given the same scenario. This meant that a number of n-grams appeared frequent only because the speakers were talking about the same thing. For the non-native recordings, the place names in the scenario were modified after every 10 speakers.

Frequent trigrams in JL1 speech			Frequent trigrams in native speech		
Word	JL1 corpus	Native corpus	Word	Native corpus	JL1 corpus
WHERE IS THE	1 (0.58%)	–	YOU TELL ME	1 (1.20%)	6 (0.43%)
CAN I GET	2 (0.54%)	52 (0.18%)	COULD YOU TELL	2 (1.08%)	18 (0.24%)
DO YOU KNOW	3 (0.52%)	52 (0.18%)	GET TO THE	3 (0.69%)	50 (0.13%)
DOES IT TAKE	4 (0.46%)	32 (0.24%)	I'D LIKE TO	4 (0.63%)	12 (0.33%)
HOW LONG DOES	5 (0.44%)	77 (0.15%)	TELL ME HOW	5 (0.54%)	41 (0.14%)
YOU TELL ME	6 (0.43%)	1 (1.08%)	TO GET TO	6 (0.51%)	164 (0.06%)
IS THERE ANY	7 (0.43%)	130 (0.09%)	MUSEUM OF FINE	7 (0.48%)	106 (0.09%)
LONG DOES IT	8 (0.42%)	48 (0.18%)	I NEED TO	8 (0.48%)	94 (0.10%)
WHAT IS THE	9 (0.38%)	130 (0.09%)	THE MUSEUM OF	9 (0.45%)	40 (0.15%)
TO GO TO	10 (0.37%)	19 (0.27%)	OF FINE ARTS	10 (0.45%)	92 (0.10%)
HOW MUCH IS	11 (0.37%)	248 (0.06%)	SO COULD YOU	11 (0.42%)	280 (0.04%)
I'D LIKE TO	12 (0.33%)	3 (0.63%)	IT TAKE TO	12 (0.39%)	14 (0.28%)
WHAT KIND OF	13 (0.30%)	248 (0.06%)	MORE INFORMATION ABOUT	13 (0.36%)	1274 (0.01%)
IT TAKE TO	14 (0.28%)	11 (0.39%)	PLEASE TELL ME	14 (0.33%)	280 (0.04%)
GO TO THE	15 (0.26%)	41 (0.21%)	LIKE TO GO	15 (0.33%)	43 (0.14%)
MUCH IS THE	16 (0.25%)	248 (0.06%)	TAKE TO GET	16 (0.30%)	164 (0.06%)
WHAT TIME DOES	17 (0.24%)	–	I GET TO	17 (0.30%)	95 (0.10%)
COULD YOU TELL	18 (0.24%)	1 (1.20%)	COULD YOU EXPLAIN	18 (0.30%)	155 (0.05%)
CAN I BUY	19 (0.23%)	–	WILL IT TAKE	19 (0.27%)	280 (0.04%)
I GET THE	20 (0.22%)	655 (0.03%)	TO GO TO	20 (0.27%)	10 (0.37%)
HOW MUCH DOES	21 (0.22%)	248 (0.06%)	TO GET THERE	22 (0.27%)	50 (0.13%)
HOW FAR IS	22 (0.22%)	248 (0.06%)	THE INTERSECTION OF	22 (0.27%)	1274 (0.01%)
TURN TO THE	23 (0.20%)	248 (0.06%)	TELL ME WHERE	23 (0.27%)	164 (0.06%)
HOW TO GET	24 (0.20%)	56 (0.15%)	LONG WILL IT	24 (0.27%)	280 (0.04%)
WHAT TIME IS	25 (0.19%)	130 (0.09%)	I WANT TO	25 (0.27%)	37 (0.16%)

around here" and "is there any good sight points" whereas native speakers reserved "is there any" for mass nouns ("Is there any seafood on the menu?") and paired plural nouns with *are*: "Are there any restaurants nearby?"

Perplexity and Entropy

As mentioned in Section 3.1.1, the perplexity of the non-native queries was lower than the native queries with respect to a language model trained on native speech. In other words, given a two-word history, the language model was better able to predict the words in the JL1 speech than in the native speech. This observation is also true at the individual speaker level, although there is far more variance in perplexities of the JL1 speakers, as can be seen in Figure 3.2. It should be noted that the individual speaker corpora are very small ($\sim$ 750 words).

To gain an understanding of how the non-native speakers differ from each other in their use of English, I examined the Kullback-Leibler (KL) divergence (Manning and Schütze, 1999, p.72) in the frequencies of words, word trigrams, and part-of-speech trigrams. While KL divergence does not tell us exactly where the distributions of words and n-grams differ, it does give us an idea of the magnitude of the difference. KL divergence is defined as

$$D(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)}$$

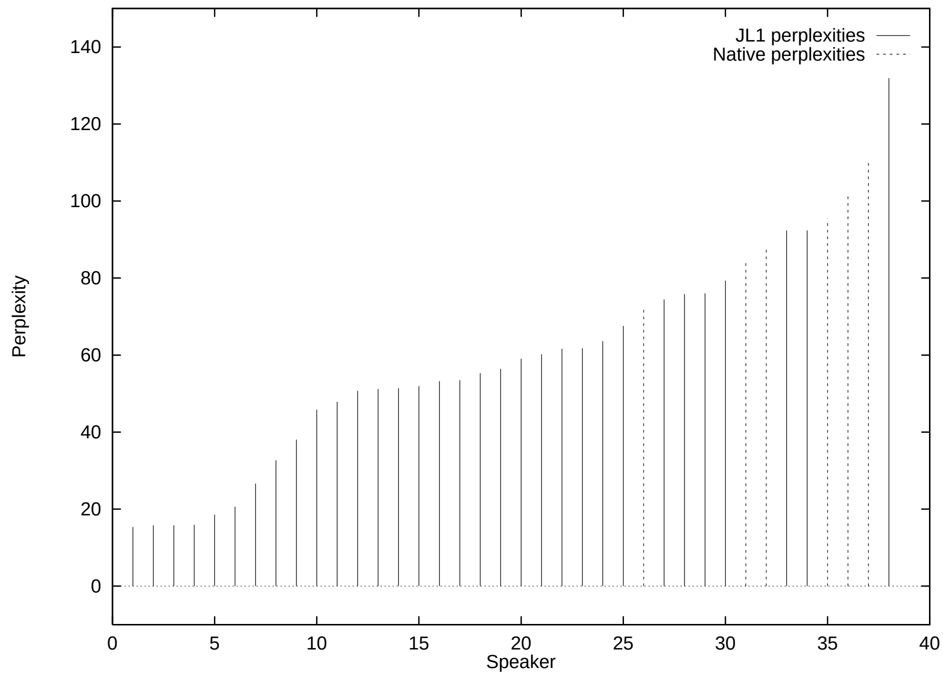


Figure 3.2: Native and non-native speaker perplexities with respect to a language model trained on in-domain native speech. Speakers are listed along the x axis in order of increasing perplexity

which represents the difference between modeling a distribution with the correct probability mass function p and the incorrect function q. To calculate word-level KL divergence, the frequency of each word type that appeared in all of the JL1 and native data, a total of 996 word types, was computed for each speaker corpus. The frequencies were normalized by the number of word tokens in a corpus to obtain a distribution for that corpus. Smoothing was then applied to distribute a probability mass of .01 across the words that did not occur in that corpus. Using these frequency distributions, I was able to measure the word-level KL divergence between two corpora.

Word-frequency-level divergence is straightforward to measure, but may say less about how the speaker uses language than the breadth of his vocabulary with respect to a fixed domain. Two native speakers well-versed in the terms commonly used in making travel arrangements, for example, may tend to use the same sorts of words and expressions in forming queries, leading to low divergence between their speech. Two non-native speakers unfamiliar with the discourse conventions in a given domain and with vocabularies limited to distinct sets of words, on the other hand, may diverge much more in their lexical choices.

Trigram-level divergence captures differences in language use better, but because of the size of the corpora there were very few trigrams with significant probability mass. Measuring divergence at the part-of-speech level reduces the number of unique types to be compared, possibly allowing a tighter model of each speaker's speech. Computation of word trigram and part-of-speech trigram KL divergence were set up as described above, with the trigram frequencies replacing the word unigram frequencies. Part-of-speech tagging was

Comparison (p-q)	Word unigram	Word trigram	Part-of-speech trigram
native-native	1.04	9.61	4.48
nonnative-native	3.06	13.67	7.25
nonnative-nonnative	1.99	12.46	6.60

Table 3.5: Kullback-Leibler divergence (relative entropy) of word and part-of-speech n-gram frequencies between native and non-native speaker corpora

done using the MXPOST tagger (Ratnaparkhi, 1996); ungrammatical sentences in the spontaneous speech did not appear to be affecting tagging accuracy.

Table 3.5 shows the average divergence for native and non-native speakers, both inter- and intra-group. When computing the intra-group divergence, divergence between each speaker corpus and all the others combined were calculated; these divergences were then averaged. The divergence between the native and non-native corpora were consistently higher than the intra-group divergences. Divergence between non-native speakers was also very high in all measurements. This is evidence that non-native speakers are more different from each other in the way they use language than native speakers are.

Vocabulary growth rate

The vocabulary growth rate measures the number of unique words that are introduced as the corpus grows. When the corpus is small, each new text (article, collection of utterances, etc.) contains many word forms, words that have not been seen before. As more text is added, the growth rate slows, since many of the words in the new texts already appear in the corpus. The vocabulary growth rate varies for different types of corpora – a corpus of bus schedule queries, for example, would have a slower growth rate than a corpus of unrestricted spontaneous speech. The difference between vocabulary growth rates in different languages can be large; for comparable corpus types, the vocabulary growth curve in English reaches saturation earlier than it does in more highly inflected languages like Spanish and agglutinative languages like Turkish.

Vocabulary growth rates are compared across languages and tasks in Figures 3.3 and 3.4. Figure 3.3 shows how the difference between English and Spanish vocabulary growth rates remains similar across tasks: in broadcast news, conversational speech, and meeting scheduling, the rate of introduction of unique words is consistently slightly higher in Spanish than in English. The vocabulary growth rate is highest for broadcast news and lowest for task-oriented dialogues (meeting scheduling). The discrepancy between the curve pairs is greatest for the most restricted tasks, possibly because gender and number agreement requirements in Spanish result in many word forms even when set phrases account for a large proportion of the dialogue.

Figure 3.4 compares vocabulary growth rates for five languages in the single task of meeting scheduling. German and Spanish have more extensive inflectional and compounding systems than English does, producing faster vocabulary growth. While only a small amount of data was available for Japanese and Korean, it is evident even from the part of the curve that is shown that the rate of introduction of new words is extremely high. These trends are highly dependent on how vocabulary items are defined, however. For languages like

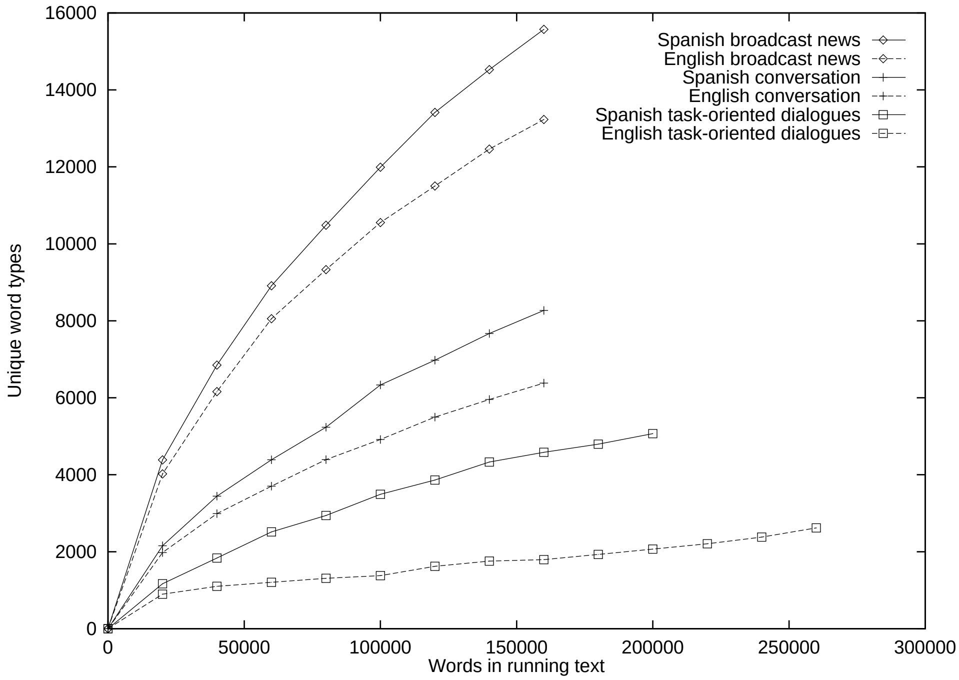


Figure 3.3: English and Spanish vocabulary growth for different tasks. The number of unique word types is shown as a function of the number of word tokens in the corpus (Geutner, 1995)

Japanese and Korean which have no spaces or few spaces in their written form, a choice has to be made during transcription about how the text units will be segmented. In the data that was used for the charts in Figure 3.4, the Japanese and Korean data was segmented at the *bunsetsu* level, which corresponds roughly to a noun or verb plus an article, but can contain modifiers or noun-verb sequences as well. Naturally, this type of segmentation results in a very high vocabulary growth rate. When the Japanese meeting scheduling data, for example, is segmented morphologically, the growth rate looks much like the English growth rate shown in Figure 3.4. However, this is not an entirely fair comparison, as the English text has not been segmented morphologically. In any event, the message to be derived from these curves is that languages with similar properties show similar vocabulary growth rates, and the differences between languages are consistent across tasks. It can therefore be surmised that native and non-native English should have similar vocabulary growth curves, and if they do not there is some fundamental property distinguishing them.

A comparison of the vocabulary growth rate in the spontaneous portion of my non-native database with native databases of similar size and content is shown in Fig. 3.5. The native data from my spontaneous database is also shown, although there is only a small amount. The two larger native corpora that I can compare the non-native transcripts to are a collection of from interactions at an information booth and a collection of hotel reservation and travel planning dialogues. In the information booth dialogues only the query side utterances were used to calculate vocabulary growth rate; for the travel dialogues the side information was not available so all utterances were used.

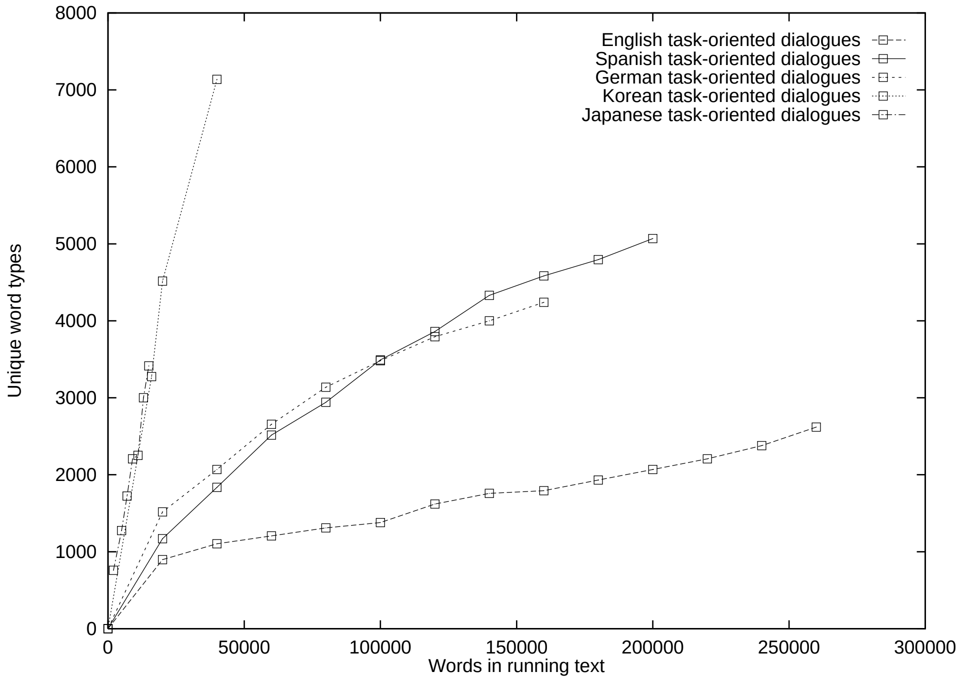


Figure 3.4: Vocabulary growth in five languages for the spontaneous scheduling task

As can be seen in Fig 3.5, the vocabulary growth curve of the non-native tourist queries is similar in shape to that of the native travel dialogues, and the native tourist queries seem to be following the same trend. Vocabulary growth in the information booth dialogues is very fast; this is probably because the locations that the travelers are asking about are not restricted, meaning that each new query may introduce not only a new proper noun but also new adjectives describing it, nearby landmarks, and other unseen words. The travel dialogues, which are scenario based, offer a better comparison for this reason; a higher vocabulary growth rate for the non-native speakers in a controlled scenario-based task tells us that the non-native speakers are using more unique words to express the same thing, while the same difference in an unrestricted task may only mean that the speakers are asking about different topics.

When calculating the vocabulary growth rates, the transcripts from all the speakers were appended in the order in which they were recorded. Because the vocabulary growth curve is fairly smooth, it does not appear to be the case that each speaker uses a radically different set of words - if they did, we would see lurches in the curve where a new set of transcripts was introduced. Fig. 3.6 shows what the vocabulary growth curve for the non-native data would look like if the utterances were introduced in random order. Although the curves are similar, the randomized curves are steeper where the corpus is small. It is likely that this indicates that the speakers are indeed using slightly different words and expressions; because the utterances are in random order, there may be utterances from many speakers in the first 2000 words of the corpus, introducing a wider variety of words early on.

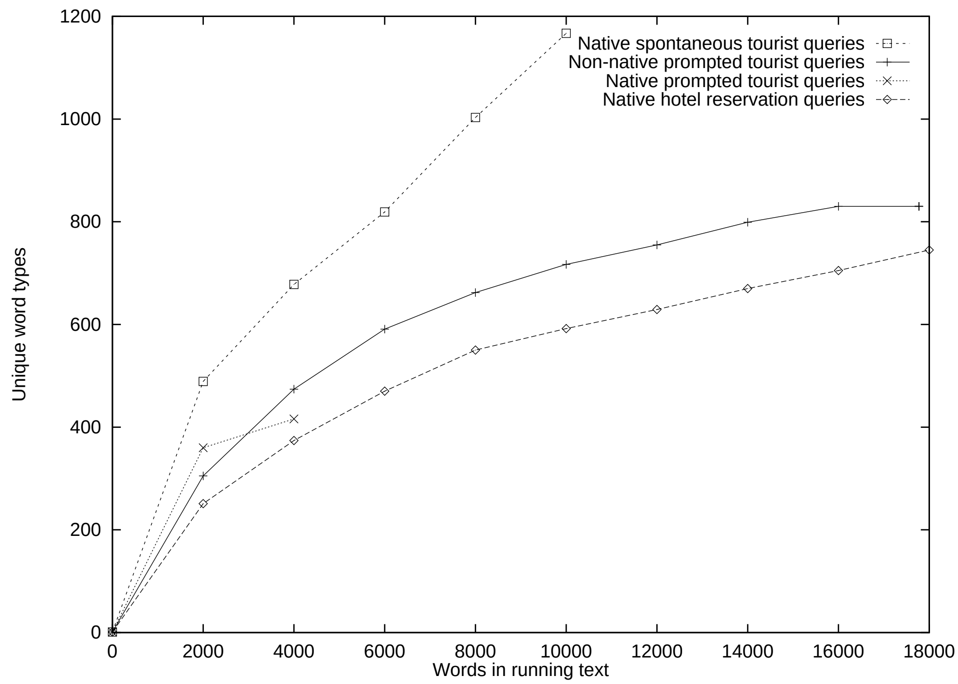


Figure 3.5: Vocabulary growth rates for native and non-native tourist domain speech.

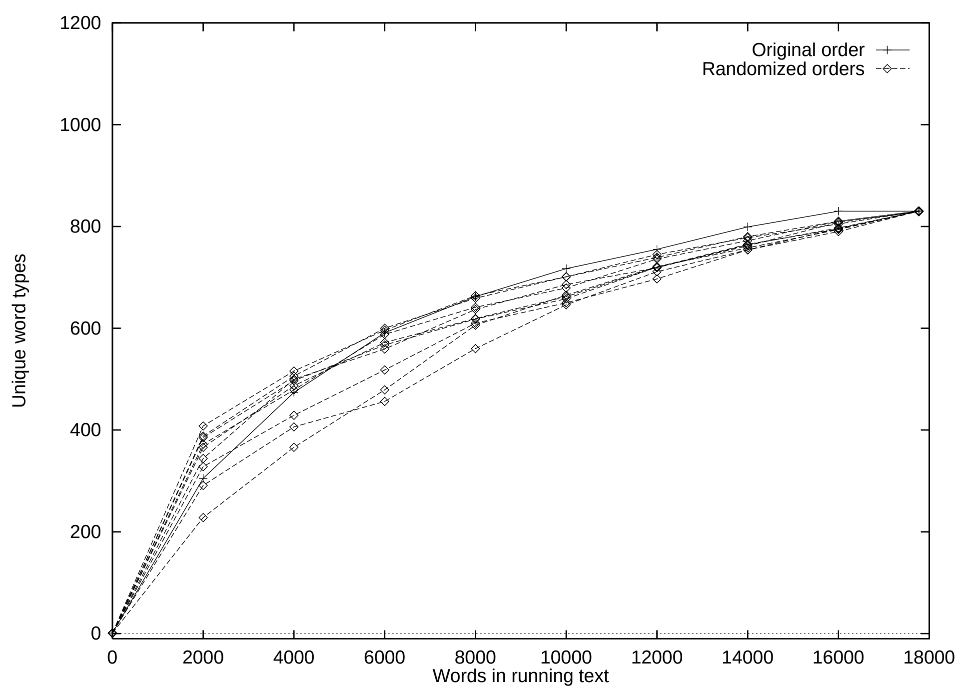


Figure 3.6: Vocabulary growth in the non-native corpus with utterances introduced in randomized orders

		Non-native		Native	
Phrase	Contraction	occurrences	percent	occurrences	percent
can not	can't	6 / 7	85.71	3 / 3	100.00
did not	didn't	3 / 4	75.00	3 / 8	37.50
do not	don't	30 / 31	96.77	9 / 10	90.00
does not	doesn't	3 / 3	100.00	2 / 2	100.00
going to	gonna	9 / 17	52.94	1 / 3	33.33
i am	i'm	35 / 42	83.33	23 / 44	52.27
i have	i've	3 / 26	11.54	0 / 15	0.00
i will	i'll	16 / 25	64.00	3 / 3	100.00
i would	i'd	59 / 67	88.06	13 / 31	41.94
it is	it's	35 / 43	81.40	4 / 4	100.00
that is	that's	17 / 23	69.57	6 / 14	73.91
there is	there's	2 / 8	25.00	4 / 5	80.00
want to	wanna	6 / 40	15.00	2 / 11	18.18
what is	what's	19 / 77	24.68	2 / 5	40.00
where is	where's	10 / 82	12.20	0 / 2	0.00
you are	you're	1 / 4	25.00	0 / 0	-
you will	you'll	6 / 13	46.15	0 / 0	-

Table 3.6: Contracted forms in native and non-native speech. The number of occurrences of each base form is given along with the number of times it is contracted (contracted / total)

Contractions

Table 3.6 shows the most common contracted words and simplified forms in the native and non-native samples. Because the native corpus is small, some of the occurrences of the base forms are very low. It is interesting, though, to see both the difference in the rates of occurrence of some contractable base forms and the rates at which the more common base forms are contracted. In most cases where there is a significant difference between native and non-native rates of contraction, it is because one speaker set or the other is using the expression in a context where it is not contractable. For example, depending on the syntactic role, "I am" can be contracted ("I'm going to the station") or not ("Can you tell me where I am?")

Notably, "I am" occurs proportionally much more frequently in the native data, yet the contraction rate is lower than in the non-native speech for the reason described above.

As noted in Section 3.5.1, the non-native speakers in my sample showed a strong preference for simple questions like "where is the train" over embedded questions such as "can you tell me where the train is," accounting for the difference in occurrence rates of "what is" and "where is", for example, in contractable contexts.

3.5.2 Speaking rate and pause distribution

Features describing the pace and fluency of speech are another point of contrast between native and non-native speakers. In Table 3.7, the word rates, silence rates, average phone durations, and average pause durations are listed for the native and non-native speakers in my data.

Because questions of timing can be highly speaker-dependent, and I wished to contrast read and spontaneous speech, these calculations were done for a small set of 12 native speakers of Japanese who both read the Snow White story and completed the spontaneous task. In addition to the native speakers of Japanese, figures for six native speakers of Chinese are also shown.

The word rate is the number of words the speaker utters per second, not including silences. Not surprisingly, the native speakers consistently speak with a higher word rate than the non-native speakers, although the effect is less pronounced for the read speech than the spontaneous speech. The other three features shown answer the question of whether this is due to quicker articulation of individual phonemes, fewer pauses between words, or both. The silence insertion rate is the ratio of silence elements to words. For example, if the speaker says

"Once upon a time <pause> in a great castle <pause>, a Prince's daughter <pause> grew up happy and <pause> contented, in <pause> spite of a <pause> jealous <pause> stepmother."

the silence insertion rate is $7/22 = .32$. The silence insertion rates for the two non-native groups are similar, and in both read and spontaneous speech are approximately twice that of the native speech. All speaker groups show a significantly higher silence insertion rate in the read speech than in the spontaneous speech.

Neither the phone durations nor the pause durations differ significantly when comparing read and spontaneous, and native and non-native speech. The difference in speaking rate, then, is almost wholly due to the number of inter-word pauses present in the non-native speech. This has clear consequences for speech recognition: because non-native speakers are relaxing the vocal apparatus between words, the cross-word coarticulatory effects present in native speech will not be as consistently realized in non-native speech. Inter-word silence is triggered by a complex collection of factors that are not necessarily related to the phonological environment, such as difficulty of and familiarity with the word, overall comprehension of the text, and fatigue. The same cross-word phoneme pair that saw a pause inserted three sentences earlier may be read with native-like elision when the words involved are easier or the sentence is shorter, meaning that modeling non-native cross-word behavior may not be as straightforward as just turning off cross-word modeling.

	word rate		**silence insertion rate**		**phone duration**		**pause duration**	
speaker	spont	read	spont	read	spont	read	spont	read
Japanese	2.42	2.33	0.17	0.49	0.11	0.11	0.10	0.09
Chinese	2.70	2.28	0.18	0.47	0.11	0.11	0.10	0.12
Native	4.01	3.84	0.10	0.22	0.08	0.07	0.10	0.11

Table 3.7: Speaking rate and pause distribution statistics for non-native speakers. The word rate is reported in terms of words per second. The silence rate is a silence-to-word ratio. Average phone duration and pause duration are measured in seconds.

3.5.3 Disfluencies

It has been observed that native spontaneous speech contains many instances of abandoned words, stutters, restarts, repetitions, filler words, and other disfluencies, some of which occur systematically enough to warrant incorporation in the language model (e.g. Shriberg and Stolcke, 1996). Disfluencies often occur when the speaker is searching for the right word or expression, or is pronouncing a word that is difficult to articulate; they can also occur when the speaker is reading aloud and comes to a word that he does not know how to pronounce, or simply trips over his tongue. Native speakers may attempt to repair prosodic errors when they reach a point in the sentence where they realize that they have used inappropriate stress placement or intonation. Non-native speakers may go back to re-read a phrase when they have stumbled over an unfamiliar word. For both native and non-native speakers, read speech is not always smooth.

Figure 3.7 shows graphically the difference in native and non-native (JL1 only) speaker rates of repair, repetition, fragments, and filler words in the read news data. A disfluency *rate* is defined as the number of times the disfluency occurs per hundred words:

$$\frac{\#\text{ of disfluencies}}{\#\text{ of words}} \cdot 100$$

The JL1 speakers show significantly higher rates of all types of disfluency that were measured. Interestingly, although the non-native retrace rate was over three times the native retrace rate, the retrace *length*, or the number of words that the speaker "rewinds" after an interruption, is similar for native and non-native speakers. This retrace rate agrees with those reported by Eklund and Shriberg (1998), who found parallel disfluency patterns in native Swedish and English speech.

3.5.4 Reading errors

Although in a read speech task the speaker's utterance is supposed to match what is written on the page, there are often many discrepancies. This is particularly problematic in applications where the search is constrained to follow an expected word sequence. In my database, the non-native speakers showed significantly higher rates of both disfluencies and reading errors.

For the purposes of this work, a *reading error* is defined as the deletion of a word that was part of the text to be read, the insertion of a word that was not in the text, or the substitution of one word for another. These errors occur in both native and non-native speech. When native speakers read aloud from text, they may absorb an entire phrase or sentence at a glance and repeat it from short-term memory. Although this conversion is almost instantaneous, the encoding and decoding process (i.e. visual to semantic to acoustic) can introduce error. A secondary source of reading error in native speech is the layout of the text on the page. Native speakers seem more likely to make errors at line boundaries and when the text is presented in very narrow columns, although this has not been formally analyzed here.

In the non-native speech samples analyzed in this dissertation, the speakers appear to read one word at a time; they often pause between words (which contributes to the high silence rate) and do not show the same

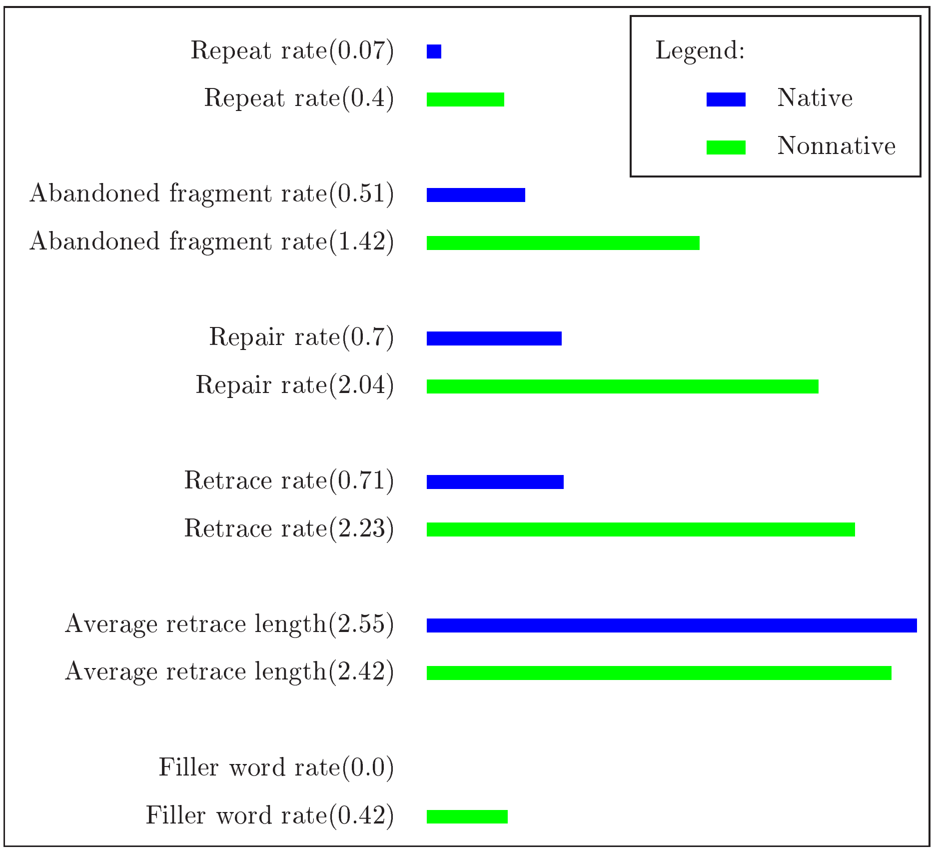


Figure 3.7: Disfluency rates for native and non-native (JL1) speakers in the CND reading task

Error type	Non-native	Native
Morphological variant	55.74%	21.27%
Orthographically similar	27.76	48.93
Semantically similar but orthographically distant	0	8.51
Misread numeral	3.91	2.1
Neologism	3.42	0
Function word substitution	3.32	12.50
A-the	1.66	6.25
Other	4.20	0
Total number of errors	1555	47

Table 3.8: Breakdown of non-native and native misread words

tendency to substitute semantically similar words or phrases. In fact, many substitutions are completely inappropriate semantically, indicating that the speaker does not understand what he is reading. While semantically inappropriate substitutions do occur in native speech – most native speakers have experienced reaching the end of a passage of text with the realization that they have no idea what they have just read! – they are much less frequent.

Of the 21,958 words in the entire native data base, there were only 8 inserted words, 9 deleted words, and 57 misread words, an average of .39 extra-text words per 100. In contrast, in the 67,669-word JL1 subset of the non-native database, speakers averaged 2.77 extra-text words per 100. A breakdown of the main categories of misread words is shown in Table 3.8. Numbers for native speakers are shown for reference, but as the number of actual native reading errors was very small, this distribution may not be representative of the actual distribution in native speech.

Substitution of a morphological variant

Although native and non-native reading errors fell into the same general categories, the errorful native sentences were far more likely to be semantically meaningful and syntactically correct than the errorful non-native sentences. For example, the following two sentences both contain examples of singular-plural substitutions.

(3.8) Native morphological substitutions

a. Doctors are studying the pill's *effect* on patients
(original text)

b. Doctors are studying the pill's *effects* on patients
(spoken)

(3.9) Non-native morphological substitutions

a. *American students* perform poorly on standardized tests
(original text)

b. *American student* perform poorly on standardized tests
(spoken)

Function word substitution

Another category of error that appeared in both native and non-native speech is function word substitution. It is easy for a native speaker who has understood the general meaning of the sentence to carelessly substitute one function word for another without changing the impact on the listener. It is also easy for a non-native speaker who is only reading words left to right without full comprehension to substitute a function word that completely changes the meaning of a sentence or even makes it meaningless..

(3.10) Native function word substitutions

a. *As* if that task were not challenging enough...
(original text)

b. *And* if that task were not challenging enough...
(spoken)

(3.11) Non-native function word substitutions

a. The amount of time students spend *on* homework is increasing
(original text)

b. The amount of time students spend *as* homework is increasing
(spoken)

In a special case of function word substitution, "a" and "the" are interchanged. This pair alone was responsible for nearly one-third of non-native function word substitutions, but more significantly, insertion and deletion of "a" and "the" accounted for half of all insertion and deletion errors. No other patterns were apparent in the types of words that were inserted and deleted. There was also a surprising number of instances of a/the substitution in *native* speech. An informal examination of a/the substitution in native and non-native speech suggests again that the native speakers will make these errors, but only when the the integrity of the sentence is preserved; this hypothesis is difficult to verify, however, as the source texts (and opportunities for ungrammatical substitution) are not the same, and grammaticality and comprehensibility judgements vary from listener to listener.

Substitution of an orthographically similar word

While both native and non-native speakers substituted orthographically similar words, native speakers again tended to choose words that preserved the integrity of the sentence, if not the meaning.

(3.12) Native orthographic substitutions

a. The politics of the *region* have always been unstable
(original text)

b. The politics of the *religion* have always been unstable
(spoken)

(3.13) Non-native orthographic substitutions

a. Environmentalists oppose construction of the Three *Gorges* Dam
(original text)

b. Environmentalists oppose construction of the Three *George* Dam
(spoken)

Substitution of a semantically similar word

Native speakers sometimes substitute semantically similar but orthographically dissimilar words; this error never occurred in the non-native sample.

(3.14) Native semantic substitutions

a. Tremendous change is anticipated over the next *few* years
(original text)

b. Tremendous change is anticipated over the next *several* years
(spoken)

Neologisms

In neologisms, non-native speakers make up a word. Sometimes these are compositions of common base forms and common endings that are inappropriate together. At other times they are unsuccessful attempts to read an unfamiliar word. This type of error did not appear in the (small) native sample.

(3.15) Native neologisms

a. But rain is nothing new for *Northwesterners*
(original text)

b. But rain is nothing new for *Northwesterns*
(spoken)

(3.16) Non-native neologisms

a. The diamonds sat *glittering* in the sand
(text)

b. The diamonds sat *glitting* in the sand
(spoken)

3.5.5 Experiment 1: Detection of non-native spontaneous speech by native judges

Experiment 1: Introduction

It is suspected that ungrammaticality and unnaturalness in non-native spontaneous speech are a factor in recognition error (e.g. Livescu and Glass, 2000). Because the statistical language models that are widely used in speech recognition are designed to find and learn patterns, a mismatch in the patterns that appear in the training and test data will contribute to suboptimal performance of the model.

How ungrammatical is non-native speech? The answer to this question depends on the definition of *grammaticality*, and even the definition of *speech*. Large bodies of work in linguistics rest on the assumption that native speakers are all competent judges of grammaticality, and indeed there are many sequences of words that any native speaker would flag as ungrammatical. Many of these studies, however, examine hypothesized sentences that may never have been uttered and are associated with no acoustic features – they are not speech. Real speech, even native speech, is full of ungrammaticalities; filler words, word fragments, and unfinished thoughts pepper spontaneous speech.

The statistical measurement of perplexity provides a measure of the predictability of a corpus of text. While predictability is not the same as grammaticality, if a language model is trained on grammatical native speech, it is not unreasonable to expect that a measurement of perplexity with respect to that model will be based to some extent on implicit grammatical constraints. What, then, does the observation that the Japanese utterances are lower in perplexity than the native utterances say about non-native speech? That it is more grammatical than native speech? Probably not. All that we can infer is that the non-native utterances contain patterns that also appeared in the training data; we have no idea whether these patterns are used appropriately in either the semantic or the syntactic context.

Another way to quantify the "non-nativeness" of an utterance is to measure the consistency with which independent native judges identify it as non-native. This method has the disadvantage of being utterance-based; a short utterance that is all wrong is given the same non-native label as a long utterance that is almost correct. We obtain a direct measurement of the distance between the native and non-native corpora, however, that is independent of a concept of grammaticality that may not be important for conveying meaning in spontaneous speech.

Experiment 1: Data

599 utterances from data sets N-A-S and NN-A-S were arbitrarily selected for this experiment, with an average of 34 utterances from 6 native and 12 non-native speakers.

Experiment 1: Method

Four native judges were asked to classify the 599 utterances. Because all of the non-native speakers were strongly accented, the judges were only allowed to see the transcripts. Judges were not told the percentage of non-native speakers in the sample. Utterances were presented to the judges in random order, varying from judge to judge. An average of 34 utterances per speaker was presented to the judges.

Experiment 1: Results

Table 3.9 shows the precision and recall of judgements from each of the native speakers. The precision measures how many of the utterances judged to be non-native actually were non-native, and the recall represents how many of the non-native utterances were identified as such. For example, for judge 1, 85% of the utterances judged to be non-native were actually non-native and 15% had been uttered by native speakers. 68% of the non-native utterances were correctly labeled as non-native, and 32% were labeled as native. The precision is much higher overall than the recall, meaning that the native judges seldom mistakenly label an utterance as non-native, but are not as good at identifying non-native utterances.

$$\text{precision} = \frac{\text{\# of times a non-native utterance was judged non-native}}{\text{\# of non-native judgements}}$$

Grader	Precision	Recall
1	0.85	0.68
2	0.87	0.54
3	0.88	0.44
4	0.89	0.46

Table 3.9: Precision and recall of native judgements of non-nativeness

	Full agreement		3/4 judges agreed		Actual
	Judged so	Actually so	Judged so	Actually so	totals
Judgements of nativeness	282	200	118	46	260
Judgements of non-nativeness	57	57	72	67	339

Table 3.10: Agreement of native judges, and corresponding actual labels of the utterances

$$\text{recall} = \frac{\text{\# of times a non-native utterance was judged non-native}}{\text{\# of non-native utterances}}$$

Table 3.10 shows how well the native judgements agreed, and for different levels of agreement how well the judgements corresponded with the actual labels. Of the 599 utterances, 260 were actually from native speakers and 339 were from non-native speakers. In 282 of their judgements, all four native judges agreed that the utterance was native, and in 57 of their judgements all judges agreed that the utterance was non-native. Of the 282 utterances that the judges fully agreed were native, only 200 actually were, while all 57 of the utterances all four judges agreed were non-native were truly non-native.

Experiment 1: Conclusion

The results in this experiment show that while native speakers seldom mis-identify a native utterance as non-native, they are only able to detect half of the non-native utterances; the other half are judged to be native. This may mean that the half of the non-native utterances judged native are grammatically correct and lexically typical of native speech. It is important to keep in mind, however, that *native* spontaneous speech is often ungrammatical and disfluent. It is likely that in many cases, the judges have no way to tell whether a speech "error" is a spontaneous effect or a non-native effect, and are therefore reluctant to mark an utterance non-native. Ungrammaticalities in native spontaneous speech may also be responsible for the false judgements of non-nativeness.

Chapter 4

Acoustic Modeling

A foreign accent, as viewed separately from features such as incorrect syntax or unusual word choice that also mark a speaker as non-native, is characterized by sound. An interdependent collection of properties, including melody, cadence, and segmental realization must be mastered for a non-native speaker to “lose” his accent. An accent, not necessarily a foreign one, is perceived when the listener detects patterns that are different from the ones he is used to hearing or identifies with unaccented speech.

In this chapter, I explore how accent is represented in the acoustic model and how the acoustic model can be adapted to better handle variation in non-native speech. Specifically, I investigate the contribution of different types of acoustic material to acoustic model improvement. Using native English data, Japanese-accented English (L2) data, and native Japanese (L1) data, I demonstrate how recognizer performance can be improved with respect to speaker idiolect, via speaker adaptation, and habits shared by speakers of a common L1, via training and adaptation to the non-native condition.

This chapter is structured as follows. In Section 4.1, I describe the baseline system on which my experiments build. In Section 4.3, I use the baseline acoustic models to find where modeling of non-native speech is poor. In Section 4.4, I document how adaptation to the speaker and condition can improve recognizer performance. In Sections 4.5 and 4.6, I present experiments in system training with L1 and L2 data. I summarize improvements in acoustic modeling in Section 4.7.

4.1 Baseline system

All recognition experiments described in this dissertation used the Janus Recognition Toolkit JRTk (Finke et al., 1997). Recognition experiments are done exclusively on the CND read speech database, specifically data sets N-E-R, NN-E-R, NN-T-R, and NN-X-R. The baseline system for CND used acoustic models trained on Broadcast News data and an interpolated language model combining broadcast news text (150M words) , written news text (10M words), written CND archive text (1M words), and children’s literature text (1M words). Interpolation weights were estimated using arbitrarily selected subsets of the training and cross-validation data sets NN-T-R and NN-X-R. Language modeling will be discussed further in Section 4.1.3.

CMU/ISL's Broadcast News (ISL-BN) system selected because it was the most robust available, having been trained on a large amount of data that varied in speech type and recording condition while remaining within the news domain. Pilot tests of several systems showed that the BN system offered the best initial baseline. Because there are some consistent differences between the BN task and the children's news task, the BN system was adapted somewhat for optimal performance on the Children's News (CND) task. This section describes the initial configuration of the system, the measures taken to maximize performance on CND , and my verification that any mismatch between the system and the task does not compromise my interpretation of overall results.

4.1.1 Baseline acoustic models

The acoustic models for the broadcast news system were trained on approximately 66 hours of data recorded from radio-broadcast news programming. The acoustic data was not limited to clean broadcast speech, but also included spontaneous broadcast speech (known as F0 speech), speech over telephone channels (F1), speech in the presence of background music (F2), speech under degraded acoustic conditions (F3), and speech from highly proficient non-native speakers (F4), all conditions that occur from time to time in radio news (Garovolo et al., 1997).

The baseline recognizer is a quinphone system with 2000 codebooks sharing 6000 distributions.

a quinphone system: the allophonic models take into account the two phones preceding and the two phones following each base phone.

with 2000 codebooks: 2000 allophonic groups are recognized; each allophonic group is modeled with Gaussian mixtures described by the same *means* and *covariances*

sharing 6000 distributions: each allophonic group is a collection of allophones that can be described by associating different *weights* with the means and covariances that model the parent allophonic group. There are a total of 6000 sets of weights in the system.

Vocal tract length normalization and cepstral mean subtraction are applied at the speaker level. Linear discriminant analysis (LDA) is used to find the most discriminative of the MFCC, delta, and power features and reduce the dimensionality of the feature fector describing each frame. This recognizer has an overall WER of 19.7%, with a WER on the clean (F0 only) subset of the test data of 9.4%. System details of ISL-BN and the Broadcast News test set are summarized in Table 4.1.

Number of codebooks	2000
Number of distributions	6000
Total number of Gaussians	104,746
Polyphone window	5 phones (2 preceding and 2 following)
Features used	MFCC, delta, delta-delta, power
Dictionary size	40,000
Language model type	trigram; Kneser-Ney backoff; cutoff=2
Language model training corpus	160 million words
Language model perplexity	155
OOV rate	1.1
Number of test speakers	81
Average number of utterances per speaker	5.8
WER (F0)	9.4%

Table 4.1: System details for the baseline system and the Broadcast News test set

4.1.2 Experiment 2: Determining the error due to system mismatch

Introduction

This experiment addresses the questions of channel mismatch and speaker variability. It should be noted that the only potential source of channel mismatch is the unique features of the recording device and environment; there is no difference in bandwidth or sampling rate between the BN and CND data. However, it is possible that the ISL-BN acoustic models perform better on BN speech than locally-recorded CND speech because the channel used in recording the evaluation data is more similar to those found in the training data. If this is the case, we would need to be concerned that any improvements we see from adaptation do not come from better modeling of the non-native condition but rather better modeling of the channel conditions. This experiment is not meant to be an exhaustive evaluation, but rather an informal confirmation that any channel mismatch is not severe enough to invalidate future experimental results.

Data

To set an initial error rate for system mismatch experiments, a 484-word segment of NPR acoustic data was selected. This segment will be known as NPR1, and is approximately equal in length to the test article that all CND speakers read. The NPR1 text is given in Appendix A.2. This data was read by a single announcer (speaker PA1) during a single broadcast under F0 conditions.

So that speech from the professional BN announcer could be directly compared to speech from a locally recorded speaker, a graduate student (speaker LS) was asked to read the NPR1 text. This student also read

	Read by	LM score	WER
NPR1-PA1	professional announcer 1	102.6	6.4
NPR1-LS	local speaker	102.6	7.4
NPR2-PA2	professional announcer 2	112.6	22.8
NPR2-LS	local speaker	112.6	14.7
CND1-LS	local speaker	115.3	13.2

Table 4.2: Comparison of recognizer performance on BN and CND data, after unsupervised adaptation, using the ISL-BN language model

evaluation article CND1 that was read by all native and non-native test speakers (see Section 3.4).

Because the NPR1 and CND1 texts differed substantially in language model score, the local speaker was asked to read a second BN passage (NPR2) that was taken from an on-the-scene segment and received a score from the ISL-BN language model that was much closer to that given to CND1. This text is given in Appendix A.3. This text was originally spoken by a second BN announcer (PA2).

Method

Because the non-channel-related conditions of the NPR1 recording (speaker, speech mode, environment) could not be duplicated, it was necessary to approximate the conditions using a local speaker and assess the error introduced by the approximation. This experiment therefore addresses two potential sources of error.

1. Speaker variability: local speaker vs. BN speakers

2. Channel mismatch: local and BN recordings of BN texts

It will not be possible to find an exact value for channel mismatch. However, based on these two comparisons, we can draw conclusions about the severity of the mismatch and the likely effect on further experiments. Corresponding results from text CND1 are given here for reference only; the issue of language model mismatch will be discussed in greater detail in Section 4.1.3.

Results

Language model score and WER for NPR1, NPR2, and CND1 spoken by speakers PA1, PA2, and LS are given in Table 4.2.

Speaker LS is not recognized quite as well as speaker PA1 reading the same text. This difference could be due either to channel mismatch or speaker variability. The difference (6.4 vs. 7.4) is not large, and we also see from Table 4.2 that the ISL-BN system performs substantially better on speaker LS than speaker PA2 (14.7 vs. 22.8) when those two speakers are reading the same text.

Conclusions

The observation that ISL-BN recognizer performed nearly as well on local speaker LS as professional announcer NPR1, and much better on speaker LS than speaker NPR2, suggests that the effect of channel

Test set	Language model	
	BN baseline	Interpolated
NPR1-PA1	7.4	8.9
NPR2-LS	14.7	16.2
CND-LS	13.2	12.7

Table 4.3: Measurements of WER for local speaker 1 comparing baseline BN and interpolated language models on baseline BN and CND test sets

mismatch is much smaller than the effect of speaker variability. The principal conclusion that I will draw from this experiment is that while there may be a slight mismatch in the acoustic channel, the effect after speaker adaptation is not severe enough to compromise the interpretation of future experimental results.

4.1.3 Language modeling

The BN language model is a trigram model using Kneser-Ney backoff (Kneser and Ney, 1995) with a trigram frequency cutoff of 2 (trigrams that only occured once in the training corpus were treated as unseen). The training data consisted of 150 million words of transcribed broadcast news text and 10 million words of written news text.

This is a very large and robust language model. However, slightly higher WER rates found in Experiment 2 (see Table 4.2) for the CND data compared to BN data for the same speaker suggested that there might be a small mismatch between the type of language used in the adult-oriented BN text and the child-oriented CND text. This potential mismatch was addressed by interpolating two independent trigram language models with the larger BN language model. These two new language models were built from CND archive text and non-CND news written for children. Context-independent interpolation weights were estimated from the training and cross-validation corpora NN-X-R and NN-T-R. This interpolated language model is used in a final rescoring pass of the word lattice for a 5.5% relative decrease in WER for the six-speaker native test set. The interpolation results in a relative reduction in perplexity on the CND test data of 16%.

Table 4.3 shows that interpolating the language models decreases WER for local test speaker LS on CND data and increases WER for both speaker LS and professional anchor PA1 on BN data.

Language model parameters

There are two user-specified parameters that are used in JRTk when incorporating the language model scores into the search: the language model weight `lz` and the word insertion penalty `lp`. These parameters can have a significant effect on the recognition outcome, and it was my observation that the optimal values for non-native speakers were quite different from those for native speakers.

Table 4.4 shows the effect the language model parameter settings have on recognition accuracy for native and non-native speakers. These figures represent the true optimal parameter values on native and non-native

Parameter settings	**Native**	**Non-native**
`lz`=36; `lp`=18 (optimal for native speakers)	17.5	78.8
`lz`=70; `lp`=90 (optimal for non-native speakers)	39.7	63.1

Table 4.4: Comparison of WERs when the language model weight `lz` and word insertion penalty `lp` are set to maximize performance for native and non-native speakers

test sets N-E-R and NN-E-R (see Table 3.3 for a description of data sets); the values actually used in the recognition experiments presented in this dissertation were calculated for an independent cross-validation set and resulted in slightly different WER measurements.

The higher optimal `lz` value for non-native speakers indicates that the system performs best when relying more heavily on the language model than is necessary for native speakers. This is not an unexpected observation, as the acoustic model does not provide as useful information as it does for native speakers. The higher optimal `lp` value may indicate that non-native speakers are inserting noises and epenthetic phones that are recognized as distinct words without a high penalty for inserting words.

For many types of experiment, the language model parameters are strictly fixed for simplicity of comparison. Because one of the goals of this work is to discover the relationships between different non-native adaptation techniques, I sometimes recalculate the language model parameter settings for optimal performance. These recalculations are always done on the independent cross-validation set NN-X-R.

New word handling in the language model

In order to eliminate variability due to out-of-vocabulary (OOV) error, all words in the test sets are added to the dictionary. A class-based component of the interpolated language model allows these OOV words to be added to the language model with the same probabilities as in-vocabulary words which have similar meanings.

4.1.4 Pronunciation dictionary

The CND dictionary is based on a 20,000-word dictionary developed for the Broadcast News task. With this dictionary, the out-of-vocabulary (OOV) rate on the CND articles is approximately 5%. In order to eliminate variability due to OOV error, all words that appear in the test utterances are included in the dictionary. Pronunciations for unusual proper names and other words of non-English origin are given in the CND text and can be entered into the dictionary as-is; pronunciations of other OOV words were taken from a much larger pronunciation dictionary.

All dictionary adaptation experiments described in Sec. 5 were built on top of this baseline dictionary.

Number of codebooks	2000
Number of distributions	2000
Total number of Gaussians	104,746
Polyphone window	5 phones (2 preceding and 2 following)
Features used	MFCC, delta, delta-delta, power
Speaker adaptation	supervised MLLR on 50 utterances
Dictionary size	26,110
Language model type	trigram; Kneser-Ney backoff; cutoff=2
Language model interpolation	BN, children's news, children's stories
Language model training corpus	161.2 million words
Language model perplexity	300
OOV rate	0
Number of native test speakers	6
Number of non-native test speakers	10
Average number of utterances per speaker	38
WER (F0)	18.0%

Table 4.5: System details for the ISL-CND system and the CND test set

4.1.5 Overall CND performance and conclusions about the baseline system

Baseline recognizer performance for one speaker was given in Section 4.1.2. In this section the baseline performance for the CND system (ISL-CND) on the native test set that will be used throughout this dissertation is given.

The CND native test set consists of six speakers, all reading the same article. ISL-CND uses the interpolated language model and domain-adapted dictionary described in Sections 4.1.3 and 4.1.4. Details of this system and the test set are given in Table 4.5.

Performance for all six native test speakers is listed in Table 4.6. Although the average WER is higher for CND than for BN, for the reasons discussed throughout Section 4.1 I have concluded that this discrepancy is due to inherent characteristics of the speakers and the task and not any mismatch or flaw in the acoustic and language modeling.

Speaker	204	205	206	207	240	241	average
WER	20.5	15.2	20.1	20.8	18.5	12.7	18.0

Table 4.6: Baseline recognizer performance on the six native CND test speakers

Establishing that the acoustic mismatch error is small and that baseline performance matches that of the currently best-performing speech recognizers allows us to have confidence that optimizations that are made for non-native speakers are due to better modeling of non-native speech and not to general system

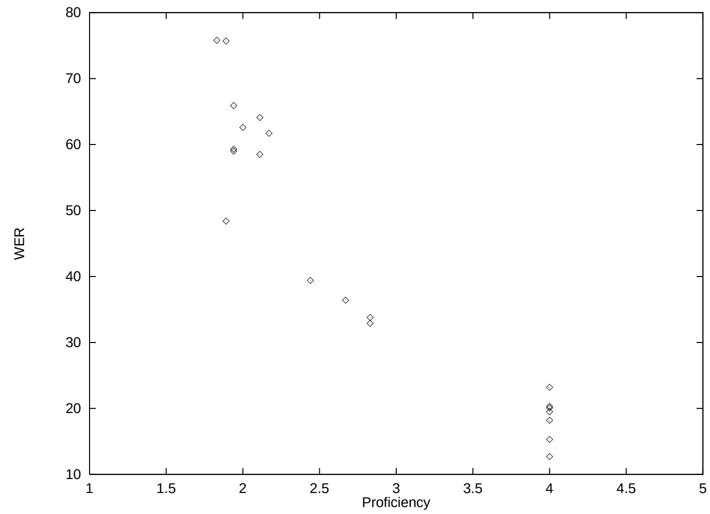


Figure 4.1: WER plotted against SPEAK proficiency score. Native speech is given a score of 4; noticeably non-native speech, even if completely intelligible, can score no higher than 3

improvements. The matter of speaker variability, however, will not be ignored; as will be seen, variability among non-native speakers is extreme, and exploring the interaction between speaker characteristics and modeling techniques will be a theme throughout this thesis.

Figure 4.1 puts the baseline performance of the ISL-BN system in the context of proficiency, showing WER for the native and non-native test sets as well as a group of four higher-proficiency speakers. The native test set N-E-R is that shown in table 4.6 and is the one that will be used in all acoustic modeling experiments unless otherwise specified. The non-native test set NN-E-R is a proficiency-controlled set of 10 speakers; as discussed in Chapter 3, this set of speakers all scored between 1.83 and 2.17 on a scale from 0 to 4 using the SPEAK assessment (SPE, 1987). We can see three clear clumps in the figure. Although there is variation among the native speakers (those with a proficiency score of 4), recognizer performance for all native speakers is better than that for any non-native speakers. Recognition of the four high-proficiency speakers is better than that of any lower-proficiency speaker. It is these lower-proficiency speakers that are the focus of this dissertation.

4.2 Significance testing

All improvements reported in this dissertation are statistically significant unless it is specifically stated that the improvement is insignificant. The NIST statistical test package released with the scoring package SCTK (NIST, 2000) was used to measure statistical significance; specifically, a matched-pairs test was used to evaluate sub-utterance-level differences in recognizer performance.

4.2.1 Basic steps in significance testing

The procedure for testing significance of any change consists of the following steps.

1. Establish the null hypothesis, H_0, and the alternate hypothesis, H_a. In the case of measuring recognizer improvement, the null hypothesis says that improvements we are seeing are a result of chance.

2. Specify a test statistic (function) Y that discriminates between H_0 and H_a.

3. Specify the "extreme" value (one-sided or two-sided) of Y in the direction of H_a. To show an improvement in error rate, the small extreme supports H_a.

4. Calculate the probability (p-value) of seeing Y at and beyond its observed value.

5. If the p-value is less than a fixed value (0.05, 0.01, e.g.), reject the null hypothesis. In the case of measuring recognizer improvement, this represents the conclusion that the results are not due to chance.

4.2.2 Special considerations for speech recognizer evaluation

Many people think of significance testing in the context of an experiment in which an experimental group that has been exposed to some sort of process is compared to a control group that has not. In such a scenario, the null hypothesis H_0 is that any differences between the two groups are coincidental and the process had no real effect. When we compare an improved speech recognizer to a baseline recognizer, we are doing something slightly different. We generally want to test the recognizer on a fixed test set, so that differences in WER can be attributed solely to differences in the algorithm or model. However, this means that there is no experimental group; the exact same set of utterances is processed by both the baseline recognizer and the new recognizer. In this situation, we are not concerned with inherent variation between two data sets that might make the process appear to have an effect, but rather with the external validity of the single data set. Although upon first consideration this may appear to simplify the problem, a more sophisticated statistical approach is actually required (Gillick and Cox, 1989) than would be if each recognizer were tested on an independent test set.

4.2.3 Test statistics

t-test

The t-test is useful when one wants to take into account the *magnitude* of the difference between the two systems. Additionally, it incorporates the variance among the samples in the normalization term, so data with less variance is more significant.

$$t = \frac{\overline{\chi} - \overline{\mu}}{\sqrt{\frac{s^2}{N}}}$$

for:

χ the sample mean

μ the real mean

s^2 the sample variance

N the sample size

However, this t-test does not take into account the variance in the real distribution, which is important when one is comparing two systems. Therefore, the following variation is used:

$$t = \frac{\overline{\chi} - \overline{\mu}}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

for:

χ the mean of the error rates of system 1

μ the mean of the error rates of system 2

var_1 the variance in the error rates of system 1

var_2 the variance in the error rates of system 2

n_1 the number of samples from system 1

n_2 the number of samples from system 2

The t-test for recognizer evaluation makes two crucial assumptions:

1. the distribution of outputs (error rates) is normal
2. the outputs of the system are independent

It has been argued that the latter does not hold in the case of speech recognizer evaluation (Gillick and Cox, 1989).

Matched pairs test

The matched pairs test can be used when the independence assumption does not hold. It has been said that this is the case in speech recognition, when the errors made in recognizing word w_i can affect how word w_{i+1} is recognized.

The matched pairs test is a way of formulating a two-sample problem as a one-sample problem, by making the sample points differences between outputs of the two systems instead of the outputs themselves. The data is segmented such that the errors made in one segment are independent of the errors made in the neighboring segments. In speech recognition, utterances can usually be the segments. The p-value then answers this question: if the average difference in performance of the two systems is zero, what is the chance that random sampling would result in an average as far from zero (or further) as observed in this experiment?

The matched pairs test is executed as follows. For n segments, define

$$Z_i = N_A^i - N_B^i, i = 1, 2, \ldots, n$$

where:

N_A^i the number of errors in the i'th segment for system A

N_B^i the number of errors in the i'th segment for system B

Estimate the mean and variance of the Z_i's:

$$\hat{\mu}_Z = \sum_{i=1}^{n} \frac{Z_i}{n}$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \hat{\mu}_Z)^2$$

Then define a variable W:

$$W = \frac{\hat{\mu}_Z}{\hat{\sigma}_Z / \sqrt{n}}$$

and determine whether the probability of W being the observed value is greater than your significance level α.

Since the distribution of the means of differences of error rates tends to a normal distribution, and the number of segments is large (greater than 50), the probability can be approximated using a normal distribution. That is, if $f(x)$ is the normal distribution, x is W and $y = f(x)$ is your p-value.

4.2.4 Significance testing in this dissertation

In this dissertation, I used a two-tailed matched pairs test to measure statistical significance. When I state that a result is significant or highly significant, I mean that it is significant using this test at the $p < 0.005$ level. In a few instances, I refer to a result as being "barely" significant. By this I mean $.05 > p > .01$.

4.3 Isolating problematic sounds

In chapter 3, a number of differences between native and non-native speech that can be expected to affect recognizer performance were quantified. In this section, I present a complementary analysis, examining how well the baseline acoustic models capture the phonological properties of non-native speech.

4.3.1 Phonetic confusion

Phonetic confusion is a measure of how often an individual phone sounds like a different phone. This gives an indication of how accurate the acoustic models are with respect to the input speech. An analysis of phonetic confusion can also provide candidates for phone-substitution-based lexical modeling.

Unfortunately, phonetic confusion figures derived from recognizer output can be difficult to interpret. Failure to accurately recognize a phone may be because the pronunciation is not correct, but it could also be the result of a flaw in the acoustic model. In this thesis, I wish to address the former case and therefore will attempt to isolate confusions that are common only in recognition of non-native speech.

A phonetic confusion matrix is built by calculating, for each phone in the phone inventory, how frequently it was misrecognized as each other phone in the inventory. Depending on the objective of the analysis, confusions can represent either segmental or framewise comparisons. For example, let us say that sentence (1) was misrecognized as something more like (2).

(1) THEN THEY SWIM UPSTREAM IN A FIERCE WRONG WAY STRUGGLE TO THEIR BIRTHPLACE

(2) THEN THEY SWIM UP STREAMING FEARS RUNWAYS TRAVELED TO THEIR BUS PLACE

Isolating the words "upstream in a fierce wrong way struggle to" for more detailed examination, we can identify the errors /n/ → /ŋ/, /ə/ → /ŋ/, /s/ → /z/, /ɔ/ → /ə/, /ŋ/ → /n/, /ʌ/ → /æ/,/g/ → /v/, and /t/ → /d/ in a phone-by-phone comparison:[1]

ə	p	s	t	r	i	m	ɪ	n ə	f	i	r	s	r	ɔ	ŋ	w	eɪ̆	s	t	r	ʌ	g	ɫ̩	t	u
ə	p	s	t	r	i	m	ɪ	ŋ	f	i	r	z	r	ʌ	n	w	eɪ̆	z	t	r	æ	v	ɫ̩	d t	u

leading to the following phonetic confusion matrix shown in Table 4.7, where the prescribed phones are shown vertically and the recognized phones are shown horizontally.

These confusions would be said to have been generated through a segmental breakdown of word recognition. While this sort of breakdown is simple to do and is a useful method for finding potential pronunciation variants, it does not represent confusion due to phone insertion and deletion well. For example, in the misrecognition UPSTREAM IN A FIERCE → UP STREAMING FEARS, the /ə/ sound in the word "a" is effectively absorbed in the model for /ŋ/. In the matrix given above, the mapping /ə/ → /ŋ/ is given equal weight to the mapping /ə/ → /ɔ/. This is not strictly appropriate, however. A more accurate estimate of phonetic confusions can be found by either calculating mappings on a frame-by-frame instead of a segmental basis or restricting the word recognition so that the source of phone insertions and deletions is known.

Framewise estimation of confusion

To generate a framewise estimation of phonetic confusions using word recognition output, the active phones in the input speech and recognizer output are compared for each 10-ms window.

[1] For simplicity of illustration, the recognized phone string shown here is more accurate than it actually was. In actual experiments, no language model was used, and the confusions were much higher. The words corresponding to the phone strings are provided only for illustration.

w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{1}$
u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{1}$	0
r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{4}{4}$	0	0
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{1}$	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{1}$	0	0	0	0
ɫ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{1}$	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{2}{2}$	0	0	0	0	0	0
ɪ	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{1}$	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{1}$	0	0	0	0	0	0	0	0
z	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
t	0	0	0	0	$\frac{1}{4}$	0	0	0	0	$\frac{3}{4}$	0	0	0	0	0	0	0	0	0	0	0
s	0	0	0	0	0	0	0	0	$\frac{1}{3}$	0	0	$\frac{2}{3}$	0	0	0	0	0	0	0	0	0
ŋ	0	0	0	0	0	0	$\frac{1}{1}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	$\frac{1}{1}$	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	$\frac{1}{1}$	0	0	0	0	0	0	0	0	0	0
d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ə	0	0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ʌ	0	$\frac{1}{1}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
æ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	0	0	$\frac{1}{1}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ɔ	æ	ʌ	ə	d	g	n	ŋ	s	t	v	z	f	ɪ	i	ɫ	m	p	r	u	w

Table 4.7: Example of a phoneme confusion matrix

(*. . . upstream in a. . .*)

. . . ʌʌʌʌʌʌʌʌʌʌʌpppppppppssssssssssstttttttttrriiiiiiiiiiiiiimmmmmmmɪɪɪɪɪɪɪɪɪɪɪnnnnnnnnəəəə. . .

. . . ʌʌʌʌʌʌʌʌʌʌʌpppppppppssssssssssstttttttttrriiiiiiiiiiiiiimmmmmmmɪɪɪɪɪɪɪɪɪɪŋŋŋŋŋŋŋŋŋŋŋŋŋ. . .

0ms 100ms 200ms 300ms 400ms 500ms 600ms 700ms 800ms 900ms 1000ms 1100ms

Now, instead of the partial matrix

$$
\begin{array}{c|ccc}
\text{ŋ} & 0 & \frac{1}{1} & 0 \\
\text{n} & 0 & 0 & \frac{1}{1} \\
\text{ə} & \frac{1}{2} & 0 & \frac{1}{2} \\
\hline
 & \text{ə} & \text{n} & \text{ŋ}
\end{array}
$$

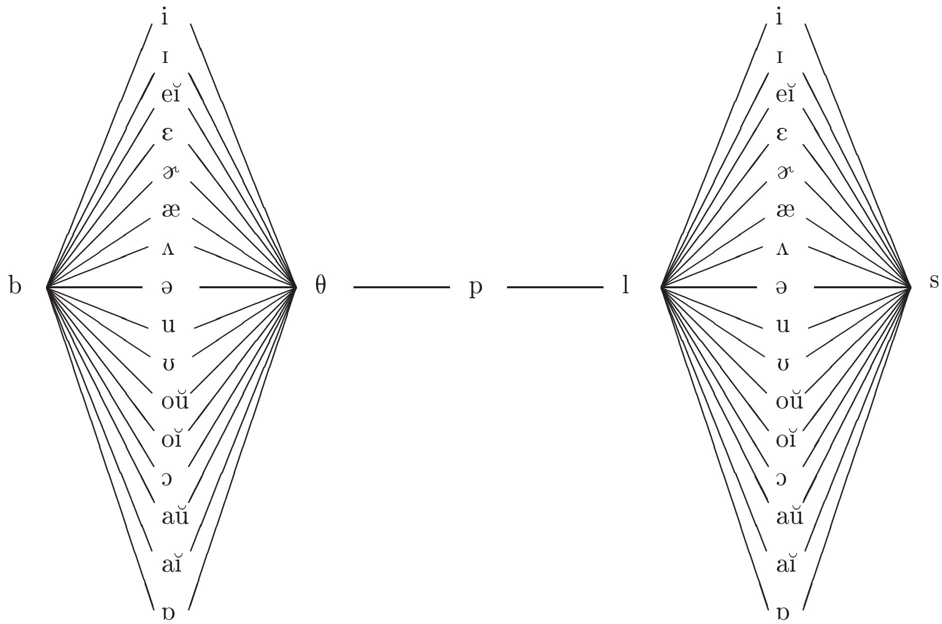
we have the following.

$$
\begin{array}{c|cccc}
\text{ɪ} & 0 & 0 & \frac{1}{11} & \frac{10}{11} \\
\text{ŋ} & 0 & \frac{21}{21} & 0 & 0 \\
\text{n} & 0 & 0 & \frac{8}{8} & 0 \\
\text{ə} & \frac{11}{15} & 0 & \frac{4}{15} & 0 \\
\hline
 & \text{ə} & \text{n} & \text{ŋ} & \text{ɪ}
\end{array}
$$

Because phone transition times in the recognition output rarely match up exactly with the actual times, using a framewise estimation introduces a large number of small confusions. The /ɪ/ $\rightarrow$ /ŋ/ confusion in the new matrix represents such a case. Although individual instances of transition mismatch are typically very short, overall, they can introduce a significant amount of noise, particularly for phones whose average duration is itself not long. This is one of the disadvantages of framewise estimation of phonetic confusion.

Underspecified alignment

Another way to ensure a straightforward (and not necessarily one-to-one) mapping between phones in the input speech and the recognizer output is to force the system to generate a specific number of phones. In underspecified forced alignment, the user can specify some features of the word that is to be recognized and then allow the recognizer to "fill in" the missing features to generate a full phonetic specification for that word. For example, it may be known that vowels are a major source of phonetic confusion. In order to find the vowel pairs that are most confusible, the user may opt to fix the consonants and only allow the recognizer a choice where a vowel sound is expected. The recognizer would be presented with a phone network, the best path through which it can calculate via a Viterbi search:



In this example, the word is "birthplace" and the canonical pronunciation is [bɚθpleĭs]. However, imagine that the speaker actually says something more like [bɑːθpu̥ɯrɛs]. Running an underspecified forced alignment of the input speech to the network would tell us which model sequence best matches the speaker's pronunciation, in this case perhaps [bɒθplɛs].

Underspecified forced alignment, then, can be used to generate strings of phonemes similar to those generated through phonetic expansion of word recognition output. The former offers several advantages. First, the user can restrict sources of variation according to the objective of the study. Second, there is no interference from the language model; only the acoustic match is optimized. Third, framewise estimation of confusion can be accomplished without introducing noise due to phone transition time mismatch. Finally, one-to-many and many-to-one relationships beween canonical and empirical phone sequences (representing epenthesis, simplification, e.g.) can easily be explored.

An underspecified forced alignment of sample sentence (1) produces the following mapping (the top line is the fully specified forced alignment result, the same as the one given above).

(*. . . upstream in a. . .*)

. . . ʌʌʌʌʌʌʌʌʌʌʌpppppppppsssssssssssttttttttttrriiiiiiiiiiiiimmmmmmmmɪɪɪɪɪɪɪɪɪnnnnnnnnəəəə. . .

. . . ɒɒɒɒɒɒɒɒɒɒɒɒppppppuuusssssssssssstttttttttrriiiiiiiiiiiiimmmmmmmmiiiiiiiiinnnnnnnnnəəəə. . .

In this case, the underspecification allowed a choice between all vowels, between consonants with the same place of articulation, between nasal consonants, and between /l/ and /r/. In addition, epenthetic vowels were allowed post-consonantally. The fact that in the presence of a choice the alveolar nasal is correctly recognized but the preceding high front vowel is not indicates that acoustically, the confusion is between /i/ and /ɪ/, not between /n/ and /ŋ/. The mappings in the word-recognition-based example were influenced by the words in the lexicon, the coarticulatory relationship between /ɪ/ and /ŋ/, and the high frequency of the morpheme "-ing," among other factors.

Figure 4.2 shows phonetic confusion in the training data[2] estimated via underspecified forced alignment. The size of the bubble at each point represents the magnitude of the confusion. For example, confusion between /u/ and /ʊ/ is high for non-native speakers. It is also high for native speakers, however. The non-native speech is characterized primarily by greater degrees of acoustic confusion between the same pairs of phones that are confusible in native speech.

Unrestricted phoneme recognition

A third method of generating a phonetic transcription of input speech is phoneme recognition. In normal LVCSR, information about the words and word sequences that are meaningful in a language is used to help identify phones. Normal native speech is full of departures from the prescribed pronunciation. For example, the alveolar nasal in "one-way struggle" can be highly labialized in anticipation of the labiovelar approximant. With the knowledge that "one" is an English word and that "one-way" is a common word sequence in English, the human listener may perceive the nasal as an /n/ when acoustically and articulatorily it is closer to an /m/. Higher-level linguistic knowledge contributes greatly to successful recognition of connected speech, and word-based recognition generally produces a far more accurate sequence of phones than phoneme recognition.

Nevertheless, phoneme recognition can be a useful tool for exposing idiosyncrasies in the production of words. In unrestricted phoneme recognition, the decoder is run with a uniform language model[3] and with a lexicon containing only phonemes. If there are 46 phonemes, there would be 46 "words" in the lexicon. The result of the search is the sequence of phones representing the acoustic models that best matched the input speech at each point in time. Phoneme recognition hypotheses can then be used in the same way as word recognition hypotheses or underspecified alignment hypotheses for segmental or framewise estimation of phonetic confusion.

Figure 4.3 shows phonetic confusion estimated via a framewise comparison of phoneme recognition hypotheses. Although the phoneme recognition error is similar for native and non-native speakers (52.1% and 57.2% respectively), the native confusions seem to be distributed more evenly across phoneme pairs while the non-native confusions are concentrated in certain "catchall" phones. Specifically, /ɪ/,/t/, and /silence/ tend to be hypothesized inappropriately by the recognizer.

The frame-by-frame values for canonical pronunciation and phoneme recognition output for a non-native speaker's realization of the phrase "upstream in a" are juxtaposed below. This is the same sequence that

[2] "Training data" in this case is the part of the CND database designated for further acoustic model training (N-T-R and NN-T-R). This set of data was not involved in training of the acoustic models used to generate phonetic transcriptions.

[3] Syllabic and phonotactic constraints can be introduced by assigning likely phone sequences higher probabilities in the language model.

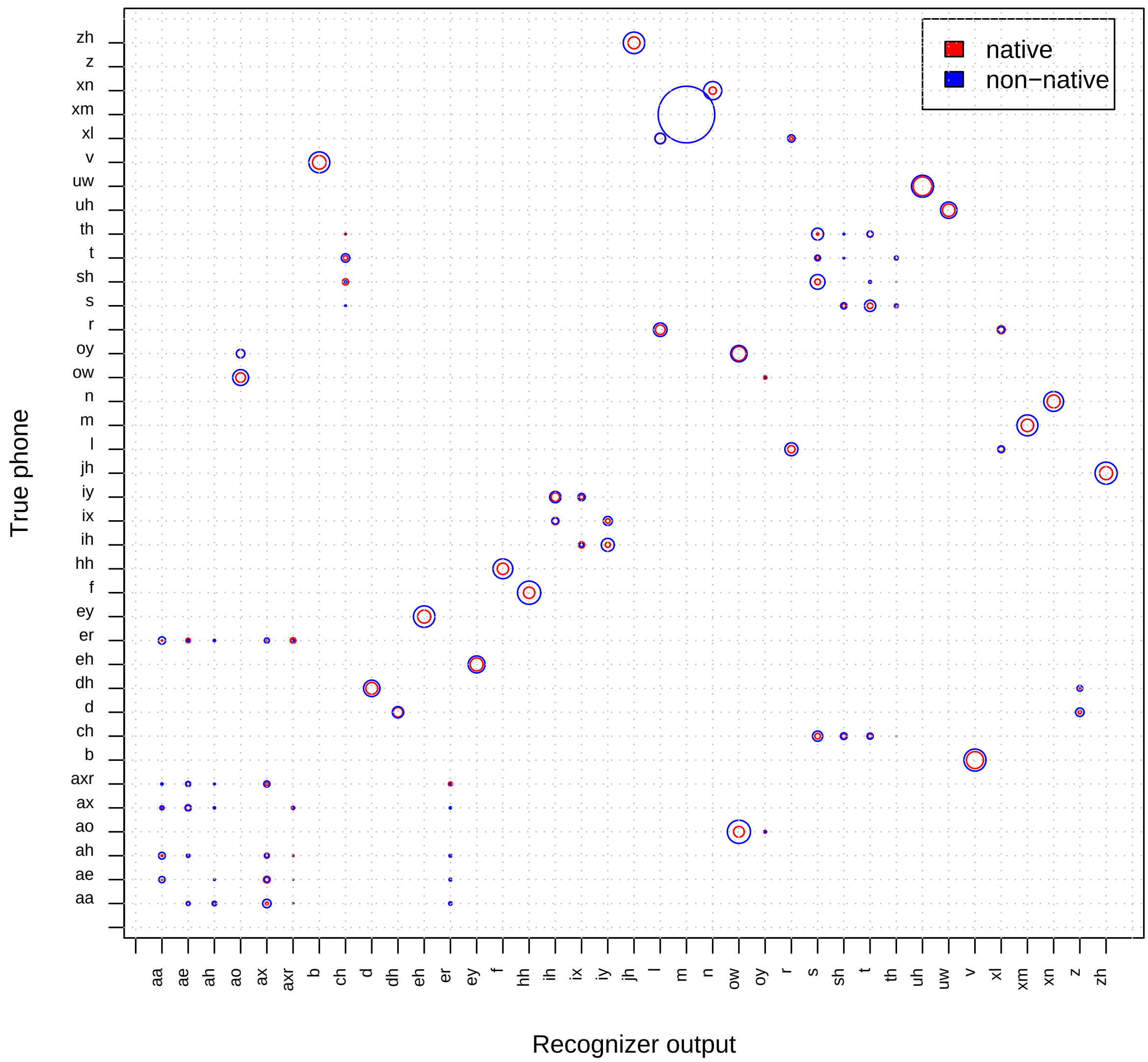


Figure 4.2: Phoneme confusions in underspecified alignment

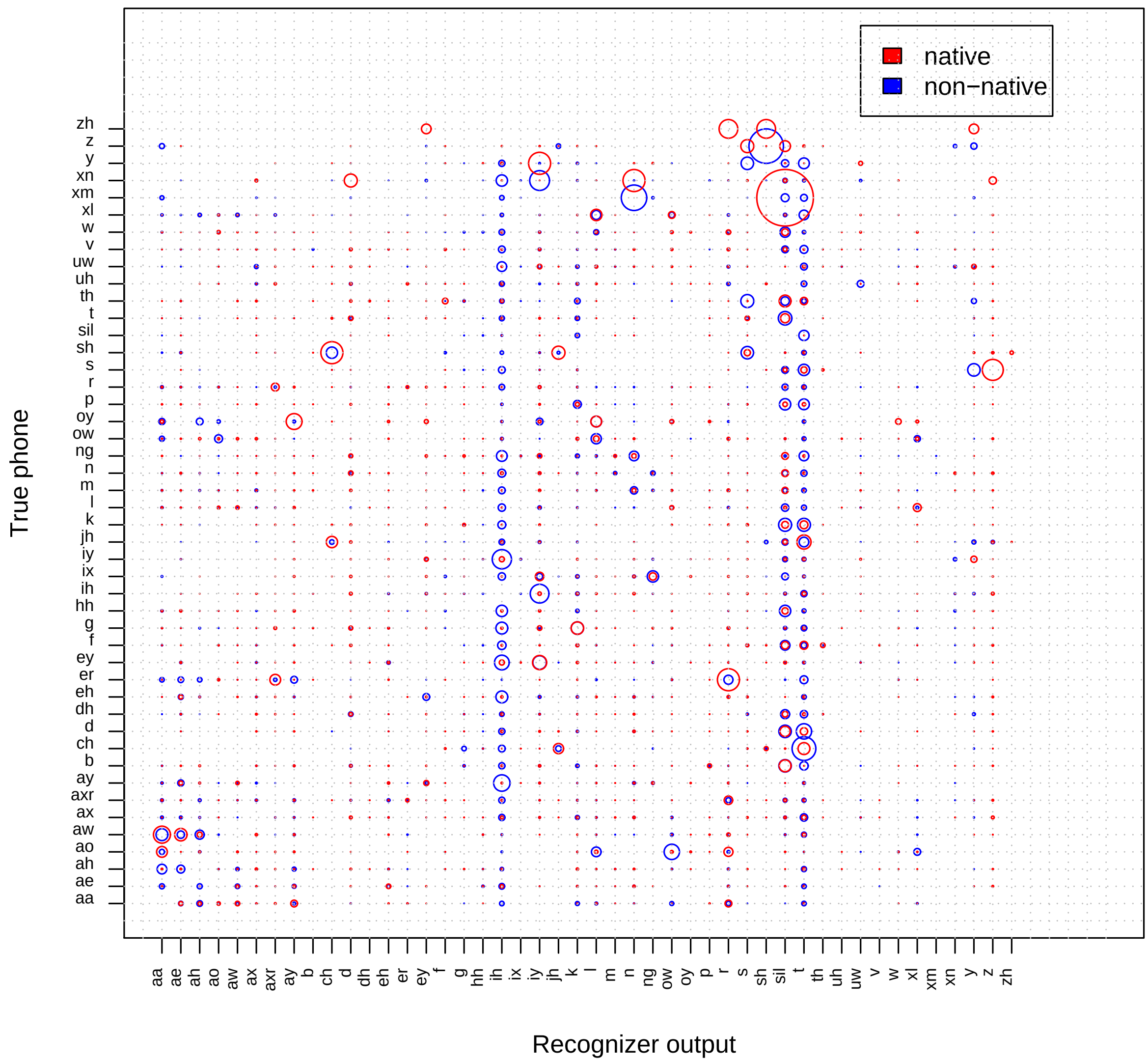


Figure 4.3: Phoneme confusions in unrestricted recognition

has been used to illustrate word recognition based and unrestricted alignment based phoneme generation.

(. . . *upstream in a*. . .)

. . . ʌʌʌʌʌʌʌʌʌʌʌppppppppppsssssssssssttttttttttrriiiiiiiiiiiiimmmmmmmɪɪɪɪɪɪɪɪɪnnnnnnnnnəəəə. . .

. . . kkkkkɒɒɒɒɒɒppppppppppppssssssssʧʧʧʧʧʧʧʧʧʧʧʧʧiiiiiiiiiiiiiiiiiiiiiiiiiiiiiinnnnnnnnnnnnnɛɛ. . .

Both the valuable information and undesirable noise that phoneme recognition hypotheses contain are apparent in this example. The first substitution, /ʌ/ → /k/, is peculiar. The speaker does produce a pronounced glottal stop at the onset of the word "upstream," which would probably not be present in smooth native speech and may be the source of the recognizer's perception of a voiceless velar stop. Indeed, this phenomenon may partially explain the surprisingly high rate of substitution of voiceless stops for vowels. The sequence /tr/ is recognized as /ʧ/, which is plausible as this combination can be palatalized in native speech as well. Quality of the phoneme recognition degrades toward the end of the phrase, however, where the final consonant in "upstream" and the initial vowel in "in" are lost altogether.

In Section 4.6 and Chapter 5 phonetic confusion will be used to predict phone substitutions. Both framewise confusion through unrestricted phone recognition and segmental confusion through underspecified alignment will be used.

Context-dependent vs. context-independent models

In the previous paragraphs, I have discussed methods that can be used to generate phonetic transcriptions for estimation of phonetic confusion. It is also important to consider the type of acoustic model that is being matched to the input speech. The models used in ordinary LVCSR are usually context-dependent, that is, they model the acoustics of a given phone in a given context. If they were trained on native speech, however, they may not accurately reflect the phonetic contexts that trigger variation in non-native speech.

In all of my calculations of phonetic confusion, phonetic transcriptions were generated using context-independent models. While the context-independent models are associated with an increased word error rate, analysis of phonetic confusion and comparison of confusion in native and non-native speech is more straightforward with context-independent models. I also wished to avoid allowing phonotactic and coarticulatory patterns found in native speech to influence the match of models to non-native speech.

4.3.2 Polyphone coverage

One of the reasons that modifications to the dictionary may not work well is that the new phonemic transcriptions can include phone sequences that were not in the training data. For example, if the pronunciation /daĭlɛkuto/ is proposed as a variant for the word "direct," the sequence /ɛkuto/, which never appeared in the training data, is introduced. Even if the variant is an accurate reflection of the speaker's pronunciation,

(... *swim*)				*upstream*							*in*				*a*				*fierce*				(*wrong*...)
		ə	p	s	t	r	i	m			ɪ	n			ə			f	i	r	s		
	m	**ə**	p	s						m	**ɪ**	n	ə	n	**ə**	f	ə	**f**	i	r			
	m	ə	**p**	s	t					m	ɪ	**n**	ə				ə	f	**i**	r	s		
		ə	p	**s**	t	r												f	i	**r**	s	r	
			p	s	**t**	r	i												i	r	**s**	r	
				s	t	**r**	i	m															
					t	r	**i**	m	ɪ														
						r	i	**m**	ɪ														

Figure 4.4: Illustration of how polyphones are defined for the utterance fragment "... (swim) upstream in a fierce (wrong-way struggle)..."

because no polyphone model was ever trained for this sequence, the trained model for the canonical pronunciation might match the input speech better than the generic model that serves as a backoff model for the unseen variant polyphone.

I have found that the polyphone coverage, or percent of polyphones in a test data set that appeared in the training corpus, is much lower for non-native speakers than for native speakers. To calculate polyphone coverage, a reference corpus is generated, in this case by aligning the training data to the manual transcriptions using the baseline dictionary. Twelve percent of the words in the baseline dictionary have variant pronunciations listed, averaging 1.2 variants per word with variants. As part of the alignment process, the variant that most closely matches the actual pronunciation is identified, yielding a more accurate phonetic representation than a non acoustically derived phonetic expansion of the words in the manual transcription would. The number of polyphones in this reference corpus is then calculated. In the ISL-BN recognition system, each phone in the data is associated with a polyphone comprising that phone and the two preceding (one if the phone is word initial) and two following (one if the phone is word final) phones. In the case of utterance-initial and utterance-final phones, no preceding/following phones are included in the polyphone sequence. The breakdown of an example utterance fragment into polyphones is shown in Figure 4.4.

Table 4.8 lists the polyphones associated with each phone that appears in the example utterance fragment. Four phones appear more than once, and for those phones multiple polyphones are listed. There are a total of fourteen polyphones in this example. There are 5.5 million polyphone tokens and 4.1 million polyphone types in the reference corpus. Of the fourteen polyphones in the example, only eight, or 57%, were among the 4.1 million polyphone types that appeared in the reference data. This utterance fragment, then, has a polyphone coverage of 57%.

ʌ	p	s	t	r	i	m	ɪ	n	f
m ʌ p s	m ʌ p s t	ʌ p s t r	p s t r i	s t r i m	t r i m ɪ	r i m ɪ	m ɪ n ə	m ɪ n ə	ə f i r
n ʌ f		i r s r		f i r s r	ə f i r s				

Table 4.8: Polyphones associated with each phone that appears in the utterance fragment "... (swim) upstream in a fierce (wrong-way struggle)..."

4.3.3 Experiment 3: Polyphone coverage after phone substitutions

Introduction

To find how the polyphone coverage is affected by phone insertions and deletions common in non-native speech, I generated several experimental corpora for which polyphone coverage was measured. In each case, the input speech was aligned to the manual transcriptions using the variant-sensitive procedure described above. The reference corpus used for all conditions was the transcribed NN-E-R corpus.

In this experiment, three variables are adjusted: phonetic expansion dictionary, speaker nativeness, and acoustic model type. Insights into the polyphonic makeup of non-native speech will come from comparing coverage of non-native speech before and after introduction of non-native variants in the dictionary. These results cannot be accurately interpreted, however, without examining how the same changes in the dictionary affect coverage of native speech, and whether alignment using context-dependent models yields significantly different polyphones from those generated using context-independent models.

Data

Polyphone coverage measurement requires a test corpus and a training corpus. The percentage of polyphones in the test corpus that also occur in the training corpus is the polyphone coverage of the test corpus. For this experiment, the test corpora were the shared articles from N-E-R and NN-E-R; the training corpus was the unique articles read by each speaker in NN-E-R. Because the transcribed training corpus will be expanded phonetically based on the canonical pronunciations in the dictionary, the fact that the articles were originally read by non-native speakers does not affect the estimation. Phonetic expansion of the test corpora will be discussed below.

Method

Potential non-native variation in pronunciation was allowed by augmenting the baseline dictionary with variants generated using several complementary methods. One set of variants was produced using information about the phonotactic structure of the speaker's native language. Another set was based on the phonetic confusion measurements presented in Section 4.3.1. Hand-coded variants were also added, along with variants derived from native-language representations of loanwords from English. These dictionaries are described in greater detail in Section 5. The expanded dictionary is very large (1.13 million words); the number of

	CI models				CD models	
	baseline dictionary		expanded dictionary		expanded dictionary	
	native	non-native	native	non-native	native	non-native
Polyphone tokens	92.1	93.7	65.4	46.9	73.8	52.8
Polyphone types	92.1	93.4	61.7	42.6	69.4	48.2

Table 4.9: Polyphone coverage of native and non-native speech

base words is the same as in the baseline dictionary, but instead of 12% of the words having variants listed, 99% are associated with variants, averaging 48 variants per word. If one were attempting decoding with this dictionary, the search space would be enormous. Because I am doing alignment, however, the word sequence is known and the recognizer is only asked to determine which of a given list of phone sequences best matches the input speech. By allowing variants generated by a variety of methods, I maximize the probability that a model sequence that truly matches the input speech is found. Comparisons of the different methods and the contribution of variant pronunciations to recognition accuracy are discussed in Section 5.

Polyphone coverage (the percentage of polyphones in the test corpus that also occurred in the training corpus) was measured for the baseline and expanded dictionaries using the context-independent models and for the expanded dictionary using the context-dependent models.

Results

Table 4.9 shows polyphone coverage for native and non-native speakers. We can see that the polyphone coverage of the non-native data is much higher when the non-native pronunciations are forced to conform to canonical pronunciation standards (66.7% coverage with the baseline dictionary) than when more flexibility to identify the true phone sequence is allowed (43.1% coverage with the expanded dictionary). This says that the non-native speakers are producing phone sequences for which polyphones would not have been trained. However, we can also see from Table 4.9 that coverage of *native* speech decreases (79.7% to 63.0%) when the alignment is not restricted to canonical pronunciation standards.

Conclusions

Pronunciations that were intended to be representative of non-native speech are registering as the closest match for native as well as non-native realizations of the words. There are several possible explanations for this. First, poor quality in the acoustic models may be causing the wrong variant to be selected. Second, the native speaker may actually be pronouncing the words in a way that is closer to the selected "non-native" variant than the canonical pronunciation. Third, the variant may have been one that was derived from phoneme recognition output, and might reflect internal bias in the acoustic model more than true L1-conditioned variation.

All three of these hypotheses are probably correct in some cases. One might consider evaluating the first by comparing context-dependent and context-independent results using the expanded dictionary; because

	CI models		CD models	
	native	non-native	native	non-native
Hand-coded	8.0	9.8	9.4	13.4
Phoneme recognition	75.0	61.4	74.3	57.0
Underspecified alignment	9.0	5.0	8.3	6.2
Linguistically motivated	7.5	22.5	7.6	21.8
Derived from L1 representations of loanwords	0.5	1.3	0.4	1.7

Table 4.10: Source of pronunciation variants selected during alignment

the context-dependent models are more accurate than the context-dependent models, if the problem is with model quality we should see a significant decrease in the number of non-native variants that match to native speech, which indeed we do. However, this is not a fair comparison, as the context-dependent models enforce precisely the constraints that I wish not to be bound by in my investigation of the "true" realizations of words in speech. The second and third hypotheses can be investigated by looking at the variants that were chosen. Distributions of variant types selected using context-dependent and context-independent models are given for native and non-native speakers in Table 4.10. The most striking differences are that linguistically-motivated variants are selected more often for non-native speakers than for native speakers, and that variants derived through phoneme recognition are selected more often for native speakers than for non-native speakers. This suggests that many of the variants identified in native speech are tied to the way phones are modeled in the recognizer, supporting the third hypothesis. We also have from Table 4.10 clear evidence that the linguistically-motivated variants capture non-native speech phenomena.

4.3.4 Implications for acoustic modeling

Knowledge of the distribution of sounds and the relationship between prescribed and recognized phones in native and non-native speech will guide us as we strive to improve acoustic modeling of the non-native condition. We have seen that confusion between numerous phones is higher for non-native speakers than for native speakers. The pairs /ɪ,i/, /ɔ,o/, and /θ,s/ are satisfying to see highlighted in the confusion matrices as these are substitutions one might predict from either a linguistic analysis of Japanese or experience listening to Japanese natives speaking English. By the same token, however, the absence of pairs like /ɚ,ɒ/ and /l,r/ in the matrix is disappointing. It was observed in experiment 3 that compared to the native test set N-E-R, a large number of the phone sequences that appear in non-native test set NN-E-R do not occur in the corresponding training data.

In the following sections, I will describe a number of methods for improving performance of the acoustic models on Japanese-accented English, including some specifically intended to counteract problems of phonetic confusion and polyphone coverage. Approaches like MLLR adaptation and Viterbi training with accented data will address issues of phonetic confusion, poor overall modeling of non-native speech, and some insertion and deletion of phones. Discrepancies in the polyphones found in native and non-native speech require more sophisticated modeling, and I will present results for training and adaptation of the polyphonic decision trees.

4.4 Adaptation

In speaker adaptation, acoustic models that have been trained for general speech are adjusted so that they better model the speech characteristics of a specific condition. Adaptation does not have to be limited to individual speaker adaptation; general models can be specialized to compensate for differences in acoustic environment or the characteristics of a group of speakers. Non-native speakers with strong accents are natural candidates for adaptation because of the magnitude and consistency of many deviations from standard native pronunciation.

Acoustic adaptation can be applied in either the feature space or the model space. Feature-space methods include cepstral mean subtraction and vocal tract length length normalization, both of which are applied in the ISL-BN system. Adaptation techniques commonly applied in the model space include maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation.

Pilot experiments on the non-native data indicated that adaptation would be crucial if a level of recognizer performance on which further experiments would be meaningful were to be achieved. In this section I compare applications of MLLR and mixed-style, or simplified MAP, adaptation, using both native-language and accented data. I discuss both the differences between the two approaches and experimental results of applying them for non-native speech.

4.4.1 Model-space adaptation

The two types of adaptation that I discuss in this section operate by modifying the parameters of the acoustic model, specifically the means of the Gaussian mixture models that represent each phonetic state. This section focuses on using adaptation to estimate a better general model of Japanese-accented English before individual speaker adaptation is applied to further specialize the model.

MLLR is an example of what is called transformation-based adaptation. In transformation-based adaptation, a single transformation operation is applied to all models in a transformation class. The transformation function is estimated from a small amount of held-out data. In the Janus implementation of MLLR, the optimal number of transformation classes is determined dynamically.

In mixed-style adaptation, the model parameters are re-estimated individually. Using held-out adaptation data, sample mean values are calculated. An updated mean is then found by shifting the original value toward the sample value. If there was insufficient adaptation data for a phone to reliably estimate a sample mean, no adaptation is performed. The degree of shift toward, or interpolation weighting factor of, the sample value is globally applied to all transformations. This is where mixed-style adaptation differs from true MAP adaptation, in which interpolation weights are estimated separately for each transformation. Because similar gains have been observed in MAP and mixed-style adaptation (Soltau, 2001), I will use the simplified form. All references to MAP adaptation in this dissertation therefore describe not true MAP adaptation, but mixed-style adaptation.

Both MLLR and MAP adaptation are popular and effective in boosting LVCSR performance (Woodland, 1999). Because transformation-based adaptation defines a transformation function for the entire class, it can calculate an updated mean even for phones that did not appear with critical frequency in the adaptation data. For this reason, it can be effective when not much data is available. However, a transformation function that is optimal for the class may not be optimal for all individual models, and with MLLR one runs the risk of applying the function improperly and shifting some means *away* from the observed sample value. This does not happen with MAP adaptation, as each parameter is adapted separately. When the adaptation data is representative of the test data, MAP adaptation performance improves as the amount of adaptation data increases. With only a small amount of adaptation data, however, MLLR tends to provide the better model (Doh, 2000).

4.4.2 Experiment 4: Adaptation to the non-native condition

Introduction

There were two questions that I sought to address through adaptation exeriments.

1. Does L1 material provide better adaptation data than accented L2 data?

2. Does MAP adaptation perform better than MLLR adaptation for non-native speech?

The first question is important for two reasons. First, collecting L1 data is sometimes easier than collecting accented L2 data. For well-represented L1s like Japanese and Spanish, L1 acoustic corpora might already be available. And by using L1 data to adapt, the potential combinatorial problem of having to collect speech data for each L1-L2 pair can be avoided. Second, L1 data might provide a more consistent representation of non-native speech than L2 data does. If the variation in phonetic realization is very great in the accented L2 speech, new sample means may not be very meaningful, and adapting to them may degrade rather than improve the model. The best performance might be achieved by first adapting to consistent data that is representative of the accented speech and then adapting to individual idiosyncrasies in the realization of specific phones. The problem with this argument, of course, is that it assumes a regular mapping between L1 and accented L2 phones, a suggestion that has been disputed in e.g. (Brière, 1966).

The second question asks whether transformation-based or Bayesian adaptation is more appropriate for non-native speech. One might speculate that because the non-native data is highly variable the risk of improperly applying transformation functions would be high, suggesting that MAP adaptation would be the better choice as long as there is enough adaptation data. This is only a hypothesis, however, so one would like to address the question empirically.

In these experiments, I use the 10-speaker proficiency-controlled set of Japanese-accented English (NN-E-R) as the test set.

Data

The L1 data that was used for these experiments was Japanese read news from the Nikkei Shimbun. This data was selected because it was similar in task and topic to the Japanese-accented English data. The data collection methods and environments were identical.

The accented L2 adaptation data was drawn from the training set of Japanese-accented read news data (NN-T-R).

The test data was the proficiency-controlled non-native set NN-E-R.

Experiment 4.1: MLLR adaptation

Method

Prior to individual speaker adaptation, MLLR adaptation based on speech from varying amounts of adaptation speech was applied. First, the number of adaptation speakers was varied; as with individual speaker adaptation, 50 utterances from each speaker were used. Performance was calculated for 3, 5, 10, and 15 adaptation speakers. Second, the number of speakers was fixed, but the number of utterances from each speaker was varied. Performance was calculated for 240, 444, 811, and 1296 words, evenly drawn from 10 adaptation speakers. These numbers approximate the number of words in the 3, 5, 10, and 15-speaker adaptation sets.

Results

Figure 4.5 shows the results of applying MLLR with L1 and L2 data. While adaptation with accented L2 data leads to improved performance, adapting with L1 data results in a performance degradation that increases with the amount of adaptation data used.

The benefit from adaptation with larger amounts of adaptation data is clear, at least up to the 10-speaker level. One might wonder whether it is the variety among speakers or simply the number of adaptation utterances from one speaker that contributes most to the gain. Figure 4.6 shows how performance changes when the adaptation utterances are distributed evenly over 10 different speakers. As we can see from Figure 4.6, the curves are steeper when the adaptation words are *not* distributed evenly across speakers; one may conclude that the effect of increasing amounts of adaptation data is stronger when the amount of speech

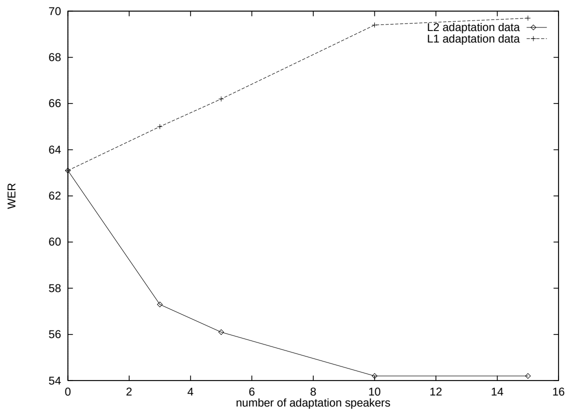


Figure 4.5: MLLR adaptation using L1 and L2 adaptation data and varying numbers of adaptation speakers

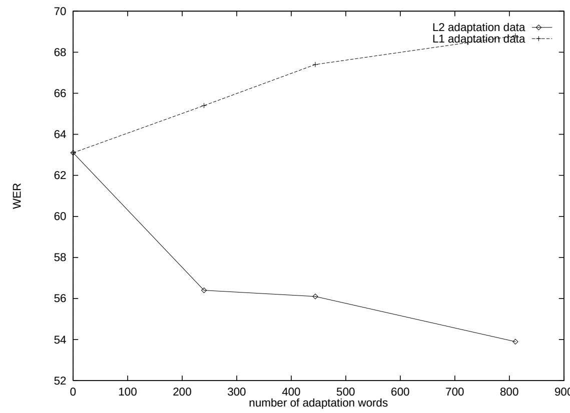


Figure 4.6: MLLR adaptation with with L1 and L2 adaptation data and varying numbers of adaptation words

from each adaptation speaker reaches a critical level. When the adaptation words are distributed evenly across speakers, the benefit is not seen as quickly because there is initially more diversity in the adaptation data set.

Experiment 4.2: MAP adaptation

Method

The MAP adaptation implementation used in these experiments is an approximation to the standard algorithm in which the original means are shifted toward the sample means using a single experimentally-determined interpolation weight, instead of calculating the shift individually for each senone. This method has been found to produce equivalent or better results than the traditional implementation (Soltau, 2001).

Results

Performance after MAP adaptation is shown in Figure 4.7. On the horizontal axis is the degree of shift toward the sample mean (the interpolation weight). When the interpolation weight is 1, the adapted mean is identical to the sample mean. When the interpolation weight is 0, the adapted mean is identical to the prior mean (i.e., there is no adaptation).

As with MLLR adaptation, we see a degradation in performance when adapting with L1 data. When adapting with L2 data, we see that the optimal interpolation weight is 0.75.

Conclusions

A comparison of MLLR and MAP adaption is given in Table 4.8. MAP adaptation performs significantly better than MLLR adaptation, at least when the amount of adaptation data is large.

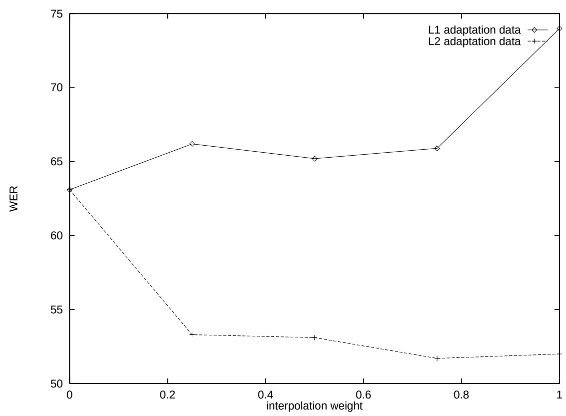


Figure 4.7: MAP adaptation using L1 and L2 adaptation data and varying interpolation weights

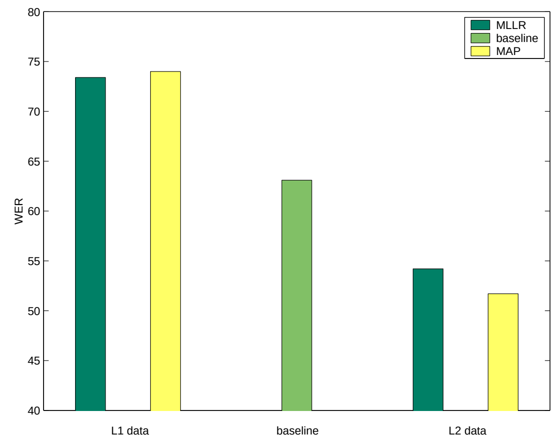


Figure 4.8: Comparison of MLLR and MAP adaptation for 15 adaptation speakers

4.4.3 Adaptation for proficient speakers

The consistent observation in these experiments that adapting with L1 data results in a performance degradation is disappointing, as it reinforces the conclusion seen elsewhere in LVCSR and NLP that clever modeling cannot compete with plenty of well-matched data. It also contrasts with the results cited in (Liu and Fung, 2000a). It was my initial hypothesis that the lower proficiency levels in my test set were responsible; proficient speakers may have a strong accent, but if their speech is stable, it may be easier to attribute consistent mispronunciations to interference from L1. Specific interference from L1 for less proficient speakers, on the other hand, may not influence articulation as much as other effects encountered along the learning curve do.

Unfortunately, this does not appear to be the case. Figures 4.9 and 4.10 show adaptation results for four proficient speakers. We see the same trend as for the less proficient speakers; using L1 data to adapt to the non-native condition results in a performance degradation while L2 data improves performance, and the degradation/improvement grows with the amount of adaptation data. While the improvements are small, the degradation is even more severe than it is for the less proficient speakers, both for varying numbers of adaptation speakers and varying numbers of adaptation words distributed evenly across adaptation speakers.

4.4.4 Conclusions from adaptation experiments

The clear result from experiments performed on the data collected for this dissertation is that adaptation to the non-native condition is successful when accented L2 adaptation data is used and harmful to overall WER when L1 adaptation data is used. This trend holds for both the lower-proficiency speech that is the target of this research and the type of high-proficiency speech that has been more widely studied (although the sample of high-proficiency speech available for this research was small).

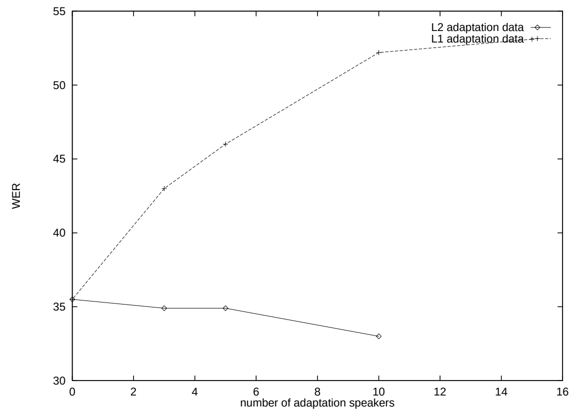


Figure 4.9: MLLR adaptation for proficient speakers varying number of adaptation speakers

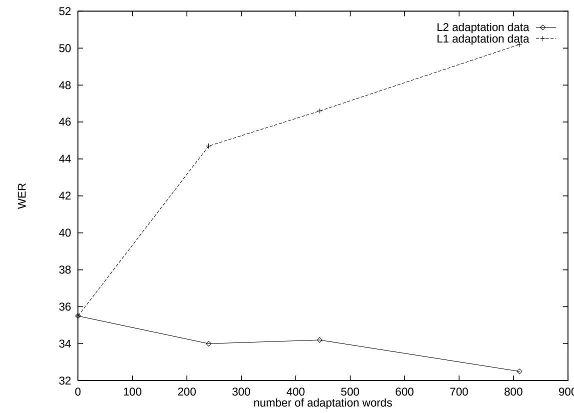


Figure 4.10: MLLR adaptation for proficient speakers varying number of adaptation words

4.5 Training

In Section 4.4, I compared methods and data sources for adaptation, and found that the greatest WER reduction comes with first using MAP to adapt to the non-native condition, and then applying MLLR again to those adapted models to adapt to the current speaker. In this section I show how WER can be further reduced through retraining of the system using L1-dependent data. I investigate whether better results can be achieved with L1 data or accented L2 data, and present a number of variations on the standard training procedure that improve recognition performance.

In discussions of recognition system development, I will focus on two phases: building the decision tree that describes allophonic variation, and refinement of the parameters that describe the probability of a certain sound being associated with a certain acoustic model. The first may be referred to as *clustering*, and the second as *training*. For clarity, I will use the term *building* to refer to the process of creating a new recognition system from scratch, a process which is sometimes also called training.

4.5.1 Experiment 5: Building a system with accent-dependent data

Introduction

It was shown in Section 4.4 that while using accented data for adaptation improves recognition performance, adapting with L1 data results in a performance degradation. In speaker adaptation, the model inventory is kept the same, but the expectation of what a model sounds like is shifted towards what has been seen in the limited set of adaptation speech. The L1 data does not have the chance to make its maximal contribution, as the model inventory is based on the polyphones found in native speech; two allophones that are quite different in L1 may be used to update the same model if the two contexts do not trigger variation in English.

By rebuilding the system based on the contexts that are meaningful in L1, we can use the L1 data to its full advantage.

I will compare a system built with a mixture of L1 and native English data with a system built with a mixture of accented L2 and native English data. The large amount of native data contributes to the robustness of the model, while the smaller amount of L1-specific data ensures that L1-specific phone sequences and phone realizations are seen during clustering and training.

Data

The L1-specific (native Japanese and Japanese-accented English) data used in this experiment was the same as that used for adaptation experiments described in Section 4.4.2.

The L1 data that was used for these experiments was Japanese read news from the Nikkei Shimbun. This data was selected because it was similar in task and topic to the Japanese-accented English data. The data collection methods and environments were identical. Approximately 3 hours of this data was used for training.

The entire training set of Japanese-accented read news data (NN-T-R) was used for this experiment. This set totals approximately 3 hours of speech from 15 speakers.

The test data was the proficiency-controlled non-native set NN-E-R.

Method

The procedures for building the two systems were identical. Both were bootstrapped from the baseline system, with initial labels written using those acoustic models. For each system, a new Linear Discriminant Analysis (LDA) matrix was computed, with codebook and distribution parameters then calculated by k-means and trained for seven epochs. The result of this process was a context-independent system. To incorporate phonetic context, a new model was created and trained for each polyphone whose frequency was above a certain threshold. A decision tree was then grown to find polyphones whose central phones are similar and can be used to train the same model. LDA, kmeans, and Viterbi training were applied again to complete the context-dependent system.

Before decoding the test data, optimal language model parameters were found using cross-validation data, so the language model parameters used in testing the two systems were not the same. Speaker-adapted weights were estimated by applying MLLR on 50 utterances of unseen adaptation data from each speaker.

Results

Figure 4.11 shows the WER reduction achieved by rebuilding the system with L1-dependent data. Results are shown both for the test set average and the individual speakers. The baseline WER is given as a line plot to make it easy to see for which speakers the rebuilt system results in a degradation. Overall, there is no

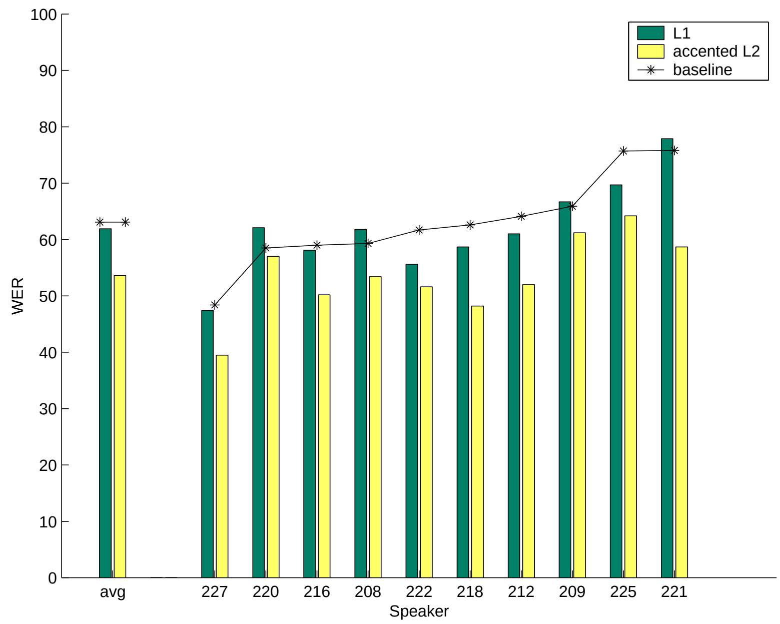


Figure 4.11: WER reduction from rebuilding the system with L1 and accented L2 data

significant difference between the system rebuilt with L1 data and the baseline system. The improvements in the system rebuilt with accented data, however, are highly significant ($p < .001$, matched pairs t-test).

4.5.2 Retraining

Given the observed positive contribution of incorporating accented data in system building, it was of interest to determine whether the effect can be approximated by limiting the specialization to clustering or training. I began with the retraining case, which is the more straightforward of the two. To retrain using the accented data, two Viterbi training iterations were run on the fully trained baseline acoustic models described in Section 4.1. To clarify the effect this has, let us briefly review the Viterbi training process.

Review of Viterbi Training

As described in e.g. (Rabiner, 1990), a hidden Markov model consists of possible states $S = s_1 \cdots s_N$ and observations $O = o_1 \ldots o_M$ and parameters π, A, and B defined as follows:

π the initial state distribution

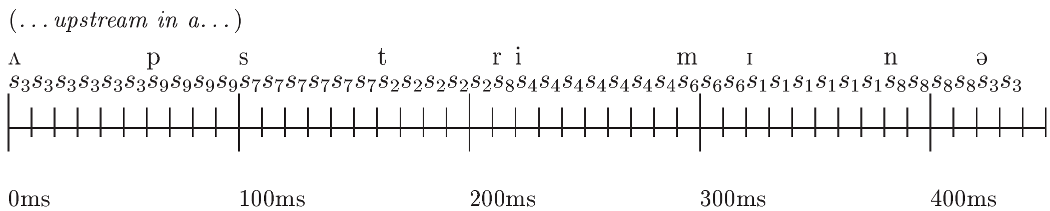
A the state transition probability distribution

B the observation symbol probability distribution

As HMMs are used for speech recognition, an observation o corresponds to an acoustic event that is heard, and the states s_i correspond to phonological units. In this explanation, I will assume that the unit of representation is the phoneme.

In Viterbi training, the values of π, A, and B are iteratively refined to more accurately predict the initial state, transitions between states, and association of states with observations (phonemes with sounds). This is accomplished by first using the current parameters to estimate the most probable sequence of states, and then updating the parameters based on the number of times each state and observation were seen.

The Viterbi algorithm (Forney, 1973) is used to find the state sequence $q_1 \cdots q_T$ that best matches the acoustic sequence given the model parameters. In training, the word sequence, and therefore the prescribed phoneme sequence, is known. However, the exact time alignment of states must be established. For the word sequence "upstream in a," discussed in Section 4.3.1 (assuming a somewhat faster speaking rate), the true time alignment might look like the following.



There are a number of factors that make arriving at the correct alignment difficult, including poor initial modeling of some phones, noises, silences between words, and phone transitions that don't fall at 10ms intervals. The alignment typically gets more accurate with each training iteration, because the model used to estimate it improves.

After an alignment has been found, the model parameters are updated so that the model is optimal given the new counts.

4.5.3 Experiment 6: Retraining with non-native data

Introduction

When a system that has been built on native speech is trained with non-native data, the updates to the model parameters will reflect the sound-state mappings that are present in the data. If the non-native speakers are consistent in their deviations from native speech, the model shift should result in better recognition. If the non-native data is inconsistent, however, using it to train the model can result in a general degradation of the model.

In Section 4.4, I showed that recognition improves with speaker adaptation. By training using the accented data, I am essentially extending this approach, updating not only the mixture means but also the mixture weights and covariances (the full representation of the observation model B). We also benefit from the iterative component of the training process. Based on the improvements that were seen with adaptation on accented data, one would expect that the model does improve with training on accented data.

Data

The accented L2 (Japanese-accented English) data used in this experiment was the same as that used for adaptation experiments described in Section 4.4.2 and rebuilding experiments described in Section 4.5.1.

The entire training set of Japanese-accented read news data (NN-T-R) was used for this experiment. This set totals approximately 3 hours of speech from 15 speakers.

The test data was the proficiency-controlled non-native set NN-E-R.

Method

The baseline acoustic models described in 4.1 were trained two additional forward-backward iterations using only the 3 hours of accented data.

Results

Table 4.11 shows the results of training two epochs on the same 15 training speakers (representing 3 hours of acoustic data) that were used for adaptation experiments. The improvement in overall WER was highly significant as measured by the matched-pairs test described in Section 4.2.

Speaker	**baseline WER**	**retrained WER**
208	64.8	42.9
209	65.0	74.2
212	74.0	54.2
216	59.6	40.8
218	64.6	36.4
220	64.7	59.1
221	92.2	38.6
222	57.4	36.5
225	77.3	53.9
227	53.6	34.8
AVG	67.3	47.2

Table 4.11: Improvements in WER for the retrained system

Conclusion

Retraining in only the final phase with the accented data results in a significant drop in WER, yielding the best performance so far.

4.5.4 Experiment 7: Model interpolation

Introduction

In an effort to decrease word error further, I experimented with model interpolation. As the retrained acoustic models (from here on called *non-native models*) were trained on a small amount of data, there is a danger of overfitting, a problem which has been addressed by smoothing the models via interpolation with a more robust model (e.g. Huang et al. (1996)). In the native and non-native model sets, there is a one-to-one mapping between senones (atomic acoustic units, generalized sub-triphones in ISL-BN; c.f. Hwang (1993)) representing the same phonetic context. In the native model, the mixtures of Gaussians are based on many training samples, while in the non-native model, the mixtures of Gaussians are probably overfitted to the non-native training data. My goal is to move the non-native distribution towards the native distribution to the point of maximum robustness.

Data

No acoustic data was involved in this experiment. The two model sets that were interpolated were the baseline model set and the retrained model set generated from Experiment 6.

The test data was the proficiency-controlled non-native set NN-E-R.

Method

To achieve the goal of moving the non-native distribution towards the native distribution to the point of maximum robustness, I interpolated each element of the corresponding native and non-native mean and covariance vectors as well as the distribution weights. Specifically, for each non-native senone S^A in a system with R mean vectors in each codebook and an underlying feature space dimensionality of N, the mean vector μ, the covariance matrix C, and the distribution weight vector d are interpolated with those of the native senone S^B to create senone model S^C:

$$\forall i \in R.\forall j \in N.\mu_{ij}^C = \frac{\mu_{ij}^A w + \mu_{ij}^B(1-w)}{2}$$

$$\forall i \in R.\forall j \in N.C_{ij}^C = \frac{C_{ij}^A w + C_{ij}^B(1-w)}{2}$$

$$\forall i \in R.d_i^C = \frac{d_i^A w + d_i^B(1-w)}{2}$$

Where w is the experimentally determined weighting factor.

The new covariances were calculated in this way in order to find a medium between the smaller variances in the native models and the larger variances in the non-native models. It was not my intent to re-calculate them to represent the variance across all native and non-native samples. The counts that are stored to record the number of times each senone was seen in the training data were also updated.

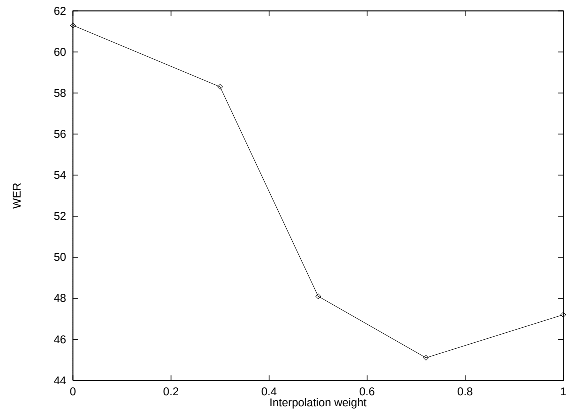


Figure 4.12: Results for interpolation with different interpolation weights. A weight of 0 represents performance with the original acoustic models. A weight of 1 represents performance with the new models.

Results

Figure 4.12 shows the effect on word error rate of interpolating with different weights w. The optimal weighting factor was found to be .72; this contrasts with the result in Witt and Young (1999), which found the optimal weighting factor to usually be less than .5 with a similar interpolation scheme.

Conclusion

The model interpolation yields an improvement of 6.25% relative over the retrained models, which is significant. The fact that an improvement is achieved at all indicates that there is a small overtraining effect with the retraining; the retrained models are slightly overspecialized toward the specific speakers in the training set NN-T-R, and interpolating these models back with the baseline models adds robustness that leads to better performance on unseen test speakers.

4.6 Clustering

Non-native speakers are known to have difficulty acquiring context-conditioned phonetic contrasts when the L2 phoneme is perceived as corresponding to an L1 phoneme that is not subject to, or does not trigger, the same variation (Flege, 1993). For example, in English, the word-final stop contrasts /p,b/, /t,d/, and /k,g/ are distinguished not only by voicing but also by length of the preceding vowel. This effect is so profound that even when the final phone itself displays the correct voicing characteristics, if the length of the preceding vowel is inappropriate the final phone can easily be mistaken for its voiced/voiceless counterpart. Japanese, on the other hand, exhibits context-conditioned variation that does not occur in English; voiceless consonants can trigger devoicing of the following high vowel and come consonants undergo heavy palatization preceding /i/. If the Japanese speakers are carrying these allophonic relationships over into their English articulation, and failing to observe those appropriate in English, the context decision tree that was built on native speech may not represent very accurately the environments that are phonologically critical for them.

It is not a certainty, however, that the native decision tree will not arrive at an acceptable model for a segment of non-native speech, or that a decision tree trained on non-native data will specify a better model. To understand why, let us consider the decision tree growing process.

4.6.1 Review of phonetic clustering

The purpose of phonetic clustering in JRTk is to find the phonetic units that behave similarly in an environment and pool examples of them to build a single model. The phonetic unit that the ISL-BN system uses for this is the sub-phone: the beginning, middle, and end of a phoneme are recognized as separate units. Number and consistency of training examples contribute to the quality of the model; the clustering procedure uses information about the phonetic environment to group acoustic samples in the way that maximizes both consistency and number of training examples in each group. Modeling at the sub-phone level allows data for the middle part of a phone, which may show little effect from neighboring phones, to be pooled, while the beginnings and ends may be more appropriately modeled separately as features like voice onset and release vary according to context.

JRTk uses a decision tree to find the optimal groupings and classify input speech samples in decoding. Questions about the previous and following two phonemes are asked to find the split that creates the best two new data subsets. Figure 4.13 shows what the tree might look like. In the case of the phone /l/, the most important question (measured in terms of entropy reduction) is whether or not the current phone occurs at a word boundary (`0=wb?`). Because JRTk represents both word ends and word beginnings as word boundaries, a second question is asked to determine whether the current /l/ is word final (`+1=wb?`). When the answer to this question is no (`n`), the tree stops asking questions, indicating that differences in realization of word-initial instances of /l/ are not significant enough to warrant specialized modeling. All acoustic samples of word-initial /l/ are "bucketed" together to build a single model, designated model 48.

Model 73 is also defined fairly early in the tree. This model represents instances of /l/ that are preceded by a /u/ but are neither word-final nor word-penultimate. We can see from the number of counts in model 73's bucket that occurences of this context in the training data were relatively rare. It is likely that the samples were bucketed together at this point not because they were similar but because their number had approached the minimum required for creating a model.

4.6.2 Native trees and non-native input

In the previous section, I alluded to the two reasons that training data samples are bucketed together to render an acoustic model: similarity and sparsity. As long as test speakers exhibit the same characteristics as the training speakers with respect to these two features, the acoustic models will describe their speech as well as they did the training speakers. What happens when the training and test data is mismatched?

The English word "Pacific" is familiar to many Japanese speakers. It is lexicalized in Japanese, occuring, for example, in the name of a popular sports league. It is phonologically simple, and its realization in

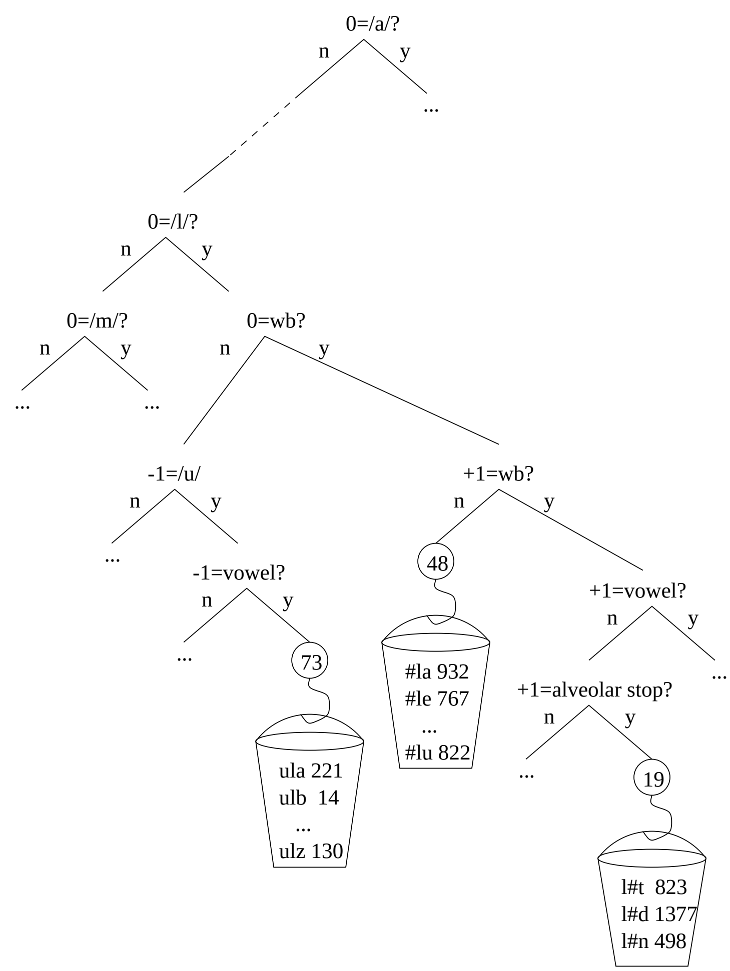


Figure 4.13: Example fragment of a phonetic clustering tree. At each leaf node, a bucket is shown holding the contexts found in the training data that were assigned to that node and their respective counts

the non-native training data was very consistently [paʃifikɯ̥], contrasting with the likely native realization [pəˈsɪfɪk˺]. It is easy to see how the native acoustic models, which are designed to distinguish pairs like [i,ɪ], [s,ʃ], and [k,k˺], will not give the intended phone sequence a high score in decoding. Let us explore training the phonemes /ɪ,s,k/ with Japanese-accented samples of the word "Pacific" and the native-based decision tree.

/ɪ/ and /i/ are highly contrastive in English. Although each exhibits allophonic variation, notably in duration for /i/ and reduction for /ɪ/, these symbols do not normally describe the same phonetic event.[4] They are acoustically very close, however, and many questions are asked in the decision tree in order to properly model /ɪ/ in particular. In Japanese-accented English, nearly all instances of /ɪ/ are realized as [i]. The pool of training data for /ɪ/ may be split up necessarily as samples are assigned to contexts that are meaningful in English, but there is no failure on the part of the system to identify environments in which /ɪ/ undergoes allophonic alternation, and this is a relatively common phone which should not suffer greatly from data splitting. When speaker adaptation is subsequently applied, a general /ɪ/ → [i] mapping should be learned.

In Japanese, /s/ preceding /i/ is always realized as [ʃ]. It can be very difficult for Japanese natives to produce the English phone sequences [si] (and by extension [sɪ]). Because English speakers do observe a contrast in this environment, it might be thought that the English decision tree would not isolate this allophone of /s/ for specialized modeling, bucketing acoustic samples that are close to /ʃ/ in with more pure /s/ examples. However, in the baseline decision tree, the questions `+1=syllabic`, `+1=front-vowel`, and `+1=high-vowel`, describing the phonetic context __{/i,ɪ/} are the first to be asked. Although /ʃ/-/s/ substitution before /i/ is an error that can significantly decrease intelligibility and lead to confusion in the search, it is not modeled inappropriately in the decision tree, and individual speaker adaptation should address the realizational problem.

The epenthetic [ɯ̥] appearing after /k/ presents a different kind of problem. This is not a substitution error, but rather an insertion error, resulting in the new polyphone /fɪku#s/ (assuming that the next word is "salmon"). A pre-/u/ context is recognized in the decision tree, but more specific modeling of the full context was not deemed necessary in native-based clustering. It is possible that this /k/, and the following /u/ (if the epenthetic vowel is to be modeled as /u/), will benefit from clustering with non-native data.

Although accented speakers vary in allophonic distribution in ways that native speakers do not, it is not necessarily the case that non-native-based clustering will help, as I have attempted to illustrate. In the following sections, I address this question empirically, describing two methods for incorporating non-native data in the clustering process.

[4] I have thus far avoided characterizing the phone inventory used in recognition as phonetic or phonemic. It is in fact inconsistent. Certain phonetically distinct sounds, such as [k,k˺,kʰ], are transcribed as the phonemic /k/ with the expectation that contextual clustering will assign them to different models. In other cases, allophonic variants are assigned full phonemic status. Morphophonological variation is always represented phonetically.

Training data source	**Hours of data**	**Quinphone coverage**		**WER**
		types	tokens	
Native	60	92%	92%	63.1%
Non-native	3	50	57	78.4
Non-native (cheating)	3	91	99	40.8

Table 4.12: Effect on WER of re-growing the tree with non-native data

4.6.3 Re-growing the tree

By re-growing the tree from scratch with a sufficient amount of non-native data, one would expect to capture important patterns of allophonic distribution in Japanese-accented English. I did not find the three hours of training data to be sufficient for this task, however. The number of polyphones is small; only 10% of the polyphone types (46% of tokens) in the full native training data set appear in the small non-native training data set. By contrast, 92% of the polyphone types (and 92% of tokens) in the non-native data appear in the 60 hours of native data.

Table 4.12 shows how recognizer performance degrades when the tree is trained with only the small non-native data training set. The WER figures represent performance after post-clustering LDA, kmeans, and training on the 3 hours of non-native data. Results from clustering with native and non-native training data are contrasted with the result from a cheating experiment, in which training speakers' readings of the evaluation article were included in the training data. When all of the evaluation polyphones (although the coverage does not actually reach 100%, as reading errors and disfluencies add new polyphones for each speaker) are represented by multiple examples in the training data, word error decreases dramatically. Although the new decision tree may handle the polyphones it has seen in sufficient quantity in the non-native data more appropriately than the native tree would, the overall system suffers greatly from the loss of the robustness that the native tree provides.

4.6.4 Experiment 8: Decision tree adaptation

Introduction

In order to include questions relevant to non-native speech in the decision tree without rebuilding it from scratch, I adapted the Polyphone Decision Tree Specialization (PDTS) (Schultz and Waibel, 1999) method for porting a decision tree to a new language. This method was originally designed to support multilingual recognition systems that use data from a number of different languages to train models representing a broader range of phonemes than would occur in one language. Each time a new language is added, it brings with it phonemes and polyphones that have not yet been seen by the system. PDTS allows questions to be asked about these new polyphones in the decision tree and new model mixture weights to be trained for them without discarding the questions about the polyphones that the new language shares with the old one.

Data

The accented L2 (Japanese-accented English) training data used in this experiment was the same as that used for adaptation experiments described in Section 4.4.2 and rebuilding experiments described in Section 4.5.1.

The entire training set of Japanese-accented read news data (NN-T-R) was used for this experiment. This set totals approximately 3 hours of speech from 15 speakers.

The test data was the proficiency-controlled non-native set NN-E-R.

Method

While I am not working with a new language, phone substitution, elision, and epenthesis in non-native speech can introduce many new polyphones, as was shown in Section 4.3.2. To use the PDTS method, I first identified new polyphones by aligning the training utterances using the expanded dictionary described in Section 4.3.2. Included in the dictionary were variants generated from linguistic rules, free phoneme recognition, and underspecified alignment. The recognizer selected the best acoustic match for each word during alignment, generating a list of new polyphones. The new polyphones were then integrated into the decision tree, with branches pruned back to the point where the new polyphone data could be inserted, and re-grown with new specialization where the new data showed sufficient internal diversity or divergence from the native data.

Results

Although I observed a large performance gain from PDTS on cross-validation data, only a small improvement over the baseline was seen for test data, as shown in Table 4.13. The cross-validation data is used to find the optimal language model settings before evaluation on the test set. Recognizer performance on this data set is normally an accurate predictor of recognizer performance on the test data, as verified by periodic spot checks. However, as we can see from table 4.13, the cross-validation data was quite positively affected by PDTS where the test data was negatively affected. This trend held for varying pruning thresholds, the number of polyphone samples necessary in the adaptation data to justify a new branch. It is difficult to understand why this should be the case; cross-validation, test, and adaptation speaker sets are all mutually disjoint, and the test utterances used for evaluation on both cross-validation and test speakers were not included in the adaptation data. Because all evaluation speakers are reading the same article, there is no dependency on the number of new polyphones. A check of the language model parameters on the test data confirmed that the settings that were selected as optimal during cross-validation were also optimal for the test data. The cross-validation speakers did have slightly lower proficiency ratings than the test speakers, so one possible (and intuitively plausible) explanation would be that PDTS is more effective for lower-proficiency speakers.

	Cross-validation data	Test data
Baseline	61.6	63.1
Baseline dictionary	59.6	60.3
Expanded dictionary	56.5	65.9
Expanded dictionary and higher threshold	54.9	64.9

Table 4.13: System performance after PDTS

Conclusion

The question of why PDTS performed better for the cross-validation speakers than for the test speakers is interesting, but somewhat tangential as the standard methodology for evaluation in LVCSR bars us from investigating individual differences in performance between test and cross-validation speakers. More relevant is the question of why PDTS did not perform better in the main evaluation on the test speakers. As has been mentioned earlier in this section, PDTS only grows a new set of branches for polyphones that did not appear in the training data. While we do not have new polyphones in the prescribed pronunciation as we would if adding a new language, I demonstrated in Section 4.3.2 that the phonetic realization of words in non-native speech contains polyphones that are not found in native speech and would not have been considered in building the decision tree. In this respect, there is a potential for seeing the same sort of improvement that Schultz and Waibel (1999) observed when adding Portuguese to a multilingual system. However, in the case of a new language, adaptation and test speakers are native speakers and can be expected to exhibit consistency in allophonic variation – this is the premise supporting the entire decision tree clustering approach that has worked so well in LVCSR. When speakers are not natives or proficient non-natives, they may not share tendencies to similar environmental influence as they individually approach articulation of English. The observation that performance with the baseline dictionary, in which only native polyphones are considered, is stronger than with the expanded dictionaries, which allow the newly-trained polyphones, is evidence to support the hypothesis that although new polyphones do exist in non-native speech their realizations are not consistent enough across speakers to benefit from specialized modeling.

By examining only allophonic behavior in contexts that are not found in English, PDTS also does not take into account variation in the many contexts that *are*. For example, there is quite some variability in Japanese natives' realization of English /f/ and /h/. [f] does not occur in Japanese other than in loanwords. /h/ preceding /u/ is realized as a bilabial fricative, and depending on the speaker may sound to a native GA listener as either [f] or [h]. Loanwords that originally contained /f/ may be realized with either [f] as in [ofisɯ] ("office") and [h] as in [terehon] ("telephone"). Confusion in nativization of loanwords, speaker variability in realization of [ɸ] in Japanese, and reduced articulatory performance when concentrating on speaking English all contribute to a general inconsistency in production of /f/ and /h/; the transcriptions contain a number of such substitutions as "feet/heat," "who'd/food," and "follow/hollow." The baseline decision tree for /f/, however, buckets together all contexts in which the following segment is a rounded

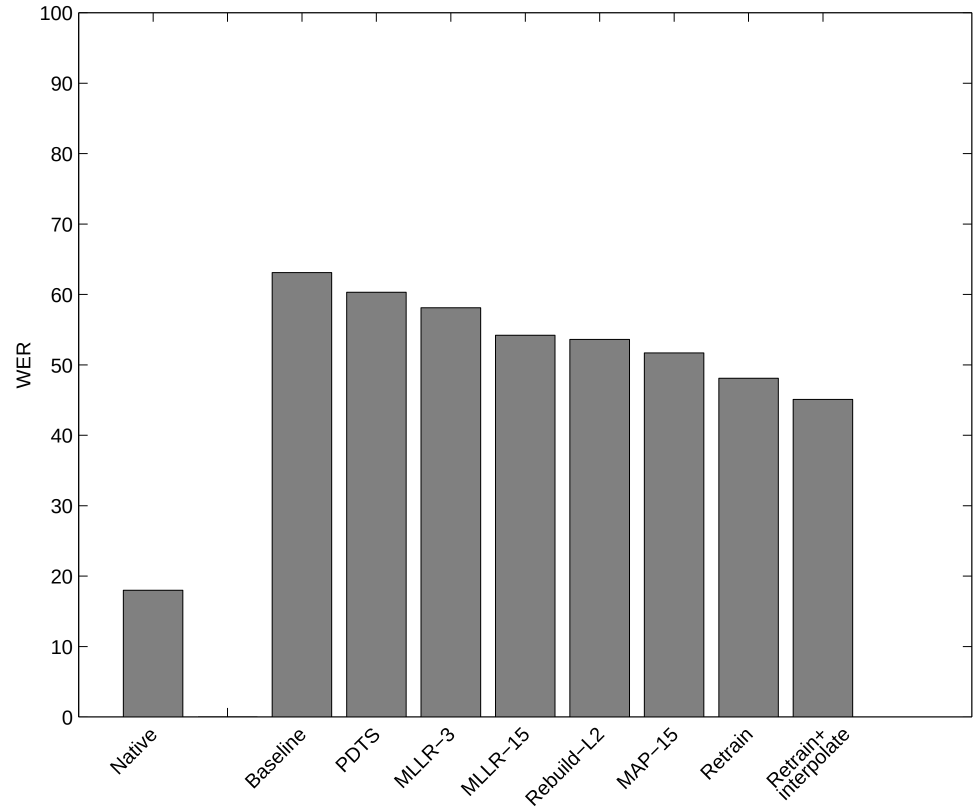


Figure 4.14: Summary of acoustic modeling results

vowel that does not precede /w/,/r/, or /ɚ/. No distinction is made between high round vowels and low round vowels. The decision tree would never learn through PDTS that /f/ behaves differently before /u/ than before /ɔ/ because the polyphones occur in English and have already been accounted for in the decision tree.

4.7 Summary of acoustic modeling results

In this chapter, I have shown how application of acoustic model training and adaptation techniques contributes to increased recognition accuracy on non-native speech. A summary of the individual contributions of each method is shown in Figure 4.14.

The baseline word error rate for the proficiency-controlled set of non-native test speakers was 63.1% after MLLR speaker adaptation. Adapting the allophonic decision tree to the non-native condition (PDTS) reduces WER to 60.3%. Acoustic model adaptation to the non-native condition via MLLR adaptation on three adaptation speakers (MLLR-3) prior to test speaker adaptation reduces WER to 58.1%. MLLR adaptation to the non-native condition with 15 adaptation speakers (MLLR-15) reduces WER to 54.2%. Rebuilding the system from scratch with accented data (Rebuild-L2) reduces WER to 53.6. MAP adaption with 15 (MAP-15) speakers reduces WER to 51.7%. Additional training iterations using 3 hours of non-native speech (Retrain) reduces WER to 48.1%. Finally, interpolation of the retrained models with the baseline models with an interpolation weight of .3 reduces WER to 45.1%, a 29% relative reduction in error over the baseline.

Among the techniques that I did not find to work well on this data were rebuilding the system with L1 data, adaptation with L1 data, and PTDS with additional training.

Phonetic confusion is much higher in the non-native data than in similar native data, with the most confusible phone pairs in non-native speech being /m̩,m/, /ɔ,o/, /f,h/, /b,v/, and /u,ʊ/. A number of these confusions are also significant in native speech; /u,ʊ/ and /b,v/ were the most confusible pairs for native speakers. Other confusions that were notably higher in non-native speech include /ʤ,ʒ/, /ɛ,eĭ/, and /ʃ,s/.

As discussed in Chapter 3, non-native speakers make use of a variety of strategies as they build their competence in spoken language. Phonological simplification, such as insertion of vowels to break up consonant clusters and failure to observe complex allophonic patterns in the second language, can introduce phone sequences that never occurred in the training data and were not incorporated into the polyphone decision tree. Although a flexible alignment of non-native utterances to reference text revealed that there are indeed many new polyphones in the non-native speech, of the decision tree to the non-native speech resulted in only a small improvement in recognition accuracy. Possible explanations include that environmental influence is not consistent across speakers or within one speaker's articulation; that differences in allophonic alternation in environments that exist in both English and Japanese are more significant than expected; and that phone insertions, deletions, and substitutions are effectively absorbed in the course of speaker adaptation.

Chapter 5

Lexical Modeling

The lexical model specifies how phones combine to make words. By modifying the native lexical model we can represent segmental substitutions, insertions and deletions frequent in non-native speech. If speakers of a common native language are known, or are found, to systematically substitute[1] one phone sequence for another, this substitution can be incorporated in the lexical model for a more accurate representation of the phonemic realization of words.

There are several problems with lexical modeling that make it not as straightforward a solution to adapting to foreign accents as it might seem. First, a more accurate phonemic representation may not be linked to an increase in recognizer accuracy. Second, context-sensitive speaker adaptation is very effective in learning speaker-dependent deviations in phonetic realization, and independently modifying the phonemic representation may counteract the benefits of adaptation. And third, whether substitutions accented speakers appear to make are true phonemic substitutions is an open question, as discussed in Section 2; neither human perception nor recognizer error is an unbiased indicator of the underlying form of non-native speech.

Nevertheless, lexical modeling is a non-data-intensive, linguistically intuitive approach to adapting to non-native speech that has been applied with success in alignment-based tutoring applications (Auberg et al., 1998) and limited domains (Livescu and Glass, 2000) and for new varieties of native speech (Humphries and Woodland, 1997). Direct modification of the lexical model also seems appropriate for L2 words that have been nativized in L1, although one must be wary of arbitrarily assigning L1-L2 phone mappings.

In this chapter, I compare data-driven and linguistically-motivated methods for finding probable phonemic representations of English words in Japanese-accented speech.

[1]Throughout this chapter, I will use the term *substitute* to refer to replacement of one phone sequence with another, subsuming the insertion case and the deletion case.

5.1 Background

There are two primary considerations in lexical modeling: specifying probable phone sequence transformations and incorporating them, for optimal recognizer performance, in the search. Transformations can be specified either by predicting, based on linguistic evidence, likely mappings between L1 and L2 phones, or by inferring mappings from recognizer output. Both methods have been found to be successful in different contexts. Fung and Liu (1999) based mappings between English and Cantonese on average formant frequencies in native speech. Auberg et al. (1998) and Kawai (1999) selected mappings based on the minimal pairs that were to be taught in their language tutoring systems. Humphries and Woodland (1997) found that British phone representations of American speech could be derived from unrestricted phoneme recognition of American data using a British system. Similar data-driven approaches to transformation inference have been used by Huang et al. (2000) for Mandarin dialects, Amdall et al. (2000) for proficient non-native speakers of English, and Suzuki et al. (2000) Japanese-accented English.

Once a description of potential variation has been completed, the list of actual variants for base lexical forms that will be allowed in the search must be compiled. Let us take as an example the English word "abroad." Generating all combinations of the sample phonemic substitutions /ə/ → /ɒ/,/b/ → /bu/, /r/ → /l/, /ɔ/ → /o/, and /d/ → /do/, all reasonable for Japanese-accented English, yields 31 variants:

/ɒblɔd/	/ɒbrɔd/	/ɒbulɔd/	/ɒburɔd/	/əblɔd/	/əbrɔdo/	/əbulɔdo/	/əburɔdo/
/ɒblɔdo/	/ɒbrɔdo/	/ɒbulɔdo/	/ɒburɔdo/	/əblɔdo/	/əbrod/	/əbulod/	/əburod/
/ɒblod/	/ɒbrod/	/ɒbulod/	/ɒburod/	/əblod/	/əbrodo/	/əbulodo/	/əburodo/
/ɒblodo/	/ɒbrodo/	/ɒbulodo/	/ɒburodo/	/əblodo/	/əbulɔd/	/əburɔd/	

A thoughtful implementation of potential Japanese-English transformation rules, allowing commonly observed substitutions only in contextually plausible positions, generates an average of 40 variants per base word in the lexicon. This is not a tractable search space for the recognizer, both in terms of sheer size and in terms of confusability; new variants are very similar to existing words, and discriminating between them becomes an extremely difficult task.

Effective prioritization of variants, then, is critical. Humphries and Woodland (1998) suggest using a decision tree to choose the most probable variants given phonemic context, with a maximum of four variants per word. Amdall et al. (2000) select transformation rules based on log likelihood in an adaptation set, pruning the list using a pruning heuristic. Livescu and Glass (2000) rank rules by maximum likelihood in training data and determine a pruning factor by evaluating performance on development data.

In the next two sections I will document the response of the recognizer to a number of prioritization and pruning methods for linguistically-motivated and data-driven modeling of the non-native data set. All recognition experiments use the best-performing acoustic models described in Chapter 4.

5.1.1 Terminology and phonetic symbols

The symbols that I use to represent sounds in speech will be familiar to users of IPA representations. This symbol set will help to facilitate a common understanding of the transformations that I describe. It is important to be clear, however, that the symbols actually used in the lexicon represent something slightly different. In this chapter, I use phonetic symbols to illustrate four different things. The canonical pronunciation of a word is an abstraction which will be described using IPA symbols delimited by slashes, that is, a standard phonemic specification. The realization of a word in speech will be described using IPA symbols delimited by brackets, a phonetic specification. Transformations actually applied to the lexicon will also use IPA symbols, but without delimiters, so as not to imply that the symbols in the lexicon correspond to any precise IPA specification. In discussions of the internal representation or output of the recognizer, I will use the ARPABET symbol set, which is described in Appendix C.

5.2 Linguistically-motivated modeling

Acquisition of non-native phonology, as noted in Chapter 2, has been very well-studied, in terms of both the general acquisition process and the specific case of Japanese-accented English.

5.2.1 Some phonological properties of Japanese-accented English

While we have conflicting reports of the nature of phonetic production in non-native speech, literature in ESL describes consistent trends in the English of Japanese natives that can be used for empirical evaluation.

Epenthesis

Japanese has a strict (C)V syllable structure, the only exceptions being /n/, which can be syllabic, and geminate consonants. Vowel length is phonemic. Epenthesis of the vowels /i,o,u/ to simplify consonant clusters and force open syllables is common in Japanese-accented speech. These intrusive vowels have been shown to affect intelligibility (Tajima et al., 1997), and frequency of epenthesis has not been found to be linked to familiarity with or nativization of the word (Tajima et al., 2000). Because vowels are often devoiced following a voiceless consonant in Japanese (Akamatsu, 1997), epenthetic vowels in Japanese-accented English can be very subtle.

Full-quality vowels

Japanese has a five-vowel system, with vowels realized in positions similar to the first, second, fifth, seventh, and eighth cardinal vowels [i,e,a,o,u]. Vowels are always full quality, and sequences of vowels are not diphthongized. The system of vowel reduction in English is not easily acquired by Japanese speakers, which can significantly affect intelligibility as full quality in vowels is linked to a perception of stress for native English listeners (Giegerich, 1992).

Confusion stemming from nativization and orthography

English words represented in the Japanese syllabary are ubiquitous in Japan. Loanwords are frequent, the native syllabary is sometimes used to make the introduction to formal study of English more gentle, and movie posters, newspapers, and karaoke screens are all likely to contain foreign words and names rendered in the Japanese script. This easy dependency on a familiar orthography facilitates fossilized mappings of English sounds to Japanese ones. Moreover, because some English sounds are represented by the same Japanese characters (/l/ and /r/, /v/ and /b/, /ɒ/ and /ʌ/ among others), speakers may not only have trouble with the phonetic distinction but also with remembering which the original phone was. The former situation may be addressed with adaptation to the speaker's idiosyncratic realization of the target phone, but the inconsistency introduced by the latter may well be best addressed by allowing multiple variants in the lexicon.

Observation of Japanese allophonic patterns

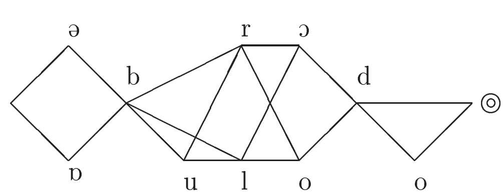
There are some notable allophonic alternations in Japanese that are not found in English. For example, /s/ is realized as [ʃ] preceding /i/. While speakers with formal exposure to English are generally aware that the phonetic distinction between [si] and [ʃi] is contrastive, production is often a problem.

5.2.2 Transformation rules

Based on research in ESL for Japanese natives, a set of context-sensitive transformation rules was compiled. For each word in the lexicon, an arc was added to the pronunciation network for each applicable substitution. For the example in Section 5.1 of the word "abroad," we have the following base pronunciation network.

◎ —— ə —— b —— r —— ɔ —— d —— ◎

If we recognize the potential substitutions /ə/ → /a/,/b/ → /bu/, /r/ → /l/, /ɔ/ → /o/, and /d/ → /do/, we obtain the following pronunciation network, which generates all of the variants listed in Section 5.1.



Rule	**Word**	**Canonical pronunciation**	**Sample realization**
r → l	reason	/rizən/	[lizan]
∅ → o / {t,d} __ #	adult	/ədʌlt/	[adaɾɯto]
j → ∅ / $ __ i	year	/jir/	[iɚ]

Table 5.1: Sample transformation rules. The symbol $ represents a syllable boundary

Pronunciation networks are created in this way for each base word in the lexicon. A full list of transformation rules is given in Appendix B; several examples are shown in Table 5.1, each with a phonetic transcription of an instance of a word in the training data in which the transformation was observed.

However accurate the rules, the application is not foolproof because the base lexicon contains a number of transcription inconsistencies. For example, the syllable-initial /r/ in words like "generator" is often transcribed as a syllable-final /ɚ/, leaving a vowel at the head of the next syllable. This means that the rule ɚ → ɒ / __ $, which generates the appropriate variant /sɚʧɪŋ/ → [sɒʧɪŋ] for "searching" also generates the inappropriate variant ʤɛnɚeĭtɚ → ʤɛnɒeĭtɚ for "generator." Also, compound words appear to behave differently at component boundaries than the same phone sequence would at an ordinary syllable boundary, and this sort of compositional information is not available in the lexicon. However, because this lexicon is only used for bootstrapping the variant extraction process, it does not appear that the spurious paths have a negative effect.

5.2.3 Associating probabilities with transformations

Having established which transformations would be allowed, I next explored ways of assigning probabilities to individual transformations and transformation combinations. Enumerating all paths through the new pronunciation networks yields 915,672 realizations for 22,761 words, compared with 26,110 realizations in the baseline lexicon. Using this very large lexicon, I aligned the acoustic data from training set NN-T-R (which will not be used for further training) to the transcripts. All variants were assigned equal initial probabilities, so the one representing the closest acoustic match was selected during alignment.

This process generates a list of realizations that occurred in the recorded data. There are several ways to interpret the list.

Word-based interpretation

In a word-based interpretation, variants that were selected during forced alignment are added to the test lexicon. This approach has the disadvantage of not generalizing to words that were not encountered in the alignment data. However, it has the advantage of ensuring that all new variants are plausible, which is not necessarily the case when applying transformation rules to new words. I tested two implementations of word-based transformation:

1. l → r
2. r → l
3. ɪ → i

Table 5.2: Rules applied in dictionary **R1**

> **W1** Variants that represented more than 20% of occurrences of the base word in the alignment data were selected for the test lexicon
>
> **W2** Variants that occurred more than twice in the alignments were selected for the test lexicon

These thresholds were determined by two criteria: keeping the dictionary size to less than 60,000, and not exceeding an average of three pronunciation variants per word. Implementation **W1** is biased toward infrequent words; if a word appears only twice in the training data and one instance is a variant, that variant will exceed the minimum frequency threshold of 50% and be added to the lexicon. The nature of my task makes this bias particularly strong. Because many words only occur in one article, and no two training speakers read the same article, if a speaker's pronunciation of that word is idiosyncratic the probability of the variant matching his speech will be high. Implementation **W2** is biased toward frequent words.

Rule-based interpretation

In a rule-based interpretation, instead of adding the exact variants that were selected during alignment, one finds the rules that were most frequently invoked to generate the variants selected during alignment and apply them to the test dictionary. This method generalizes easily to new data, but because it operates on all words in the test lexicon only a few transformations can be implemented without exceeding the optimal lexicon size.

> **R1** Rules that applied more than 500 times in the training set were applied to the baseline test lexicon to generate new variants for testing

The selected rules are given in Table 5.2. Because the application of just these three rules expanded the lexicon size to 60,244, no variations on this implementation were tested.

Phone-based interpretation

In a phone-based interpretation, one examines the individual phone substitutions that occurred in the words which were selected during alignment and use them to generate a new lexicon. This method has the same generalization benefit as the rule-based approach. With the additional information about the phonetic environment, however, the application of transformation rules can be restricted based on context.

1. ɪ → i	9. ɚ → ɒ	17. m → mu	25. v → vu	33. k → ku	41. f → fu
2. r → l	10. eĭ → ei	18. z → zu	26. ʤ → ʤi	34. p → pu	42. ɒr → ɒ
3. l → r	11. t → to	19. j → i	27. s → su	35. ŋ → ngu	43. θ → θu
4. ə → ɒ	12. v → b	20. dz → z	28. s → ʃ	36. d → ʤ	44. ɔĭ → oi
5. ð → z	13. ɔ → o	21. d → do	29. w → u	37. ŋ → n	45. wʊ → u
6. ə → i	14. θ → s	22. aĭ → ai	30. aŭ → au	38. g → gu	46. w → o
7. æ → ɒ	15. ʌ → ɒ	23. ʧ → ʧi	31. ʃ → ʃi	39. r → ɒ	47. u → ʌ
8. l → lu	16. ɚ → ɒ	24. ʊ → u	32. ji → i	40. ʒ → ʤ	48. eĭ → ɛi

Table 5.3: Top context-independent phone substitutions in alignment data

In the word-based approach I did not need to find the base-to-variant alignments because the forced alignment result gives us precisely this information. When we replace the words in the aligned utterances with their phonetic expansions, we have instances of both insertion and deletion in the empirical phone sequence, and must re-establish the alignment at the phone level. Because the variant candidates were generated by the phonological transformation rules, I knew which canonical phone sequences could potentially experience a deletion. Of these, only two appeared with significant frequency in the training data: dz → z and ji → i . I elected to treat the sequences /dz/ and /ji/ as single units, allowing them to align to [z] and [i] respectively to allow deletions. Specifically, all instances of syllable-final `D Z` and syllable-initial `Y IY` were replaced with the symbols `D_Z` and `Y_IY` in both the canonical expansions and the empirical expansions. Similarly, allowable insertions were represented by new symbols, so that there were effectively no insertions or deletions. These expansions were then aligned using the NIST `sclite` scoring package (NIST, 2000). It was necessary to resolve some alignment errors by hand:

Text:	solar power is the key
Initial alignment:	`s ow l axr p aw axr IH **** z DH ax k iy`
	`s ow l axr p aw axr IY Z_UW z ** ax k iy`
Correct alignment:	`s ow l axr p aw axr IH z   DH ax k iy`
	`s ow l axr p aw axr IY Z_UW Z ax k iy`

These cases were rare, however, and easy to detect automatically; this example, `Z_UW` was listed among the insertions, but because ∅ → /zu/ was not one of the original transformation rules, its appearance indicated an alignment ambiguity.

A context-independent implementation of the phone transformations derived from these alignments would expand the lexicon very quickly, as in implementation **R2**. Because the top rules in **R2** were all context-independent, applying the top three phone transformations found in the phone-level interpretation yields essentially the same lexicon. The top context-independent phone transformations found in phone-level

Number of substitutions	Lexicon size
1	34645
2	47886
3	60275
4	92682
5	93047
10	186735

Table 5.4: Growth of the lexicon with the application of context-independent substitutions

analysis are given in Table 5.3. Table 5.4 shows how the lexicon size expands with the number of substitutions applied.

I tested two lexicons generated using context-independent phone substitution probabilities.

P1 Only the most frequently occurring phone substitution was applied to the base lexicon to generate the test lexicon

P2 The top two most frequently occurring phone substitutions were applied to the base lexicon to generate the test lexicon

Context-dependent substitution frequencies were calculated for both three- and five-phone windows. The most frequent context-dependent substitutions are given in Tables 5.5 and 5.6. The influence of word frequency is obvious when looking at substitutions given the broader context. The first five can clearly be attributed to occurrence of "the," "fifty/fifteen," "were," "with," and "dollars." This is not necessarily a bad thing, as better modeling of frequent words would be expected to have a greater effect than better modeling of rare words. It is only mentioned so that the bias is understood.

Two applications of context-dependent phone-level substitution were implemented.

P3 Phone substitutions that occurred more than seven times in the context of a given 3-phone window were applied to generate the test lexicon

P4 Phone substitutions that occurred more than once in the context of a given 5-phone window were applied to generate the test lexicon

In the 5-phone window case, the pruning was not necessary to limit lexicon size, but was applied for smoothing purposes.

Implementations **P3** and **P4** estimate probability of a phone substitution in context based on frequency. A contrasting implementation for the wider context used decision tree learning of phone substitutions.

1. ɪ → i / l __ ŋ	9. ɪ → i / k __ ŋ	17. j → i / # __ u	25. ɪ → i / w __ n
2. ɪ → i / r __ s	10. ə → ɒ / m __ n	18. ji → i / # __ r	26. ɪ → i / w __ ð
3. ɪ → i / r __ z	11. ɚ → ɒ / t __ #	19. l → r / ɒ __ ɚ	27. ɪ → i / s __ k
4. ɚ → ɒ / w __ #	12. ɪ → i / w __ θ	20. v → vu / ə __ #	28. ə → ɒ / z __ n
5. l → r / # __ æ	13. θ → s / # __ r	21. r → l / ɛ __ i	29. æ → ɒ / # __ n
6. m → mu / aĭ __ #	14. z → zu / ə __ #	22. l → r / ɪ __ i	30. r → l / # __ i
7. ɪ → i / d __ s	15. l → r / ɚ __ i	23. æ → ɒ / p __ t	31. θ → s / # __ aŭ
8. z → zu / ɪ __ #	16. v → b / # __ aĭ	24. ɪ → i / ʃ __ p	32. dz → z / n __ #

Table 5.5: Most frequent substitutions conditioned on a 3-phone window

1. ə → ɒ / #ð __ ##	9. ɪ → i / #s __ ks	18. d → do / æn __ ##	26. ɪ → i / ## __ t#
2. ɪ → i / #f __ ft	10. ɪ → i / #ʃ __ p#	19. ɪ → i / tr __ p#	27. r → l / fɔ __ ##
3. ɚ → ɒ / #w __ ##	12. ji → i / ## __ rz	20. ɪ → i / #b __ gæ	28. ð → z / wɪ __ ##
4. ɪ → i / #w __ θ#	13. ə → ɒ / aŭz __ nd	21. ð → z / ## __ i#	29. l → r / wɪ __ ##
5. l → r / dɒ __ ɚz	14. æ → ɒ / əp __ tr	22. θ → s / ## __ aŭz	30. ɪ → i / #w __ l#
6. l → r / ri __ i#	15. v → vu / #ə __ ##	23. ə → ɒ / ## __ v#	31. ɪ → i / ## __ n#
7. l → lu / wɪ __ ##	16. ə → ɒ / #k __ nt	24. ɪ → i / #ð __ s#	32. ð → z / ## __ ə#
8. z → zu / #ɪ __ ##	17. ɪ → i / #w __ ð#	25. ɪ → i / ## __ z#	11. v → b / ## __ aĭb

Table 5.6: Most frequent substitutions conditioned on a 5-phone window

> **P5** Phone substitutions were predicted with a decision tree trained on transformations observed in a 5-phone window of context.

The publicly available C4.5 package (Quinlan, 1993) was also used to learn likely transformations. C4.5 requires two input sources: a specification of attributes that should be considered in making a decision about the transformation, and a set of training data that provides the values for those attributes and the correct class for a series of training examples. In my application of C4.5, I allowed five attributes: the canonical identity of the phone whose surface form is to be predicted, and the canonical identities of the two preceding and two following phones. I chose to learn transformations on a word-by-word basis; although the identities of the phones in preceding and following words were available, I did not use them for predicting the surface form of the phone. There were two reasons for this decision. First, as was discussed in Chapter 3, inter-word pauses are twice as frequent in the non-native speech database as in the native speech database (see Table 3.7), and cross-word coarticulatory effects are not strong. Speakers tend to pronounce words one by one, as they have learned them. Second, when we are building the new pronunciation networks from the baseline lexicon, we have no cross-word context to work with. One can only make predictions based on the phones that make up each word. There would be no reason to use attributes for decision tree growing that we know will not be available for classification. C4.5 does allow a wild card value for attributes, which could be used at word boundaries; I elected to specify a boundary phone value instead, so that word-initial and word-final effects could be considered.

For each word in the training data, then, there were as many training examples provided as phones in the canonical pronunciation. In order to simplify estimation of deletion, phone sequences that could undergo simplifying elision were represented as a single symbol, as described earlier in this section. For the word "abroad," with the canonical form [əbrɔd] and an empirical realization of [abulod], the training data was specified as follows.

Two preceding phones		Canonical form	Two following phones		Surface realization
<s>	<s>	ə	b	r	ɒ
<s>	ə	b	r	ɔ	bu
ə	b	r	ɔ	d	l
b	r	ɔ	d	<s>	o
r	ɔ	d	<s>	<s>	d

5.2.4 Experiment 9: Linguistically-motivated lexical modeling

Introduction

The preceding section described a number of methods for augmenting the pronunciation networks. In this experiment, I test recognition with all eight methods to see if any result in an improvement in recognizer performance.

Data

In these experiments, the test data remains fixed, as for previous experiments, to the proficiency-controlled test set NN-E-R, while the pronunciation lexicon is varied. The eight test lexicons described so far in this section are summarized below.

W1 Variants that represented more than 20% of occurrences of the base word in the alignment data were selected for the test lexicon

W2 Variants that occurred more than twice in the alignments were selected for the test lexicon

R1 Rules that applied more than 500 times were applied to the baseline test lexicon to generate new variants for testing

P1 Only the most frequently occurring phone substitution was applied to the base lexicon to generate the test lexicon

P2 The top two most frequently occurring phone substitutions were applied to the base lexicon to generate the test lexicon

P3 Phone substitutions that occurred more than seven times in the context of a given 3-phone window were applied to generate the test lexicon

P4 Phone substitutions that occurred more than once in the context of a given 5-phone window were applied to generate the test lexicon

P5 Phone substitutions were predicted with a decision tree trained on transformations observed in a 5-phone window of context.

Method

Applications of all eight methods were tested through acoustic rescoring of the word lattice created for each utterance during initial decoding. One of the disadvantages of lexical modeling is that adding pronunciation variants to the lexicon increases confusability in the search. *Lattice adaptation* is a technique which uses

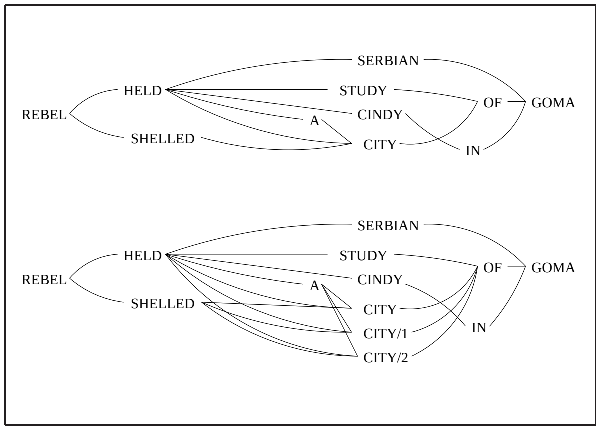


Figure 5.1: Lattice segment for the text "THE REBEL HELD CITY OF GOMA" before (top) and after (bottom) adding pronunciation variants for the word "CITY"

a word transition lattice to constrain the search space before pronunciation variants are added. The new lattice is then rescored at the acoustic level. For this data, it was determined experimentally that adding new links results in better performance than replacing links.

Figure 5.1 shows a segment of a lattice before and after incorporation of pronunciation variants for the word "city." For each link bound to the word "city" (standard GA form [sɪɾi]), links for the two pronunciation variants CITY/1 [siti] and CITY/2 [ʃiti] are added.

Speaker adaptation

MLLR adaptation based on 50 adaptation utterances was applied for each speaker. All lexical adaptation methods were tested with and without allowing the new pronunciations for acoustic adaptation. It might be thought that allowing the new pronunciations would always result in better adaptation; if a new pronunciation is found to be the best acoustic match before adaptation, it might make the most sense to adapt the phones specified in its path to the acoustic adaptation data than the phones specified in the canonical path for an even better acoustic match. On the other hand, allowing the new pronunciations might be viewed as counterproductive to adaptation; yes, the new phone sequence might be a better initial match, but the point of adaptation is to learn an individual speaker's preferred realization of the phone he is trying to pronounce, which is most likely the canonical phone. If adaptation examples are siphoned off to update a different phone, the true target model does not learn the idiosyncratic realization as well, and the alternate model may be confused by the adaptation example if it is not consistent with other examples for which it is the true base model.

Results

I did not observe any statistically significant ($p < .01$, 2-tailed t-test) changes from applying any of the eight methods described above to generate alternate lexicons. Table 5.7 gives recognizer performance for each case.

Lexicon	WER		Lexicon size	% of test words	Variants
	baseline adapt	lexmod adapt		with new variant	per word
baseline	45.1	45.1	26110	N/A	1.17
W1	45.5	46.1	27180	14.9	1.19
W2	46.8	46.6	26229	9.2	1.19
R1	46.8	47.1	60244	48.8	2.40
P1	45.9	47.0	34628	20.0	1.55
P2	45.9	46.6	47862	37.7	2.03
P3	46.4	46.9	31595	18.6	1.46
P4	44.9	45.1	31152	20.5	1.53
P5	45.6	46.0	31200	12.0	1.28

Table 5.7: Lexicon statistics and recognizer performance for rule-based lexical modeling. Separate WER figures are given for decoding with new pronunciations allowed (lexmod adapt) and excluded (baseline adapt) in speaker adaptation

5.3 Data-driven modeling

Although linguistically-motivated lexical modeling is attractive from a theoretical point of view, and is the only option when adapting to a new speaker group for which no acoustic adaptation data is yet available, it assumes a model of human speech that may conflict with what is meaningful for the recognizer. The recognizer does not yet have the sophisticated ability of a human to perceive sounds in the context of syntax and semantics and a myriad of sociolinguistic factors. It is not completely inferior, however; its model of how sounds map to phonetic units is complex, identifying 118 distinct realizations of /t/, for example, where a linguist might only recognize five or six.

In this section I describe experiments in *data-driven* lexical modeling. In data-driven modeling, the recognizer is involved from the start, telling us which phones it perceives when presented with an acoustic stream.

5.3.1 Initial mappings

In Section 5.2, I obtained initial phone mapping candidates via an analysis of Japanese and Japanese-accented English. In this section, I will describe how similar mappings were obtained using phoneme recognition.

An initial phoneme recognition pass was done using context-dependent acoustic models, a uniform phone language model, and a phone lexicon. Phoneme recognition error was 67.2% for the non-native test speakers. Segmental alignment of the phone recognition hypotheses to the phone expansions of the reference text yielded a list of frequent substitution, insertion, and deletion errors. The ten most frequent of each type of error is given in Table 5.8.

Substitutions	Insertions	Deletions
s → z	∅ → SIL	ə → ∅
n → ŋ	∅ → t	n → ∅
ˈɪ → i	∅ → d	t → ∅
t → d	∅ → p	r → ∅
t → p	∅ → n	ˈɪ → ∅
ə → u	∅ → i	l → ∅
i → ɪ	∅ → ə	d → ∅
ˈɪ → ɪ [2]	∅ → r	æ → ∅
ə → ɛ	∅ → z	m → ∅
t → SIL	∅ → garbage	k → ∅

Table 5.8: Most frequent substitution, insertion, and deletion errors as found by aligning phone recognition hypotheses to phone expansion of reference text

Although we now have what seems to be a plausible list of substitutions, insertions, and deletions, using it to predict errors that will be seen in individual lexical items is tricky. For example, in the utterance fragment "American kids spend more time...," there are a number of deletion errors that are not obviously attributable to phonological effects.

ə	m	ɛ	r	ɪ	k	ə	n	k	ɪ	d	z	s	p	ɛ	n	d	m	ɔ	r	t	aĭ	m
	ð	ɛ		ɚ	k		ŋ				ə	s	p	ɛ	n				l	t	aĭ	m

A framewise alignment of the type discussed in Section 4.3.1 would provide us with a straightforward mapping, but it is not clear that this mapping is what one would want for lexical modeling. Rules like ə → ð / # __ m and dmɔ → ∅ / n __ r[3] cannot be said arise from anything other than poor acoustic modeling (listening to the acoustic data confirms that these phonemes are indeed articulated), and it is not the role of lexical modeling to compensate for such inadequacy. Rather, I focused on substitutions that could be ascribed to some sort of phonological interference.

I elected to use these initial segmental mappings to bootstrap an underspecified alignment pass. Underspecified alignment is described in detail in Section 4.3.1. With this method, we allow the system to find the best match among a list of plausible substitutions (including insertions and deletions) that were detected during phoneme recognition, while enforcing structure on the alignment in the form of the approximate number of phones that are to be identified. Features of plausible substitutions were defined as follows.

1. Having the same or a similar place or manner of articulation as the canonical phone

[2]The ISL-BN recognizer treats stressed and unstressed /ɪ/ as separate phonemes.

[3]It should be noted that this is an extreme example illustrating the problem of deletion errors. Insertion and deletion errors were in general well balanced in this data.

2. Representing deletion in the initial phone sequence that creates an open syllable

3. Representing insertion in the initial phone sequence that creates an open syllable

4. Sharing at least one vowel feature with the canonical phone (both high vowels, for example)

5. Representing decomposition or monophthongal realization of a diphthong

6. Having a possible mapping to the same Japanese orthographic symbol as the canonical phone

If none of these features were present, the substitution was not allowed. Underspecified alignment generates a new surface phone sequence for which a mapping to the canonical form is easily derived. With this mapping, we can duplicate experiments carried out for linguistically-motivated modeling to understand which approach, if either, leads to an improvement in recognizer performance.

5.3.2 Experiment 10: Data-driven lexical modeling

Introduction

Testing for data-driven lexical modeling closely paralleled that of rule-based modeling. The purpose of this experiment was to determine whether data-driven modeling results in an improvement in recognizer performance where linguistically-motivated modeling does not.

Data

In these experiments, the test data remains fixed, as for previous experiments, to the proficiency-controlled test set NN-E-R, while the pronunciation lexicon is varied. The eight test lexicons described so far in this section are summarized below.

Lexicons associating different probabilities with substitutions were defined as follows.

D1 The top two most frequent context-independent substitutions were applied to generate the test lexicon

D2 The top three most frequent context-independent substitutions were applied to generate the test lexicon

D3 Phone substitutions that occurred more than seven times in the context of a given 3-phone window were applied to generate the test lexicon

D4 Phone substitutions that occurred more than once in the context of a given 5-phone window were applied to generate the test lexicon

Lexicon	WER		Lexicon size	% of test words	Variants
	baseline adapt	lexmod adapt		with new variant	per word
baseline	45.1	N/A	26110	N/A	1.17
D1	44.9	45.0	37436	26.7	1.57
D2	45.5	45.0	51847	42.7	2.16
D3	45.2	45.5	58267	58.7	2.57
D4	45.8	45.6	45108	52.6	2.19

Table 5.9: Lexicon statistics and recognizer performance for data-driven lexical modeling. Separate WER figures are given for decoding with new pronunciations allowed (lexmod adapt) and excluded (baseline adapt) in speaker adaptation

As with the linguistically-motivated experiments, new pronunciation paths were added via lattice adaptation, and MLLR adaptation was applied both allowing and excluding the new pronunciations.

Method

The testing method was the same as that described in Section 5.2.4. A word lattice was generated during an initial decoding pass using the baseline lexicon; pronunciation variants were added to the lattice and an acoustic rescoring pass was run to generate the final hypothesis.

Results

Recognizer performance with data-driven lexical modeling is summarized in Table 5.9. As with the linguistically-motivated lexical modeling, there is no significant difference in recognition accuracy for any of the new lexicons.

5.4 Conclusions from lexical modeling experiments

This investigation of lexical modeling for low-proficiency Japanese speakers of English has not found that any of a number of approaches contributes significantly to improved recognizer performance. I now examine why this is the case, first considering in more detail the lexical modeling approaches mentioned in Section 5.1.

Amdall et al. (2000) report an improvement from 29.2% WER to 28.3% for Wall Street Journal (LDC, 1994a) using a data-driven lexical modeling approach. This represents a 3% absolute improvement. Our generation of lexicion **D3** is similar to the method they describe. Both approaches use unrestricted phone recognition to obtain initial context-dependent phone mappings. Confusability constraints are then enforced, in the form of phonotactic constraints in our case and restriction to the single most probable phone substitution in a given triphone context in their system. A phone substitution candidate derived this way is called, in their terminology, a "rule." I will borrow their usage in this comparison; this usage should not be confused with the phone substitution rules operating on phone class abstractions and variable-length contexts used to generate lexicon **R1**. Rule firing frequency constraints are applied in both Amdall's method and mine; rules occurring in the training data fewer than 6 and 7 times respectively are not considered in testing. The test set (WSJ) is read news, just as mine is. The primary differences, then, are my use of maximum likelihood instead of log-likelihood in calculating substitution probabilities, and the higher overall proficiency of the speakers.

Livescu and Glass (2000) report an improvement from 20.9% to 18.8% for the JUPITER weather query system. This represents a 10% relative improvement. The JUPITER task is quite different from ours: it is a spontaneous task but highly restricted in domain (lexicon size 2000 compared to 26000 for our task); no attempt is made to control or estimate the proficiency of speakers; and the goal, as in Amdall's system, is to adapt to non-native speech in general as opposed to one speaker group in particular. It is difficult to compare our implementations directly, as JUPITER uses a FST-based decoder and encodes pronunciation variants in the form of a phoneme confusion FST that is composed with the existing acoustic, lexical, and language FSTs, but Livescu and Glass's generation of phone substitution candidates is similar to our methods **D1** and **D2**. They first obtain initial context-independent phone confusions by aligning reference transcripts with underspecified alignment output, allowing variable-length substitutions. This confusion matrix is represented as an FST, which can be pruned to optimize recognition accuracy; Livescu and Glass found, however, that the best performance came with no pruning. This last result is the most striking difference between our experiences. They found that the lexicon size was increased to only 1.5 times its original size from adding all confusions discovered through an underspecified alignment based confusion estimation method almost identical to ours; our lexicon size expanded to 36 times its original size. This may be because their speakers showed less variation in pronunciation; it could also be that their inital acoustic models were more tolerant of deviant pronunciations. The specific vocabulary may play a role as well; if the words used in the weather query task are mostly common and familiar words, the speakers may be able to pronounce them more successfully than in the read news tasks.

Fung and Liu (1999) report an improvement from 30.8% to 26.7%, for the undescribed "HKTIMIT" recognition task. Fung and Liu use a purely knowledge-based approach, working with linguists to identify sounds that do not occur in Cantonese and probable substitutes from the English phone set. This method parallels our linguistically-motivated variant generation process. A total of 43 transformation rules are identified in our system, compared to 28 in Fung and Liu's, but it may be the case that Fung and Liu use context-independent rules, in which case each rule would apply to more instances in the lexicon. Fung and Liu see their lexicon size double with the application of their rules; they therefore require no pruning to maintain a manageable lexicon size and confusability level. We do not know, however, exactly how the rules are applied; if the realization estimated to be the most probable for each word is simply added to the lexicon, a doubling in size of the lexicon is to be expected. Our lexicon grows as quickly as it does because alternate links are added to the word pronunciation networks for each possible substitution; pruning of this network to identify the likely paths through the entire network based on training data is an integral part of our method. Because we do not know the specifics of the HKTIMIT task, it is difficult to compare our results directly, but the original TIMIT task (LDC, 1994b) is a read speech task covering 2342 unique phonetically-engineered sentences. The lexicon size is 6100. Although there was no formal or informal evaluation of speaker proficiency, because the speakers were all college students in a bilingual environment one can assume that their exposure to English is fairly extensive. In fact, one of the motivations of Fung and Liu's work is that code-switching is frequent among students at the university, and an ASR system deployed there will need to be able to handle both Cantonese and Cantonese-accented English.

Humphries and Woodland (1998) successfully used lexical modeling of accent variation in WSJ to recognize American-accented speech with a British recognizer. They report an improvement of 21.3% to 18.6%, a relative gain of 13%. (This is their result after speaker adaptation; they share our observation that pronunciation modeling is more effective with unadapted, or lower-quality, acoustic models.) As in our method, they begin with an unrestricted phone recognition pass, aligning the results to the reference transcript to generate phoneme confusions in context. Rather than run an additional plausibility-constraining pass, as was done by Amdall and by Livescu as well as in our system, Humphries moves directly to a decision tree clustering phase. This approach most closely resembles our method **D4**, in which we found the most likely substitution using decision tree clustering *after* enforcing plausibility constraints on the initial phoneme confusion matrix via underspecified alignment. We did not work with decision trees in data-driven lexical modeling because results from linguistically-motivated modeling indicated that decision-tree-based pruning did not produce significantly different results from maximum-likelihood-based pruning. We also used the C4.5 algorithm (Quinlan, 1993), where Humphries and Woodland used the CART algorithm (Brieman et al., 1984).

It is my interpretation, based on these observations and experience with phonetic transcription of the CND data, that the speakers in the present study are at a phase in their development of spoken English in which deviations from standard English pronunciation are very complex. As they build their articulatory skills, they are inconsistent in phonetic realization where speakers with more experience, however heavily accented, have developed idiosyncratic articulatory habits. Training and adaptation, which model speech at a fine sub-segmental level, are more appropriate than even context-sensitive segmental modeling. With this in mind, it is probably not insignificant that the speakers have all been in the United States for only a short time after having extensive formal study of English. It would not be unreasonable to think that their spoken English is undergoing complex changes as they are suddenly exposed to many new varieties of English, and work to transfer an academic knowledge of the language to a physiological competence. The success of lexical modeling for native speech would support this hypothesis, as native speakers are more consistent in their phonetic realizations of words than non-native learners are.

Another factor that may play a role in the effectiveness of lexical modeling is recognition task. Although experiments with recognition of spontaneous speech for my speakers clearly indicate that spontaneous speech is a harder problem for LVCSR, it may be better suited for lexical modeling. In read speech such as that in my task and in Amdall's, the speakers cannot choose the words they speak. They cannot avoid words that are difficult to pronounce, and may struggle with words that are new to them. In query-based tasks such as Livescu and Glass's JUPITER weather query system, speakers approach the system with something they want to know, and can rely on words and fixed phrases that are familiar to them. Read speech, while only mildly affected by a speaker's command of syntax and semantics of the language and as such "easier," may not be a strong candidate for either rule-based or data-driven modeling at the lexical level.

Chapter 6

Hypothesis-Driven Accent Classification

In order to take advantage of the techniques for modeling non-native speech described in previous chapters, the system must know that the speaker is non-native. A nativeness decision can be either binary, classifying the speech sample as native or not, or multilateral, associating the speech sample with a specific native language or language group. In this chapter, I demonstrate that 1) high-accuracy nativeness classification can be implemented and 2) it improves overall system performance significantly, as measured by the matched-pairs test discussed in Section 4.2.

6.1 Problem Description

There are many features distinguishing native and non-native speech, as has been discussed in Chapter 3. The key decisions in designing a classifier are which of those features to use and what classification algorithm will make best use of the selected features in the data that is available. These are not independent decisions. For example, the first formant frequency (F1) of specific phonemes in context may allow very accurate discrimination, but if those contexts do not appear frequently in the training data, it may be impossible to build a robust model to classify them. Decision tree learning may be theoretically possible given the amount of data available, but if the target function does not have discrete output values, specifying the splitting questions may be difficult.

Another consideration in designing a classifier for speech recognition is the recognizer itself. The most accurate classification may not result in the best recognizer performance. It might be best to treat the most proficient non-native speakers as native speakers for the purposes of acoustic model selection. Proficiency may also not be well correlated with recognition accuracy; the best overall system performance may be achieved by classifying some of the less proficient speakers as native.

In formulating the accent classification problem, I concentrated on finding properties of non-native speech that can be easily and reliably extracted and that give a meaningful result for speech recognition. These criteria led to the development of a hypothesis-driven approach, using naive Bayes classification for both binary and multilateral discrimination.

This chapter is structured as follows. In Section 6.2 the principles behind hypothesis-driven classification are described. Section 6.3 provides an overview of Bayesian classification. The software package used for classification experiments is also described here. Experimental design and results, including end-to-end system results with classification-based model switching, are presented in Section 6.4. Finally, a discussion of the discriminative features in this formulation of the accent classification problem is presented in Section 6.5.

6.2 Hypothesis-driven Classification

This approach to accent classification, or more properly L1 classification, bases the classification decision on recognizer hypotheses of what was said. The hypothesis can be either a word hypothesis, generated using a word-based lexicon and language model, or a phone hypothesis, generated using a lexicon made up only of phones and optionally a language model (effectively a phonotactic model).

Determining the nativeness of the speaker is framed as a document classification problem. For each training speaker (native and non-native), a set of training utterances is defined and recognizer hypotheses are generated. This data set is not unlike a set of articles, each written by a different writer, originating from two different publications. If differences in the individual preferences of writers are overshadowed by differences in the stylistic themes of their publications, it is possible to categorize documents according to source using statistical algorithms, as was shown in (Argamon-Engelson et al., 1998). I extend this idea to nativeness classification, asking a classifier to decide whether a set of utterances is representative of native speech based on a training corpus of native and non-native speech "documents."

There are two important advantages in formulating the problem this way. First, one may build on a large body of research in machine learning and document classification. My choice of naive Bayes classification is based on consistently strong performance in document classification tasks (Lewis, 1998) and favorable comparison to other classification techniques when class distributions are not radically skewed (Yang and Liu, 1999).

Second, by using the recognition hypothesis instead of acoustic features, one takes the behavior of the recognizer into account without relying on an acoustic score whose interpretation may not be straightforward. Other resources that have been successfully used in accent discrimination include acoustic features, such as F0 (Fung and Liu, 1999), and score from a set of competing L1-specific acoustic models (Teixeira et al., 1996). Using competing acoustic models requires building the models, which is expensive in terms of both computation and data; a more troublesome issue with this approach, however, is that a Viterbi score from an HMM built from one set of data is not necessarily comparable to a score from an HMM built from another set of data. Acoustic features, while very discriminative, may not capture the most meaningful distinctions from the point of view of the recognizer. If the goal of the classification is solely to determine whether a speaker is native or non-native, acoustic features may offer the best basis for discrimination. I assume, however, that the nativeness classification will be used to trigger specialized modeling, and that a native recognizer may respond better to some non-native speakers than a non-native model will. In these cases, implicit modeling of recognizer behavior in the classification engine may lead to more appropriate, although not necessarily more strictly accurate, classification.

The question of appropriate versus accurate classification is largely moot given the target population, as non-native acoustic models performed better than native models on all target speakers. However, it may become more important as the proficiency range of LVCSR system users broadens.

A further advantage of hypothesis-based classification is that the recognizer itself may be treated as a black box. This permits the algorithm to be implemented without access to the internal workings of the recognizer, an option which may be attractive to users of commercial software packages or researchers in other areas of NLP who are not interested in manipulating recognizer components.

6.3 Bayesian Classification

Bayesian classification is well suited to the task of L1 categorization for several reasons. Bayesian learning methods support probabilistic hypotheses, which allow a nativeness threshold to be set or the result to be incorporated with other sources of information. Bayesian classification incorporates the marginal probability of the class, so knowledge of the distribution of speakers likely to use the system can help to improve classification accuracy. Bayesian models also handle conflicting examples gracefully, and are not as vulnerable to data sparsity problems as methods like decision tree learning that iteratively partition training data.

6.3.1 Bayes decision theory

The objective in Bayes decision theory is to minimize the probability of decision error. For example, if there are two possible outcomes ω_i and ω_j, and it is known that ω_i occurs three-quarters of the time and ω_j occurs one quarter of the time, always guessing ω_i will result in the lowest decision error rate. The policy of always guessing ω_i is called a *decision rule* and can be stated as:

$$\text{Decide } \omega_i \text{ if } P(\omega_i) > P(\omega_j)\text{; otherwise decide } \omega_j. \tag{6.1}$$

If information beyond the basic occurrence probabilities is available, that information can be incorporated in the decision rule. For example, if ω_i represents warm weather and ω_j represents cold weather, the a priori probability of warm weather may be higher, but if it is snowing out, one can guess that the weather is probably cold. If $\mathbf{x}$ represents snow falling, we can amend the decision rule to be:

$$\text{Decide } \omega_i \text{ if } P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})\text{; otherwise decide } \omega_j. \tag{6.2}$$

In order to minimize decision error, Bayes decision theory calls for selecting the course of action that results in the smallest expected loss, or *risk*. Each possible course of action α_i is associated with a risk R:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$$

where $\lambda(\alpha_i|\omega_j)$ is the loss associated with choosing course of action α_i. Specifically, Bayes decision theory prescribes selection of the state ω that that maximizes the *a posteriori* probability $P(\omega_j|\mathbf{x})$, a course of action that will minimize the risk R.

6.3.2 Naive Bayes classification

In classification problems, the states ω_j are classes and the feature vectors $\mathbf{x}$ are properties of the data, for example, word distributions in text classification tasks. A Bayes classifier uses Bayes decision rule to determine which class the present data belongs to. Restating Rule 6.2 in terms of classes c_i and utterances $\mathbf{u}$ gives

$$\text{Decide } c_i \text{ if } P(c_i|\mathbf{u}) > P(c_j|\mathbf{u})\text{; otherwise decide } c_j \tag{6.3}$$

or more generally

$$\text{Decide } c_i \text{ if } P(c_i|\mathbf{u}) > P(c_k|\mathbf{u}) \text{ for all } k \neq i \tag{6.4}$$

Although we probably do not know the conditional probabilities $P(c|\mathbf{u})$, we can calculate them using Bayes rule.

$$P(c_i|\mathbf{u}) = \frac{P(\mathbf{u}|c_i)(P(c_i)}{P(\mathbf{u})} \tag{6.5}$$

Because the probabilities of the utterances are constant across classes, Equation 6.5 can be simplified as

$$P(c_i|\mathbf{u}) = P(\mathbf{u}|c_i)(P(c_i) \tag{6.6}$$

The task of the classifier, then, is to assign an utterance to a class $\hat{c}$ such that

$$\begin{aligned}\hat{c} &= \underset{c_i}{\operatorname{argmax}}\, P(c_i|\mathbf{u}) \\ &= \underset{c_i}{\operatorname{argmax}}\, P(c_i)P(\mathbf{u}|c_i) \end{aligned} \tag{6.7}$$

Expanding the notion of a set of utterances $\mathbf{u}$ to a set of word attributes a_i (indicating presence or absence of a word, perhaps, or word counts), we have the following.

$$\hat{c} = \underset{c_i}{\operatorname{argmax}} P(c_i)P(a_1, a_2 \cdots a_n|c_i) \tag{6.8}$$

A naive Bayes classifier is a special kind of Bayes classifier. The naive Bayes assumption is that the attributes used for description are all conditionally independent (Manning and Schütze, 1999). If a feature vector $\mathbf{u}$, which represents an utterance, is thought of as a set of individual word features, the naive Bayes assumption says that their occurrences are independent. This is, of course, not strictly true; grammatical constraints and lexical relationships certainly influence the presence and order of words. However, the assumption simplifies the model, and the decision made can still be optimal (Domingos and Pazzani, 1997), approaching the performance of neural network and decision tree learning models (Mitchell, 1997). Applying the naive Bayes assumption brings us to

$$\hat{c} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i) \prod_j P(a_j|c_i) \tag{6.9}$$

where the c_is are classes that are members of a class set C and the a_js are word-level attributes.

6.4 Experiments

In this section, I describe the design of a hypothesis-driven Naive Bayes classifier and the methodology used to evaluate it. I compare classification based on hypotheses and transcriptions, on read and spontaneous speech, on words and phonemes, on words and parts of speech, and on phonemes and phone classes. I find that not only is L1 classification based on recognizer hypotheses possible, it is *more* accurate than classification based on manual transcriptions of native and non-native speech.

Three experiments in accent classification are described in this section: word-based binary classification of the speakers in test sets N-E-R and NN-E-R as native or non-native; word-based binary and multilateral classification of native English, Japanese, and Chinese speakers; and phone-based binary classification of the speakers in test sets N-E-R and NN-E-R. Because the general methodology and materials are the same for all three experiments, they are discussed here; text data and individual experimental results are described in discussions of each experiment.

6.4.1 General methodology

In order to frame accent detection as a document classification problem, files containing utterance text are created for each speaker. The utterance text can be transcriptions of utterances or recognizer hypotheses (word-level or phone-level).

Classification of both read and spontaneous speech was evaluated. For spontaneous speech experiments, data sets N-A-S, NN-A-S, and C-A-S were used. Read speech experiments examined classification on data sets N-E-R and NN-E-R. Each speaker in this data set reads 3 articles, one of which was common to all speakers, as described in Sections 3.1.2 and 3.4. Four train/test conditions were evaluated in read speech experiments:

A Train and test on shared article

B Train and test on disjoint articles

C Train on disjoint articles; test on shared article

D Train on shared articles; test on disjoint article

For conditions A and B, leave-one-out training and testing was done in order to maximize the size of the training set. That is, for each speaker in the combined N-E-R and NN-E-R sets, a classification model was trained on all the other speakers to discriminate between native and non-native documents. Accuracy of that model was then tested on the held out speaker. Overall classification accuracy was calculated by averaging accuracy for all leave-one-out tests.

For conditions C and D, there was no need for leave-one-out testing as training and testing were done on separate data sets.

The baseline accuracy to which classification accuracy should be compared is calculated by dividing the number of test speakers in the most common training class by the total number of test speakers. This is the accuracy that would be achieved by a model that always guesses the most common class found during training. For example, in the N-E-R and NN-E-R sets there are 8 native and 10 non-native speakers. Always guessing “non-native” would yield a baseline accuracy of 56% (10/18). Baseline accuracies are listed for each experiment.

6.4.2 Materials

To carry out the experiments described in this section, I made use of publicly available classification and part-of-speech software packages. These are described here, along with the configuration of the recognizer that was used for the classification experiments.

Text classification

The Rainbow statistical text classification package (McCallum, 1996) was used for all classification experiments. Rainbow implements a naive Bayes classifier for text, with a number of features specialized for text applications. In running Rainbow, no feature selection[1] was used. Token unigrams, bigrams, and in some cases trigrams were treated as independently occurring features. Punctuation and capitalization were removed from the transcriptions to make them consistent with the hypotheses.

Preliminary experiments showed that words commonly considered stopwords, such as function words, contributed significantly to discrimination. Therefore, in all of the experiments described in this chapter, no list of stopwords to exclude was defined.

Because the data set was relatively small, the training and test sets were defined by a random partitioning of the full data set into 70% training and 30% testing. This random partitioning was repeated 20 times and classification accuracy was averaged over the 20 trials for each experiment. The full data set for each experiment consisted of exactly one "document" from each of the speakers.

Recognizer

The recognizer used to generate the hypotheses was the ISL-BN system described in Section 4.1. Baseline WER on native speech was 18.0% in the CND read news task and 63.1% on non-native speech. The choice to use a system that performs poorly on non-native speech was motivated by the expectation that a nativeness classification will be used to trigger specialized non-native modeling, and that the initial processing will be done with the standard native acoustic models.

Part-of-speech tagging

In some of the experiments that will be described, words in the utterance sets were replaced by their parts of speech using the publicly available MXPOST toolkit (Ratnaparkhi, 1996). MXPOST is a maximum entropy tagger that achieves 96.6% accuracy on unseen Wall Street Journal articles. Because the data set evaluated in (Ratnaparkhi, 1996) is similar to ours in both content and genre, I assume that tagging accuracy on the CND database is similarly high.

Read and spontaneous speech

For this thesis, both read and spontaneous speech were collected from the non-native speakers, and both were used in the investigation of L1 classification. Upon first consideration, it may be thought that spontaneous speech is easier to classify than read speech because the differences in word choice contribute to the decision.[2] However, the ultimate goal is to use recognizer hypotheses for classification, and recognition errors introduce noise that may diminish this effect somewhat. I wished to both establish whether spontaneous speech can

[1] In discussions of text classification, the term *feature selection* refers to limiting the vocabulary used for classification. Common feature selection techniques include using only content words and using only words that appear with high frequency. Feature selection typically improves classifier performance, so results may have been even higher with judicious feature selection.

[2] Although, as noted in Section 6.3.2, I am effectively ignoring word order and syntax in my classification model, the *presence* and *frequency* of individual words and n-grams strongly influences the classification decision, as will be discussed in following sections.

indeed be classified with greater accuracy than read speech can and analyze the differences in the word features that contribute most to L1 classification of these two types of speech.

A second reason for electing to study classification of both read and spontaneous speech was that read speech can be restricted in a way that allows one to control variables such as vocabulary, difficulty, and content. By having speakers all read the same text, one can isolate the contribution of recognition error to classification accuracy. One can also evaluate the classifier in ways that are not possible when the data is spontaneous, by comparing training on a single article that is read by all speakers with training on a disjoint set of articles, for example.

Finally, a number of important speech recognition applications and tasks target speech that is read. Language tutoring applications, in which speakers are often asked to read specific words and sentences, and speaker-dependent enrollment, in which users must read aloud from text to allow the system to adapt to their voice, are two examples. In these cases, a nativeness classifier would need to base its decision only on differences in the way the speakers read the same pieces of text. Precisely this situation will be addressed in the "train/test on a common article" evaluation.

Transcriptions and recognizer hypotheses

In order to understand the performance of hypothesis-driven classification, it is important to subject classification of manual transcriptions to the same evaluations. If classification of hypotheses is less accurate than classification of transcriptions, one can predict that L1 classification will improve as recognition technology develops. If recognition accuracy is very poor, it may also only be meaningful to evaluate classification on transcriptions. If, on the other hand, classification of hypotheses is *more* accurate than classification of transcriptions, we are given evidence of a synergistic relationship between the recognition and classification processes.

A comparison of classification on hypotheses and transcriptions tells us more than just which is more accurate. We also learn about the words and types of words that are important in detecting non-native speech in these two data types. While the objective of integrating L1 classification in this thesis work is to improve the overall performance of the recognition system, the same type of classification can be used in text-based natural language processes such as language modeling and parsing. The value of a thorough examination of L1 classification of both recognition hypotheses and transcripts, then, clearly extends beyond the immediate context of speech recognition for low-proficiency non-native speakers.

6.4.3 Experiment 11: Word-based classification of read speech

In word-based classification experiments, the features used as input to the classifier were word identities and parts of speech. There are several reasons for looking at parts of speech as well as word identities. First, it reduces the size of the feature set, allowing more robust modeling and handling of unseen words. Second, it allows one to make generalizations about the types of words that are important in discrimination. Third, it compensates somewhat for recognition error. Finally, it increases the experimental validity of using unique renditions of a single article, read by all speakers, for training and/or testing.

Data

In read speech experiments, data sets N-E-R and NN-E-R were used for training and testing in the four configurations described in Section 6.4.1. The baseline classification accuracy of classification on this data set is 56%, achieved by always guessing that the speaker is not native, the state with the highest *a priori* probability (cf. Equation 6.1).

Each of the four experimental conditions reveals unique properties of the data and its classification potential. When training and testing on the common article, a high classification accuracy shows that even when the printed words were exactly the same, reading errors made by native and non-native speakers were enough to identify them. When training and testing on unique articles, a high classification accuracy shows that the classifier is extremely robust, and that patterns that mark non-native speech are independent of the words and phrases in the text. High classification accuracy when the training articles are all the same and test articles were all unique shows that the patterns found in non-native readings of one text are so discriminative that they generalize to detect non-nativeness in a wide variety of texts. And high classification accuracy when the training articles are all different and the test articles are all the same shows that non-native speakers display consistent (found in all renditions of the test article) and text-independent (learned from a set of disjoint articles) idiosyncrasies in reading.

Documents were created for each speaker consisting of either transcriptions of a reading of an article or recognizer hypotheses of a reading of an article. For evaluating classification based on part of speech, the words in the documents were replaced by their part of speech as assigned by MXPOST (Ratnaparkhi, 1996).

For this experiment, an additional document set was created to evaluate the hypothesis that a 21% vs. 58% WER is in and of itself detectable. In order to establish whether the classifier is modeling the way the recognizer responds to non-native speech or simply the higher word error, I artificially raised the word error rate of the native speech. This was accomplished by adding white noise to the signal until the word error rate was close to that of the non-native speech (56%).

Results

Table 6.1 shows results of training and testing a naive Bayes classifier under the four conditions described above. Classification accuracies are given for both transcriptions and recognizer hypotheses. The most striking result is that classification of hypotheses is consistently more accurate than classification of transcriptions. This is strongly counterintuitive, as the recognizer is generally viewed as a noisy channel that would be expected to *mask* non-native patterns. Yet the effect is consistent and highly significant ($p < .005$) as measured by a matched-pairs test.

Although the classification accuracy for the noise-added hypotheses decreases, it is still much higher than

Condition	word-identity	POS
train and test on common article (trans)	83%	74%
train and test on common article (rec)	**94**	**100**
train and test on common article (high-WER rec)	66	77
train and test on disjoint articles (trans)	41	40
train and test on disjoint articles (rec)	47	**77**
train on disjoint articles; test on common articles (trans)	56	56
train on disjoint articles; test on common articles (rec)	56	**95**
train on common articles; test on disjoint articles (trans)	56	56
train on common articles; test on disjoint articles (rec)	56	**83**

Table 6.1: Classification accuracy of read speech for two-way classification of Japanese and American English speakers reading texts in English. Baseline is 56%.

the baseline, suggesting that there is something special about the recognition errors made on non-native speech. The observation that classification in the non-noise-added case is based to some degree on features of high-WER speech, as opposed to non-native speech, should not be thought of as indicating that such classification is invalid. If a high word error rate is a feature of non-native speech, using it as a basis for classification is not illegitimate. It only indicates that WER plays a significant role in discriminating between recognizer output for native and non-native speakers.

Another important observation is that classification based on parts of speech outperforms classification based on word identities in almost all cases. This is particularly true when disjoint articles are involved, a condition under which word-identity classification never exceeds the baseline and is often considerably worse. When training on a disjoint set of articles and testing on the common article, the classifier detects non-native speech with 95% accuracy using parts of speech, compared to 56% (baseline) accuracy when using words. This is evidence, as discussed above, that the same patterns that are found in all speakers' renditions of the common article are present in different speakers' readings of disjoint articles. Under the same conditions, however, the classifier performs no better than the baseline when the input is transcriptions instead of recognizer output; differences in word distribution among the disjoint articles overshadow the non-native effects in the transcriptions.

6.4.4 Experiment 12: Word-based classification of spontaneous speech

This experiment examines classification of spontaneous speech. The recognition accuracy on the spontaneous speech was so poor that classification of recognizer hypotheses was not evaluated. Recognition of spontaneous speech has not been a focus of this thesis, and I did not optimize the recognizer for performance on this task.

However, classification results on transcriptions are quite interesting, and are included here.

Data

The domain was tourist-domain queries; speakers were prompted in their native language to ask questions of an agent about specific sights and events, as described in Section 3.1.2. For spontaneous classification experiments, data from all speakers in sets N-A-S, NN-A-S, and C-A-S (6 English, 31 Japanese, and 6 Chinese natives) was used.

The proper names that appear in the queries are unique to each native speaker group, biasing classification based on word identities. When recording, each speaker was given a scenario that included local sight and event names, information such as ticket prices that should be obtained, and a general description of the situation. The scenarios were changed after each 5 to 10 speakers. Scenarios were given to more than one speaker so that multiple examples of non-native pronunciations of unfamiliar words would appear in the data. Scenarios were changed regularly to maximize the phonetic breadth of the data. This balance is appropriate for data collection for LVCSR, but was not the best for classification. Using part-of-speech tags instead of word identities, therefore, was not just desirable for better classification but was necessary for a fair evaluation. Because my concern about bias was limited to proper nouns, I performed a third type of evaluation in which only nouns were replaced with their parts of speech. Examples of the word-identity, part-of-speech, and noun-only part-of-speech documents are:

Document type	**Example sentence**						
word-identity	What	is	the	business	hours	of	Tiffany
POS	WP	VBZ	DT	NN	NNS	IN	NNP
POSNoun	What	is	the	NN	NNS	of	NNP

Results

Table 6.2 shows results of L1 classification based on words in spontaneous speech. Classification accuracy is shown for various combinations of the three speaker groups (native English, Japanese, and Chinese). Because baseline classification accuracy is estimated by always choosing the most common class, and the number of speakers in each class in the training and test data varies for the different configurations, baseline accuracy for each configuration is specified in Table 6.2. For example, for a three-way native/japanese/chinese decision, we had 31 Japanese, 6 native, and 6 Chinese speakers in the training set. The total training set size is 43 speakers. If the most common class is always guessed, the accuracy of the classifier will be 31/43, or .72.

Nearly all experimental classification accuracies are significantly higher than the corresponding baseline.

In most cases, the mixed word-POS (the POSNoun column in the table) data is most accurately classified. For binary native/non-native decisions, classification was nearly perfect. Accuracy decreased somewhat for a three-way decision; interestingly, it was also for this condition that replacing nouns with their part-of-speech tags did not significantly improve classification accuracy.

These results may prompt one to ask *why* classification is most accurate with mixed word-POS data. It

Classes	baseline	word-identity	POS	POSNoun
Native/Japanese	83%	90%	84%	**97%**
Native/Chinese	50	100	100	**100**
Native/Japanese/Chinese	72	90	74	**89**
Native/Japanese/Chinese	72	89	83	89 ($n \leq 3$)
Native/all non-native	72	87	76	**96**
Native/all non-native	72	96	90	98 ($n \leq 3$)
Japanese/Chinese	83	93	86	**100**
Japanese/Chinese	83	86	80	**100** ($n \leq 3$)

Table 6.2: Classification accuracy of spontaneous speech. Baseline classification accuracies for the different conditions are given in the table. Figures annotated with ($n \leq 3$) indicate that trigrams, and not just unigrams and bigrams, were used for classification.

would not be unusual to expect that since the noun replacement was done to compensate for a bias in the data, this configuration would result in accuracies somewhere between those of pure word and pure part-of-speech based classification. The answer may be that the mixed condition provides just enough generalizability while exploiting the discriminative power of specific non-noun word sequences. This intuition is supported by an analysis of features important in classification; singular nouns are highly indicative of non-native speech, while certain personal pronouns and associated verb forms such as "you" and "am" are indicative of native speech. The former association would not be apparent if only word identities were used, and the latter would be hidden if all words were replaced by their parts of speech. The actual word and part-of-speech sequences that contributed most to discrimination will be discussed in detail in Section 6.5.

6.4.5 Experiment 13: Phone-based classification of read speech

Phone-based classification experiments mirrored the word-based classification experiments for read speech. Only hypotheses were evaluated because phone-level manual transcriptions of all the data were not available. Whereas for word-based classification word identities were replaced with their parts of speech for a more general model, for phone-based classification phone identities were replaced with the symbols C (for consonants) and V (for vowels). Because the feature set in this latter case only had two members, the classifier was permitted to consider sequences of length up to 5.

Data

In phone-based classification experiments, the features used as input to the classifier were phone identities and classes (vowel or consonant). Phoneme hypotheses for data sets N-E-R and NN-E-R were generated by

Condition	phone	phone class
A	100	86
B	**92**	**80**
C	88	71
D	76	82

Table 6.3: Classification accuracy of read speech. Baseline is 58%.

the ISL-BN recognizer that produced the word hypotheses, with the standard lexicon replaced by one in which each phoneme was treated as an independent word, and the word language model replaced by a phone trigram language model. This may not be the most accurate phoneme recognizer, but it did not require any additional training of acoustic models[3] and was completely sufficient for the task, as will be evident.

Results

Results for phone-based classification are shown in Table 6.3. Accuracies of phone-identity classification are higher than those for phone class (C/V) classification except when the training data was the common article and the test data was disjoint articles. This suggests that a phone-based model built from multiple examples of a limited set of phone contexts does not generalize well, although performance of that same model is perfect on new renditions of the common article.

The biggest difference between word-based and phone-based classification is seen when training and testing articles are all disjoint. With data like this, the best performance of word-based classification is 77%, using part-of-speech tags. Classification of phone identities is much more accurate, at 92%.

6.4.6 Conclusions from classification experiments

The results in these experiments show that classification of recognizer hypotheses can be extremely accurate for both binary and multilateral decisions. The test condition that is most likely to be of general interest for application to speech recognition is condition B, in which all training and test articles are disjoint.

6.4.7 Accent-dependent recognition

The objective of L1 classification, of course, is to trigger a switch in the way speech is processed. For native speakers, and possibly non-native speakers with certain characteristics, standard acoustic models, language models, and lexicon would be used. If the speaker is found to be non-native, specialized modeling would be invoked. In this section, I describe how L1 classification would fit into an LVCSR system, showing

[3]Although no new acoustic models were trained for the phoneme recognizer, because phones were treated as individual words and the internal representation in the recognizer suppresses cross-word contexts of distance greater than one, the contextual models are no longer quinphone models but rather triphone models.

Non-native Speakers			Native Speakers		
Speaker	native models	non-native models	Speaker	native models	non-native models
221	82.2	**59.5**	206	**20.1**	51.6
227	47.0	**39.1**	202	**22.8**	53.4
222	58.8	**50.5**	201	**26.3**	59.1
208	61.6	**47.5**	203	**29.7**	63.6
218	59.3	**46.6**	204	**20.3**	62.1
216	62.8	**47.0**	240	**18.5**	54.2
220	62.6	**53.0**	207	**19.4**	62.5
225	77.0	**59.4**	205	**15.0**	49.5
212	66.5	**52.7**			
209	64.7	**64.7**			
AVG	64.3	52.0	AVG	21.5	57.0

Table 6.4: Performance of native and non-native acoustic models on native and non-native speakers, given in terms of WER

how recognition accuracy would improve with optimal classification and demonstrating that my method approaches this level of performance.

Gold standard

The gold standard for accent classification in an accent-dependent recognition system is measured by calculating overall system performance given optimal classification performance. At this point, I am only considering an acoustic model switch, so optimal classification performance would mean identifying a speaker as native if and only if that speaker is recognized better by the native acoustic models.

Recognition accuracy of native and non-native acoustic models is shown in Table 6.4. The optimal result for each speaker is highlighted. In this case, native speakers are always recognized best by the native models and non-native speakers are always recognized best by the non-native models. If the best-performing model set is always used, the overall WER for all 18 speakers will be 38.7%, compared with 45.6% if the native models are always used. This is the gold standard for overall system performance which I hope to approach with automatic classification.

System implementation and evaluation

To implement on-the-fly accent-dependent recognition, I used the output of my naive Bayes L1 classifier to determine whether to use acoustic models optimized for native or non-native speech for a final recognition pass. Ideally, in such a system one would like to use disjoint sets of utterances for classifier training and

	Pure native	**Non-native**	**Gold-standard switching**	**Hypothesis-driven switching**
WER	45.6	54.2	38.7	40.3

Table 6.5: Overall recognizer performance when L1 classification is used to switch to non-native acoustic models

testing, so I will use the phone-based classification, which achieved the best performance for disjoint articles. The algorithm for running accent-dependent recognition is as follows.

1. Generate a set of initial phone hypothesis using native context-dependent acoustic models, a lexicon with entries representing phonemes, and a language model built from phoneme distributions in the language model training corpus.

2. Pass the set of hypotheses through a classifier that has been trained on phoneme hypotheses of native and non-native speech

3. If the hypothesis is classified as native, re-recognize the speech with a word lexicon and a word language model

4. If the hypothesis is classified as non-native, re-recognize the speech with customized acoustic models, a word lexicon, and a word language model.

This process can be streamlined by generating word hypotheses in step 1 and classifying based on those hypotheses; if the speaker is judged to be native, the initial hypothesis will become the final hypothesis. Because the classification accuracy for word tokens is not as high as for phoneme tokens when testing on disjoint sentence sets, one could boost system performance either by using a common set for classification or biasing the classifier to prefer false negatives to false positives. I have found that falsely identifying native speakers as non-native is more harmful than falsely identifying non-native speakers as native; the mismatch between the native speech and the non-native acoustic models is severe.

Table 6.5 shows the performance of the on-the-fly accent-dependent recognition system, comparing it with the gold standard described above. One native speaker was incorrectly classified as non-native; all other classifications were correct.

6.5 Discriminative Features in Non-native Speech

In order to understand the classifier's behavior, it is helpful to look at the individual word, part-of-speech, phone, and phone class n-grams that contribute most to successful discrimination. Rainbow provides this in the form of a list of tokens that have a high probability of being found in documents in class A and a low probability of being found in documents in class B. This term is known as the *log-odds ratio*.

Words	
Native	Non-native
NMFS	the;the
the;NMFS	in;in
nineteen;hundreds	the
hundreds;now	in
hundreds	that
habitats;and	habitat;and
'll;grow	fishers

Parts of speech	
Native	Non-native
noun(pl)	noun(sing)
determiner	preposition
noun(pl);preposition	preposition;preposition
adjective;noun(pl)	noun(sing);noun(sing)
gerund;particle	particle;preposition
noun(s);verb(3s)	cardinal#;cardinal#
noun(pl);modal	verb(past)

Table 6.6: Most discriminative word and part-of-speech n-grams in transcripts of read speech, sorted by log-odds score

6.5.1 Transcriptions of read speech

Table 6.6 shows the words and parts of speech that were important in discriminating between native and non-native transcripts of the shared article, sorted by log-odds score. The top word indicating native speech was "NMFS," which was an acronym for the National Marine Fisheries Service. The native speakers always read this smoothly, while the non-native speakers often repeated and misread letters. The top n-gram for the non-native speakers, on the other hand, was a repetition of the determiner "the." Non-native speakers frequently repeated words in their reading, possibly because they were unfamiliar with the next word. The term "nineteen hundreds" also played an important role in identifying native speech. This token was written in numerals in the text ("1900s"), and non-native speakers often did not know how to read it aloud. Whether a speaker read "habitats" or "habitat" (the correct word was "habitats") was another clue to nativeness class. Reading errors involving singular-plural confusion were extremely common in the non-native speech, and relatively rare in the native speech.

The singular-plural distinction was also important in discriminating based on part of speech. A number of plural nouns was found to be the primary indicator of nativeness. It is important to keep in mind at this point that speakers were all reading the same article; the fact that plural nouns were found to be indicative of native speech does not necessarily indicate a preference on the part of native speakers for plural nouns, but rather a tendency of non-native speakers to misread plural nouns as singular in a text where plural nouns were frequent.

6.5.2 Recognizer hypotheses of read speech

Table 6.7 shows the important word and part-of-speech n-grams in discriminating between recognizer hypotheses of the shared read article. The most striking difference, and the one most encouraging for further work in classification of recognizer output, is the word "salmon." This was an article about salmon popula-

Words	
Native	Non-native
the	that
salmon	and
will	to
with	it
salmons	we
the;NMFS	someone
habitats	some

Parts of speech	
Native	Non-native
noun(pl)	verb(past)
noun(pl);preposition	personal pronoun
adjective;noun(pl)	noun(sing)
noun(pl);modal	coordinating conjunction
adjective	"to"
determiner;adjective	noun(s);verb(past)
determiner;noun(pl)	personal pronoun;verb(past)

Table 6.7: Discriminative word and part-of-speech n-grams in recognizer hypotheses of read speech

tions, so this token appeared many times. In the native speech, it was generally recognized correctly. In the non-native speech, however, it was usually not, but was rather misrecognized as "some," "someone," and "simon," among other words. Misrecognized native productions of the word "salmon," on th e other hand, did not tend to be misrecognized this way, but rather as the plural "salmons," which, incidentally, is not the correct plural form and did not appear in the article but was allowed in the search because it was produced on occasion by non-native speakers.

Turning to the part-of-speech-based classification in the right-hand part of Table 6.7, we can see that plural nouns continue to play a role in nativeness decisions. This is true for the noisy native data set as well as the baseline native data set. The top token on the non-native list is the past tense verb. It is not obvious why this form is so indicative of non-native speech. Past tense verbs also help to identify non-native speech in transcripts, indicating that non-native speakers are indeed on occasion reading past tense forms inappropriately, but the association is much stronger in the recognizer output. My hypothesis is that the non-native speakers move less smoothly from word to word, and that epenthetic vowels, unnatural consonant releases, and inter-word human noise are taken by the recognizer to be a past tense ending.

6.5.3 Spontaneous speech

Discriminative tokens for spontaneous speech are given in Table 6.8. The word tokens include tokens representing singular, plural, and proper nouns, avoiding overtraining on specific place names. Because this is spontaneous speech, we are no longer looking at reading errors, but rather genuine preferences in word usage for the different speaker groups. The non-native data set consists of speakers of both Chinese and Japanese.

Nouns, specifically singular, non-proper nouns, are a strong indicator of non-nativeness. I have observed a tendency on the part of the non-native speakers to form sentences around noun phrases, saying, for example, "what is the price of the ticket of the show" where a native speaker might say "how much does the show cost." Native speakers use more personal pronouns in their queries to the agent, as evidenced both by the importance of the personal pronoun in the part-of-speech-based classification and related verb

Words	
Native	Non-native
am	noun(s)
proper noun	the
can;you	the;noun(s)
more	is;the
more;noun	is
give;me	noun(s);noun(s)
give	how

Parts of speech	
Native	Non-native
"to";verb(base)	noun(sing)
preposition	wh-adverb
personal pronoun	verb (3s)
verb(base)	verb(3s);determiner
adjective;noun(pl)	determiner
adjective(comp.);noun(s)	wh-adverb;verb(3s)
noun(sing);modal	determiner;noun(sing)

Table 6.8: Discriminative word and part-of-speech n-grams in transcriptions of spontaneous speech

Phones	
Native	Non-native
dh	ih
th	hh
er	ao
axr	iy
ax	ow
ax;th	aa
ch	ih;ih
xn	ng
jh	ae
dh;ey	hh;ih

Phone classes	
Native	Non-native
CCC	V
CC	VV
CCCC	VCCV
C	VC
CCCCC	CVV
CCCCV	CV
VCCCC	VVC
CVCCC	VCCVC
CCCVC	CVCCV
CCCV	CVVC

Table 6.9: Discriminative phone and phone class n-grams in phoneme hypotheses

forms like "am." Sentences like "I'm interested in seeing the Empire State Building, can you give me more information" are common in the native data, where non-native speakers showed a strong preference for simple constructions like "how do I go to the Empire State Building." This tendency also partly explains the importance of wh-adverbs (how, when, where, why) in identifying non-native speech.

6.5.4 Discriminative phone sequences

Phone identities

Table 6.9 shows the phone unigrams and bigrams that were most discriminative in this test case. Most of the phones indicative of native speech are ones that are known to be difficult for non-native speakers, particularly speakers of Japanese. R-colored vowels, reduced vowels, and the interdental consonants are classic examples; when running phoneme recognition with no lexical model, these phonemes are simply not found in Japanese-accented speech. Instead, simple vowels like [ɑ] and [i] are hypothesized with great frequency.

There are two surprising entries in this table, however. First, the voiced affricate [ʤ]. /ʤ/ is common in Japanese, and while a narrow phonetic transcription would make distinctions between the realizations in English and Japanese (Akamatsu, 1997), the differences are not at all obvious to the untrained ear. This phone is not one of the ones that would first come to mind when compiling a list of common pronunciation errors made by Japanese natives, perhaps because native English speakers are not sensitive to the kinds of deviations in this phone in Japanese-accented English (as they might be in German-accented English). The recognizer, however, apparently does perceive a significant difference, which is a small piece of evidence to support automatic, rather than linguistically-motivated, modeling of pronunciation errors. The controverting evidence is that word-level substitution and deletion errors involving /ʤ/ are not frequent, and /ʤ/ does not show a high confusability with any one particular phone.

The other puzzling observation is that [æ] is indicative of Japanese speech. This phoneme is not found in Japanese, and many Japanese speakers have a tendency to substitute a back-central low vowel. This substitution often does not affect intelligibility. Here we have the reverse of the [ʤ] situation: a phone which one might predict would consistently undergo substitution, and might be better represented by another phone in the phoneme inventory. This is not merely linguistic conjecture; both trained phoneticians and ordinary transcribers marked many instances of [æ] as having been mispronounced as [ɑ]. Nevertheless, the recognizer finds this phone more frequently in recognizer hypotheses of non-native speech than in hypotheses of native speech. This difference may be related to the tendency of native speakers to neutralize this phoneme, and others, in unstressed syllables and weak forms of words. It could be that native realizations of both [æ] and [ɑ] in fluent speech are often reduced to the point that they sound like [ə] to the recognizer, if a phoneme is detected at all. Non-native realizations, on the other hand, may be of fuller quality.

Table 6.10 shows results of phoneme recognition on native and non-native realizations of the word "can," from the sentences "Humans and salmon can peacefully coexist" and "Industry can be barred from using land." There is often no vowel recognized in the native realizations of this weak-form word; in the one case that there is it is a reduced vowel. However, in the non-native realizations, there is nearly always a full vowel recognized by the phoneme recognizer, almost always /æ/.

While this is an interesting problem, the verification of my hypothesis will be left to future exploration.

Phone classes

The consonant-vowel strings that are hypothesized are not at all surprising when considering the two groups I am attempting to distinguish. Frequent consonants and consonant clusters are clear indicators of native speech, while frequent vowels and CV-type syllables are indicators of Japanese-accented speech.

The reader may be surprised by the long sequences of consonants that were found to be indicative of native

Humans and salmon **can** *peacefully coexist*		*Industry* **can** *be barred*	
Native	**Non-native**	**Native**	**Non-native**
k n	k ae n	k	k ae m
k m	n iy ae n	k	k ae m
k m	t iy ae n	k	k ae m
k n	k ae m	k n	k n
k n	k ae m	k	k ae n
k ax	n ih ae n	k	k ae
k n	k eh m	k	sh ae m

Table 6.10: Phoneme recognition on native and non-native realizations of *can*

speech. It is important to remember that the phone hypotheses represent what the recognizer perceives, and not necessarily what the user intended to utter. 5-consonant sequences are not supposed to be common in English. However, it is easy for short or reduced phones to be absorbed by the models of the surrounding sounds. Native speakers are also notorious for not obeying the articulatory and phonological rules of their languages; modeling pronunciation variation in fluent native speech is the subject of a growing body of research (Finke and Waibel, 1997; Liu and Fung, 2000b; Nakajima et al., 2000; Nock and Young, 2000). What we see from the list of discriminative phone class sequences in Table 6.9 is that because of properties of the recognizer and of the speech, a reasonably constrained phoneme recognizer finds phone class sequences in native speech that it does not in non-native speech, and those sequences are highly discriminative.

6.6 Application to language tutoring

Although this method was originally designed to *classify* speakers as native or non-native, one could imagine also using it to offer feedback to a speaker who is learning to speak a language. Those features that are found to be most discriminative in terms of deciding whether a speaker is native or non-native could be thought of points that the user might wish to improve. Rather than offer feedback on specific productions of phones, such a system would first identify general problem phones or phone sequences for each speaker, and then present exercises to the user that target those phones.

To determine the pronunciation problems that are most damaging for the user, he would first be asked to read from a text. Underspecified recognition hypotheses for that text would then be treated as the sole non-native training document to be contrasted with the native training documents. Because we can assume that the speaker is non-native, building a robust classification model is not as important as identifying phone realizations that distinguish the speaker from the native training set.

Taking underspecified alignment results for the shared set of articles (this is a case in which having all training and test speakers read from the same text is desirable, as a decision will be based only on phone

221	227	222	208	218	216	220	225	212	209
ax	ax	ax	ax	ax	ax	ax	ax	ax	ax
dh	axr	ih	ah	ih	dh	ae	axr	ae	axr
axr	dh	ae	dh	ah	ih	ih	dh	ay	ih
ey	ah	axr	axr	axr	axr	axr	ae	ey	dh
ah	th	dh	th	ey	ah	aw	th	iy	er
ay	ey	ah	er	ix	ae	dh	ah	axr	th
eh	er	ey	ey	dh	th	hh	eh	eh	ey
th	ix	xl	ng	ay	er	eh	ix	dh	xl
xl	ih	th	ix	xl	ey	th	er	aw	ae
sh	ae	uh	ae	er	sh	ah	hh	ah	aw

Table 6.11: Phones which are found to be most problematic for each speaker using the classification-based method

realization and not on phone distribution), I measured native/non-native classification accuracy on phone unigrams to be 100% for a 20-trial held-out test. Underspecified alignment hypotheses, then, appear to offer a sound basis for classification. If we look at the phones that are found to be discriminative when the non-native model is built from the speech of a single speaker, we can see which realizations are most damaging to him in terms of differentiating his speech from that of native speakers. Table 6.11 shows the ten most problematic phones for each speaker as calculated using this method.

This is only an idea for an application of accent classification; to establish its validity as a pedagogical tool one would have to measure how well the "damaging phones" identified by the classifier correlate with human perception of accent, and also determine whether speakers' pronunciation improves with use of the system. However, attacking the problem of pronunciation tutoring by finding the areas that most mark a speaker as non-native contrasts with the more common approach of analyzing individual articulations, and is an interesting direction for future work.

6.7 Summary and conclusions

In this chapter I have shown that high-accuracy text-based nativeness classification can be implemented and improves overall system performance significantly. A text-based classification method, one that operates on the recognizer hypotheses as opposed to acoustic features, was chosen because of both its novelty and the potential for its application in situations when access to acoustic features is not desirable. For example, the output of the classifier described in this chapter could be used to switch to a non-native grammar for parsing, or to separate native from non-native utterances in language modeling; it could also be paired with off-the-shelf recognition software that does not allow access to acoustic features. Although similar methods have been used to identify the author and source of publication of a written text, somewhat similar tasks, to my knowledge, this is the first time naive Bayes based text classification techniques have been applied to classification of spoken language.

The method described here performs well in classifying transcriptions of spontaneous speech for both 2-way (native/non-native, Japanese/Chinese, etc.) and 3-way (Native/Japanese/Chinese) distinctions. Perhaps more surprising is that both transcriptions and hypotheses of read speech can also be classified with high accuracy. In the CND task, all articles were originally written by native speakers; The fact that the classifier can identify the reader as native or non-native shows that the types of reading errors made by native and non-native speakers are highly discriminative. The most interesting observation is that hypotheses are classified *more* accurately than transcriptions. This clearly says that the recognizer is responding differently to native and non-native speech.

Stopwords, those extremely common words that are often excluded from consideration in classifying native-produced text, were found to be extremely discriminative. Pruning the classification vocabulary to a list of only 70 words brought accuracy of the most difficult task, classification of speakers as native or non-native when each reading unique articles, to 87%.

Phoneme recognition hypotheses were in general a better source of input data than word recognition hypotheses. If a mandatory two-pass process in an option, the recognizer can be used to first produce a phone hypothesis, and then re-recognize the utterance at the word level with the appropriate acoustic models.

Once a speaker has been classified as native or non-native, the system can re-recognize the utterance using customized acoustic models. For our test set (N-E-R + NN-E-R), automatic model switching yielded a relative improvement of 9% over using native acoustic models for all speakers.

Chapter 7

Conclusion

Non-native speech is very diverse. Even restricting this study to a specific L1 group, proficiency level, task, and mode of speech, we have seen tremendous intra- and inter-speaker variation in the production of spoken language. As speakers traverse the learning curve, they experiment with sounds and words, sometimes generating common patterns and sometimes generating one-of-a-kind events that defy classification. Because accurate recognition depends on finding and modeling speech patterns, this diversity poses a substantial challenge for LVCSR.

The results presented in this dissertation show that while there are many elements of non-native speech that remain difficult to model, a small amount of acoustic data can be put to effective use in decreasing recognition error for non-native speakers. In this chapter, I summarize major results and contributions and discuss promising directions for extensions of this work.

7.1 Summary

In this section, the principal results and observations from each chapter in the main body of the dissertation are outlined.

Chapter 3 Characterization of non-native speech

- Native and non-native speech can be distinguished using a number of qualitative measures, including
  - Word frequency
  - N-gram frequency
  - Perplexity
  - KL divergence (more variability in non-native speech than native speech)
- Vocabulary growth rate for non-native speakers higher for native speakers, both individually and in the aggregate
- Use of contractions is different for different L1 groups
- Frequent pauses in non-native speech account contribute to a significantly slower overall speech rate and inhibit cross-word coarticulation

- Reading errors are frequent in non-native speech; 2% of words in speech do not match the source text
- Substitutions in reading are most frequently morphological variants for non-natives and orthographically similar words for natives
- Native judges show high recall but low precision detecting non-nativeness in transcribed utterances

Chapter 4 Acoustic modeling

- Context-dependent models perform better than context-independent models for low-proficiency LVCSR
- Optimal language model settings for native and non-native speech are significantly different
- Phonetic confusion occurs in the same pairs as native speech, but is more extreme
- Polyphone coverage decreases for both native and non-native speakers when non-canonical pronunciations are allowed both for native and non-native speakers
- MAP adaptation performs better than MLLR for adaptation to the non-native condition with a large adaptation data set
- Accented L2 data is a better source of adaptation data than L1 data
- Additional forward-backward iterations with L2 data give the greatest performance gains, at 30% relative word error rate reduction
- Interpolation of retrained models with baseline models improve performance further

Chapter 5 Lexical modeling

- In a large-vocabulary system, adding pronunciation variants to the lexicon before decoding can severely degrade recognizer performance. Acoustic rescoring after adding variants to the lattice results in superior recognition accuracy
- For the task and speakers that this dissertation centers on, neither data-driven nor linguistically-motivated approaches to variant derivation contribute to significantly reduced WER. This may be true of lower-proficiency speech in general
- Allowing variant pronunciations that are associated with a particular L1 group during speaker adaptation does not appear to significantly affect the quality of the adaptation

Chapter 6 Accent classification

- Naive Bayes classification can be used to make accurate bilateral and multilateral decisions about the speaker's L1
- Recognition output is more reliably classified than transcripts
- Spontaneous speech is more reliably classified than read speech
- Using a mixture of words and part-of-speech tags maximizes classification accuracy
- Phone-based classification outperforms word-based classification when training and test texts are all disjoint
- Classification results can be used to switch between native and non-native acoustic models for a significant reduction in overall WER

7.2 Major contributions

Primary contributions of this work can be summarized as follows.

A characterization of low-to-mid proficiency Japanese-influenced English. Native speakers of Japanese are of great interest in non-native speech recognition; they represent a large potential audience for language-learning software, and comparatively low speaking proficiencies for equivalent study of English makes their speech a greater challenge for LVCSR than that of many other L1 groups. The properties of speech known to be important for LVCSR have not been thoroughly examined for this group, however. This dissertation provides an extensive analysis of linguistic features such as syntax, lexical choice, fluency, and inter-speaker variation, comparing read and spontaneous speech, for lower-proficiency native speakers of Japanese.

A frame of reference for characterizing language use in other non-native speaker groups. While this dissertation focuses on one speaker group, the metrics used for speech characterization are general and similar analyses can be performed for any native language or proficiency level. Limited three-way comparisons between native speakers of English, Japanese, and Mandarin are provided to demonstrate how multilingual analysis could be approached.

A controlled study of speech errors and LVCSR performance for a specific L1 background, English proficiency, speech mode and task. It is known that non-native speech varies widely, and that variation has a negative effect on recognition accuracy. Most examinations of non-native LVCSR, however, target either high-proficiency speakers or a range of speaker proficiencies. By controlling these variables, this dissertation is able to provide strong statements about the character of the data and its response to statistical modeling and recognition.

An evaluation of adaptation and training methods and data sources for non-native speech recognition. Through a comparison of adaptation methods, training data sources (L1 vs. L2), and training data amounts, this dissertation shows how compensation for foreign accent can be expected to improve with different modeling techniques.

Significant improvements in LVCSR performance for low-proficiency read speech. The experiments described here resulted in a 30% relative improvement in recognizer accuracy, closing nearly half of the gap between performance on native and non-native speech.

A comparison of linguistically-motivated and data-driven approaches to pronunciation modeling for non-native speech. Although this dissertation did not find that lexical modeling improved recognition significantly for this data set, it provides a detailed comparison of variant generation and pruning techniques that can be used as a basis for pronunciation modeling for other proficiencies and L1 groups.

A novel and accurate method for detecting non-native utterances. Acoustic and lexical modeling experiments were designed to maximize recognizer performance for a L1-specific recognition system. If this recognizer is then to be used in conjunction with a native system or other L1-specific systems, a model-switching strategy must be employed. The method presented in this dissertation is extremely accurate in binomial and multinomial classification of both recognizer hypotheses and transcriptions.

7.3 Future directions

The research presented in this dissertation only begins to address the complex problem of modeling the diverse population of non-native speakers. While I have tried to explore the issues that I did choose thoroughly, there were many tempting paths that I chose, in the interest of time, not to follow. A few are listed below.

7.3.1 Allophonic modeling

Although the implementation of allophone tree adaptation discussed in Chapter 4 was not effective for this data set, I believe that allophonic modeling has a great deal of promise. A more sophisticated allophonic adaptation method may be able to capture L1-specific alternations in phonetic environments that occur in both L1 and L2. An allophonic model that encodes L1-dependent variation is particularly appropriate for systems that target a specific speaker group; one might expect that the influence of environment on phonetic realization, of which most speakers are unaware, is the least likely to be affected by speaker-internal inconsistency. If allophonic alternations are indeed conditioned on the same contexts when speaking L2 as when speaking L1, adaptation of all polyphones, and not just those that are introduced through phone insertion, deletion, and substitution, may contribute to a decrease in WER.

7.3.2 Speaker dependency

Speaker adaptation, which has been found to greatly improve recognizer performance, targets speaker-specific effects in the acoustic model. Speaker dependency in the lexical model, however, has not been addressed. Experiments in lexical modeling suggest that although global modeling does not improve recognizer performance, individual speakers are modeled better by some methods than others, and adapting the lexical model based on speaker-dependent properties may result in an increase in recognition accuracy.

7.3.3 Extension to other languages

In order to present a controlled study of L1-dependent LVCSR, only native speakers of one language were targeted in this dissertation. The overhead involved in collecting acoustic data for multiple languages, and ensuring relative uniformity of language background and skill among speakers, also prevented the investigation from extending the range of L1s beyond the limited study of Mandarin natives presented in Chapter 3. Whether the same adaptation methods are effective for speakers of other languages, and if not what that tells us about both L1-specific influences on L2 and the nature of non-native speech in general, has been left to future exploration.

7.3.4 Language modeling

Adaptation of the language model, which describes likely sequences of words, has not been addressed in this dissertation. It was observed, however, that speakers of certain L1s show common patterns in sentence construction. It is possible that recognizer performance could be improved by incorporating these patterns in the language model, and language model adaptation is a natural extension of this work.

7.4 Illustrative examples

With all of the word error rate figures, performance charts, and adherence to LVCSR evaluation conventions that prevent us from making simple observations about the specific errors seen in the test set (thereby avoiding, to use a timely analogy, "teaching to the test"), it is easy to lose track of what the changes in performance that we are seeing really mean. The examples represent the recognition result, after adaptation, of one randomly chosen utterance for some of the models that have been discussed in this dissertation. Misrecognized words are shown in italics.

Reference	environmentalists the government and ordi= ordinary folks team up to save the northwest's won= wondrous wild salmon
Baseline	*and that meant that* the *state department* and *the court that* ordinary folks *teamed* up to *say that no scientist but under a flight attendant*
PDTS[1]	*environment that against* the government and *a quarter* ordinary folks *teamed* up to save the northwest's *one hundred flight attendant*
MLLR-3	*environment that against* the government and *all that* ordinary folks *teamed* up to *see that and also based on wonders like exxon*
MLLR-15	*environment that* the *state department* and *the court that* ordinary folks *teamed* up to *see that northwest one wonders like exxon*
Rebuild-L2	*environment baptist* the government and *order* ordinary folks team up to save the northwest's *one hundred slide on*
MAP-15	*environment the list* the government and *called* ordinary folks team up to save the northwest's *one wonderful like son*
Retrain +interp	*and that meant the least* the government and *ordered* ordinary folks *teamed* up to save the northwest's *one wonders like son*
Lexical modeling	*environmental risks* the government and *all that* ordinary folks team up to save the northwest's *one wonders* wild *son*

The progression through better and better stages of modeling is evident from these examples. We move from a hypothesis that really gives the reader no clue as to what the speaker was trying to say to a hypothesis that is extremely close, showing evidence of confusion surrounding similar phones (/l,r/ and /t,k/ in "environmentalists"/"environmental risks"), word fragments ("ordi=" recognized as "all that," "won=" recognized as "one"), and unusual words ("wondrous," "salmon").

While there are still clearly problems that remain to be resolved, the experiments in this dissertation show how much ground can be covered with a small amount of data and techniques that are for the most part widely used. We may never be able to coax native-level performance out of the recognizer for low-proficiency non-native speech, but this work suggests that speech recognition for non-native speakers is a realistic goal, and outlines analysis and adaptation methods that will contribute to reaching it.

[1] The abbreviations used here are the same as those given in the summary of acoustic modeling results in Figure 4.14 on page 96

Appendix A

Data collection and speaker proficiency evaluation

A.1 SPEAK rating criteria

Pronunciation

0 Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be unintelligible.

1 Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be occasionally unintelligible.

2 Some consistent phonemic errors and foreign stress and intonation patterns, but speaker is intelligible.

3 Occasional nonnative pronunciation errors, but speaker is always intelligible.

Grammar

0 Virtually no grammatical or syntactical control except in simple stock phrases.

1 Some control of basic grammatical constructions but with major and/or repeated errors that interfere with intelligibility.

2 Generally good control in all constructions with grammatical errors that do not interfere with overall intelligibility.

3 Sporadic minor grammatical errors that could be made inadvertently by native speakers.

Fluency

0 Speech is so halting and fragmentary or has such a nonnative flow that intelligibility is virtually impossible.

1 Numerous nonnative pauses and/or a nonnative flow that interferes with intelligibility.

2 Some nonnative pauses that do not interfere with intelligibility

3 Speech is smooth and effortless, closely approximating that of a native speaker.

Comprehensibility

0 Overall comprehensibility too low in even the simplest type of speech.

1 Generally not comprehensible because of frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control.

2 Comprehensible with errors in pronunciation, grammar, choice of vocabulary items or infrequent pauses or rephrasing.

3 Completely comprehensible in normal speech with occasional grammatical or pronunciation errors.

A.2 NPR1

Relief workers have returned to the rebel held city of Goma in eastern Zaire in a second attempt to distribute food to starving refugees.
The B. B. C.'s Alan Little reports the first convoy was held up at the border.
The B. B. C.'s Alan Little reporting from Goma town.
The government of Zaire says the entry of aid convoys from Rwanda represents a violation of territorial sovereignty.
The supplies carried by the convoy are the first to reach eastern Zaire from rwanda since last spring.
Canada has offered to lead a multinational military force to help ease the refugee crisis in Zaire.
A third gasoline storage tank has erupted in flames at a petroleum storage facility near Mexico City.
At least a dozen people have been injured and about twenty four hundred have been forced to evacuate their homes.
The fire was triggered by an explosion in two storage tanks holding more than four million gallons of gasoline.
Firefighters have been spraying a curtain of water in an effort to contain the blaze.
More snow is falling this morning in northern Ohio and other parts of the great lakes region tens of thousands of homes remain without electricity.
From member station W. C. P. N. in Cleveland Joe Smith reports.
A Delta airlines jetliner slid off a runway at Cleveland's snowy Hopkins international airport last night.
No one was injured.
It was the second such incident at the airport in as many days.
This is N. P. R. news.
Meetings in advance of an economic summit in Cairo have failed to produce any breakthrough in negotiations on the withdrawal of Israeli troops from the west bank city of Hebron.
Secretary of state Warren Christopher met with Palestinian leader Yasser Arafat.
The state department says the session yielded no discernible progress.
President Clinton met meets with congressional leaders today in a search for common ground.
Ways to balance the budget are at the top of the agenda.
Mr. Clinton insists the two sides are not that far apart.
He has suggested the administration and congressional republicans pick up where they left off before the political conventions.
Twenty students were arrested Monday during a protest at the University of California Riverside.
They were demonstrating against an affirmative action proposition that was passed by California voters last week.
For member station K. C. L. U. Jeff Barry reports.
Court action begins in orange county California today on O. J. Simpson's effort to regain custody of his two youngest children eleven year old Sidney and eight year old Justin.
The children have been living with the parents of their mother Nicole Brown Simpson since she was murdered two and a half years ago.
The question for the court is whether they will stay with their grandparents or move in with their father.
This is national Public Radio News from Washington.

A.3 NPR2

The guardians of the electronic stock market NASDAQ who have been burned by past ethics questions are moving to head off market fraud by toughening the rules for companies that want to be listed on the exchange Marketplace's Philip Boroff reports.
As part of the proposals penny stocks will be eliminated from NASDAQ These trade for literally pennies.
Less than a dollar a share.
They are the stocks of speculative companies.
On wall street they are the longest of the long shots.
Some penny stocks grow into established corporations.
Others are shell companies.
Incorporated firms without assets or prospects.
Some of these are sold by small unsavory brokerage firms that dump them upon gullible investors.
David Whitcomb is a Rutgers University finance professor and frequent NASDAQ critic.
That is the real change it is reducing the status of cheap stocks so.
that at least NASDAQ is not giving them its seal of approval.
Also these companies will no longer appear in newspapers on NASDAQ's list.
And Whitcomb says investors may be less prone to buy them if they are not listed in the paper.
NASDAQ officials say they are not only trying to fight fraud by raising listing standards they are doing a periodic tuneup of their market Which they hope will help promote public confidence In New York.
I am Philip Boroff for Marketplace.
Today the Dow Jones industrial average gained thirty eight and three quarter points.
Details when we do the numbers.
Later on tonight's program life in the fast lane.
And coming up next a fast food Godzilla joins the burger wars in Japan.
I am David Brancaccio this is Marketplace.
American popular culture whether it is rock and roll fashion or Hollywood movies has long been an important export Even though statisticians have a hard time measuring its value.
Take fast food.
When the first American style burger joint opened in London's fashionable Regent street some twenty years ago it was mobbed.
Now it is Asia's turn.
As the people in the far east get richer they are anxious to try and able to afford burgers fries pizza and chicken The latest entrant is Burger King.
Its C. E. O. Robert Lowes arrived in Japan to launch a belated effort to grab a share of the country's annual five billion dollar burger market.
Marketplace's Tokyo bureau chief Jocelyn Ford reports.
Asia is in the midst of a fast food rush and Burger King is the latest American chain to try to get a bite of the booming business.
So far this year Burger King has opened five outlets in the Tokyo area and it plans to expand to thirty five by the end of the year.
The world's number two burger chain is twenty years and over a thousand outlets behind its archrival McDonald's but Burger King C. E. O. Robert Lowes says better late than never.
The fastest growing markets in the world today are essentially Latin America and Asia Pacific.
I find it very difficult to comprehend any company who desires to be one of the better global companies in the business that it is in ignoring those markets.
Lowes says in Asia the demographic recipe is right for growing the fast food market.
As the economies develop as you know more women work in the workplace it demands more convenience and while sometimes the dual income families they are making more money I think they want the higher quality products.
There is a long list of American fast foods that have successfully made the long march to Asia.
From Seoul to Singapore hungry consumers can grab pizzas burgers and chicken.
But the welcome mat is not always out.
In some markets there has been a nationalist backlash to western fast food joints.

Take India for example.
Earlier this year a farmer's group ransacked a K. F. C. outlet and McDonald's was met with protests by farmers when it opened its first lamb burger restaurant last month.
Patricia Horvath is an analyst with U. B. S. securities in Tokyo.

A.4 CND1

A SAFETY NET FOR SALMON

Environmentalists, the government and ordinary folks team up to save the Northwest's wondrous wild salmon

Pacific salmon have never had it easy. Sure, the fish begin life gently enough, wiggling around in sun-dappled creeks and pools with their brothers and sisters. When they are bigger, they set off downstream to the ocean, where they'll grow up.

But after a few years in the ocean, life gets tough. Something in their nature tells them that it's time to go home. Salmon find their way back to the mouth of the river that carried them to the ocean. Then they swim upstream in a fierce, wrong-way struggle to their birthplace. Some travel hundreds of miles! They jump against the currents and waterfalls that once carried them out to sea. After they reach their birthplace, females lay eggs to continue the circle of life. Then the salmon die of exhaustion and old age. What a way to go!

As if that journey weren't challenging enough, the people who share the salmon's habitat have made life even harder. Pollution, overfishing and habitat destruction threaten salmon species with extinction. Only 100 years ago, the rivers of Washington State and Oregon were just jumping with salmon. But in most places, fishermen today catch one-third fewer Chinook salmon than they did in the early 1900s. Now the government has decided to get serious about rescuing these silvery symbols of the wild Northwest.

Save The Salmon: It's The Law

Last week the National Marine Fisheries Service (NMFS) announced that nine kinds of salmon and related fish would be protected under the Endangered Species Act. One of those, the Upper Columbia Chinook salmon, is now listed as endangered. The other eight fish are considered threatened.

The 26-year-old Endangered Species Act is one tough environmental law. When a species is protected under the act, industry can be barred from using land where the endangered animal might be found. Ordinary citizens also face strict rules about using protected habitat.

But there's no way to keep people away from the salmon habitats, and that makes protecting the salmon really tricky. At least 5 million people, including those in the big cities of Portland and Seattle, live near the Columbia and Willamette river systems, where threatened salmon swim. No use of the Endangered Species Act has ever affected so many people.

The new rules will change the way people farm, fish, harvest timber, build homes, use water and chemicals, and work (see chart). Restoring salmon populations to healthy levels will be an upstream struggle for everyone in the area. But so far, Northwesterners say they are up for the challenge.

"The salmon are an important part of our lives," said Seattle Mayor Paul Schell last week. "We understand that preserving our environment has a direct impact on our quality of life and our economy." Will All This Effort Be Worth It?

Over the next two months, public officials, Native American leaders, businesses and environmental groups will come up with plans for meeting the law's requirements. The NMFS must approve these plans, but some groups are already taking steps to help the salmon. They want to get a head start on the far-reaching and costly changes that will be needed.

Rollie Schmitten, director of the NMFS, hopes that the people of Washington and Oregon will continue to support efforts to save the salmon, even when it means making expensive changes in how folks live and work. He says that in the end, their personal sacrifices and higher taxes will pay off. "Humans and salmon can peacefully coexist and even enhance each other's quality of life," says Schmitten.

After all, both species thrive on the same things: clean water, green shade trees and a safe place to come home to.

A.5 TFK1

After a few years in the ocean, life gets tough. Something in their nature tells them that it's time to go home. Salmon find their way back to the mouth of the river that carried them to the ocean. Then they swim upstream in a fierce, wrong-way struggle to their birthplace. Some travel hundreds of miles. They jump against the currents and waterfalls that once carried them out to sea. After they reach their birthplace, females lay eggs to continue the circle of life. Then the salmon die of exhaustion and old age.

Last week the National Marine Fisheries Service announced that nine kinds of salmon and related fish would be protected under the Endangered Species Act. One of those, the Upper Columbia Chinook salmon, is now listed as endangered. The other eight fish are considered threatened. The 26-year-old Endangered Species Act is one tough environmental law. When a species is protected under the act, industry can be barred from using land where the endangered animal might be found. Ordinary citizens also face strict rules about using protected habitat.

Over the next two months, public officials, Native American leaders, businesses and environmental groups will come up with plans for meeting the law's requirements. The NMFS must approve these plans, but some groups are already taking steps to help the salmon. They want to get a head start on the far-reaching and costly changes that will be needed. Rollie Schmitten, director of the NMFS, hopes that the people of Washington and Oregon will continue to support efforts to save the salmon, even when it means making expensive changes in how folks live and work. He says that in the end, their personal sacrifices and higher taxes will pay off. "Humans and salmon can peacefully coexist and even enhance each other's quality of life," says Schmitten.

A.6 Japanese prompts

.

Scenario 2. Restaurants

あなたはガイドブックが勧めている the Lemongrass Grill に食事に行こうとしています。the Lemongrass Grill について以下の項目を尋ねてください。

- 何料理の店か
- 値段
- 営業時間
- 予約は必要か
- the Plaza Hotel からの距離
- the Plaza Hotel へ帰る時に乗るべきバス・地下鉄など

A.7 Snow White

SNOW WHITE AND THE SEVEN DWARVES

Once upon a time in a great castle, a Prince's daughter grew up happy and contented, in spite of a jealous stepmother. She was very pretty, with blue eyes and long black hair. Her skin was delicate and fair, and so she was called Snow White. Everyone was quite sure she would become very beautiful. Though her stepmother was a wicked woman, she too was very beautiful, and the magic mirror told her this every day, whenever she asked it.

"Mirror, mirror on the wall, who is the loveliest lady in the land?" The reply was always; "You are, your Majesty," until the dreadful day when she heard it say, "Snow White is the loveliest in the land." The stepmother was furious and began plotting to get rid of her rival.

Calling one of her trusty servants, she bribed him with a rich reward to take Snow White into the forest, far away from the Castle. Then, unseen, he was to put her to death. The greedy servant, attracted to the reward, agreed to do this deed, and he led the innocent little girl away. However, when they came to the fatal spot, the man's courage failed him and, leaving Snow White sitting beside a tree, he mumbled an excuse and ran off. Snow White was all alone in the forest.

Night came, but the servant did not return. Snow White, alone in the dark forest, began to cry bitterly. She thought she could feel terrible eyes spying on her, and she heard strange sounds and rustlings that made her heart thump. At last, overcome by tiredness, she fell asleep curled under a tree.

At last, dawn woke the forest to the song of the birds, and Snow White too, awoke. She found a path and walked along it, hopefully. On she walked till she came to a clearing. There stood a strange cottage, with a tiny door, tiny windows and a tiny chimney. Everything about the cottage was much smaller than it ought to be. Snow White pushed the door open.

"I wonder who lives here?" she said to herself, looking round the kitchen. "What tiny plates! And spoons! There must be seven of them, the table's laid for seven people." Upstairs was a bedroom with seven neat little beds. Going back to the kitchen, Snow White had an idea.

"I'll make them something to eat. When they come home, they'll be glad to find a meal ready." That evening, seven tiny men marched home singing. But when they opened the door, to their surprise they found a bowl of hot soup on the table, and the whole house very clean. Upstairs was Snow White, fast asleep on one of the beds. The chief dwarf shook her gently.

"Who are you?" he asked. Snow White told them her sad story, and tears came to the dwarves' eyes. Then one of them said, as he noisily blew his nose:

"Stay here with us!"

"Hooray! Hooray!" they cheered, dancing joyfully round the little girl. The dwarves said to Snow White:

"You can live here and keep house while we're down at work. Don't worry about your stepmother leaving you in the forest. We love you and we'll take care of you!" Snow White gratefully accepted their hospitality, and the next morning the dwarves set off for work. But they warned Snow White not to open the door to strangers.

Meanwhile, the servant had returned to the castle, with the heart of a deer. He gave it to the cruel stepmother, telling her it belonged to Snow White, so that he could claim the reward. Highly pleased, the stepmother turned again to the magic mirror. But the mirror replied: "The loveliest in the land is still Snow White, who lives in the seven dwarves' cottage, down in the forest." The stepmother was very angry.

"She must die! She must die!" she screamed. Dressing herself as an old woman, she put a poisoned apple with the others in her basket. Then, taking the quickest way into the forest, she crossed the swamp at the edge of the trees. She reached the bank unseen, just as Snow White stood waving goodbye to the seven dwarves on their way to work.

Snow White was in the kitchen when she heard the sound at the door: KNOCK! KNOCK!

"Who's there?" she called.

"I'm an old woman selling apples," came the reply.

"I don't need any apples, thank you," she replied.

"But they are beautiful apples and so juicy!" said the velvety voice from outside the door.

"I'm not supposed to open the door to anyone," said the girl.

"And quite right too! Good girl! If you promised not to open up to strangers, then of course you can't buy. You are a good girl indeed!" Then the old woman went on.

"And as a reward for being good, I'm going to make you a gift of one of my apples!" Without a further thought, Snow White opened the door just a tiny crack, to take the apple.

"Isn't that a nice apple?" Snow White bit into the fruit, and as she did, fell to the ground in a faint: the effect of the terrible poison left her lifeless instantly.

Now chuckling evilly, the wicked stepmother hurried off. But as she ran back across the swamp, she tripped and fell into the quicksand. No one heard her cries for help, and she disappeared without a trace.

Meanwhile, the dwarves came out of the mine to find the sky had grown dark and stormy. Loud thunder echoed through the valleys and streaks of lightning ripped the sky. Worried about Snow White, they ran as quickly as they could down the mountain to the cottage.

There they found Snow White, lying still and lifeless, the poisoned apple by her side. They did their best to bring her around, but it was no use.

They wept and wept for a long time. Then they laid her on a bed of rose petals, carried her into the forest and put her in a crystal coffin.

Each day they laid a flower there.

Then one evening, they discovered a strange young man admiring Snow White's lovely face through the glass. After listening to the story, the Prince (for he was a prince!) made a suggestion.

"If you allow me to take her to the Castle, I'll call in famous doctors to waken her from this strange sleep. She's so lovely, I'd love to kiss her!" He did, and as though by magic, the Prince's kiss broke the spell. To everyone's astonishment, Snow White opened her eyes. She had amazingly come back to life! Now in love, the Prince asked Snow White to marry him, and the dwarves reluctantly had to say good bye to Snow White.

From that day on, Snow White lived happily in a great castle. But from time to time, she went back to visit the little cottage down in the forest.

A.8 Example of a transcript of read speech

3.26	5.41	#rustle# Storming Disney's Kingdom
6.07	14.33	/br/ *Anastasia leads the charge as rival studios <;1 &studio> move in on toon town #rustle#
16.00	23.65	{-/A headstr=/- a <;1 &an> headstrong} Russian princess will try to win your heart this week
23.80	30.53	Fox Animation Studios <;1 &studio> is /ls/ {-/re=/- *releasing} its first cartoon feature /br/ Anastasia
30.81	51.24	Like Disney Studios' best loved hits /br/ the movie features a beautiful ((heroine)) <;del a> devilish ((villain)) /ls/ cute animal sidekicks catchy songs /br/ a plot that *rewrites history and an all star cast doing voices
51.82	60.31	Fox wants to {-/pro=/- prove} #rustle# that a cartoon movie doesn't have to come from Disney in order to be a winner with kids
60.98	74.19	/br/ The movie which opens November twenty first /br/ is based on the true story of a royal princess who disappeared in the nineteen /br/ seventeen revolution in Russia
74.69	84.31	/ls/ The {-/part=/- partly} computer animated backgrounds of great cities and snowy landscapes are {-/brea=/- ((breathtaking))}
84.33	89.26	/ls/ Anastasia herself is a smart /br/ lovable *heroine
90.07	99.76	But it takes more than #pause# gorgeous cartooning and #pause# /ls/ good storytelling to make a hit animated movie these days
100.12	105.47	/br/ Will kids and parents buy Anastasia toys games and videos <;1 &video> too
106.01	109.09	Will they go see the movie more than once
109.37	119.04	Will <;ins the> Fox's #pause# film sell as many action figures and fast food /br/ meals as The Little Mermaid or ((Aladdin))
119.29	129.03	/br/ Anastasia's pro= producers who spent about /br/ fifty three million dollars making the movie have their fingers crossed
129.72	139.03	I really hope it will {-/com=/- /ls/ compete} with the best Disney pictures says Fox movie chief Bill <;1 &billy> {-/*mechanic/- Mechanic}
140.62	155.62	The company that has #begin rustle# ruled the animation #pause# kingdom for sixty years does not plan to sit still #end rustle# while a little princess grabs for the cartoon movie throne
156.36	161.36	Disney will try to lure kids away from Anastasia this month
161.46	171.41	/br/ Its new Robin Williams movie Flubber /br/ and nineteen eighty nine's The Little Mermaid will compete against Fox's film in theaters
172.18	174.75	/br/ Are we going to make it easy for them /ls/
174.73	178.83	/ls/ No says Disney movie group chairman Richard Cook
179.22	181.03	/br/ Are we going to compete
181.03	181.79	You bet
182.10	185.95	Don't be fooled by the pretty songs and scenery
185.97	187.25	This is war
188.83	190.97	/br/ How A Mouse Became A Giant
191.53	201.63	It all started in nineteen thirty seven <;1 &seventh> #pause# with a movie about a fair skinned beauty and seven short guys
202.90	210.80	The first movie length cartoon was Walt Disney's /br/ eighty two minute Snow White and the Seven Dwarfs
211.46	216.05	The animated musical delighted audiences all over the world
216.51	222.84	Disney {-/went/- went on} to make more than thirty animated features and had little competition
223.30	233.22	The company has sold millions of dollars' worth of toys /br/ games *clothes and videos /br/ based on its popular {-/c=/- characters}

A.9 Example of a transcript of spontaneous speech

```
1.23    4.93    /ls/ alright /uh/ where is the the [Empire State Building ((IY M P AY ER S
                T EY T B IH L D IY NG))] /ls/ located
5.57    7.72    /uh/ how much is admission fee
7.8     10.84   /uh/ how long do you think it take to look around
11.36   17.12 is there any /n/ /n/ any good place to see /ah/ near [Empire State Building
                ((IY M P AY ER S T EY T B IH L D IY NG))]
17.15   20.75   is there any good restaurant around in the around there /h#/
21.4    28.76   {-/how long i=/- (how long is)} how long is it from the [Edison ((EH D IH S AH N))]
                [Hotel ((HH OW T EH L))] to to the [Empire State Building ((EH M P AY ER
                 S T EY T B IH L D IH NG))]
32.80   37.4    /uh/ what kind of restaurant is it #noise# the [Chelsea Bistro and Bar
                ((CH EH L S IY B IY S T R OW EH N B AA R))]
37.58   42.06   /h#/ and how much do I expect to pay for the the restaurant /h#/
42.78   45.24   /uh/ what is the business hour
45.66   48.5    /uh/ do you think I need a reservation
49.09   55.15   /uh/ how long is it from here to the /uh/ from the [Edison ((EH D IH S AH N))]
                [Hotel ((HH OW T EL))] to that to that restaurant /h#/
55.45   65.73   /uh/ do you have any suggestions *pause* {+/when I finish/+ *pause* when I
                finish} my dinner *pause* to go back to the [Edison ((EH D IH S AH N))]
                [Hotel ((HH OW T EH L))] say like bus or /uh/ subway #noise#
67.55   71.90   how long does it how long does it take to go to [Long Island
                ((L AO NG AY L IH N T))]
72.0    76.41   /uh/ what kind of /uh/ transportat= transportation is available /h#/
76.79   81.35   /ls/ /uh/ do you know how much it is like for the bus or train /h#/
81.6    84.63   /uh/ is there anything interesting in [Long Island ((L AO NG AY L IH N T))]
84.96   86.06   /ls/ /uh/
86.45   88.64   is there any good restaurants
89.59   95.01   /uh/ do you what time is the last bus last ship or last train
96.58   101.17  where is the [Rockettes ((R AO K IH T))] [Rockettes ((R AO K IH T S))] located
                #noise# /um/
103.36  110.33  {+/how long/+ *pause* how long} is it from the [Rockettes ((R AO K IH T S))]
                to the most /uh/ to the nearest transportation
110.63  112.57  when does the show begin #click#
112.61  114.88  when does the show /uh/ finish
115.38  117.0   /ls/ how much is the ticket
117.28  119.96  /h#/ /uh/ {+/how can I/+ how can I} buy the ticket
```

A.10 Speaker Demographics

A.10.1 Speakers completing the read task

Speaker ID	Gender	L1	Age	Years studying English	Years immersed in English	Proficiency score
201	f	english	30	—	—	4
203	f	english	19	—	—	4
204	m	english	23	—	—	4
205	m	english	37	—	—	4
206	f	english	19	—	—	4
207	m	english	20	—	—	4
240	m	english	25	—	—	4
241	m	english	26	—	—	4
242	m	english	32	—	—	4
208 *	f	japanese	29	7	1.5	1.94
209 *	m	japanese	29	8	0.75	1.94
210	f	japanese	33	8	0	1.83
211	f	japanese	57	8	0	1.11
212 *	m	japanese	31	8	2	2.11
213	f	japanese	29	8	0	1.06
214	f	japanese	29	8	0	1.00
215	f	japanese	25	6	2	2.00
216 *	m	japanese	36	10	0.33	1.94
217	m	japanese	27	6	9	2.83
218 *	f	japanese	26	10	3	2.00
219	f	japanese	34	8	7	2.67
220 *	m	japanese	31	10	1.5	2.11
221 *	f	japanese	31	10	1.5	1.83
222 *	f	japanese	23	10	0.50	2.17
223	f	japanese	26	10	3.5	2.44
224	f	japanese	32	5	11	2.83
225 *	m	japanese	31	10	2.5	1.89
226	f	japanese	30	10	7	2.05
227 *	f	japanese	29	17	0.67	1.89
228	f	japanese	25	9	0.67	2.00
229	f	japanese	26	8	0.50	2.17
230	m	japanese	25	10	0.50	1.44
231	f	japanese	31	7	1.3	1.22
232	f	japanese	28	8	1.5	2.00
233	m	japanese	33	10	0.17	1.89
234	f	japanese	34	10	0.50	2.00
235	f	japanese	31	8	0	1.33
236	f	japanese	36	8	1	1.00
237	m	japanese	40	6	0	1.00
239	m	japanese	40	8	1.75	1.33

Speakers with an asterisk (*) by their names were part of the proficiency-controlled test set.

A.10.2 Speakers completing the spontaneous task

Speaker ID	Gender	L1	Age	Years studying English	Years immersed in English
009	m	english	19	—	—
010	f	english	19	—	—
012	f	english	30	—	—
102	f	english	25	—	—
108	m	english	26	—	—
105	m	english	22	—	—
106	m	english	41	—	—
806	m	taiwanese	24	0	14
801	f	mandarin	24	10	1
802	f	mandarin	28	15	1
808	f	mandarin	29	17	1
805	f	mandarin	—	1	9
804	m	mandarin	24	10	0
803	m	mandarin	30	7	1
807	m	mandarin	27	13	1
001	m	japanese	27	6	3
002	f	japanese	27	10	2
003	f	japanese	31	10	0.42
004	f	japanese	31	15	8
005	f	japanese	29	7	1.5
006	f	japanese	28	13	1.5
007	f	japanese	31	15	8
008	f	japanese	29	7	1.5
011	m	japanese	31	8	2
013	f	japanese	28	13	1.5
014	f	japanese	22	8	.5
015	f	japanese	21	7	1
016	f	japanese	21	7	3
017	f	japanese	31	3	0.58
018	m	japanese	21	9	0
019	m	japanese	22	10	0
020	m	japanese	21	15	0
021	m	japanese	26	8	2
022	f	japanese	26	7	1
023	m	japanese	27	25	0.42
024	m	japanese	26	6	8
025	f	japanese	29	7	1.5
026	m	japanese	29	8	2
027	m	japanese	29	8	0.75
028	f	japanese	29	7	0.67
029	f	japanese	25	8	0.25
030	f	japanese	42	12	17
031	f	japanese	30	6	0.58
032	m	japanese	30	15	1
033	f	japanese	20	3	4
034	m	japanese	19	10	0
035	f	japanese	19	6	0.25
036	f	japanese	28	8	0.25
037	m	japanese	35	3	0
038	f	japanese	20	3	1
039	m	japanese	23	8	1
040	m	japanese	23	0	18

Appendix B

Phonological transformation rules

This appendix lists the transformation rules used to produce the dictionaries described in Section 5.2.2. In the first column are the rule tags. The rule is given in the second column. The symbols used in the rules represent the units in the lexicon, i.e., the base phone is the one that was in the canonical transcription and the surface phone is the one that will be added. An example of a word that is affected by the transformation is shown for each rule, with the canonical lexicon entry and a phonetic transcription of an instance of that word in the training data in which the transformation was observed. Because the units of representation are different, the surface symbol on the right side of the rule may not match its counterpart in the phonetic transcription. The phonetic transcription is shown to give as accurate a portrait as possible of the actual realization. Distinctions that were not phonemic in the original lexicon, such as [a,ɒ] and [o,oː], are suspended in the new lexicon. Some global transformations were added to resolve transcription inconsistencies in the lexicon.

This list shows rules for *adding* paths to the pronunciation networks. These are not replacement rules.

MA-3	s → ʃ / __ {i,ɪ,ɪ}	citizen	/sɪtɪzən/	[ʃitizun]
MA-3	h → f / __ {u,ʊ}	hood	/hʊd/	[ɸuːd]
MA-3	f → h / __ {ɔĭ,o,ɔ}	telephone	/tɛləfon/	[tɛlɛhoːn]
MA-4	w → ∅ / __ {u,ʊ}	woman	/wʊmən/	[uːman]
MA-5	w → u / __ {i,ɪ,ɪ}	wish	/wɪʃ/	[uɪʃː]
MA-5	w → u / __ {eĭ,ɛ}	wedding	/wɛdɪŋ/	[uɛdɪŋgu]
MA-5	w → u / __ {ɔĭ,o,ɔ}	water	/wɔtɚ/	[uɔːtaː]
MA-6	d → ʤ / __ {i,ɪ,ɪ}	candidate	/kændɪdeĭt/	[kjanʤideit]
CC-1	d → ∅ / __ z$	needs	/nidz/	[niːzu]
CC-2	∅ → o / {t,d} __ C	handmade	/hændmeĭd/	[hɒndomɛido]
CC-3	∅ → i / {ʤ,ʧ,ʃ} __ C	hitchhiking	/hɪʧhaĭkɪŋ/	[hiʧːihaikingu]
CC-5	∅ → u / {p,b,f,v,θ,ð,s,z,ʒ,g,k,m,l,ts} __ C	difficult	/dɪfɪkəlt/	[difikaluto]
CC-4	w → u / {p,b,f,v,θ,ð,s,z,ʒ,g,k,m,l,ts} __	swam	/swæm/	[suam]
CC-6	w → u / {ʤ,ʧ,ʃ} __ V	Schweitzer	/ʃwaĭtsɚ/	[ʃuaitsaː]
CC-6	w → u / {t,d} __ V	twelve	/twɛlv/	[tuɛlubu]
CC-7	w → ∅ / __ {u,ʊ}	woman	/wʊmən/	[uːman]
CC-8	w → u / $ __ {ɔĭ,o,ɔ}	wove	/wov/	[uovu]
CC-8	w → u / $ __ {eĭ,ɛ}	wedding	/wɛdɪŋ/	[uɛdɪŋgu]
CC-8	w → u / $ __ {i,ɪ,ɪ}	weekend	/wikɛnd/	[uikuɛnd]
CC-9	j → ∅ / $ __ i	year	/jir/	[iɚ]
CC-12	ŋ → n / __ {k,g}	Bangkok	/bæŋkɔk/	[baŋkok]
CC-13	ŋ → ngu / __ $	Hemingway	/hɛmɪŋweĭ/	[hɛmɪŋgueː]

FV-1	∅ → i / {ʤ,ʧ,ʃ} __ #	bridge	/brɪʤ/	[buriʤːi]
FV-2	∅ → u / {p,b,f,v,θ,ð,s,z,ʒ,g,k,m,l,ts} __ #	reptile	/rɛptaĭl/	[rɛputaĭːlu]
FV-3	∅ → o / {t,d} __ #	adult	/ədʌlt/	[adaɾɯto]
FV-4	ŋ → ngu / __ #	swimming	/swɪmɪŋ/	[suimiŋgu]
RL-1	r → ∅ / {ɔĭ,o,ɔ} __ $C	morphology	/mɔrfɒləʤi/	[moːfoloʤi]
RL-2	r → ɒ / {ɔĭ,o,ɔ} __ $V	moreover	/mɔrovɚ/	[moɒoːvɒː]
RL-3	r → ɒ / {ɔĭ,o,ɔ} __ #	more	/mɔr/	[moɒ]
RL-4	r → ∅ / {aĭ,aŭ} __ C	Arkansas	/ɒrkənsɔ/	[ɒːkɒnsoː]
RL-7	r → ɒ / V __ C	cart	/kɒrt/	[kɒːt]
RL-8	r → ɒ / V __ $	heirloom	/ɛrlum/	[eɒlum]
RL-9	r → ɒ / V __ #	gear	/gir/	[giɒ]
RL-11	ɚ → ɒ / __ $	searching	/sɚʧɪŋ/	[saːʧingu]
RL-12	ɚ → ɒ / __ C	searching	/sɚʧ/	[saːʧi]
RL-13	ɚ → ɒ / __ #	sir	/sɚ/	[saː]
MD-1	aĭ → ɒi	like	/laĭk/	[laik]
MD-2	ɔĭ → oi	boy	/bɔĭ/	[boi]
MD-3	eĭ → ɛi	make	/meĭk/	[meik]
MD-4	aŭ → ɒu	house	/haŭs/	[haus]
MD-5	æ → jɒ / k __	cash	/kæt/	[kjat]
MP-13	θ → s	breath	/brɛθ/	[blɛs]
MP-14	ð → z	then	/ðɛn/	[zɛn]
MP-15	v → b	never	/nɛvɚ/	[nebɚ]
MP-16	ɾ → t	water	[wɔɾɚ]	[uɔːtaː]
MP-18	ʒ → ʤ	measure	/mɛʒɚ/	[meʤa]
MA-1	l → r	place	/pleĭs/	[prɛs]
MA-2	r → l	reason	/rizən/	[lizan]
MP-1	{ɔĭ,o,ɔ} → o			
MP-2	{eĭ,ɛ} → ɛ			
MP-3	{i,ɪ,ɪ} → i			
MP-6	{æ,ɒ,ʌ,ə} → ɒ			
MP-9	{u,ʊ} → u			

Appendix C

ARPABET-IPA mappings

NOISES

+BR	breathing
+HU	human noise
+NH	non-human noise
+SM	lip smack
+TH	throat clearing
+LA	laughter
+F	semantic noise (um, uh)

DIPHTHONGS

AW	aŭ
AY	aĭ
EY	eĭ
OY	ɔĭ

VOWELS

AA	ɒ
AE	æ
AH	ʌ
AX	ə
AO	ɔ
EH	ɛ
ER	ˈɚ
AXR	ɚ
IH	ɪ
IX	ɪ
IY	ˈi
OW	o
UH	ʊ
UW	u

CONSONANTS

B	b	K	k	SH	ʃ
CH	ʧ	L	l	T	t
D	d	M	m	TH	θ
DH	ð	N	n	V	v
DX	ɾ	NG	ŋ	W	w
F	f	P	p	Y	j
G	g	R	r	Z	z
HH	h	S	s	ZH	ʒ
JH	ʤ				

SPECIAL PHONES

Syllabic continuants		Unreleased stops	
XL	ɫ̩	PD	p˺
XM	m̩	TD	t˺
XN	n̩	KD	k˺

References

[Ahlen et al.1997] Sondra Ahlen, Brian Connelly, Michelle Corkadel, Rob Malkin, Anuj Vaidya, and Rodolfo Vega. 1997. Data collection scenarios for c-star travel domain. Technical Report CMU-LTI-97-153, Carnegie Mellon University.

[Akamatsu1997] Tsutomu Akamatsu. 1997. *Japanese Phonetics: Theory and Practice.* Lincom Europa, Newcastle.

[Amdall et al.2000] Ingunn Amdall, Filipp Korkmazskiy, and Arun C. Surendran. 2000. Joint pronunciation modeling of non-native speakers using data-driven methods. In *Proc. ICSLP*, Beijing.

[Argamon-Engelson et al.1998] Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-based text categorization: What newspaper am I reading? In *AAAI Workshop on Learning for Text Categorization.*

[Auberg et al.1998] Stefan Auberg, Nelson Correa, Victoria Locktionova, Richard Molitor, and Martin Rothenberg. 1998. The Accent Coach: An English Pronunciation Training System for Japanese Speakers. In *Proc. Speech Technology in Language Learning (STiLL).*

[Beaugendre et al.2000] Frédéric Beaugendre, Tom Clase, and Hugo van Hamme. 2000. Dialect adaptation for mandarin chinese. In *Proc. ICSLP.*

[Beebe1987] Leslie M. Beebe. 1987. Myths about interlanguage phonology. In Georgette Ioup and Steven H. Weinberger, editors, *Interlanguage Phonology: The Acquisition of a Second Language Sound System*, Issues in Second Language Research. Newbury House, Cambridge, MA. Originally presented at the National TESOL Convention, San Francisco, 1980.

[Bell1984] Allan Bell. 1984. Language style as audience design. *Language in Society*, 13:145–204.

[Bernstein et al.1990] Jared Bernstein, Michael Cohen, Hy Murveit, Dimitry Rtischev, and Mitchel Weintraub. 1990. Automatic evaluation and training in english pronunciation. In *Proc. ICSLP*, Kobe.

[Bratt et al.1998] Harry Bratt, Leo Neumeyer, Elizabeth Shriberg, and Horacio Franco. 1998. Collection and Detailed Transcription of a Speech Database for Development of Language Learning Technologies. In *Proc. ICSLP.*

[Brieman et al.1984] L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees.* Wadsworth, Inc.

[Brière1966] Eugene Brière. 1966. An investigation of phonological interference. *Language*, 42(4):768–796.

[Briggs1986] Charles Briggs. 1986. *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research.* Cambridge University Press, Cambridge.

[Burger et al.2000] Susanne Burger, Karl Weilhammer, Florian Schiel, and Hans G. Tillmann. 2000. Verbmobil data collection and annotation. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, pages 539–552. Springer, July.

[Byrne et al.1998] William Byrne, Eva Knodt, Sanjeev Khudanpur, and Jared Bernstein. 1998. Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English. In *Proc. Speech Technology in Language Learning (STiLL).*

[Clark and Swinton1979] John L. D. Clark and Spencer S. Swinton. 1979. An Exploration of Speaking Proficiency Measures in the TOEFL Context. TOEFL Research Report 4, Educational Testing Service.

[Corder1967] S. P. Corder. 1967. The significance of learners' errors. *International Review of Applied Linguistics*, 5(4):161–170.

[Cucchiarini et al.1998] C. Cucchiarini, H. Strik, and L. Boves. 1998. Quantitative assessment of second language learners' fluency: an automatic approach. In *Proc. ICSLP*, Sydney.

[Cucchiarini et al.2000] Catia Cucchiarini, Helmer Strik, Diana Binnenpoorte, and Lou Boves. 2000. Towards an Automatic Oral Proficiency Test for Dutch as a Second Language. In *Proc. ESCA Workshop on Incorporating Speech Technology in Language Learning (InSTIL)*, Dundee.

[Dalby et al.1998] Jonathan Dalby, Diane Kewley-Port, and Roy Sillings. 1998. Language-Specific Pronunciation Training Using the HearSay System. In *Proc. Speech Technology in Language Learning (STiLL).*

[Dickerson1974] Lonna Dickerson. 1974. *Internal and External Patterning of Phonological Variability in the Speech of Japanese Learners of English.* Ph.D. thesis, University of Illinois.

[Doh2000] Sam-Joo Doh. 2000. *Enhancements to Transformation-Based Speaker Adaptation: Principal Component and Inter-Class Maximum Likelihood Linear Regression.* Ph.D. thesis, Carnegie Mellon University.

[Domingos and Pazzani1997] Pedro Domingos and Michael Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130.

[Ehsani et al.1997] Farzad Ehsani, Jared Bernstein, Amir Najimi, and Ognjen Todic. 1997. SUBARASHII: Japanese Interactive Spoken Language Education. In *Proc. Eurospeech.*

[Eklund and Shriberg1998] Robert Eklund and Elizabeth Shriberg. 1998. Crosslinguistic Disfluency Modeling: A Comparitive Analysis of Swedish and American English Human-Human and Human-Machine Dialogs,. In *Proc. ICSLP.*

[Ellis1997] Rod Ellis. 1997. *Second Language Acquisition.* Oxford University Press.

[Eskenazi and Hansma1998] Maxine Eskenazi and Scott Hansma. 1998. The fluency pronunciation trainer. In *Proc. Speech Technology in Language Learning (STiLL).*

[Eskenazi1996] Maxine Eskenazi. 1996. Detection of foreign speakers' pronunciation errors for second language training - preliminary results. In *Proc. ICSLP.*

[Eskenazi1997] Maxine Eskenazi. 1997. Issues in Database Creation: Recording New Populations, Faster and Better Labelling. In *Proc. Eurospeech.*

[Ferguson1989] Charles A. Ferguson. 1989. Language teaching and theories of language. In James E. Alatis, editor, *Georgetown University Round Table on Languages and Linguistics 1989.* Georgetown University Press.

[Finegan1994] Edward Finegan. 1994. Sociolinguistic perspectives on register. chapter Introduction. Oxford University Press.

[Finke and Waibel1997] Michael Finke and Alex Waibel. 1997. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proc. Eurospeech*, Rhodes.

[Finke et al.1997] Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries, and Torsten Zeppenfeld. 1997. The JanusRTk Switchboard/Callhome 1997 Evaluation System. In *Proc. the LVCSR Hub5-e Workshop.*

[Flege1993] James Emil Flege. 1993. Production and perception of a novel, second-language phonetic contrast. *J. Acoust. Soc. Am.*, 93(3):1589–1608, March.

[Forney1973] G. D. Forney. 1973. The Viterbi Algorithm. *Proc. IEEE*, 61:268–278, March.

[Franco et al.1997] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen. 1997. Automatic Pronunciation Scoring for Language Instruction. In *Proc. ICASSP*, Munich.

[Fung and Liu1999] Pascale Fung and Wai Kat Liu. 1999. Fast Accent Identification and Accented Speech Recognition. In *Proc. ICASSP.*

[Garovolo et al.1997] John S. Garovolo, Jonathan G. Fiscus, and William M. Fisher. 1997. Design and Preparation of the 1996 HUB-4 Broadcast News Benchmark Test Corpora. In *Proc. DARPA Speech Recognition Workshop.*

[Garrett1995] Nina Garrett. 1995. ICALL and Second Language Acquisition. In V. M. Holland, J. D. Kaplan, and M. R. Sams, editors, *Intelligent Language Tutors: Theory Shaping Technology.* Lawrence Erlbaum, Mahwah, NJ.

[Geutner1995] Petra Geutner. 1995. Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems. In *Proc. ICASSP*, Detroit.

[Giegerich1992] Heinz J. Giegerich. 1992. *English Phonology: An introduction.* Cambridge University Press.

[Gillick and Cox1989] L. Gillick and S.J. Cox. 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proc. ICASSP.*

[Goldman-Eisler1958] Frieda Goldman-Eisler. 1958. Speech PRoduction and the Predictability of Words in Context. *Quarterly Journal of Experimental Psychology*, 10:96–106.

[Grant1999] Linda Grant. 1999. Form to meaning: Bridges in pronunciation teaching. *TESOL Matters*, 9(6).

[Huang et al.1996] X.D. Huang, Mei-Yuh Hwang, Li Jiang, and Milind Mahajan. 1996. Deleted interpolation and density sharing for continuous hidden markov models. In *Proc. ICASSP*, Atlanta.

[Huang et al.2000] Chao Huang, Eric Chang, Jianlai Zhou, and Kai-Fu Lee. 2000. Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. In *Proc. ICSLP*, Beijing.

[Humphries and Woodland1997] J. J. Humphries and P. C. Woodland. 1997. Using Accent-Specific Pronunciation Modeling for Improved Large Vocabulary Continuous Speech Recognition. In *Proc. Eurospeech*, Seattle.

[Humphries and Woodland1998] J. J. Humphries and P. C. Woodland. 1998. The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training. In *Proc. ICASSP*, Seattle.

[Hwang1993] Mei-Yuh Hwang. 1993. *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition.* Ph.D. thesis, Carnegie Mellon University.

[Imperl1999] Bojan Imperl. 1999. Clustering of context-dependent speech units for multilingual speech recognition. In *Proc. the ESCA workshop on Multi-lingual Interoperability in Speech Technology (MIST).*

[Jackson1932] J.H. Jackson. 1932. *Selected Writings, vol. II.* London.

[James1980] Carl James. 1980. *Contrastive Analysis.* Longman, London.

[James1998] Carl James. 1998. *Errors in Language Learning and Use: Exploring Error Analysis.* Addison Wesley Longman Ltd., Harlow, Essex, England.

[Kawai and Hirose1997] Goh Kawai and Keikichi Hirose. 1997. A CALL System Using Speech Recognition to Train the Pronunciation of Japanese Long Vowels, the Mora Nasal and Mora Obstruents. In *Proc. Eurospeech*, Rhodes.

[Kawai1999] Goh Kawai. 1999. *Spoken Language Processing Applied To Nonnative Language Pronunciation Learning.* Ph.D. thesis, University of Tokyo.

[Kawai2000] Hayao Kawai, Chair. 2000. Japan's Goals in the 21st Century. Technical report, Office for the Prime Minister's Commission on Japan's Goals in the 21st Century.

[Kim et al.1997] Y. Kim, H. Franco, and L. Neumeyer. 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In *Proc. Eurospeech*, Rhodes.

[Kneser and Ney1995] Reinhard Kneser and Hermann Ney. 1995. Improved Backing-off for m-gram Language Modeling. In *Proc. ICASSP*, pages 181–184.

[Köhler1999] Joachim Köhler. 1999. Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks. In *Proc. the ESCA workshop on Multi-lingual Interoperability in Speech Technology (MIST).*

[Labov1972] William Labov. 1972. *Sociolinguistic Patterns.* University of Pennsylvania Press.

[Labov1984] William Labov. 1984. Field methods of the project on linguistic change and variation. In *Language in Use: Readings in Sociolinguistics*, pages 28 – 66. Prentice-Hall.

[Lamble et al.1999] D Lamble, T Kauranen, M Laakso, and H Summala. 1999. Cognitive load and detection in thresholds in car following situations: safety implications for using mobile (cellular) telephones while driving. *Accident Analysis and Prevention*, 31:617–623.

[Langlais et al.1998] Philippe Langlais, Anne-Marie Öster, and Björn Granström. 1998. Automatic Detection of Mispronunciation in Non-native Swedish Speech. In *Proc. Speech Technology in Language Learning (STiLL).*

[LDC1994a] LDC. 1994a. http://www.ldc.upenn.edu/Catalog/docs/csr2s/readme.txt.

[LDC1994b] LDC. 1994b. http://www.ldc.upenn.edu/readme_files/timit.readme.html.

[LDC1996a] Linguistic Data Consortium, 1996a. *The 1996 Hub-4 Annotation Specification for Evaluation of Speech Recognition on Broadcast News.*

[LDC1996b] Linguistic Data Consortium, 1996b. *Transcription conventions for Spanish Callhome.*

[LDC1997] LDC. 1997. http://www.ldc.upenn.edu/Catalog/LDC97S44.html.

[Lee1990] Kai-Fu Lee. 1990. Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition. In *Proc. ICASSP.*

[Lewis1998] David Lewis. 1998. Naive (Bayes) at forty: The independence assumption. In *Proc. ECML*.

[Lippi-Green1997] Rosina Lippi-Green. 1997. *English with an Accent: Language, Ideology, and Discrimination in the United States*. Routledge, London.

[Liu and Fung2000a] Wai Kat Liu and Pascale Fung. 2000a. MLLR-based accent model adaptation without accented data. In *Proc. ICSLP*.

[Liu and Fung2000b] Yi Liu and Pascale Fung. 2000b. Modeling pronunciation variations in spontaneous mandarin speech. In *Proc. ICSLP*.

[Livescu and Glass2000] Karen Livescu and James Glass. 2000. Lexical Modeling of Non-native Speech for Automatic Speech Recognition. In *Proc. ICASSP*.

[Manning and Schütze1999] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

[Mayfield Tomokiyo and Burger1999] Laura Mayfield Tomokiyo and Susanne Burger. 1999. Eliciting Natural Speech from Non-Native users: Collecting Speech Data for LVCSR. In *Proc. the ACL-IALL Joint Workshop in Computer-Mediated Language Assessment and Evaluation in Natural Language Processing*.

[Mayfield Tomokiyo and Jones2001] Laura Mayfield Tomokiyo and Rosie Jones. 2001. Naive Bayes Detection of Non-native Utterances. In *Proc. NAACL*, Pittsburgh.

[Mayfield Tomokiyo2000] Laura Mayfield Tomokiyo. 2000. Linguistic Properties of Non-native Speech. In *Proc. ICASSP*.

[McCallum1996] Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow.

[Millar et al.1994] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody. 1994. The Australian National Database of Spoken English. In *Proc. ICASSP*.

[Mitchell1997] Tom M. Mitchell. 1997. *Machine Learning*. McGraw-Hill, Boston.

[Nakajima et al.2000] Hideharu Nakajima, Yoshinori Sagisaka, and Hirofumi Yamamoto. 2000. Pronunciation variants description using recognition error modeling with phonetic derivation hypotheses. In *Proc. ICSLP*.

[Neumeyer et al.1996] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. 1996. Automatic Text-independent Pronunciation Scoring of Foreign Language Student Speech. In *Proc. ICSLP*.

[NIST1999] NIST. 1999. http://www.nist.gov/speech/tests/bnr/bnews_99/bnews_99.spec.htm.

[NIST2000] NIST. 2000. http://www.nist.gov/speech/tools/.

[Nock and Young2000] Harriet Nock and Steve Young. 2000. Loosely coupled HMMs for ASR. In *Proc. ICSLP*.

[Oller1974] D. K. Oller. 1974. Toward a general theory of phonological processes in first and second language learning. In *Proc. the Western Conference on Linguistics*, Seattle.

[Pawley and Syder1983] Andrew Pawley and Frances Hodgetts Syder. 1983. Two Puzzles for Linguistic Theory. In J.C. Richards and R.W. Schmidt, editors, *Language and Communication*, pages 191–226. Longman, London.

[Price1998] Patti Price. 1998. How Can Speech Technology Replicate and Complement Good Language Teachers to Help People Learn Language? In *Proc. Speech Technology in Language Learning (STiLL)*.

[Quinlan1993] J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

[Rabiner1990] Lawerence R. Rabiner. 1990. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*. Morgan Kaufmann, San Mateo, California.

[Rampton1987] Ben Rampton. 1987. Stylistic variability and not speaking 'normal' english: some post-labovian approaches and their implications for the study of interlanguage. In Rod Ellis, editor, *Second Language Acquisition in Context*. Prentice-Hall.

[Ratnaparkhi1996] Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proc. Empirical Methods in Natural Language Processing Conference*.

[Ronen et al.1997] Orith Ronen, Leonardo Neumeyer, and Horacio Franco. 1997. Automatic Detection of Mispronunciation for Language Instruction. In *Proc. Eurospeech*.

[Schultz and Waibel1999] Tanja Schultz and Alex Waibel. 1999. Language Adaptive LVCSR through Polyphone Decision Tree Specialization. In *Proc. the ESCA workshop on Multi-lingual Interoperability in Speech Technology (MIST)*.

[Schwartz et al.1997] Richard Schwartz, Hubert Jin, Francis Kubala, and Spyros Matsoukas. 1997. Modeling Those F-Conditions - Or Not. In *Proc. the 1997 DARPA Speech Recognition Workshop*.

[Shriberg and Stolcke1996] Elizabeth Shriberg and Andreas Stolcke. 1996. Word Predictability after Hesitations,. In *Proc. ICSLP*.

[Soltau2001] Hagen Soltau. 2001. Personal communication.

[SPE1987] 1987. Guide to SPEAK. Produced by the Test of English as a Foreign Language Program, Princeton, NJ.

[Stolcke and Shriberg1996] Andreas Stolcke and Elizabeth Shriberg. 1996. Statistical Language Modeling for Speech Disfluencies,. In *Proc. ICASSP*.

[Suzuki et al.2000] Tadashi Suzuki, Jun Ishii, and Kunio Nakajima. 2000. A generating method of English pronunciation dictionary for Japanese-English recognition. In *Proc. ICSLP*, Beijing.

[Tajima et al.1997] K. Tajima, R. F. Port, and J. Dalby. 1997. Effects of temporal correction on intelligibility of foreign-accented English. *J. Phonetics*, 25:1–24.

[Tajima et al.2000] Keiichi Tajima, Donna Erickson, and Kyoko Nagao. 2000. Factors affecting native Japanese speakers' production of intrusive (epenthetic) vowels in English words. In *Proc. ICSLP*, Beijing.

[Tarone et al.1983] Elaine Tarone, Andrew D. Cohen, and Guy Dumas. 1983. A closer look at some interlangauge terminology: a framework for communication strategies. In Claus Færch and Gabriele Kasper, editors, *Strategies in Interlanguage Communication*, Applied Linguistics and Language Study. Longman.

[Tarone1978a] Elaine Tarone. 1978a. The phonology of interlanguage. In J.C. Richards, editor, *Understanding Second and Foreign Language Learning: Issues and Approaches*. Newbury House, Rowley, MA.

[Tarone1978b] Elaine Tarone. 1978b. The phonology of interlanguage. In Richards (Tarone, 1978a), page 78.

[Teixeira et al.1996] Carlos Teixeira, Isabel Trancoso, and António Serralheiro. 1996. Accent identification. In *Proc. International Conference on Spoken Language Processing*, Philadelphia.

[Towell and Hawkins1994] R. Towell and R. Hawkins. 1994. *Approaches to Second Language Acquisition*. Multilingual Matters, Clevedon.

[Van Dyke1997] Julie A. Van Dyke. 1997. A process model of learning definiteness in ESL. Master's thesis, Carnegie Mellon University.

[van Leeuwen and Orr1999] David A. van Leeuwen and Rosemary Orr. 1999. Speech Recognition of Non-native Speech Using Native and Non-native Acoustic Models. In *Proc. the ESCA workshop on Multi-lingual Interoperability in Speech Technology (MIST)*.

[Ward1990] Wayne Ward. 1990. The CMU Air Travel Information Service: Understanding Spontaneous Speech. In *Proc. DARPA Speech and Natural Language Understanding Workshop*.

[Wardhaugh1970] Ronald Wardhaugh. 1970. The contrastive analysis hypothesis. *TESOL Quarterly*, 4(2):281–95. Cited in (James 1998).

[Wardhaugh1998] Ronald Wardhaugh. 1998. *Sociolinguistics*. Blackwell Publishers, Oxford.

[Witt and Young1997] Silke Witt and Steve Young. 1997. Language Learning Based on Non-Native Speech Recognition. In *Proc. Eurospeech*, Rhodes.

[Witt and Young1999] Silke Witt and Steve Young. 1999. Offline Acoustic Modeling of Non-native Accents. In *Proc. Eurospeech*.

[Wolfson1976] Nessa Wolfson. 1976. Speech Events and Natural Speech: Some Implications for Sociolinguistic Methodology. *Language in Society*, 5:188 – 209.

[Woodland1999] P.C. Woodland. 1999. Speaker adaptation: Techniques and challenges. In *Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Keystone, Colorado.

[Yang and Liu1999] Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *22nd Annual International SIGIR*, pages 42–49, Berkeley, August.