DISSERTATION SUMMARY

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

Laura Mayfield Tomokiyo

May 2, 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% (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.

1 Introduction

Not so many years ago, dialogues like the following were the target of intensive basic research in speech recognition and natural language understanding.

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 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 both an interesting research problem and 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 because speech recognition of new learners is not reliable.

It seems clear that native speakers are able to identify non-native speakers based on features like accent, syntax, and fluency. Children can pinpoint 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 the 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 used in this dissertation, this is an extremely fine-grained representation. Acoustic events are modeled on a sub-phonetic level, and tens of times as many variations are recognized as would be in a traditional phonetic analysis. 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 simplification 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 success for 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 fit together 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, representing explicitly 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 lexical model.

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 also 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 are 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 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.

2 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. 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.

Although 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, we have seen few detailed studies contrasting native and non-native speech patterns with respect to features that are important to LVCSR. In this chapter, I provide such an analysis, describing the differences between the native and non-native speech samples I have collected and the methods used to quantify them. This analysis is important for speech recognition, but has implications for other areas of natural language processing such as parsing and discourse processing.

2.1 Database composition

Spoken data was collected primarily from native speakers of Japanese, with a few native speakers of Mandarin included for comparison. English proficiency of all speakers was evaluated using 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). Using recruitment and elicitation strategies that were refined during pilot data collection experiments, a total of 58 native Japanese, 8 native Mandarin, and 10 native English speakers were recorded completing read and spontaneous tasks in English. The read task involved reading aloud from the Children's News Database (CND), a collection of news articles written for children covering current events in the years 1998-2000. This database was selected over more common databases such as Wall Street Journal because of the extreme difficulty speakers had in reading adult-oriented texts. In the spontaneous task, speakers were prompted for queries in the tourist domain. In addition, speakers read from the story of Snow White. A breakdown of this database is given in Table 1.

This dissertation focuses on read speech and lower-proficiency speakers. For the recognition experiments that will be described in later sections, a test set of 10 native Japanese speakers was defined meeting these criteria. The analysis presented in this section, however, covers all speakers, of both higher and lower proficiency, and in read and spontaneous tasks.

Native language	Prompted		Story		News (CND)	
	# 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 1: General information about the non-native speech database

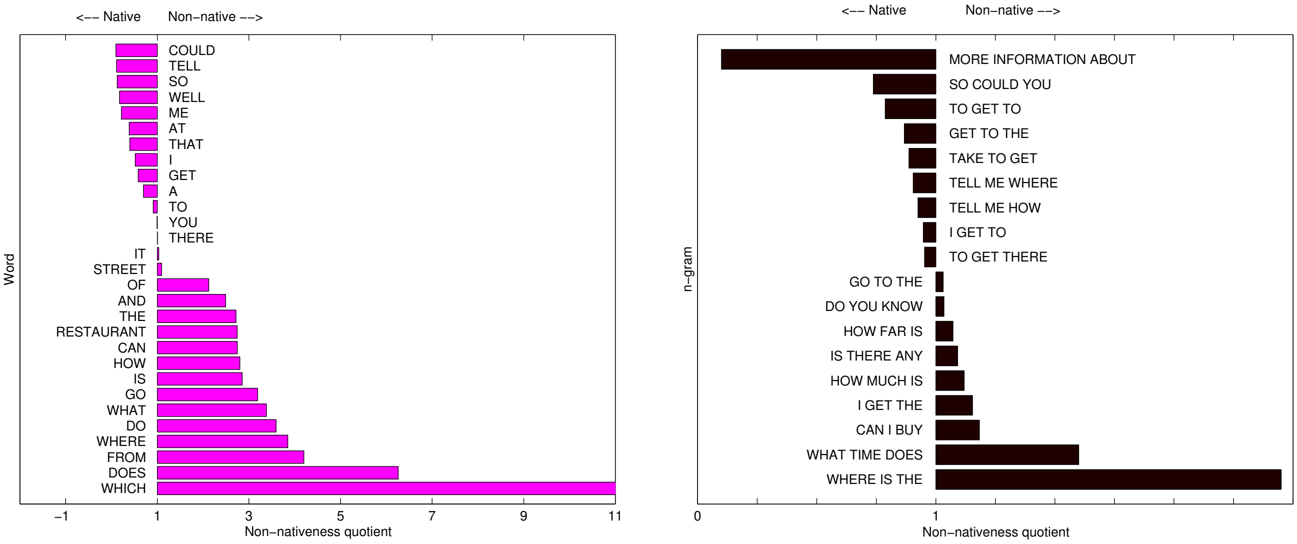

Figure 1: Non-nativeness quotients for words and trigrams

2.2 Data characterization

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.

2.2.1 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. Figure 1 illustrates how different the *non-nativeness quotients* (NNQ) of a number of words and n-grams are in spontaneous speech.

$$\text{NNQ} = \frac{\text{normalized frequency of token in non-native speech}}{\text{normalized frequency of token in native speech}}$$

Perplexity and divergence are two other statistical measures which can be used to distinguish native and non-native spontaneous speech. Perplexity encodes the predictability of an utterance or corpus. The perplexity of the non-native spontaneous speech in this database is generally lower than that of native speech when measured with respect to a large native language model. Perplexity calculated in this way does not tell us, however, the extent to which non-native speakers differ from each other; furthermore, because perplexity is an averaged term, it does not tell us whether the text is uniformly predictable or contains regions of high suprise balanced by regions of low surprise. These questions can be addressed by measuring KL divergence. Non-native divergence from other non-native speakers is greater than native divergence from other native speakers for word unigrams, word trigrams, and part-of-speech trigrams, as shown in Table 2.

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 2: Kullback-Liebler divergence (relative entropy) of word and part-of-speech n-gram probabilities between native and non-native speaker corpora

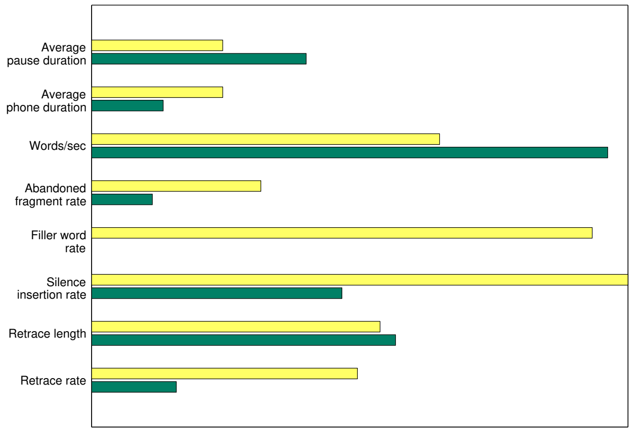

Figure 2: Disfluency rates for native and non-native (JL1) speakers in the CND reading task. This chart shows proportional differences between native and non-native speakers in different categories; the scale is not uniform across categories.

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: Breakdown of non-native and native misread words

2.2.2 Fluency and disfluencies

Features describing the smoothness and pace of speech can also be used to characterize non-native spoken language. Non-natives in general speak more slowly than natives do, with more pauses between words, more word fragments and repetitions, and more repairs. Differences between native and non-native read speech are illustrated in Figure 2.

2.2.3 Reading errors

In read speech, reading errors are of greater significance for non-native speakers than for native speakers. In the CND database, native speakers averaged .39 errors (substitutions, insertions, or deletions of words) per 100 words; the native Japanese speakers averaged 2.77 errors. A breakdown of the types of reading error is given in Table 3.

2.2.4 Other measurable features

This section has only described a few of the characteristics of non-native speech that are quantifiable and relevant for speech recognition. Others include vocabulary growth in spontaneous speech (faster for *non-*native speakers, both individually and in the aggregate), use of contractions (distributed differently for native and non-native speakers), and human-judged "nativeness" (native speakers identify non-native transcripts with high precision).

3 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,

	BN	CND native	CND non-native
Without adaptation	7.4	31.3	92.0
With adaptation	6.4	18.1	63.1

Table 4: Baseline WER on Broadcast News (BN) native, CND native, and CND non-native speech

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.

This section explores 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.

3.1 Baseline system

All recognition experiments described in the dissertation used the Janus Recognition Toolkit JRTk (Finke et al., 1997). Recognition experiments are done exclusively on the CND read speech database. 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 a subset of the CND archive for cross-validation. Performance of this system, with and without speaker adaptation, is given in Table 4, which highlights the need for special consideration of non-native speech. All recognition experiments assume that individual speaker adaptation is applied unless otherwise specified; all references to baseline performance in this chapter refer to the 63.1% value given in Table 4.

3.2 Adaptation

Experiments in adaptation focused on comparing L1 and accented L2 adaptation material. In other words, I sought to establish whether, for a native speaker of Japanese, Japanese-language (L1) or Japanese-accented English (L2) acoustic data allows better adaptation. Collecting accented L2 data is problematic because a new set of data must be collected for every L1-L2 pair; if it were to be found that L1 data is acceptable for adaptation the data collection process would be streamlined, as acoustic data already exists for many L1s. However, to use L1 data for training and adaptation of an L2 system, a mapping between L1 and L2 phonetic units must be defined. Whether this mapping is based on linguistic information about L1 and L2, the modeling of L2 in the recognizer, or a combination, it assumes a direct, invariant influence of L1 on L2 articulation that does not reflect the complex nature of adult phonological acquisition.

Figures 3 and 4 show performance of MAP and MLLR adaptation algorithms using L1 and L2 data. For both algorithms, using L1 data for adaptation results in a performance degradation while L2 data provides an improvement, both of which increase with the amount of adaptation data.

3.3 Training

Two types of training were explored: full rebuilding of the system, with new labels, LDA, and allophonic decision tree, with L1 and accented L2 data *added* to the native data used to train the baseline system; and additional training iterations using accented L2 data to update the model means.

Figure 5 shows the change in word error rate (WER) when the system is rebuilt with L1 and accented L2 data. Although we do not see the degradation with L1 data that we did with adaptation, there is no significant change in overall WER when L1 data is used for system development. L2 data, on the other hand, contributes to a significant improvement in system performance. Figure 5 shows performance of individual speakers as well as overall performance, and we can see that some speakers are more positively affected by the retraining than others. In no case, however, did an L1-trained system perform better than an L2-trained system.

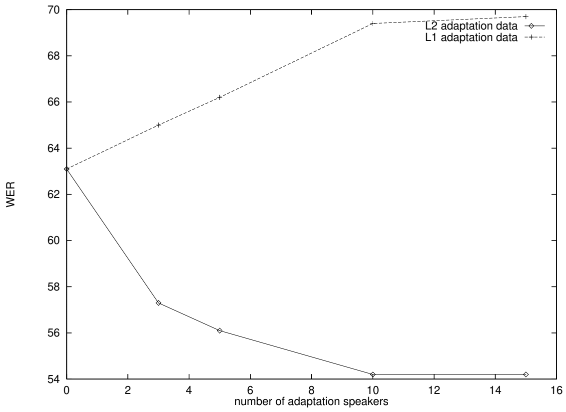

Figure 3: MLLR adaptation with varying numbers of adaptation speakers

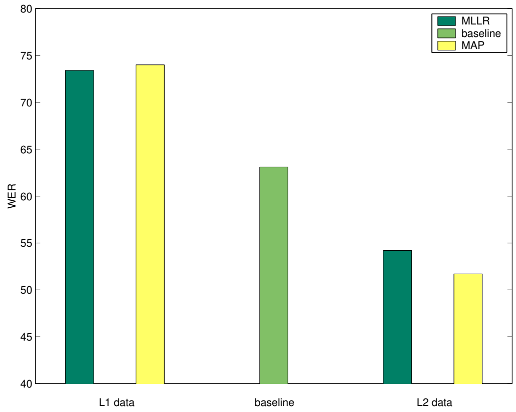

Figure 4: Comparison of MLLR and MAP adaptation with 15 adaptation speakers

Figure 6 shows the improvement in WER when the new data is incorporated only in the final stage of training. Specifically, after 7 forward-backward training iterations with the baseline native data, an additional 2 forward-backward training iterations were run with 3 hours of accented L2 acoustic data. In this experiment, the retrained models were interpolated with the baseline models in order to avoid overtraining to the training speakers. The curve in Figure 6 represents the change in performance with varying interpolation weights. On the right side of the curve, we have the fully specialized models, retrained with only accented L2 data. On the left side, we have the baseline models, performing at 63.1% as described in Section 3.1. The slight overtraining effect is apparent in this figure, as the optimal performance comes with an interpolation weight of 0.72, backing off slightly from the fully specialized models.

3.4 Other acoustic modeling results

A number of results from the dissertation have not been discussed in this summary. Context-dependent models were found to perform better than context-independent models for this low-proficiency LVCSR task. Optimal language model settings for native and non-native speakers are significantly different. Phonetic confusion occurs in the same pairs as in native speech, but with greater magnitude. Polyphone decision tree specialization (PDTS), which has been successful in adapting to new languages, was not found to be as effective for adapting to new accents.

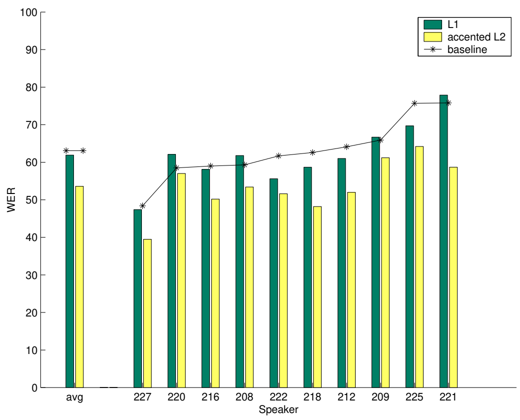

Figure 5: System performance after rebuilding the system with L2 and accented L2 data

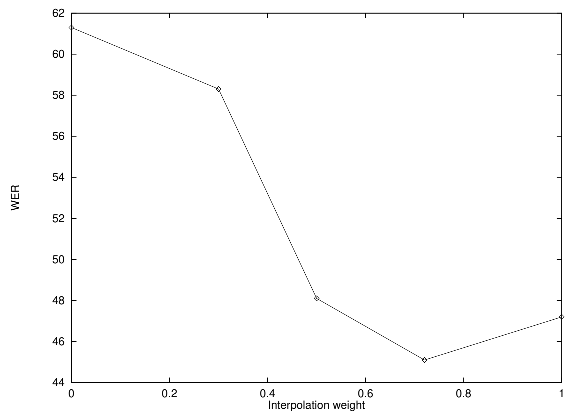

Figure 6: System performance after interpolating models retrained on L2 data with the baseline models

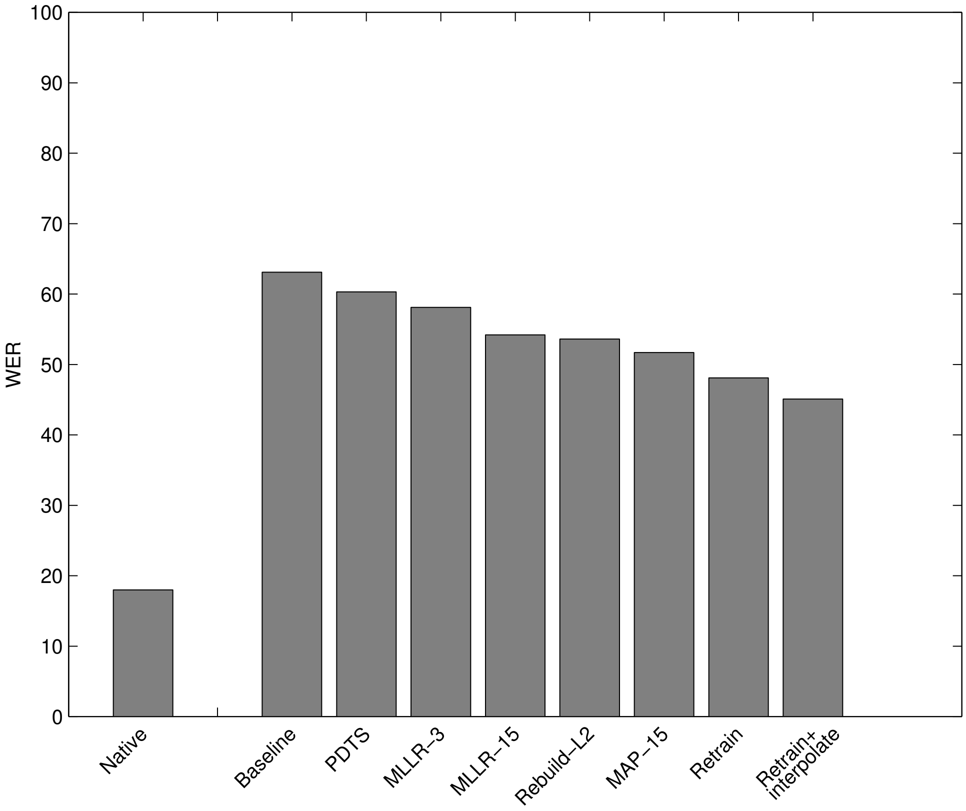

Figure 7: Summary of acoustic modeling results

3.5 Acoustic modeling summary

This section has highlighted 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 7.

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.

4 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

[1]Throughout this section, I will use the term *substitute* to refer to replacement of one phone sequence with another, subsuming the insertion case (substituting two or more phones for a single phone) and the deletion case (substituting one phone for two or more).

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; 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 and others, 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.

This section compares data-driven and linguistically-motivated methods for finding probable phonemic representations of English words in Japanese-accented speech.

4.1 Lexical specification and prioritization

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. Once the transformations have been specified, they must be prioritized, as the search cannot handle an unlimited number of variants for each lexicon entry.

4.1.1 Linguistically-motivated variant prediction

For linguistically-motivated prediction, a list of context-dependent transformation rules was developed based on observations from ESL for Japanese natives and information about the phonological structure of Japanese. Applying all of these rules to the lexicon leads to generation of 915,672 pronunciation variants. To prune this list, the acoustic model training data was aligned to its transcripts using the expanded lexicon; the variants that were selected as the best acoustic match during alignment were analyzed to determine what transformation rules and direct substitutions would be incorporated in the test lexicon. The following lexica were defined:

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

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

R1 Transformation rules that occurred 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.

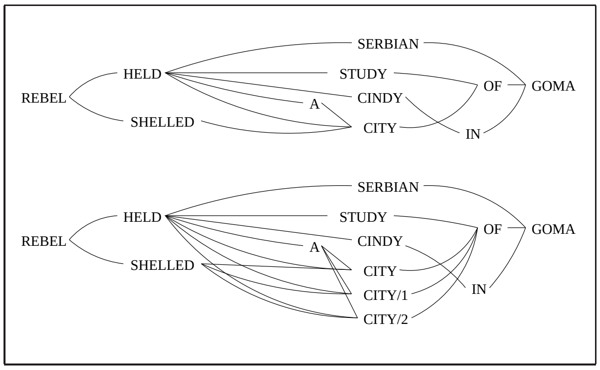

Figure 8: Lattice segment for the text "THE REBEL HELD CITY OF GOMA" before (top) and after (bottom) adding pronunciation variants for the word "CITY"

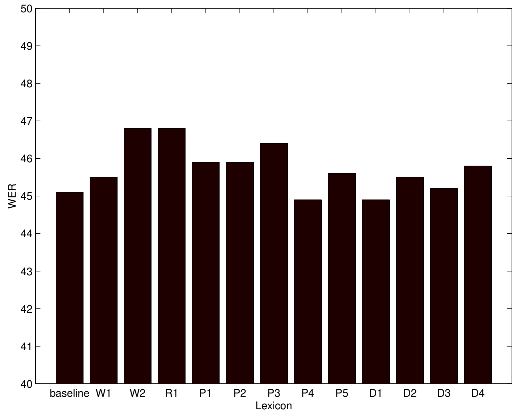

Figure 9: Lexical modeling results

4.1.2 Data-driven variant prediction

In data-driven modeling experiments, the alignment to potential variants was bootstrapped via phoneme recognition, rather than information from ESL and Japanese phonology. Potential transformations were then prioritized to create the following test lexica:

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

5 Testing and Results

To test these lexica, the *lattice adaptation* method was used. In lattice adaptation, a word lattice is first created using the baseline lexicon. For each word node in the lattice, a new node is added with the same links and score for each variant of that word in the expanded lexicon. An acoustic rescoring pass is then run on the new lattice. This method allows incorporation of a larger number of relevant variants than full decoding with an expanded lexicon would, because the search space is greatly constrained before the introduction of confusable variants. Lattice adaptation is illustrated in Figure 8, which 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.

Lexical modeling experiments were run with the best acoustic models yielded by acoustic modeling experiments, namely those retrained with accented L2 data. The baseline accuracy for lexical modeling experiments is therefore 45.1% WER. Although initial experiments on suboptimal acoustic models had suggested that lexical modeling would be successful, for the final, acoustically-optimized system, neither linguistically-motivated nor data-driven approaches to lexical modeling resulted in a significant decrease in WER. Results are shown in Figure 9.

5.1 Discussion

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 physical competence.

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 et al., 2000), which observed only small gains from lexical modeling for proficient non-native speech, 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 the weather query system described in (Livescu and Glass, 2000), speakers approach the system with something they want to know, and can rely on words and fixed phrases that are familiar to them. Isolated phone recognition, as discussed in (Fung and Liu, 1999), is a very simple task that lends itself to modeling at the segmental level because speech production is also segmental. Read speech, while only mildly affected by a speaker's command of syntax and semantics of the language, may not be a strong candidate for either rule-based or data-driven modeling at the lexical level.

6 Speech 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. This section summarizes a high-accuracy nativeness classification method that improves overall system performance significantly.

6.1 Introduction

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

Condition	Word		POS		Stopwords		Phone	Phone class	Word+phone
	trans	hypo	trans	hypo	trans	hypo	hypo	hypo	hypo
A	83%	94%	74%	100%	49%	99%	100%	86%	100%
B	41	47	40	77	43	89	92	80	86
C	56	56	56	95	56	56	88	71	41
D	56	56	56	83	56	56	76	82	47

Table 5: Word- and phone-based classification accuracy of read speech. Conditions A-D are as defined in Section 6.2. Baseline accuracy (achieved by always guessing the most common class) is 56% .

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.

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.

6.2 Experiments

Classification experiments were run on both hypotheses and transcripts; in NLP systems with a recognition component, performance on transcripts is often used for individual evaluation of other components as interference from recognition error can be extreme. Word-based, phone-based, and part-of-speech-based classification of both read and spontaneous speech was explored, although only read speech classification is discussed in this summary.

The publicly available Rainbow text classification package (McCallum, 1996) was used for the naive Bayes engine. Stopwords were included, as they were found to be effective predictors of nativeness; in some experiments *only* stopwords on the long SMART (Buckley, 1985) list were used as the vocabulary for classification. Part-of-speech tags were generated using the MXPOST (Ratnaparkhi, 1996) toolkit. The recognition system used to produce the original hypotheses was the 63.1% baseline CND system; non-native utterances were re-recognized using the best-performing interpolated model system described in Section 3.

In evaluating classifier performance, four test conditions were defined:

(A) train on shared article, test on shared article

(B) train on disjoint articles, test on disjoint articles

(C) train on disjoint articles, test on shared article

(D) train on shared article, test on disjoint articles

Each of these conditions has a real-world application which is not discussed in this summary.

6.3 Classification results

Table 5 shows results of classification on the four conditions described above. Classification results are always higher for the part-of-speech-tagged hypotheses than the word hypotheses. Restricting the vocabulary to stopwords greatly improves performance for matched-condition classification but not for the mixed conditions.

As can be seen from Table 5, classification of hypotheses is *more* accurate than classification of transcripts. Part-of-speech- and phone-based classification are the strongest performers, achieving high accuracy in all four conditions.

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

Table 6: Overall recognizer performance when L1 classification is used to switch to non-native acoustic models

6.4 Accent-dependent recognition

With reliable accent discrimination, we can combine standard recognition with recently proposed techniques for adapting to non-native speech to run on-the-fly accent-dependent recognition (Mayfield Tomokiyo, 2000; Fung and Liu, 1999), e.g. Ideally, in such a system we would like to use disjoint sets of utterances for classifier training and testing, so we will use the word-based classification with the stopword vocabulary, which achieved strong performance for this condition.

Table 6 shows how recognizer performance is improved when utterances identified as non-native by our classifier are re-recognized with customized acoustic models. Performance with hypothesis-driven model switching approaches the "gold standard" result, which represents recognition with perfect classification.

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 section are summarized major results and contributions and discuss promising directions for extensions of this work.

7.1 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 in performance between 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 bilateral and multilateral classification of both recognizer hypotheses and transcriptions, and provides further insight into the nature of non-native speech.

7.2 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.

7.2.1 Allophonic modeling

Although the implementation of allophone tree adaptation discussed in Section 3 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.

7.2.2 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.

7.2.3 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 discussed in the dissertation. 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.

7.2.4 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.

References

[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 and others1998] Stefan Auberg et al. 1998. The Accent Coach: An English Pronunciation Training System for Japanese Speakers. In *Proc. Speech Technology in Language Learning (STiLL)*.

[Buckley1985] C. Buckley. 1985. Implemetation of the Smart Information Retrieval Retrieval System. Technical Report 85-686, Cornell University.

[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.

[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*.

[Fung and Liu1999] Pascale Fung and Wai Kat Liu. 1999. Fast Accent Identification and Accented Speech Recognition. In *Proc. ICASSP*.

[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.

[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.

[Lewis1998] David Lewis. 1998. Naive (Bayes) at forty: The independence assumption. In *Proc. ECML*.

[Livescu and Glass2000] Karen Livescu and James Glass. 2000. Lexical Modeling of Non-native Speech for Automatic Speech Recognition. In *Proc. ICASSP*.

[Mayfield Tomokiyo2000] Laura Mayfield Tomokiyo. 2000. Lexical and Acoustic Modeling of Non-native Speech in LVCSR. In *Proc. ICSLP*.

[McCallum1996] Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow.

[Ratnaparkhi1996] Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proc. Empirical Methods in Natural Language Processing Conference*.

[SPE1987] 1987. Guide to SPEAK. Produced by the Test of English as a Foreign Language Program, Princeton, NJ.

[Teixeira et al.1996] Carlos Teixeira, Isabel Trancoso, and António Serralheiro. 1996. Accent identification. In *Proc. International Conference on Spoken Language Processing*, Philadelphia.

[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.