Handling Non-native Speech in LVCSR: A Preliminary Study

Laura Mayfield Tomokiyo

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
laura@cs.cmu.edu

Abstract

In moving towards full incorporation of CSR in applications whose users include non-native speakers, an understanding of how the system can be modified to increase its tolerance to non-native idiosyncrasies such as accented pronunciation and disfluent form is essential. While experiments geared towards restricted-use systems have suggested that extremely simple techniques are effective, preliminary analyses of LVCSR performance on low-proficiency users indicate that there is room for improvement. In this paper, we report on efforts to increase the robustness of an English LVCSR system with respect to errors common to native speakers of Japanese, using automated techniques that can be easily extended to other native speaker groups and require little acoustic training data.

1. Introduction

In large-vocabulary continuous speech recognition (LVCSR), speakers have no restrictions on the vocabulary or sentence style that they use. They are expected to speak at a normal pace without exaggerated enunciation. If the system is to be used exclusively in the context of a particular topic or speaking style, it is often optimized for peak performance over those features, but in principle, an LVCSR system places no restrictions whatsoever on the speaker's use of language.

In the absence of restrictions that can be used to build a rule-based model of what the speaker is likely to say and how he is likely to say it, a speech recognizer must rely entirely on statistical information about acoustic and linguistic properties of speech. Statistical models require an enormous amount of training data, and are finely tuned to patterns that are present in that training data. If the speech to be recognized is different in any way from the speech used for training, the user must be prepared either to accept a degradation in performance or to adapt the models to the target speech.

When the differences are very systematic, as is the case with a channel or gender mismatch, it is often possible to shift the models toward the condition with a small amount of adaptation data. For non-native speakers, however, particularly the lower-proficiency speakers that are likely users of speech-driven language learning applications, there are many types of variation, some of which may not be at all systematic, to consider.

In this paper, we first give some background by highlighting the problems of non-native LVCSR and comparing recent results from several systems on different types of non-native speech. With this in mind, we describe the task and speakers that are the target of our research. Third, we discuss some of the features of the speech in this database, particularly reading errors. Finally, we present experiments we have completed in adapting LVCSR to accented speech without acoustic training data.

2. Background

Performance of LVCSR on non-native speech is highly dependent on the domain, type of speech and speaker proficiency, as well as on the actual system specifications. The first two factors affect native speakers as well; the same system that reaches recognition accuracies of 95% on read news may drop to 65% on conversational speech. English proficiency, however, is a less important variable for native speakers.

Some of the more obvious challenges non-native speech presents for speech recognition are accented pronunciation, pronunciation errors due to unfamiliarity with a word, errors in syntax, and syntax that is unusual but not incorrect. Other difficulties include overemphasis of word boundaries (many LVCSR systems expect specific cross-articulatory behavior at word boundaries) and failure to produce appropriate allophones in context.

If a speaker is concentrating on what the best translation for a word is, whether his grammar is correct, or where his tongue is supposed to be, he has fewer cognitive cycles to apply to the task of speaking. This can affect the natural regularity of any deviations from native speech, and the most effective techniques being developed in LVCSR (speaker adaptation, statistical language modeling, pronunciation modeling) depend on regularity.

Table 1 on the next page shows a sampling of recent results on non-native speech from a variety of sites, including our own. All show large differences in performance between native and non-native speakers. These figures are not given to compare individual systems; they are merely meant to illustrate how wide the gap between native and non-native speakers can be and how much domain and speech type can affect performance.

System A is the best-performing system combination from the 1999 DARPA Broadcast News evaluation. Native speakers evaluated in system A are professional news readers and the non-native sample includes native speakers of British and Indian English as well as highly fluent professional English-language broadcasters whose native language is not English. Systems B and C are also read news systems, but the speakers are ordinary people who were asked to read a segment of news text. Speakers in system B [1] were native speakers of Dutch, a language which is phonetically similar to English and whose speakers are generally able to attain a high level of English proficiency. System C is our system. Speakers were native speakers of Japanese and tested at low-to-mid proficiency (see Section 3.1.1). Systems D and E are not read news systems. System D [2] is a conversational system and system E [3] is a telephone information system, which speakers can call to access weather information. Non-native speakers in system E were not evaluated for proficiency.

Table 1: Recognition performance (WER = word error rate) on native (N) and non-native (NN) speakers for different systems and speech characteristics.

System	Mode/domain	Speaker proficiency	WER	
			N	NN
A	Read news	Very high	8.0	14.0
B	Read news	High	22.4	68.8
C	Read news	Low	18.1	52.7
D	Conversation	Low	36.0	66.6
E	Weather query	Mixed	11.7	26.6

3. Experimental Context

3.1 Speaker characteristics

The research described in this paper focuses on native speakers of Japanese who have had substantial instruction in school but little exposure to conversational English.

3.1.1 Non-native speakers

A total of 50 speakers were recorded and evaluated for English proficiency. Evaluations were conducted by trained raters following the guidelines of the Educational Testing Service SPEAK (*Speaking Proficiency English Assessment Kit*) testing procedures [4,5]. The full SPEAK test, developed as a part of the TOEFL program, rates speakers for fluency, grammar, pronunciation and overall comprehensibility in a variety of tasks. For our proficiency assessments, only the read portion of the SPEAK evaluations were used.

Speakers were each evaluated by two raters, each at three different points in the reading. Raters assigned scores for pronunciation, fluency, and overall comprehensibility for each speaker on each passage, for a total of nine scores per speaker per rater. The SPEAK guidelines suggest that a difference in scores of greater than 0.95 represents an unacceptable level of inter-coder agreement, and speakers should be evaluated by a third rater when such disagreement occurs. In our data, the maximum individual score discrepancy between rater scores was 0.67; the average discrepancy across all categories was 0.14. A third rater was therefore not used.

Of the 50 speakers, a subset of ten was chosen for a test set. The test speakers were selected based on proficiency score (in order to minimize variability within the test set, speakers with similar proficiency scores were chosen), and the test set was gender-balanced. Recognizer performance on the test speakers was not known at the time of data set partitioning.

Table 2 shows the average scores across all categories given by the two raters for each speaker, along with the average of the two rater scores.

Table 2: Proficiency scores assigned to each speaker by the two raters R1 and R2, and averages

Speaker	R1	R2	Average
fho221	2.0	1.67	1.83
fmh227	2.0	1.78	1.89
fmo222	2.0	2.33	2.17
fnh208	2.0	1.89	1.95
fri218	2.0	2.0	2.0
mrh216	1.89	2.0	1.95
mso220	2.22	2.0	2.11
mst225	2.0	1.78	1.89
mtt212	2.0	2.0	2.0
mty209	1.89	2.0	1.95

3.1.2 Native speakers

Ten native speakers were also recorded under the same conditions. All comparisons made in this paper to recognizer performance on native speakers refer to this data set.

3.2 Task

Both spontaneous and read speech samples were collected. This section describes primarily the read portion of the data, which the experiments presented in this paper target; a detailed discussion of the spontaneous portion can be found in [6].

Speakers each read three articles from a children's news text database. Initially, we had hoped to have speakers read from Wall Street Journal (WSJ) articles, so that performance could be directly related to that well-known task. However, it was our experience that speakers found the WSJ articles too difficult to read. In pilot tests, read WSJ speech contained many reading errors and disfluencies. The number of unfamiliar words was high, making it difficult to separate pronunciation errors stemming from the foreign accent from pronunciation errors due to lack of familiarity with the lexical item. Speakers also tired quickly and were often not willing to read the full set of articles. The children's news text is similar in domain to the WSJ text, but has a reduced vocabulary and simpler grammatical constructions.

Recording was done in a quiet room with a close-talking microphone. Speakers were alone while recording. Speakers were allowed to look over the text before recording, but were encouraged to read without worrying about whether they were familiar with the meaning and pronunciation of all the words; they were instructed to read unfamiliar words as best they could. Speakers were told that it was not necessary to re-read a passage in which they had made a reading error, but that they may do so if they prefer.

4. Non-acoustic Characteristics of Non-native Speech

Non-native read speech contains errors that are not seen in native read speech. Both native and non-native speakers stumble over words, but there is a much higher incidence of disfluent reading in non-native speech, and unlike native speech, non-native speech contains more disfluencies in read than in spontaneous form.

4.1 Read vs. spontaneous

Read speech is thought to be easier to recognize than spontaneous speech because its syntax is more regular and disfluencies are less frequent. Systems that match speaker utterances to a prepared text can make do with a low-performing recognizer because, as long as the speaker really says what he is supposed to, the recognizer's task is greatly simplified. These assumptions often play a role in the design of systems such as language tutors that use speech recognition.

For non-native speakers, however, these assumptions do not necessarily hold. The incidence of disfluencies can be much higher in read speech than spontaneous speech for non-native speakers [6]. In addition, pronunciation quality can degrade when the speaker is using words and structures that a native speaker has chosen rather than ones he is familiar with.

4.2 Reading errors

While reading errors do occur in native speech, in our database they were usually either semantically appropriate or quickly caught and repaired. Of the 21,958 words in the entire native database, 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 a 67,669-word 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. Numbers for native speakers are shown for reference, but the number of actual reading errors was so small that differences in error distribution should not be considered significant.

Table 3. Breakdown of non-native and native misread words

Error type	Non-native	Native
Morphological variant	55.74%	21.27%
Orthographically similar	27.76	48.93
Semantically similar	0	8.51
Misread numeral	3.91	2.1
Neologism	3.42	0
Function word substitution	3.32	12.5
A-the	1.66	6.25
Other	4.20	0
Total number of errors	1555	47

Although native and non-native reading errors fell into the same 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.

Native text Doctors are studying the pill's *effect* on patients

Native misread Doctors are studying the pill's *effects* on patients

Non-native text *American students* perform poorly on standardized tests

Non-native misread *American student* perform poorly on standardized tests

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.

Native text *As* if that task were not challenging enough...

Native misread *And* if that task were not challenging enough...

Non-native text The amount of time students spend *on* homework is increasing

Non-native misread The amount of time students spend *as* homework is increasing

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.

Native text The politics of the *region* have always been unstable

Native misread The politics of the *religion* have always been unstable

Non-native text Environmentalists oppose construction of the Three *Gorges* Dam

Non-native misread Environmentalists oppose construction of the Three *George* Dam

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

Native text Tremendous change is anticipated over the next *few* years

Native misread Tremendous change is anticipated over the next *several* years

In errors which we term 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 native sample.

Non-native text But rain is nothing new for *Northwesterners*

Non-native misread But rain is nothing new for *Northwesterns*

Non-native text The diamonds sat *glittering* in the sand

Non-native misread The diamonds sat *glitting* in the sand

5. Handling Non-native Pronunciation in Recognition

This section describes experiments in lexical and lattice adaptation. In all recognition experiments, the JANUS speech recognizer was used with a configuration similar to that described in [8]. The standard JANUS Broadcast News language model was interpolated with two smaller language models built from collections of children's news and children's stories. Context-dependent acoustic models were used, adapted to the speaker with 50 adaptation utterances.

5.1 Lexical adaptation

One straightforward way to allow pronunciations typical of specific non-native speaker groups is to add pronunciation variants to the dictionary or pronunciation networks. This technique has been used in [9], among others. We experimented with several methods of deriving pronunciation variants to see which, if any, result in an improvement in recognition accuracy.

5.1.1 Selection of pronunciation variants

Four methods were used to predict pronunciation variants for Japanese-accented English. In the first, a rule-based mapping of phonemes was applied. This mapping operated primarily on vowels which do not occur in Japanese, such as /æ/ (mapped to /a/), and /ɪ/ (mapped to /i/). Some consonants, including /r/ and /θ/, were also mapped. In the second method, consonant cluster simplification and word-final epenthesis were added to the mapping process. Both of these methods rely on the intuition of transcribers familiar with the idiosyncrasies of Japanese-accented English to provide the most reasonable mapping. In the third method, the recognizer was used to produce mappings. For every phoneme in the training data, the recognizer was provided with a limited set of possible phonemes and asked for the one with the highest acoustic score. For example, for a segment in which the speaker was attempting to say [ɛ], the recognizer was allowed to choose between all vowel sounds. The most probable mappings for each phoneme were used, with a maximum-likelihood probability threshold determined experimentally. In the fourth method, transcribers experienced in phonetic transcription hand-coded the actual pronunciations of each occurrence of the 100 most frequent words in the training data. The most common variants for each word were included in the dictionary. This method differs from the first three in that the mappings were not phone-based, but rather word-based. The hand-coded pronunciations were meant to represent the best possible result that can be attained through automatic pronunciation variant prediction.

5.1.2 Experimental results

Table 4 shows word error rates for each of the different dictionaries.

Table 4: Recognizer performance with lexical adaptation

Variant type	WER
Baseline	52.7
Phone mappings	53.4
Cluster breakup	53.4
Recognizer-driven	54.4
Hand-coded	53.1

The difficulty in incorporating lexical adaptation in an LVCSR system is that adding "accented" pronunciation sequences often introduces phonetic sequences that do not occur in native speech and that the recognizer was not exposed to during training. If context-independent models are used, this is not a problem, and context-independent models have been reported to perform satisfactorily in some situations [10]. It has been our experience, however, that for read news, unless optimization procedures such as bucket-box intersection are also applied, context-dependent models outperform context-independent models even when lexical adaptation is used to boost the performance of the context-independent models.

The results of lexical adaptation experiments reflect this. Some variants did consistently get recognized better than their canonical counterparts did (adding [lo] as a pronunciation variant of [lɔ] for the word *law*, for example, resulted in much improved recognition of this particular word). However, overall, there was a consistent degradation observed in recognition accuracy.

5.2 Lattice adaptation

One of the disadvantages of dictionary-level adaptation is that adding pronunciation variants to the dictionary increases confusability in the search. *Lattice adaptation* is a technique that uses a word transition lattice to constrain the search space before pronunciation variants are added. The new lattice is then rescored at the acoustic level.

Figure 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 form [sɪɾi]), links for the two pronunciation variants CITY/1 [siti] and CITY/2 [ʃiti] are added.

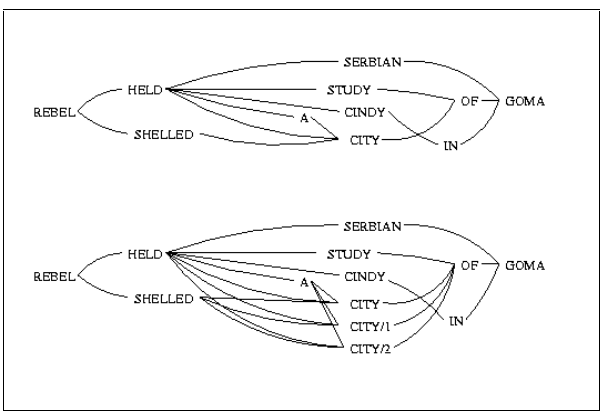

Figure 1: Lattice segment for the text "THE REBEL HELD CITY OF GOMA" before (top) and after (bottom) adding pronunciation variants for the word "CITY"

It was determined experimentally that adding new links results in better performance than replacing links.

5.2.1 Selection of pronunciation variants

Three sets of pronunciation variants were tested for lattice adaptation. The first two, representing major phonetic shifts and consonant cluster simplification, were the same as those used for lexical adaptation experiments. The third made use of a Japanese dictionary. Many English words are used in everyday Japanese, and online Japanese-language dictionaries contain accepted Japanese pronunciations for them. It is possible that these false friends are the ones that are the most often mispronounced, and using only this set results in a smaller lattice and smaller search space.

5.2.2 Experimental results

Word error rates after acoustic rescoring on an adapted lattice are shown in Table 5. The baseline here is different from that in Table 4 because the language model parameters, which are determined on a cross-validation set, were not the same.

Table 5: Recognizer performance with lattice adaptation

Variant type	WER
Baseline	55.7
Phone mappings	52.8
Cluster breakup	52.5
Japanese dictionary	52.7

The difference between using different methods of producing pronunciation variants is negligible, and the improvement from lattice adaptation over the baseline, when measured with a t-test without taking into account speaker information, is not significant. The variance between speakers is just too great. However, if a matched-pairs test is used to incorporate speaker information, the improvement with lattice adaptation becomes very significant ($p < 0.001$). This indicates that for some speakers, lattice adaptation is consistently working well, while for others it is not.

Figure 2 shows in detail how improvement from lattice adaptation varies by speaker. The four lines represent the four different types of pronunciation variant; they are connected not because of any temporal relationship but rather to show where the lines representing the "best dictionary" cross in the speaker space.

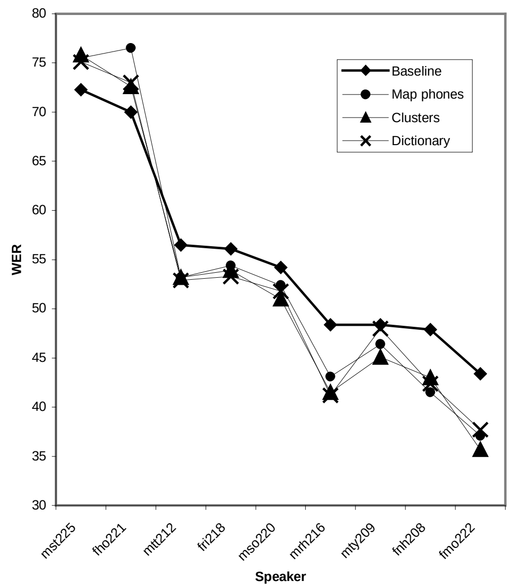

Figure 2: Lattice adaptation results broken down by speaker

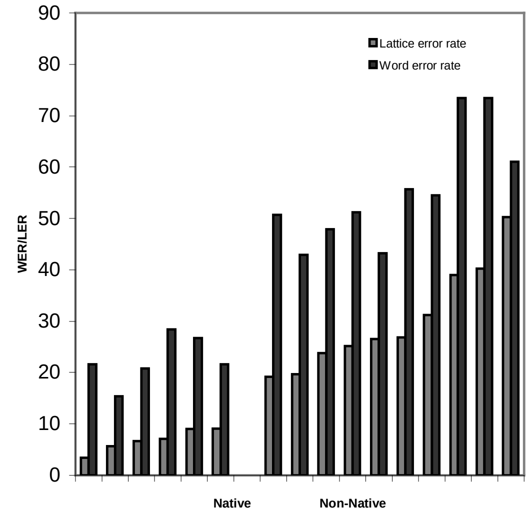

Figure 3: Lattice error rate (LER) and word error rate (WER)

5.2.3 Lattice error

In order for lattice adaptation to be successful, the word for which variants are being added must be in the lattice to begin with. This is not a trivial assumption in the case of non-native speakers; although the search space represented by the lattice is still quite large, the lattice error rate[1] is generally higher for non-native speakers than it is for native speakers.

Figure 3 shows the relationship between word error rate and lattice error rate for native and non-native speakers. The two non-native speakers with extremely high word error rates are the two speakers for whom lattice adaptation is harmful.

6. Conclusion

The non-native speakers we worked with made many reading errors, and there were noticeable patterns in the errors that were made. More than half of the errors were substitutions of morphological variants, and over half of those involved errors in number agreement. Although the native speakers made few reading errors overall, there were entire categories of native error that were missing in non-native speech (semantically similar but orthographically dissimilar words) and vice versa (neologisms). Reading errors by native speakers were much less likely to affect the comprehensibility of the sentence.

Lexical adaptation, in which pronunciation variants are added to the dictionary, was successful only in limited cases and resulted in an overall degradation in LVCSR performance. However, lattice adaptation, in which pronunciation variants are added to the lattice after initial matching with acoustic and language models, showed promise for all but the poorest-performing speakers.

7. Future Work

There are many possible extensions to this work. Given the effectiveness of lattice adaptation and the tendency of the non-native speakers in our test set to interchange singular and plural noun and verb forms, it would be interesting to incorporate syntactic variants as well as pronunciation variants in lattice adaptation.

It was our initial hope that different applications of lexical adaptation would benefit different speakers, and that probabilities of pronunciation variants could be modified depending on the speaker for improved recognition. Although the lexical adaptation experiments were disappointing, speakers did show strong trends in the type of lattice adaptation that was successful, and lattice adaptation may be a good candidate for probabilistic modeling.

Acknowledgements

Susanne Burger provided extremely helpful suggestions about transcription and annotation of non-native reading errors. This work would not have been possible without meticulous transliterations and phonetic transcriptions by Janet Feenstra, Denise Hill, and Pete Allen.

The author is supported by a Microsoft Graduate Research Fellowship.

---

[1] The *lattice error rate* is the word error rate of the most correct path through the lattice. If the correct path exists in the lattice, the lattice error rate for that utterance is 0%.

References

[1] David A. van Leeuwen and Rosemary Orr. Speech Recognition of Non-native Speech Using Native and Non-native Acoustic Models. *Proceedings of the ESCA workshop on Multi-lingual Interoperability in Speech Technology (MIST)*, 1999.

[2] William Byrne et al. Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English. *Proceedings of Speech Technology in Language Learning (STiLL)*, 1998.

[3] James R. Glass and Timothy J. Hazen. Telephone-based Conversational Speech Recognition in the JUPITER Domain. *Proceedings of ICSLP*, 1998.

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

[5] John L. D. Clark and Spencer S. Swinton. An Exploration of Speaking Proficiency Measures in the TOEFL Context. TOEFL Research Report 4, Educational Testing Service, 1979.

[6] Laura Mayfield Tomokiyo and Susanne Burger. Eliciting Natural Speech from Non-Native users: Collecting Speech Data for LVCSR. *Proceedings of the ACL-IALL joint workshop in Computer-Mediated Language Assessment and Evaluation in Natural Language Processing*, 1999.

[7] Laura Mayfield Tomokiyo. Linguistic Properties of Non-native Speech. *Proceedings of ICASSP*, 2000.

[8] Michael Finke et. al The JanusRTk Switchboard/Callhome 1997 Evaluation System. *Proceedings of the LVCSR Hub5-e Workshop*, 1997.

[9] Tilo Sloboda. Dictionary Learning: Performance through Consistency. *Proceedings of ICASSP*, 1995.

[10] Silke Witt and Steve Young. Offline Acoustic Modeling of Non-native Accents. *Proceedings of Eurospeech*, 1999.