LEXICAL AND ACOUSTIC MODELING OF NON-NATIVE SPEECH IN LVCSR

Laura Mayfield Tomokiyo

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

ABSTRACT

As non-native speakers become more frequent users of speech recognition applications, increasing the tolerance of the system with respect to non-native pronunciation and language use is important and is currently the focus of research in a variety of contexts. Dictionary modification, acoustic model adaptation, and acoustic model manipulation are a few of the techniques that have been reported successful in improving recognition of non-native speech. In this paper, we address the specific case of Japanese-accented English, describing the lexical and acoustic modeling techniques that give the best recognizer performance. We find that automatically generated pronunciation variants perform as well as hand-coded "golden" variants in reducing recognizer error, and that a significant improvement in system performance can be achieved with acoustic models retrained on a small amount of accented data.

1. INTRODUCTION

Non-native LVCSR is a very difficult task; recently reported results range from of 14% WER[1] for the high-proficiency speakers in Broadcast News to over 60% [2] for spontaneous speech. A number of methods for handling non-native speech in speech recognition have been proposed. These fall into three general categories: lexical modeling, topological modeling and acoustic modeling.

In lexical modeling approaches, the pronunciation dictionary is altered to reflect likely mispronunciations. Lexical modeling has resulted in improved recognition of native speech when actual pronunciations are different from the pronunciations that were trained, as is the case with regional accents [6] or spontaneous speech [14]. Because many common pronunciation errors made by non-native speakers are consistent and well-known, simple lexical modeling for non-native speakers can be very easily implemented. Both rule-based [1] and data-driven [8] generation of pronunciation variants for lexical modeling have been successful in reducing word error in recognition of non-native speakers.

Topological modeling, or allowing transitions between different sets of acoustic models, can also increase the recognizer's tolerance of non-native pronunciation without requiring any additional training. This technique is particularly effective in pronunciation tutoring applications, where the expected phoneme sequence is known or can be restricted, and feedback about the path through the native-non-native model space is helpful for the user [7, 11].

The class of acoustic modeling techniques is the largest. If training data and/or trained acoustic models are available, significant improvements in recognizer performance can be achieved with adaptation [13], model interpolation [18], training on accented speech, training on L1 speech, and recognition directly from L1 models [16].

For this study, we implemented a number of lexical and acoustic modeling techniques to compare their performance on one LVCSR task, for a group of speakers with similar English proficiencies from the same L1 background. We restricted our study to native speakers of Japanese in order to control for as many variables as possible; we expect, however, that the trends that we have observed will be similar for other L1 groups.

2. DATA

2.1. Speakers

For this research we recorded 40 native speakers of Japanese reading English news. Speakers were evaluated for English proficiency using the SPEAK system [15] and ten speakers of similar proficiencies were chosen for a test set. The test speaker proficiency scores ranged from 1.83 to 2.17 on a scale of 0 to 3 (a score of 3 does not represent native speech, but rather completely intelligible yet detectably non-native speech). The speakers had all had extensive schooling in English, but had been in an English-speaking environment for one year or less and reported some difficulty understanding and making themselves understood.

In addition to the non-native speakers, ten native speakers were also recorded under the same conditions. All comparisons to performance on native speech refer to this test set.

2.2. Task

Each speaker read three articles, totaling approximately 150 sentences, from a database of children's news. Of the three articles, two were unique to each speaker (to maximize phonological breadth), and one was read by all speakers (to provide a control article on which reading skill and recognizer performance can be evaluated independent of text content). Of the recorded data, data from 10 speakers was held out for language modeling experiments and the remainder was data was partitioned into several sets for training and testing as shown in Table 1. Speakers were told that they should make their best attempt to

[1] This was the ROVER result from the 1999 Broadcast News evaluation.

pronounce any unfamiliar words, and that if they made an error they could either continue or return to the beginning of the sentence. Recordings were done in a quiet room with a close-talking microphone; speakers were alone during recording.

Partition	speakers	utterances	hours
Testing	10	419	1.1
Adaptation	50 utterances per speaker		
Training	15	1343	2.8
Cross-validation	5	301	0.9

Table 1: Data set composition

2.3. Transcription

Detailed transcriptions were made of the recorded speech. Reading errors were transcribed and classified. Heavily accented words were transcribed phonetically and classified as consistent or unique for that speaker. Mispronounced words were flagged; most mispronunciations were due to unfamiliarity with the lexical item and were easily distinguishable from accented words. Transcription and distribution of speech errors are described more fully in [10]. Although phonetic transcriptions were done by linguistics students who had had training in field methods, inter-coder consistency was not high, the primary areas of disagreement being vowel identity and presence of r-coloring. For this reason, in experiments using the phonetic transcriptions for pronunciation variant generation, a subset of the training database that had been transcribed by a single transcriber was used.

3. EXPERIMENTS

We have implemented a number of lexical and acoustic modeling methods to compare their effectiveness in improving recognition accuracy on non-native speech. In this section, we describe the approaches that we tried and compare their performance on our test set.

3.1. Baseline system

The baseline system uses the JANUS speech recognition system [3] with quintphone contextual models. It has been our experience that even for strongly accented speakers, in LVCSR tasks, context-dependent models outperform context-independent models unless optimization techniques such as BBI are also used. There are approximately 150k Gaussians in the system, with 6000 distributions sharing 2000 codebooks. VTLN and speaker-based cepstral mean subtraction are applied. Performance of this system is 9.4% WER under baseline conditions on the Broadcast News domain. The language model for the children's news task combines two language models, one trained on text and transcribed broadcast news corpus data and a second trained on text written for children, with context-dependent interpolation weights. Performance on the native speakers in our test set was 18.9%; the degradation was found to be attributable to slight differences in speaker characteristics and domain.

3.2. Lexical Modeling

Although the exact implementation of lexical modeling approaches depends on the way word pronunciations are represented in the recognizer, all involve generation of *pronunciation variants* that reflect likely deviations from the expected native pronunciation of phones and words.

3.2.1. Automatic generation of pronunciation variants

One resource that has been used with success [8, 6] in pronunciation variant generation is phoneme recognition. Because unconstrained phoneme recognition is vulnerable to recognizer error, and many of the confusible non-native phones are also confusible for native speakers, we generated pronunciation variants in two passes. First, we ran an unrestricted phoneme recognition pass. We rejected all mappings below a certain confusibility threshold, all which were linguistically improbable (e.g. [ɚ] $\rightarrow$ [t] / [k]__), and all which also had high confusibility for native speakers (e.g. [s] $\rightarrow$ [z]). We then did an alignment pass, allowing the recognizer to choose from the network of potential mappings the phoneme sequence with the highest acoustic score. Each variant generated this way was then associated with a probability based on the frequency with which it was found in the training data.

3.2.2. Rule-based generation of pronunciation variants

Information about the phonological structure of Japanese and common mistakes made by Japanese learners of English was used to create a set of context-sensitive variant generation rules. As was noted in [4], rules based on linguistic knowledge are well-studied, stable, and not sensitive to recognizer bias or data sparsity. Linguistically-motivated variant generation rules can be applied selectively to reflect specific phonological effects; for some applications, knowledge about the probable linguistic basis of recognition errors is valuable. We applied variant generation rules in several stages to represent transformations such as direct phoneme mapping, allophonic variation, and consonant cluster simplification.

3.2.3. Text resources

Many English words are used in modern Japanese, and because these words are often converted to the Japanese syllabary when written, a rough dictionary of likely accented pronunciations can be compiled by extracting English-origin words from text. This is not as ad-hoc a method as it may seem; the widespread use of such Japanized words encourages fossilization of incorrect pronunciation in some common words while words with identical phonological contexts are pronounced correctly.

We used an online dictionary to select word-level pronunciation variants to add to our system. This approach has the benefit of adding a much smaller set of words (163, compared to up to 10,000 for automatic generation of variants), leading to faster recognition.

3.2.4. Results

In all lexical modeling experiments, a first recognition pass was run to create a word lattice. Pronunciation variants were then added to the lattice and an acoustic rescoring pass was done to produce the final hypotheses. It has been our experience that this technique results in better recognition accuracy than adding the words to the dictionary; supporting experiments are described in [9]. A hand-coded dictionary with transcribed pronunciations for frequent realizations of the 100 most common words is also tested as a gold standard.

Results are shown in Table 2. All improvements were found to be highly significant using both matched-pairs and Wilcoxon signed-rank tests. The improvements from incorporating variants based on phonological rules are smaller than those from incorporating automatically-generated and hand-coded pronunciations, and performance of the automatically-generated variants reaches that of the gold standard.

Pronunciation variant type	WER
Baseline	55.7
Phone mappings	52.8
Japanese dictionary	52.7
Cluster simplification	52.5
Automatically generated	50.8
Hand-coded	50.6

Table 2: Results of lexical modeling experiments

As the success of lattice adaptation naturally depends on having the correct word in the lattice, and we had envisioned that lexical modeling would occur on top of acoustic modeling, we did these initial experiments on "golden" lattices which were written after adaptation on perfect hypotheses. For this reason, the baseline performance is much better than that reported in the following sections. The lattice error rate in these lattices was 34.4%, which is high compared to the native speech (7%). However, when lattice error was measured on the retrained system to be described in Section 3.3.2, it was much lower (26%) than that of the gold-standard lattices, leading us to expect that the same or stronger performance gains will be seen when applying acoustic and lexical modeling in combination.

3.3. Acoustic Modeling

3.3.1. Speaker Adaptation

It has been reported that various forms of adaptation help tremendously for high-proficiency non-native speech [13, 17], and we have observed the same for speech of the lower proficiency levels that we are targeting. Without adaptation, the word error rate for our speakers on our task was over 90%. Supervised MLLR adaptation brought word error rate down to 67.3% for an unseen test set and 52% for the adaptation utterances. All experiments described henceforth assume speaker-level MLLR adaptation on 50 utterances, which was found to be the saturation level.

3.3.2. Retraining with accented data

With only two hours of training data, the options for acoustic modeling are limited. Although we found that we did not have enough data to grow a robust decision tree based on the non-native pronunciation, we were able to modify the baseline acoustic models by running two additional training iterations using the non-native data for dramatic improvements in word accuracy. Table 3 shows the decrease in word error rate for each speaker.

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 3: Improvements in WER for the retrained system

In an effort to decrease word error further, we 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 ([5], e.g). In the native and non-native model sets, there is a one-to-one mapping between senones 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. Our goal is to move the non-native distribution towards the native distribution to the point of maximum robustness. To achieve this, we 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:

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

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

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

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 our 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.

Table 4 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 [18] which found the optimal weighting factor to usually be less than .5 with an interpolation scheme that operates on only the mean vectors.

model weight	0	.3	.5	.72	1
WER	67.3	58.3	48.1	45.1	47.2

Table 4: 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.

4. CONCLUSIONS AND FUTURE WORK

We have described the methods that we found successful for adapting to non-native speech for a single test group of lower-proficiency native speakers of Japanese. We have found that lexical modeling adaptation is effective when applied directly to the lattice, and that while there is no significant difference between different phonologically-motivated pronunciation variant generation methods, an automatic technique performs better, with no significant difference from recognition on hand-coded pronunciations. We also found that additional training with a small amount of accented data can reduce error rate from 67.3% to 45.1%. While other methods of non-native acoustic modeling did not yield improvements in word accuracy, there was improvement in other areas; re-training with the L1 speech data, for example, led to a 25% reduction in lattice error.

We have only touched on some of the techniques that could be used for adapting to non-native speech. Many of these methods could be applied in combination, first retraining to get good lattices, for example, and then re-recognizing with an adapted lexicon. As techniques involving introduction of new polyphones have suffered from data sparsity problems, adapting the polyphone tree directly as suggested in [12] may be possible with small amounts of data. We have also not explored word-level modeling; although this is a read task and one would not expect to see significant differences in native and non-native word use, the large number of reading errors does result in an increase in perplexity for the non-native speakers.

5. ACKNOWLEDGEMENTS

Valuable advice and support from Michael Finke, Hua Yu, and all the members of the Interactive Systems Lab are gratefully acknowledged.

The author is supported by a Microsoft Graduate Research Fellowship.

6. REFERENCES

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

[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. In *Proc. Speech Technology in Language Learning (STiLL)*, 1998.

[3] Michael Finke et al. The JanusRTk Switchboard/Callhome 1997 Evaluation System. In *Proc. the LVCSR Hub5-e Workshop*, 1997.

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

[5] X.D. Huang, Mei-Yuh Hwang, Li Jiang, and Milind Mahajan. Deleted interpolation and density sharing for continuous hidden markov models. In *Proc. ICASSP*, Atlanta 1996.

[6] J. J. Humphries and P. C. Woodland. The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training. In *Proc. ICASSP*, Seattle 1998.

[7] Goh Kawai. *Spoken Language Processing Applied To Nonnative Language Pronunciation Learning.* PhD thesis, University of Tokyo, 1999.

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

[9] Laura Mayfield Tomokiyo. Handling Non-native Speech in LVCSR: A Preliminary Study. In *Proc. Incorporating Speech Technology in Language Learning (InSTIL)*, 2000.

[10] Laura Mayfield Tomokiyo. Linguistic Properties of Non-native Speech. In *Proc. ICASSP*, 2000.

[11] Orith Ronen, Leonardo Neumeyer, and Horacio Franco. Automatic Detection of Mispronunciation for Language Instruction. In *Proc. Eurospeech*, 1997.

[12] Tanja Schultz. Language Adaptive LVCSR through Polyphone Decision Tree Specialization. In *Proc. the ESCA workshop on Multi-lingual Interoperability in Speech Technology (MIST)*, 1999.

[13] Richard Schwartz et al. Modeling Those F-Conditions - Or Not. In *Proc. the 1997 DARPA Speech Recognition Workshop*, 1997.

[14] Tilo Sloboda. Dictionary Learning: Performance through Consistency. In *Proc. ICASSP*, 1995.

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

[16] David A. van Leeuwen and Rosemary Orr. 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)*, 1999.

[17] Frank Wallhoff, Daniel Willett, and Gerhard Rigoll. Frame-discriminative and Confidence-driven Adaptation for LVCSR. In *Proc. ICASSP*, 2000.

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