LEXICON ADAPTATION FOR LVCSR: SPEAKER IDIOSYNCRACIES, NON-NATIVE SPEAKERS, AND PRONUNCIATION CHOICE

Wayne Ward, Holly Krech, Xiuyang Yu, Keith Herold, George Figgs, Ayako Ikeno, Dan Jurafsky

Center for Spoken Language Research
University of Colorado, Boulder

William Byrne

Center for Language and Speech Research
The Johns Hopkins University

ABSTRACT

We report on our preliminary experiments on building dynamic lexicons for native-speaker conversational speech and for foreign-accented conversational speech. Our goal is to build a lexicon with a set of pronunciations for each word, in which the probability distribution over pronunciation is dynamically computed. The set of pronunciations are derived from hand-written rules (for foreign accent) or clustering (for phonetically-transcribed Switchboard data). The dynamic pronunciation-probability will take into account specific characteristics of the speaker as well as factors such as language-model probability, disfluencies, sentence position, and phonetic context. This work is still in progress; we hope to be further along by the time of the workshop.

1. INTRODUCTION

Many ASR researchers have suggested the idea of a *dynamic lexicon*: a lexicon with a large number of pronunciation variants whose probability is set dynamically according to various factors. ([1] *inter alia*). This paper is the preliminary description of our project to apply this idea to two domains: Switchboard (human-human native American English telephone conversations) and Hispanic English (conversations in English between native Spanish speakers with varying levels of accent). Both of these domains are known to have high error rates, and pronunciation variation is known to contribute to the difficulty of these tasks [2, 3, 4, 5].

The goal of this work-in-progress is to build a lexicon with a set of pronunciations for each word, in which the probability distribution over pronunciation is dynamically computed. The set of pronunciations are derived from hand-written rules (for foreign accent) or clustering (for phonetically-transcribed Switchboard data). The dynamic pronunciation-probability will take into account specific characteristics of the speaker as well as factors such as language-model probability, disfluencies, sentence position, and phonetic context.

Section 2 describes a preliminary experiment suggesting that a 'dynamic lexicon' is only useful if words have many pronunciations. Section 3 describes our preliminary work on automatically creating pronunciations. Section 4 reports on preliminary work on the foreign-accent accented data.

2. PILOT EXPERIMENT: DYNAMIC LEXICON WITH TWO PRONUNCIATIONS

Our first experiment was an oracle experiment designed to show whether having exactly two pronunciations for each of the 50 most frequent words in Switchboard, a very full pronunciation and a very reduced pronunciation, would improve recognition.

Our experiments were conducted using Sonic [6], a large vocabulary continuous speech recognition system with Viterbi decoding, continuous density hidden Markov models and trigram language models. Sonic's acoustic models are decision-tree state-clustered HMMs with associated gamma probability density functions to model state-durations. Our experiments used only the first-pass of the decoder, which consists of a time-synchronous, beam-pruned Viterbi token-passing search. Cross-word acoustic models and trigram language models are applied in this pass. This first experiment was run with an early version of Sonic, which had a WER of 42.9% on the 888-sentence Switchboard WS97-test set. (By comparison, WER on this test set in our current version of Sonic is 32.9%).

We used SRI's Hub-5 language model, generously made available by Andreas Stolcke. We built our 39,198-word lexicon from the Mississippi State ISIP Switchboard lexicon. Since this dictionary did not have every word in the LM, we used the CMU dictionary as a resource for any words that were in the LM but were not in the ISIP lexicon. We also included 1658 compound words ('multiwords'), of which 1393 were not in the ISIP or CMU lexicons. So for

Thanks to the NSF for partial support of this research via award #IIS-9978025.

these 1393 we included two pronunciations, full (by concatenating the pronunciations of the consituents words) and reduced (hand-written). The average number of pronunciations per word is 1.13.

We built 2 versions of this lexicon, which differed only in the pronunciations of the top 50 words. In the 'single-pron' lexicon, we allowed only one pronunciation for the most frequent 50 words. In the 'two-pron' lexicon, we included two pronunciations for each of these words, a canonical pronunciation and a very reduced pronunciation, with equal probabilities. Finally, we created a test set from 4237 Switchboard utterances which had been phonetically labeled [**?**, 7]. This allowed us to know, for each test utterance, whether the correct pronunciation of each word was canonical or reduced. From this we built a third dynamic lexicon, a 'cheating' or 'oracle' lexicon, which for each test set sentence only used the pronunciation that was present in the test set.

We then tested the three lexicons with and without retraining the acoustic models. Table 1 shows the results.

Models	Lexicon	WER
Baseline Model	single-pron	43.7
Baseline Model	oracle	41.8
Retrained Models	oracle	41.5
Retrained Models	two-pron	41.7

Table 1. Comparing lexicon performance on a 4237-utterance SWBD test set

Table 1 suggests that having two pronunciations rather than one for the 50 most-frequent words does in fact reduce WER (by 2%, from 43.7% to 41.8%). But an oracle telling us which pronunciation to use (41.5% WER) was not significantly better than just putting in both pronunciations (41.7% WER). This suggests that two pronunciations is an insufficient number for any kind of dynamic lexicon to be useful. In essence, with only two pronunciations, the recognizer was able to choose the correct pronunciation, even without a pronunciation probability.

As a result of this pilot, we determined that a dynamic lexicon would need to have large numbers of pronunciations, more than we were thought was possible to correctly write by hand. In the next two sections, we discuss how we are building pronunciations by clustering and rule-writing.

3. SWITCHBOARD EXPERIMENT: BUILDING MORE PRONUNCIATIONS AND MAPS

3.1. Baselines

Before describing our clustering work, we describe our intended baseline for the SWBD experiments. This is a 5-step extract-align-count-prune-retrain algorithm generalized from [8]:

1. Extract observed alternate word pronunciations from the ICSI labeled data.
2. Align pronunciations with training data
3. Count number of times each pronunciation occurs
4. Prune pronunciations with low counts
5. Retrain acoustic models with alignments to new dictionary
6. (Evaluate WER on test set)

We will then build a slightly more advanced clustered version of the algorithm, in which pronunciations are clustered into broad classes (Vowel Front, Vowel Back, Vowel Reduced, Consonant Labial, Consonant Dorsal, Silence) before accumulating counts. Then we keep at least one example of each broad class with sufficient count, before the align, prune, re-train and evaluate steps.

For example, the word *that* has 36 phone-level variant pronunciations; [dh ae] and [dh ae t] are the most frequent. It has 19 broad class variants, with [CC VF] and [CC VF CC] being the most frequent.

We have already aligned and counted pronunciations, both for phones and broad classes, and are currently working on pruning and then retraining acoustic models.

3.2. Building broad-class maps

In addition to building pronunciations, we are creating a new kind of pronunciation feature based on canonical-to-surface mappings, relying on a database originally produced by Eric Fosler-Lussier that aligns canonical pronunciations with surface pronunciations from the ICSI phonetically labeled data.

A mapping is a change or transduction from the canonical phone sequence to the surface phone sequence, containing a sequence of differing labels (of whatever length) anchored on each end by labels that are the same in both sequences. For the maps, in addition to the 7 broad classes, 3 word positions, b(eginning), m(iddle) and e(nd) were used. For example, in the following map pattern the sequence to the left of $\rightarrow$ is the canonical sequence, the sequence to the right is the surface sequence, and "vb:e" represents a back vowel at the end of a word:

sil cc:b vb:e cc:b $\rightarrow$ sil null vf cc

This algorithm has 4 steps:

1. Accumulate counts for all canonical-to-surface mappings in the training data:
   - with and without word boundary info,
   - with phones and with broad classes:
2. Prune low frequency maps
3. Cluster maps by co-occurrence into classes which will define speaker types

After computing counts from the training data, low frequency patterns were pruned to give the final set of map patterns. For each session side, the frequency of each of the patterns in the set was computed, including the frequency of each canonical string mapping onto itself. The patterns are currently being clustered to produce a set of classes with correlated pattern probabilities. These will define a set of speaker classes on the basis of the observed frequency of patterns. It is generally the case that relatively few patterns account for much of the data. For example, 19 broad class patterns account for about 50% of the sequence differences in the training data.

These derived speaker classes and their probability estimates will be used as features in the decision trees determining the probabilities for alternate pronunciations of words.

4. DYNAMIC LEXICONS FOR SPANISH ACCENTED ENGLISH

4.1. The Hispanic-English corpus and test sets

We are using the conversational Hispanic-English corpus developed at Johns Hopkins University [9]. This database contains about 20 hours of telephone conversations in English from 18 native Spanish speakers, 9 male and 9 female. All speakers were adults from South or Central America who had lived in the United States at least one year and had a basic ability to understand, speak and read English.

During the telephone conversations, the speakers completed four tasks: picture sequencing, story completion, and two conversational games. For the picture sequencing task, participants received half of a randomly shuffled set of cartoon drawings and were asked to reconstruct the original narrative with their partner. For the story completion, participants were given two identical copies of a set of drawings depicting unrelated scenes from a larger narrative context and were asked to answer three questions: "What is going on here?, What happened before?, What is going to happen next?" The first conversational game, Scruples, involved reading a description of a hypothetical situation and trying to resolve the conflict or dilemma. For the second game, the speaker pairs were asked to agree on five professionals to take along on a mission to Mars from a list of ten professions.

These data were divided into development, training and test sets according to speaker proficiency and gender. The development and test sets both include about 30,000 words; from four speakers in the test set, and two in the dev set, while the training set contains about 70,000 words from the remaining ten speakers, five male and five female (See Table 2). Speakers had been judged on proficiency scores based on a telephone-based, automated English proficiency test [10] We also listened to each speaker and rated their accent as heavy, mid and light. We then combined the proficiency scores with our accent ratings to distribute speakers with heavy, mid and light accents evenly into the different data sets. A range of the degree of accentedness is thus represented in each data set.

Set	Gender	Minutes	Words
Training	5 male, 5 female	546	69,926
Dev	1 male, 1 female	176	29,474
Test	2 male, 2 female	282	30,104

Table 2. Hispanic-English training and test set statistics

4.2. Baseline recognizer performance

We used the Sonic speech recognizer with our SWBD lexicon and acoustic models to establish a baseline from a system trained on native American English on Hispanic-English speech. Our SWBD system, as described earlier, consists of a 39,000 word lexicon, the SRI Hub-5 language model, and SWBD acoustic models. On the development test set of 176 minutes of speech and 29,974 words, we achieved a baseline word error rate of 62%.

4.3. Pronunciation rules for Hispanic-English

We next created lexical variants on the basis of seven phonological rules (See list below). These rules represent common characteristics of Spanish accented English, and they were determined by comparing literature about Spanish accents [11] to the Hispanic-English database and selecting the most appropriate characteristics. The seven rules are:

1. epenthetic schwa added before words beginning in /s/, as in *speak* [ax s p iy k];
2. past tense morpheme -ed pronounced /ax d/ following voiced consonants, as in *planned* [p l ae n ax d]";
3. reduced schwa vowels pronounced as they are spelled, the full vowel represented by the orthography, as in *minimum* [iy n iy m uw m]";
4. the mid-high vowels /ih/ and /uh/ become the high vowels /iy/ and /uw/;
5. /s/ and /z/ in word fi nal position are deleted;
6. the fricatives /sh/ becomes the affricate /ch/ in word initial position, and
7. the fricative /dh/ becomes the stop /d/.

Table 3 gives formal versions of the rules.

While we have not yet tested whether these rules help in improving recognition performance, we have analyzed some of the errors when the Switchboard recognizer is applied to the Hispanic English dev set, yielding some anecdotal observations that relate to the rule set. First, final consonants tend to be deleted, especially /s/, /z/, /v/ and /t/, causing substitutions of words with no final consonants, such as "know" for "not" and "how" for "have". Our phonological rules account only for the deletion of /s/ and /z/. Second, the /dh/ fricative is pronounced as both /d/ and /s/, not just as the /d/ we indicate in our rules. Another fricative that is

1. s → ax s / #___
2. d → ax d / voiced C ___ #
3. ax → aa / orthographic a
   ax → eh / orthographic 'e'
   ax → iy / orthographic 'i'
   ax → ow / orthographic 'o'
   ax → uw / orthographic 'u'
   axr → er / orthographic 'er'
4. ih → iy
   uh → u
5. s → 0 / ___#
   z → 0 / ___#
6. sh → ch / #___
7. dh → d

Table 3. Phonological Rules for Hispanic English

problematic is /f/, which is pronounced and recognized as /p/. Third, the softening of /b/ to a bilabial fricative causes substitution of words that have no stop consonant where the /b/ occurs, as in "busy" substituted with "easy". Fourth, many of the reduced vowels are pronounced and recognized as full vowels, which we expected based on the third phonological rule. Finally, hesitations seem to be nasalized, with "nn" for "uh", which causes the recognizer to substitute a short word beginning with a nasal, such as "no" or "not", for these hesitations.

4.4. Applying pronunciation count-prune-retraining

We next use the phonological rules discussed above to attempt to build a better baseline system for Hispanic English. We use the 3-step algorithm first proposed by [12]:

- apply phonological rules to the base lexicon, generating a large number of pronunciations,
- forced-align against the training set to get pronunciation counts
- prune low-probability pronunciations

Our base lexicon was the Switchboard lexicon described above , consisting of 39204 word tokens with 1.13 pronunciations per word type. We applied the 7 phonological rules in Section 4.3 to produce 'accented' pronunciations, which were then merged with the base lexicon, and redundant forms were removed. The resulting augmented lexicon consisted of 96954 word tokens with 2.8 pronunciations per word type. Next, this augmented dictionary was aligned with the reference corpus data, giving us counts of the number of times a particular pronunciation was choosen for a given word.

We are currently working on the pruning step. Once that is complete, we will proceed to retraining the acoustic models with the resulting dictionary. That will provide a 'static lexicon' baseline which we can then use to see the performance of our dynamic lexicon approach on the Hispanic-English data.

5. CONCLUSION

Our main result so far is that hand-writing very-reduced pronunciations for 50 frequent function words reduces word error rate even after using a lexicon with 1600 reduced-pronunciation multi-words, usually based on these same function words. Our other results are still too preliminary to admit of much conclusion, but we hope to have more results by September.

6. REFERENCES

[1] Eric Fosler-Lussier, *Dynamic Pronunciation Models for Automatic Speech Recognition*, Ph.D. thesis, University of California, Berkeley, 1999, Reprinted as ICSI technical report TR-99-015.

[2] Don McAllaster, Larry Gillick, Francesco Scattone, and Mike Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *ICSLP-98*, Sydney, 1998, vol. 5, pp. 1847–1850.

[3] Mitch Weintraub, Kelsey Taussig, Kate Hunicke-Smith, and Amy Snodgras, "Effect of speaking style on LVCSR performance," in *ICSLP-96*, Philadelphia, PA, 1996, pp. 16–19.

[4] Murat Saraclar, Harriet Nock, and Sanjeev Khudanpur, "Pronunciation modeling by sharing gaussian densities across phonetic models," *Computer Speech and Language*, vol. 14, no. 2, pp. 137–160, 2000.

[5] Dan Jurafsky, Wayne Ward, Zhang Jianping, Keith Herold, Yu Xiuyang, and Zhang Sen, "What kind of pronunciation variation is hard for triphones to model?," in *IEEE ICASSP-01*, Salt Lake City, Utah, 2001, pp. I.577–580.

[6] Bryan Pellom, "Sonic: The university of colorado continuous speech recognizer," Tech. Rep. TR-CSLR-2001-01, Center for Spoken Language Research, University of Colorado, Boulder, 2001, Revised April 2002.

[7] Steven Greenberg, "Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, 1999.

[8] Michael D. Riley, William Byrne, Michael Finke, Sanjeev Khudanpur, Andrei Ljolje, John McDonough, Harriet Nock, Murat Saraclar, Chuck Wooters, and George Zavaliagkos, "Stochastic pronunciation modeling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, pp. 209–224, 1999.

[9] W. Byrne, E. Knodt, S. Khudanpur, and J. Bernstein, "Is automatic speech recognition ready for non-native speech? a data collection effort and initial experiments in modeling conversational hispanic english," in *ESCA Workshop*, 1998.

[10] Ordinate Corporation, "The phonepass test," 1998.

[11] H. S. Magen, "The perception of foreign-accented speech," *Journal of Phonetics*, vol. 26, pp. 381–400, 1998.

[12] Michael H. Cohen, *Phonological Structures for Speech Recognition*, Ph.D. thesis, University of California, Berkeley, 1989.