

ISSUES IN RECOGNITION OF SPANISH-ACCENTED SPONTANEOUS ENGLISH

*Ayako Ikeno[∞], Bryan Pellom[∞], Dan Cer[#], Ashley Thornton[∞], Jason M. Brenier[∞],
Dan Jurafsky[∞], Wayne Ward[∞], William Byrne^{*}*

[∞] Center for Spoken Language Research, University of Colorado at Boulder

[#] Institute of Cognitive Science, University of Colorado at Boulder

^{*} Center for Language and Speech Research, The Johns Hopkins University

{Ayako.Ikeno, Bryan.Pellom, Daniel.Cer, Ashley.Thornton, Jason.Brenier,
Daniel.Jurafsky, Wayne.Ward}@colorado.edu, byrne@jhu.edu

ABSTRACT

We describe a recognition experiment and two analytic experiments on a database of strongly Hispanic-accented English. We show the crucial importance of training on the Hispanic-accented data for acoustic model performance, and describe the tendency of Spanish-accented speakers to use longer, and presumably less-reduced, schwa vowels than native-English speakers.

1. INTRODUCTION

Foreign accent is a crucial problem that ASR systems must address if they are to deal with spontaneous speech. In this paper, we study the problem of recognition of English spoken by speakers with strong Spanish accents. We first show that while applying an English-specific recognizer on Spanish data produces an unacceptable error rate. We then show that training a recognizer, even on only 20 hours of speech from only 14 speakers, drops the error rate from 68.5% to 39.2%. This suggests that if sufficient amounts of accented training data are available, performance can be quite acceptable. In most cases, however, it is difficult to obtain enough training data for specific accents. The rest of this paper therefore describes two experiments on the kinds of techniques that might be necessary for adapting a native speaker ASR system to model accented data.

2. CORPUS

Our experiments were based on the conversational Hispanic-English spontaneous speech corpus developed at Johns Hopkins University [2]. This corpus consists of approximately 30 hours of English conversation spoken by speakers whose native language was Spanish, and who had varying degrees of Spanish accent in their English. The 18 participants (nine male, nine female) were from Central or South America, and all had lived in the United States at least one year and had a basic ability to understand, speak, and read English. The Hispanic-English corpus had been collected in a dual-recording setup. Each participant was recorded with a wide-bandwidth close-talking microphone and simultaneously with narrowband telephone channels. During the 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.

Our experiments were performed only on the wideband speech, of which we had approximately 27 hours (we did not use a small read speech part of the corpus, nor did we use the 2 to 3 hours of as-yet untranscribed data). In addition, the audio files for 2 hours of the data contained only silence, although there were corresponding transcriptions. Those silent audio files were removed from the audio corpus, although the transcriptions were used for language model training.

These data were divided into development, training and test sets according to speaker proficiency and gender. The development and test sets each include about 2.5 hours; from two speakers in each of the two sets, while the training set contains about 150,000 words from the remaining fourteen speakers, seven male and seven female (see Table 1). Speakers had been judged on proficiency scores based on a telephone-based, automated English proficiency test [7]. We also listened to each speaker and rated their accents as heavy, mid and light. We used the accent ratings to assign one heavily accented male and one heavily accented female speaker to each of the dev and test sets.

| Data | Gender | Hours | Words |
|-------|------------------|-------|---------|
| Train | 7 male, 7 female | 20.0 | 154,903 |
| Dev | 1 male, 1 female | 2.5 | 11,731 |
| Test | 1 male, 1 female | 2.5 | 14,662 |

Table 1: Corpus Data Distribution

3. BASELINE SYSTEM

Experiments in this paper were conducted using Sonic, the University of Colorado Large Vocabulary Speech Recognition system [8]. Sonic is based on continuous density hidden Markov (CDHMM) acoustic models. Context dependent triphone acoustic models are clustered using decision trees. Features are extracted as 12 MFCCs, energy, and the first and second differences of these parameters, resulting in a feature vector of dimension 39. Cepstral mean normalization is applied during feature extraction. The search network is a reentrant static tree-lexicon. The recognizer implements a two-pass search strategy. The first pass consists of a time-synchronous, beam-pruned Viterbi token-passing search. Crossword acoustic models and 3-gram or 4-gram language models (in an approximate and efficient way) are applied in the first pass of search. The first pass creates a lattice of word ends. During the second pass, the resulting word-lattice is converted into a word-graph. Advanced language models (e.g. dialog-act and concept based, long span) can be used to rescore the word graph using an A* algorithm or to compute word-posterior probabilities to provide word-level confidence scores (although lattice rescoring is not considered in this work).

Sonic provides an integrated environment that incorporates voice activity detection (VAD), speech enhancement as well as various feature and model-based adaptation and normalization methods. The recognition architecture provides support for rapid portability to new languages. In 2002, Sonic was ported from English to the Spanish, Turkish, and Japanese languages.

Sonic has been benchmarked on several standard continuous speech recognition tasks for American English and has been shown to have competitive recognition accuracy to other recognition systems evaluated on similar data. Performance metrics are shown in Table 2.

| Task | Vocabulary Size | Word Error Rate |
|---------------------------------|-----------------|-----------------|
| TI-Digits | 11 | 0.4 % |
| DARPA Communicator | 3k | 12.6 % |
| Wall Street Journal (Nov. 1992) | 5k | 4.2 % |
| Switchboard | ~40k | 31.0 % |

Table 2: Word error rate for the CSLR Sonic Recognizer on several tasks: TI-Digits, DARPA Communicator telephone based travel planning domain, Nov'92 Wall Street Journal (WSJ) 5k test set and the Switchboard task.

Our baseline system for transcription of the Hispanic-English corpus consists of an integrated speech detection and multiple pass recognition search [9] as shown in Figure 1. During each recognition pass, a voice activity detector (VAD) is dynamically constructed from the current adapted system acoustic models. The VAD generates a segmentation of the audio session into utterance units and LVCSR is performed on each detected speech region. The resulting output (a confidence

tagged lattice or word string) is then used to adapt the acoustic model means and variances in an unsupervised fashion. The adapted acoustic models are then reapplied to obtain an improved segmentation, recognition hypothesis, and new set of adapted system parameters. The integrated adaptation procedure can be repeated several times resulting in sequential improvements to both segmentation and recognition hypotheses.

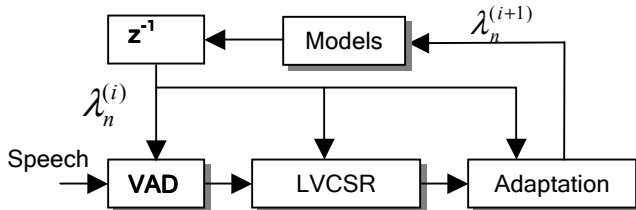


Figure 1: Diagram of Hispanic-English multi-pass recognition system.

Search is performed in a multi-pass setup using gender-dependent acoustic models in the first pass. Subsequent passes are performed using vocal-tract length normalized models. Between ASR passes, multiple regression class MLLR mean and variance adaptation is performed to adapt the system parameters to each speaker. A total of 6 regression classes are used in experiments presented in this paper.

3.1. Lexicon

Our lexicon is derived from the CMU pronunciation lexicon [4]. We also augmented our lexicon with Spanish words and ungrammatical words that occurred in our training set. Some examples are shown in Table 3.

| | |
|--------------|-------------------------------|
| BEADED | B IY DX AX DD |
| CONSIDERANDO | K OW N S IY D EY DX AA N D OW |
| FOOTS | F UH T S |
| GOYITA | G OW Y IY T AA |
| INSPIRATE | I X N S P AX R EY TD |
| PETRONILA | P EY T DX OW N IY L AA |
| RANCHEROS | DX AA N CH EY DX OW S |
| ROBERTICO | DX O B EY DX T IY K OW |

Table 3: Examples of augmented entries

Our lexicon for acoustic training also included pronunciations for the 200 most frequent word fragments (e.g., I-, M-, etc.)

3.2. Acoustic Models

Sonic's acoustic models are based on decision tree state-clustered continuous density HMMs. Iterative parameter estimation is performed using repeated Viterbi forced alignment of the training data followed by decision tree state clustering.

Acoustic models for our baseline system were estimated using the 20 hours (7 male, 7 female speakers) of data from the Hispanic-English corpus. This baseline did not rely on any kind of interpolation with native-speaker acoustic models. After Viterbi alignment, approximately 8% of our training set was discarded because it could not be aligned; mainly this was because some audio files had incomplete transcriptions (either

marked *unintelligible* or containing rare word fragments which we had not included in our training lexicon). We attempted to improve the quality of the alignments by adapting the forced aligner to each speaker using the MLLR technique, however we found that this did not result in lowered error rates in the resulting baseline system.

The final baseline system consisted of gender-dependent acoustic models for the first ASR pass. Subsequent passes utilize vocal tract length normalized models. Warping factors ranging from 0.88 to 1.12 with 0.02 increments are estimated for each training and test speaker using the hypothesis from the first-pass of the recognizer.

3.3. Language Model

Our Katz backoff trigram language model was estimated from the Hispanic-English corpus training set using the CMU Statistical Model Toolkit [3]. This language model had a vocabulary size of 4,123 words and a perplexity on the development set of 46.7.

4. ACCENT ADAPTATION: ACOUSTIC MODELS VS LANGUAGE MODEL

Speech recognition systems use both acoustic models and language models in decoding speech. Language models assign probabilities to word sequences, and acoustic models assign probabilities to words being realized as an observed acoustic feature sequence. Both types are statistical models that estimate their parameters from training data. In non-native speech, both the word sequences and the acoustic realizations of words are different than the native speech that the system was trained on. Both types of models will contribute to degraded performance. Adapting native English acoustic models to non-native English speech can significantly improve the recognition accuracy on the foreign accented speech, even with a small amount of training data. In [10] interpolating native and non-native acoustic models reduced WER from 67.3% to 45.1% for a Japanese-accented speech recognition task. We conducted a small experiment to determine the effects of using acoustic and language models trained on a different type of data than the test data. The objective in this experiment was to identify, in our foreign-accented English speech task, which factors might have more dominant influence on the system performance.

The data set was the spontaneous Hispanic-accented speech development set described earlier. Two sets of acoustic models were used, ones trained on Wall Street Journal and ones trained on the accented speech. Both training sets are wide-band speech. WSJ is speech read by native English speakers. Three language models were used, WSJ, SwitchBoard and a language model trained on the accented data. (Switchboard is spontaneous speech produced by native English speakers on a variety of topics [6])

The results shown in Table 4 indicate clearly that using both the appropriate language model and acoustic models is important for deriving the best overall performance. As the language model becomes a better match to the development data (from

read English to spontaneous English to spontaneous accented), the error rate is reduced. The accented data also matches the test

| Acoustic Model | Language Model | | | |
|----------------|----------------|------|------|--------|
| | | WSJ | SWB | Accent |
| | WSJ | 80.2 | 74.1 | 68.5 |
| Accent | 56.4 | 46.7 | 39.2 | |

Table 4: Word Error Rate % by Models

domain as well as the speaking style. However, all of the evaluations done using the non-native speech acoustic models outperform all of the evaluations with WSJ models. Re-training the acoustic models on the accented data results in significantly more accurate recognition performance; e.g., 27.4% to 29.3% gain. Using a more appropriate language model has less of an effect; e.g., 5.6% to 17.2% gain. As with language models, the accented acoustic models also are appropriate to the domain as well as the speaking style (perhaps better triphone coverage for the models).

The general trend suggests that, in this foreign-accented speech, acoustic models are even less transferable than language models, and benefit more from appropriate training data. This result differs from transferring between task domains in native English recognition. It is generally found to be the case that in native English tasks acoustic models transfer well where language models do not. This result supports the result from in [10] described above.

5. REDUCED VOWELS IN FOREIGN ACCENTED SPEECH

In our previous study [11], we found that prosodic features such as stress and pitch accent may be an important factor for the performance of an ASR system on non-native speech. Specifically, we showed that heavily-accented Spanish speakers were more likely to substitute full vowels for schwas than lightly-accented speakers. A small study of selected utterances also showed that heavily-accented speakers also had more errors in word and sentence stress than lightly-accented speakers. Our previous experiments, however, did not compare Spanish-accented speakers with native speakers. In this experiment, we follow up on our earlier finding by looking at the difference between Spanish-accented and native speech in how they deal with schwas and full vowels.

In order to determine if the results generalized to our corpus while avoiding having to hand label the stress patterns, we chose a relatively simple dependent variable to investigate: vowel durations. Our goal was to see if Spanish speakers were more likely to have full vowels (or stressed vowels) in places where native speakers had reduced vowels (schwas). We therefore looked at the average duration of schwas in the Hispanic-English corpus, and compared it to the average duration of schwas in the native-speaker Switchboard corpus. We determined the average duration of reduced vowels by performing a forced alignment to phone boundaries. Subsequently we normalized these values by the average vowel duration in accented

and native speech respectively. The ratio of these two values was calculated using equation (1) below:

$$\frac{\frac{\text{Reduced-vowel}_{\text{accented}}}{\text{Average-vowel}_{\text{accented}}}}{\frac{\text{Reduced-vowel}_{\text{native}}}{\text{Average-vowel}_{\text{native}}}} \quad (1)$$

| Phoneme | Normalized accented to native duration ratios |
|---------|---|
| AX | 1.21 |
| AXR | 1.14 |
| IX | 1.10 |

Table 5: Normalized ratios for reduced vowels. Values greater than 1 indicate longer durations for Hispanic-accented data than for Switchboard data, after normalizing for average vowel duration.

Table 5 shows that the three schwa vowels, [ax], [axr], and [ix], are longer on average in Spanish speakers than they are in native speakers, after normalizing for the length of the average vowel in the two corpora. These results show that the Spanish-accented speakers tend not to reduce vowels as much as native American English speakers do. While we have not yet applied this result in our recognizer, our hypothesis is that modifying either the lexicon or the acoustic models to represent these longer, unreduced schwas for Spanish speakers may play a key role in dealing with accent.

Is this fact about longer, unreduced schwas a fact about Spanish-accented English or a fact about all accented speakers? Previous research suggests that this result applies only to Spanish, and that different languages have different positions along the reduction scale. For example [1] compares average word duration for Mandarin-accented and Turkish-accented speakers. They point out that while Mandarin-accented speakers had longer vowels on average than native English speakers, that Turkish speakers on average had shorter vowels. [5] points out that Korean-accented speakers emphasize durational differences between (shorter) [ih] and (longer) [iy], presumably influenced by the fact that Korean has a distinction between long and short vowels. As such, those subjects might have used duration as a more important aspect of vowel identity. Thus, while these previous studies did not directly address reduced vowels, the do suggest that non-native speakers from different language backgrounds apply different strategies in pronouncing English vowels. This is an area that clearly calls for further investigation.

6. DISCUSSION

We have shown that a speech recognizer trained on native English speech performs poorly on Spanish-accented English, but that training the recognizer on Hispanic English instead drops the error rate drastically from 68.5% to 39.2%. Since accented data is not always available for training, we also performed two studies that we hope will shed light on how native-English recognizers might eventually be adapted to accented English. Our

first study showed that acoustic models are even less transferable from native to accented speech than language models, and benefit more from appropriate training data. This suggests, not surprisingly, that, when possible, acoustic training (rather than LM training) is key to better performance on accented speech. Our second study confirmed our earlier suggestions that Spanish speakers tended to use more full vowels and less schwas than native English speakers.

In our current research, we hope to apply these ideas in two ways. First, since foreign-accented acoustic data is often hard to get, we hope to try using acoustic training data in the foreign language itself. Since our first result points to the importance of acoustic training data, we hope that e.g., Spanish training data might help in building acoustic models that work well on Hispanic English. Second, we hope to build a modified Hispanic-English lexicon that directly represents the full, non-schwa vowels of the Hispanic-English speakers.

7. REFERENCES

- [1] L. M. Arslan, J. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," in *Journal of the Acoustical Society of America*, Vol. 101, No. 1, 1997
- [2] W. Bryne, E. Knodt, S. Kudanpur, J. Bernstein, "Is automatic speech recognition ready for non-native speech? A data collection effort and initial experiments in modeling conversational Hispanic English," in *Proceedings of the ESCA-ITR Workshop on speech technology in language learning*, 1998
- [3] P. R. Clarkson, R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit", in *Proceedings of ESCA Eurospeech*, 1997
- [4] CMU "The Carnegie Mellon Pronouncing Dictionary" 1996.
- [5] J. E. Flege, O.S. Bohn, S. Jang, "Effects of experience on non-native speakers' production and preception of English vowels," *Journal of Phonetics*, Vol. 25, 1997
- [6] J. Godfrey, E. Holliman, J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the ICASSP*, 1992
- [7] Ordinate Corporation, "The phonepass test," 1998
- [8] B. Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer," *Technical Report TR-CSLR-2001-01*, CSLR, University of Colorado, 2001
- [9] B. Pellom, K. Hacioglu, "Recent Improvements in the CU Sonic ASR system for Noisy Speech: The SPINE Task," in *Proceedings of the ICASSP 2003*
- [10] L. M. Tomokiyo, "Lexical and acoustic modeling of non-native speech in LVCSR," in *the Proceedings of the ICSLP Workshop*, 2000
- [11] W. Ward, H. Krech, X. Yu, K. Herold, G. Figs, A. Ikeno, D. Jurafsky, W. Byrne, "Lexicon adaptation for LVCSR: Speaker idiosyncracies, non-native speakers, and pronunciation choice," in *PMLA Workshop*, 2002

ACKNOWLEDGEMENT

Thanks to the NSF for partial support of this work via awards IIS-9733067 and IIS-9978025. Many thanks to Holly Krech.