Limitations of MLLR Adaptation with Spanish-Accented English: An Error Analysis

Constance Clarke[1,2] and Dan Jurafsky[2]

[1]Department of Psychology, University at Buffalo SUNY, Buffalo, NY, USA
[2]Department of Linguistics, Stanford University, Stanford, CA, USA
cclarke2@buffalo.edu, jurafsky@stanford.edu

10.21437/Interspeech.2006-342

ABSTRACT

We studied the effect of MLLR adaptation with Spanish-accented English to understand the strengths and weaknesses of MLLR with unseen foreign accents. We trained a global MLLR transform on 10 adaptation sentences per speaker, giving a 3.4% absolute decrease in phone error rate. We then studied the pattern of improvements across phones and phone classes. Phones that improved the least tended to be those that do not exist in Spanish. Results suggest the poorer performance is related to increased insertion and substitut*er* rates during the adaptation phase, as well as greater acoustic variability.
Index Terms: speech recognition, adaptation, MLLR, foreign accent, Spanish, error analysis

1. INTRODUCTION

Non-native speech is well-known for reducing speech recognition performance. It can result in word error rates of two to three times native error rates [1, 2]. Accent-specific acoustic models can improve performance, but are often infeasible in practice due to data collection limitations and an intractable/unknown number of potential accents.

An alternative approach is to adapt native acoustic models to non-native speech on-line during recognition. Maximum likelihood linear regression (MLLR) [3] is an adaptation technique that uses small amounts of data to train a linear transform which warps the Gaussian means so as to maximize the likelihood of the data. MLLR has produced substantial gains in non-native recognition accuracy for a variety of accents, including Spanish (7% absolute) [1], Cantonese (10%) [4], Japanese (over 20%) [5], and German (5%) [6]. Despite these gains, however, typical non-native accuracy following adaptation still falls substantially below that of native speech [1, 2, 4, 5, 6]. An important step in improving an adaptation technique is understanding its limitations. What aspects of accented speech are adaptation methods not capturing?

To begin to address this question, we performed a phone-based error analysis on the recognition results for Spanish-accented speech before and after MLLR speaker adaptation. Our goal was to determine which phones or phone classes improved the most or least with MLLR. We chose to use MLLR with a global transform because of its effectiveness with adaptation set sizes on the order of 10s of utterances. This time frame is ideal for the goal of rapid adaptation to a speaker with an unknown accent.

2. SPANISH-ACCENTED CORPUS

We used a subset of the conversational Spanish-accented English corpus collected at Johns Hopkins University [1] that included 16 speakers (8 males, 8 females) from several Spanish-speaking countries. The speakers represented a wide variety of English proficiencies. In the corpus, pairs of speakers performed collaborative tasks over the telephone, and the speech was recorded simultaneously in telephone- and wide-band formats. The present study used the wide-band speech. Speech was segmented into turns and transcribed, with a mean of a little over one hour per speaker.

3. ERROR ANALYSIS

In this study, we performed phone-recognition tests on the Spanish-accented corpus, first with native English acoustic models, and then after adapting the models with 10 utterances per speaker. We then conducted a linguistically-based analysis of the improvement in phone recognition error rate of various phone classes. We used phone recognition, rather than word recognition, in order to more directly assess the acoustic model fits and avoid the masking effects of lexical and language model constraints. We also conducted word recognition on the entire Spanish-accented data set in order to obtain a word error rate (WER) benchmark for comparison with other studies.

3.1 Recognition system

To conduct the recognition experiments, we used the SONIC large vocabulary recognizer [7]. We used only the first pass of the decoder, which consists of a time-synchronous beam-pruned Viterbi token-passing search. Crossword acoustic models and a 3-gram language model were applied in this pass.

Acoustic models: We used gender-dependent acoustic models (AMs) from the SONIC release, pre-trained with 30 hours of data (283 speakers) from the WSJ0 and WSJ1 corpora of WSJ text read by native speakers of American English. The models were context-dependent state-tied triphones.

Lexicons: For word recognition, we used the 39,000-word Mississippi State ISIP Switchboard lexicon as augmented for the Spanish-accented data set by [8] and [9]. For phone recognition, we created a phone lexicon in which each phone was a "word" with itself as the pronunciation.

Language models: For both word and phone recognition, we used cheating language models (LMs) based on the entire data set. This allowed for a reasonable level of performance

despite the task and accent mismatch between the data and acoustic models.

To create the phone LM, we replaced each word in the transcriptions with its pronunciation from the lexicon described above, and used the CMU-Cambridge Toolkit [10] to build 3-gram Katz back-off LMs with Witten-Bell discounting. The word LM had a vocabulary of 4133 words, and the phone LM had 55 "words": 50 phones (from a modified SPHINX version of the ARPAbet) plus symbols for silence and non-speech.

Word-recognition evaluation: The entire data set was used for the word-recognition test, a total of 5,530 (test) utterances and 144,539 (test) words. The recognizer used the native English AMs and the full word lexicon and LM described above. The mean WER for males was 64.4% (61.0% to 67.1%). The mean WER for females was 61.4% (43.7% to 83.1%). WER correlated strongly with English proficiency scores (r = -0.78, p < .001; scores are on a scale from 2 to 8, with higher scores signifying higher proficiency [1, 11]). These error rates are comparable to those of other accented speech studies [1, 4, 5, 6, 8].

3.2 Phone recognition and adaptation methodology

For our main experiment, we decoded the Spanish-accented speech using the phone-recognizer with the native English AMs and holding out 10 utterances per speaker for adaptation. This produced the baseline phone error rates (PERs). We then evaluated the data again following MLLR speaker adaptation.

To select the adaptation utterances, we selected all utterances with durations between 8 and 13 seconds. From these we then selected by hand the first 10 for each speaker that contained at least 50% speech. We chose to use 10 based on pilot tests showing little additional improvement with more than 10 utterances using a global transform.

MLLR adaptation was implemented in an unsupervised fashion with a global transform. Although evaluation of the full test set used phone recognition for the purpose of phone-based error analysis, recognition of the adaptation utterances during MLLR used word recognition because this is what is typically used. For each speaker, 10 adaptation utterances were first decoded with native English AMs, and the alignment output was used to adapt the AMs. Each subsequent iteration used the newly adapted models to re-recognize the adaptation data. After three iterations, phone recognition was performed on the speaker's test data (minus the adaptation data) using the adapted AMs.

The mean baseline PERs were 66.9% for male speakers and 59.4% for females. Post-adaptation PERs were 63.4% for males and 56.1% for females.

3.3 Error analysis results

We first evaluated the recognition performance for each phone. Phone-level reference transcriptions were created by replacing each word in the word transcriptions with its pronunciation in the lexicon. Sclite v2.3 was used to align the reference and hypothesis, and rates of correct identification, substitution, deletion, and insertion were calculated for each phone.

Figure 1 shows the absolute improvement in PER for each phone from baseline to post-adaptation, grouped by manner class. Two characteristics stand out. First, phones vary widely in the amount of improvement, with some benefiting by over 8% and others getting worse. In addition, /r/-variants benefited greatly from adaptation: ER, AXR, and R gained by 13.26%, 8.64%, and 6.00%, respectively.

Improvement in PER (post-MLLR – baseline) did not correlate significantly with the proportion of each phone in the adaptation materials (r = 0.18, *ns*), with the baseline PER (r =0.09, *ns*), or with PER in the adaptation utterances (r = 0.10, *ns*). These results were unexpected; more instances of a phone in the adaptation materials should result in better representation of that phone's acoustic characteristics in the global transform. Yet the correlation was actually in the opposite direction. A similar argument could be made for baseline PER.

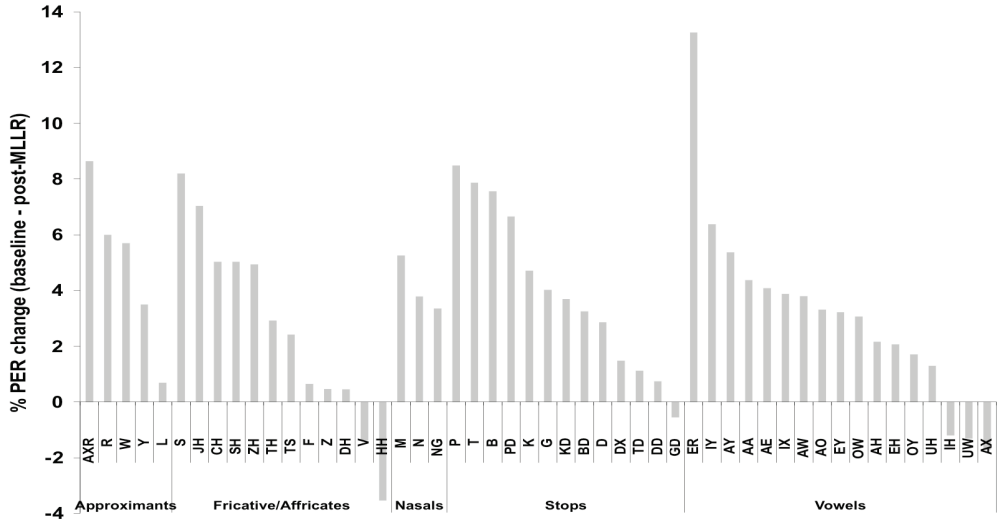

Figure 1: Improvement in phone error rate following MLLR speaker-adaptation with 10 utterances. Means represent 16 Spanish-accented speakers. (Note: stop-D, e.g., PD, represents syllable-final unreleased stop consonants.)

We grouped phones into linguistic classes to investigate how MLLR adaptation captures traditional linguistic features. Table 1 shows the baseline PER and percent improvement following adaptation for each phone class examined. To determine whether differences were statistically reliable, we entered percent improvement for each phone contrast into a separate analysis of variance (ANOVA) with phone class and gender as factors, and speaker as a random effect. Consonants improved more than vowels, $F(1, 14) = 18.54$, $p < .005$, and tense vowels improved more than lax vowels, $F(1, 14) = 23.26$, $p < .001$. There was a main effect of class for the Manner contrasts, $F(2.5, 34.7) = 3.18$, $p < .05$ (Huynh-Feldt corrected for nonsphericity), and for the Place contrasts, $F(7, 98) = 14.89$, $p < .001$. There was no effect of Voicing, and no main effects or interactions involving gender for any analysis.

Table 1: Baseline phone error rate and percent change after adaptation (post-MLLR PER – baseline PER)

Contrast	Baseline PER	Change (% absolute)
Consonants vs. Vowels†		
consonant	62.64	-3.52
vowel	62.22	-2.21
Manner*		
approximant	64.82	-4.41
fricative/affricate	60.35	-1.84
nasal	51.86	-4.14
stop	70.30	-4.17
vowel	62.22	-2.21
Voicing (Cs only)		
voiced	63.00	-3.42
voiceless	61.99	-3.69
Place of articulation (Cs only)†		
labial	53.92	-6.33
labiodental	59.08	0.34
dental	65.43	-0.86
alveolar	63.52	-3.42
postalveolar	72.36	-6.42
palatal	54.57	-3.49
velar	61.65	-4.05
glottal	70.32	3.53
Tenseness (Vs only)†		
tense	55.74	-4.04
lax	67.47	-0.72

† $p < .005$, * $p < .05$, Note: All means are weighted by phone count.

An interesting trend begins to emerge in this analysis: Phonemes that do not exist in the Spanish language seem to improve the least with adaptation. For example, lax vowels, which are not present in Spanish, improve by only 0.72% absolute compared to 4.04% for tense vowels, many of which are in the Spanish inventory. Notice also that, of the six phones that got worse after adaptation in Fig. 1, five are not in Spanish: V, HH, GD, IH, AX. (We classified a phone as "in Spanish" if it maps fairly directly onto a phoneme in Spanish, not including allophonic variants such as DH.) Based on this observation, we split the phones into two categories, Spanish and non-Spanish, and entered the absolute percent change from baseline to post-MLLR into an ANOVA. As shown in Table 2, Spanish phones improved significantly more than non-Spanish phones, $F(1, 14) = 35.43$, $p < .001$.

Table 2: Baseline phone error rates and change after adaptation grouped by whether phone exists in Spanish

		Baseline PER	Change (% absolute)
In Spanish	AA AXR B CH D EY F G IY K L M N OW P R S T UW W	57.86	-4.74
Not In Spanish	AE AH AO AW AX AY BD DD DH DX EH ER GD HH IH IX JH KD NG OY PD SH TD TH TS UH V Y Z ZH	66.91	-1.31

Note: All means are weighted by phone count.

We explored four possible explanations of the poorer adaptation results for non-Spanish phones. They related to: baseline phone error rates; error patterns produced during the decoding of adaptation utterances; variability in the realization of non-Spanish phones, and the size of the adaptation set.

Baseline phone error rate: As is clear from Table 2, non-Spanish phones had a higher baseline PER than Spanish phones. A possible explanation for the difference in improvement after adaptation is that phones with higher PERs benefit more from MLLR. However, as reported above, there was no correlation between baseline PER and percent improvement by phones.

Errors in decoding adaptation utterances: It is important to examine the errors generated in the first recognition pass of the adaptation utterances because the resulting hypotheses determine the MLLR transform. We expected a higher substitution rate for non-Spanish phones than for Spanish phones because non-native speakers often replace unfamiliar phonemes with similar sounds from their native inventory. The substitution rate was higher for non-Spanish (29.43%) than for Spanish (26.65%) phones; however, the correlation between substitution rate and PER change after adaptation was actually in the opposite direction ($r = -0.27$, $p = .06$). Therefore, the differing improvement rates are not due to more substitutions of non-Spanish phones during adaptation.

Two other error types significantly correlated with change in PER: insertion rate ($r = 0.48$, $p < .001$) and substitut*er* rate ($r = 0.33$, $p < .05$; that is, the mean percentage of time a particular phone was substituted for all target phones). Specifically, phones with higher insertion or substituter rates benefited less from adaptation. This makes sense because inserted and substituted phones are aligned with the wrong acoustic features. If non-Spanish phones have higher insertion and substituter rates, this could explain the differing improvement with adaptation; and, indeed, they do (insertion: non-Spanish 8.41%, Spanish 6.24%; substituter: non-Spanish 3.48%, Spanish 2.49%). Why are non-Spanish phones inappropriately recognized more often? One possibility is that, because the word lexicon and LM were used during unsupervised adaptation, non-Spanish phones were expected as often as in English, but did not appear as often in the data (because of non-native pronunciations). The decoder "recognizes" them, due to lexical and LM constraints, but in the wrong places. We are planning follow-up experiments to test the assumption that non-Spanish phones in the adaptation data actually are a poorer match with the acoustic models.

Variability in vowel production: Finally, we tested whether non-Spanish vowels were produced with greater variability in the test utterances. If the acoustics of these phones were highly

inconsistent, then the MLLR transform would be of little benefit. To test this hypothesis, we hand-measured the first and second formants (F1 and F2) of the Spanish vowels (AA EY IY OW UW) and of the worst-performing non-Spanish vowels (AH AX EH IH UH). Selecting tokens from the data set involved three steps. First, we aligned the full data set with the word transcriptions using SONIC. Second, for each speaker we automatically selected three tokens of each vowel, one from the first, second, and third portions of the speaker's data set. As a compromise between controlling phonetic environment and finding enough tokens, we only measured vowels in alveolar-vowel-alveolar environments (including T D TD DD DX N L S Z). Third, using Praat [12], we realigned the vowel labels by hand to begin/end halfway through the onset/offset transitions, and calculated the mean F1 and F2 for each vowel.

Several samples were discarded due to elision or transcription error, leaving 41 to 48 samples of each vowel. UH was not analyzed because only six tokens were found in the correct context. F1 and F2 standard deviations were calculated separately across males and females and used as the variability measure. Table 3 shows the results. There is no clear difference in mean variability between Spanish and non-Spanish phones. However, there was a significant correlation between F2 standard deviation and PER change for males (r = 0.71, $p < .05$). This finding is suggestive that phones produced with greater variation do not benefit as much with MLLR, but further research is needed.

Size of adaptation set: In case the Spanish/non-Spanish effect was an artifact of the small adaptation set, we re-ran the experiment using half of each speaker's data as the adaptation set (between 131 and 312 utterances), leaving the remaining half for testing, and used 55 regression classes (one per phone) to improve performance. Absolute improvement was 15.24% for Spanish phones, and 9.53% for non-Spanish. Even with a large amount of adaptation data, the effect of phone inventory remains.

Table 3: F1 and F2 standard deviations for male (M) and female (F) speakers and means (M) across vowels

		Spanish						Non-Spanish				
		AA	EY	IY	OW	UW	**M**	AH	AX	EH	IH	**M**
M	F1	93	69	62	75	49	**70**	65	87	44	52	**62**
	F2	155	190	198	207	316	**213**	214	325	196	168	**226**
F	F1	114	97	60	102	62	**87**	118	126	109	68	**105**
	F2	227	244	402	293	361	**306**	156	269	258	275	**240**

4. CONCLUSIONS

We studied the patterns of improvement in phone recognition after MLLR adaptation for Spanish-accented English. Phones varied widely in amount of improvement, from a 13% decrease in PER to an almost 4% increase. Phones with degraded performance tended to be English phones that do not exist in Spanish (e.g., lax vowels, HH, V). Our results suggest this effect is not due to baseline PER, frequency in the adaptation materials, or substitution rate during adaptation. Instead, it is associated with increased insertion and substitut*er* rates during the adaptation phase. We also found an effect of phonetic variability; more variable phones were less well modeled by MLLR (for male speakers only). Finally, English rhotic phones (R, AXR, ER) showed particular improvement from adaptation.

A remaining question is how supervised adaptation would affect the results. We ran a supervised version of the experiment with the male speakers and, surprisingly, found an increase in PER from baseline. This may be due to a large number of poorly transcribed word fragments and filled pauses in the adaptation utterances (we counted a mean of 4.4 out of 10 per speaker). Further research is needed to explore this issue as these are common in non-native speech.

The Spanish/non-Spanish effect much be checked with a native English data set to ensure it is not an inherent characteristic of these phone groups. This study is currently underway. If, however, the effect is due to production patterns of non-native phones, as we suspect, it may point to a source of MLLR's limitations in handling foreign accent. These findings thus suggest that cross-language differences in phone inventory should be an important area of focus for improvements in foreign accent adaptation.

5. ACKNOWLEDGEMENTS

This work was funded by a NIH NRSA to CC. We are grateful to W. Byrne for use of the corpus, J. Yuan, C. Boulis, and J. Brenier for helpful advice and discussion, and G. Grigoryev, C. La, and B. Lo who prepared the data and transcription files as part of a class project.

6. REFERENCES

[1] Byrne, W., Knodt, E., Khudanpur, S., and Bernstein, J., "Is automatic speech recognition ready for non-native speech? A data collection effort and initial experiments in modeling conversational Hispanic English," in *ESCA Conference on Speech Technology in Language Learning*, 1998, pp. 37-40.

[2] Tomokiyo, L. M. and Waibel, A., "Adaptation methods for non-native speech," in *Proceedings of Multilinguality in Spoken Language Processing*, 2001.

[3] Leggetter, C. J. and Woodland, P. C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[4] Liu, W. K. and Fung, P., "MLLR-based accent model adaptation without accented data," in *Proceedings of ICSLP*, 2000, 738-741.

[5] Tomokiyo, L. M., "Lexical and acoustic modeling of non-native speech in LVCSR," in *Proceedings of ICSLP*, 2000, pp. 346-349.

[6] Wang, Z., Schultz, T., and Waibel, A., "Comparison of acoustic model adaptation techniques on non-native speech," in *Proceedings of IEEE ICASSP*, 2003, pp. 540-543.

[7] Pellom, B. and Hacioglu, K., "Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task," in *Proceedings of IEEE ICASSP*, 2003.

[8] Ward, W., Krech, H., Yu, X., Herold, K., Figgs, G., Ikeno, A., Jurafsky, D., and Byrne, W., "Lexicon adaptation for LVCSR: Speaker idiosyncrasies, non-native speakers, and pronunciation choice," in *Proceedings of PMLA*, 2002, pp. 83-88.

[9] Grigoryev, G, La, C., and Lo, B., "ASR training for Spanish-accented English speech," Stanford class project, 2005.

[10] Clarkson, P. R. and Rosenfeld, R., "Statistical language modeling using the CMU-Cambridge Toolkit," in *Proceeding of ESCA Eurospeech*, 1997, pp. 2707-2710.

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

[12] Boersma, P. and Weenink, D., "Praat (Version 4.3.27)," March 2006, http://www.praat.org.