ADAPTATION METHODS FOR NON-NATIVE SPEECH

Laura Mayfield Tomokiyo and Alex Waibel

Language Technologies Institute
Carnegie Mellon University

ABSTRACT

LVCSR performance is consistently poor on low-proficiency non-native speech. While gains from speaker adaptation can often bring recognizer performance on high-proficiency non-native speakers close to that seen for native speakers [12], recognition for lower-proficiency speakers remains low even after individual speaker adaptation [2]. The challenge for accent adaptation is to maximize recognizer performance without collecting large amounts of acoustic data for each native-language/target-language pair. In this paper, we focus on adaptation for lower-proficiency speakers, exploring how acoustic data from up to 15 adaptation speakers can be put to its most effective use.

1. INTRODUCTION

As speakers learn a new language, they trace unique paths through acquisition of phonology, vocabulary, grammar, pragmatics, and even social aspects of spoken communication. The variability that this complexity engenders poses a serious problem for speech recognition. Once speakers reach a certain level of proficiency, their pronunciation may become fossilized, with the most noticeable features of their accent influenced by their native phonological system. In the early stages of learning, however, speakers experiment with new sounds, which results in phonetic realizations that are inconsistent and often distant from both the target phone and any native language phone that one might expect to influence it.

Study of the nature of non-native speech has suggested that perception of a phoneme is influenced by the phonetic contrasts that are meaningful in the speaker's native language (L1) [6] and that production is related to perception for allophonic contrasts [5]. However, it has also been observed that articulation of target language (L2) phones cannot be reliably traced to a related, interfering phone in the speaker's native language [1]. It has often been an assumption in efforts to adapt to non-native pronunciation in speech recognition that a speaker's realization of an L2 phone will fall "somewhere between" the average native realization and realization of an L1 phone that the speaker perceives as being similar to it. While adaptation based on this assumption has been successful for high-proficiency speech and simple tasks (e.g., [8, 14, 12],), both the speech degradation due to high cognitive load and the variability in articulation discussed above make recognition of lower-proficiency speech in LVCSR tasks a very hard problem [2].

In this paper, we concentrate on a specific group of lower-proficiency speakers, quantifying characteristics of their speech and comparing methods of adapting to it in LVCSR. Working with a controlled group of native speakers of Japanese, we investigate phonological properties of speech, fluency and disfluency, and reading errors in a read news task. We then discuss the effectiveness of training and of mixed-style and MLLR adaptation to the non-native condition, examining the contribution of L1 and L2 data to the adaptation process.

2. DATA

In this section, we describe the language background and proficiency evaluation of the speakers, the task and recording conditions, and the recognition system used for adaptation experiments.

2.1. Target speakers

The speakers in this study were all native speakers of Japanese. All had had 6-8 years of formal study of English and had lived in an English-speaking country for 6-12 months. All reported difficulty in making themselves understood, and rated their confidence in conversational speaking between 1.5 and 2.5 on an informal scale of 0 to 4. These speakers can be described as having a good grasp of the formal properties of English but limited productive ability.

In addition to informal evaluations, speaker proficiency in the test set was controlled with respect to scores on the formal SPEAK assessment [13]. All test speakers scored between 1.89 and 2.17 on the read speech portion of this test, which gives scores on a scale of 0 to 3 for identifiably non-native speech. Speakers assigned to the training set ranged from 1.44 to 2.83.

There were 10 test speakers, 15 training speakers, and 8 native speakers in this database.

2.2. Task

Two sets of speakers were recorded for this research. The primary group of interest, which included all test speakers, was recorded speaking English. A second group of speakers was recorded speaking their native language of Japanese.

2.2.1. Accented L2 data

Accented data, that is, recordings of native Japanese speakers speaking English, is referred to as *L2 data* because English is the speakers' L2.

Speakers completed a read news task in which they read aloud three articles from a children's news archive. This task was designed to mirror well-known tasks such as Wall Street Journal, which was determined during preliminary data collection to be too difficult for our speakers.

Of the three articles, one was common to all speakers and the other two were unique to each speaker. Article length averaged 50 sentences. The training/adaptation set represented approximately 3 hours of acoustic data.

Recording was done in a quiet room using a close-talking headset and a DAT recorder. Speakers were alone in the room while recording.

2.2.2. L1 data

Native-language data, that is, recordings of native Japanese speakers speaking Japanese, is referred to as *L1 data* because Japanese is the speakers' L1.

The L1 data that was used for model adaptation and training was taken from the Globalphone database [11] and consists of recordings of native Japanese speakers reading news articles from the Nikkei Shimbun in Japanese. Although the content of this newspaper is more difficult than that in children's news, the reduced cognitive load required for reading one's native language means that the difficulty of the L1 and L2 tasks was similar for the native Japanese speakers.

Speakers recorded an average of 15 minutes of speech. Recording was done in a quiet room using a close-talking headset and a DAT recorder. For consistency with the accented L2 data, 3 hours of this speech distributed across 15 speakers was used for training and adaptation.

2.3. Recognition system

All experiments described in this paper used the JRTk speech recognition toolkit [4] with fully continuous context-dependent acoustic models and a trigram language model. Context-dependent models were determined experimentally to perform better than context-independent models for this speaker set and task. Vocal tract length normalization and cepstral mean subtraction are applied at the speaker level. Linear discriminant analysis (LDA) is used to find the most discriminative of the MFCC, delta, and power features and reduce the dimensionality of the feature vector describing each frame. WER figures always represent accuracy after speaker-dependent MLLR adaptation on 50 utterances. Performance of this system on Broadcast News F0-condition speech is 9.4%. Because of differences in speaking style (informal vs. professional anchor) and language modeling (the broadcast news model was adapted to children's news, but is still not optimal for the task), performance on local native speakers on the children's news task is significantly higher, at 19.2%.

3. CHARACTERIZING LOW-PROFICIENCY ENGLISH

Learning to speak a new language is a journey that doesn't always follow a straight line from L1 to L2. For many speakers, reaching proficiency is a matter of years of trial and error. In this section, we discuss some of the features of non-native speech of the proficiency level we are targeting.

3.1. Reading errors

Reading errors, which are commonly assumed not to occur often enough to greatly affect system performance, were frequent in our data. Nearly 3% of the words that were read by the non-native speakers were not the words on the page, as compared to 0.4% for native speakers.

In addition, the types of reading errors that were made were distributed quite differently in native and non-native speech. Substitution of a morphological variant was by far the most common reading error in non-native speech. Singular-plural substitution represented over 60% of these morphological errors. Non-native reading errors were more likely to affect the syntactic integrity of the sentence; for example, the sentence "Doctors are studying the pill's *effect* on patients" is meaningful whether the word *effect* is singular or plural, whereas the sentence "American *student* perform poorly on standardized tests" is made syntactically incorrect by the speaker's substitution of *student* for *students*. A more detailed breakdown of reading errors in this data can be found in [9].

3.2. Phonological properties

A segment of the non-native data collected in this project was phonetically transcribed by experienced transcribers. Although a number of expected transformations (e.g., /ɪ/→ [i]) were verified during this process, the principal observation was that the number of realizations that could not be transcribed using the union of the standard American English and Japanese phone sets was great. Transcribers required an extensive set of supplemental diacritics, representing r-coloring, centering, and palatization, among other things, to begin to capture the data. There was also a great deal of intra- and inter-speaker inconsistency. One speaker, for example, consistently pronounced [ʌ] as [ɚ] – but only in the second half of one article. For some reason, he made the decision to try this pronunciation out, and then abandoned it when he began the next reading.

Diversions from standard American English phonology were also found in recognizer-driven analysis. Phoneme-level recognition of the data revealed both common insertions, deletions, and substitutions and high overall levels of phoneme confusion, consistent with observations from manual analysis. In an experiment designed to uncover lexical variants, it was found that when phone-level insertions, deletions, and substitutions are considered, 57% of the polyphones (5-phone sequences) in the test data were not seen in the training data, compared to 92% for native speech.

3.3. Fluency

The low-proficiency speakers targeted in this paper read far more slowly and haltingly than native speakers do. Frequent inter-word pauses, stumbling over words, and multiple repetitions of sequences of words have implications for both acoustic and language modeling. In particular, it has been our experience that no complex cross-word modeling is necessary for the lower-proficiency speakers because words are usually articulated one at a time, with pauses in between them.

feature	mean		std. dev.	
	N	NN	N	NN
pause duration	9.56s	17.14s	3.16	7.33
phone duration	0.08s	0.12s	0.01	5.36
pause:word ratio	1:10	1:3	0.05	0.08
words/second	3.80	2.15	0.26	0.29
repair rate	0.57	2.25	0.33	1.42
repeat rate	0.07	0.34	0.07	0.23
retrace rate	0.58	2.35	0.35	1.26
retrace length	2.55	2.57	1.04	2.29
filler word rate	0.01	0.16	0.02	0.32
partial word rate	0.45	1.52	0.20	1.05

Table 1. Comparing fluency-related statistics for native (N) and non-native (NN) speakers in the reading task

Figure 1 gives statistics for fluency (and disfluencies) for the low-proficiency non-native speakers targeted in this paper. The non-native speech is clearly more disfluent than the native speech, as measured by such diverse features as speaking rate, ratio of silence to words, and number of repaired and abandoned words. The only feature that appears to be similar for native and non-native speakers is retrace length, or the number of words a speaker "rewinds" when correcting himself. It could be that this span is influenced by the syntax of the text, which is the same for both native and non-native speakers; it has also been suggested that retrace length is constant across languages [3].

4. SPEAKER PROFICIENCY AND RECOGNIZER PERFORMANCE

In this paper, we specifically target lower-proficiency speakers. Our premise is that these speakers may need processing different from that applied to higher-proficiency speech in order to raise recognition accuracy to an acceptable level. This assumption is based on the intuition that lower-proficiency speakers are somehow harder to understand, as well as the observation that these speakers are diverse and inconsistent in their articulation. To support our assumption, let us quantitatively examine the correspondence between proficiency and recognizer performance.

Figure 1 shows how word error rate (WER) varies with speaker proficiency. We see three distinct clusters. The cluster on the far right represents native speech; native speakers automatically receive a SPEAK score of 4. The center cluster represents speakers who scored between the test set cutoff of 2.17 (the lowest actual score in this group was 2.44) and the maximum non-native score of 3. The test speakers targeted in this paper fall into the leftmost cluster. Although there is some variation in recognizer performance within the clusters, speakers in the lower-proficiency group clearly are recognized with less accuracy than those in the other two.

5. ACOUSTIC MODEL ADAPTATION

In this section, we discuss offline adaptation to the non-native condition prior to individual run-time speaker adaptation. We strive to answer two questions:

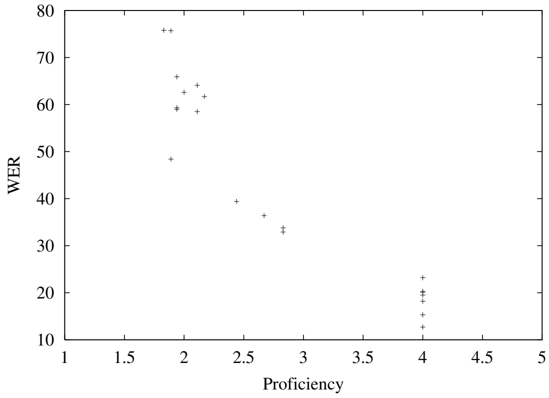

Fig. 1. Correspondence between speaker proficiency and recognizer performance in terms of word error rate (WER)

- Does L1 material provide better adaptation data than accented L2 data?
- Does mixed-style adaptation perform better than MLLR adaptation for non-native speech?

It was observed in [8] that MLLR adaptation with L1 LVCSR data gave similar improvements in accuracy to adaptation with accented L2 data when only isolated L2 phone data was available. In this paper, we explore a matched condition: same-domain LVCSR data is used for both L1 and L2 adaptation material.

In order to use the L1 data described in Section 2.2.2 for adaptation of English acoustic models, the Japanese lexicon had to be converted to the English phone set. Data-driven and IPA-based approaches to this problem have been studied (e.g. [11, 14, 8]); we used a combination. Pronunciation networks for each Japanese word were created with each Japanese phone replaced by a set of parallel transitions representing substitutions of related English phones and phone sequences. "Related" was defined to mean sharing all but one phonological feature. Therefore, any phone that differed only in place of articulation, or manner, or voicing, or vowel height, was added to the network. A forced alignment pass was then run on this network to find the path with the most likely match. Context-sensitive (considering preceding and following phone) global mappings were assigned based on the substitutions selected most often during alignment.

5.1. Mixed-style adaptation

In mixed-style training, adapted model parameters are estimated separately for each of the "styles" (in this case, L1 and L2), and then interpolated using a global interpolation weight. This is, in effect, a simple form of MAP adaptation, where an optimal weighting factor is determined experimentally rather than separately for each Gaussian based on the *a priori* distribution of the Gaussian parameters. It has been our experience that this method produces results that are similar to or slightly better than conventional MAP. If it is likely

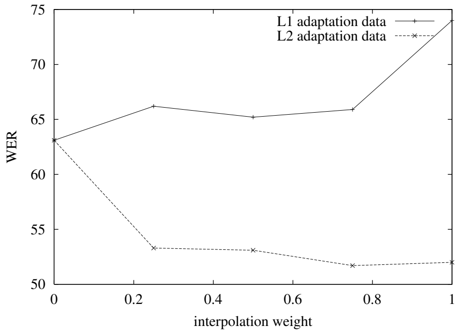

Fig. 2. Mixed-style adaptation using L1 and L2 adaptation data

	L1 data	L2 data
0 speakers (baseline)	63%	
3 speakers	68.1	58.1
15 speakers	73.4	52.5

Table 2. MLLR adaptation using varying amounts of L1 and L2 adaptation data (figures represent WER)

that the adaptation data represents the test data well, it can be heavily weighted for interpolation. As with MAP adaptation, this method performs better as the amount of adaptation data increases, as if individual parameters cannot be reliably estimated from sparse sample data no adaptation is performed. In this experiment, 15 adaptation speakers were used.

Figure 2 shows system performance after mixed-style adaptation with both L1 and L2 data. On the horizontal axis is the interpolation weight. When the interpolation weight is 1, the adapted mean is identical to the sample mean. When the interpolation weight is 0, the adapted mean is identical to the prior mean (i.e., there is no adaptation).

A clear degradation can be seen from adapting with L1 data, while the positive contribution of the accented L2 data can be seen rising steadily as the interpolation weight increases.

5.2. MLLR

In MLLR adaptation, transformation classes are defined, and model parameters of the entire class are shifted in the same direction. While this clustering allows MLLR adaptation to provide a general transformation with a small amount of adaptation data, there is a risk of shifting an individual parameter *away* from observed sample value, which is avoided in mixed-style adaptation.

Results of MLLR adaptation with L1 and L2 data are shown in Table 2. As with mixed-style adaptation, we see a degradation with the introduction of L1 acoustic material. The effect is more extreme with more adaptation speakers, indicating that sample means from the L1 data are not representative of the means in the actual accented test speech. Adaptation with accented L2 data, on the other hand, significantly improves performance over the baseline.

Results are given for 3 and 15 adaptation speakers. It is clear that the effectiveness of adaptation increases

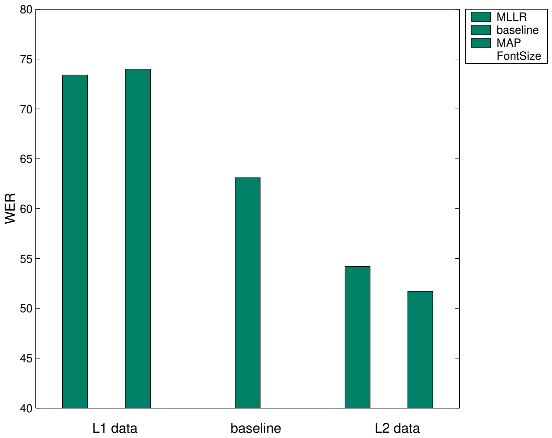

Fig. 3. Comparison of MLLR and MAP adaptation for 15 adaptation speakers

with the amount of adaptation speech. There are two reasons for this: more examples of sample values allow a more reliable estimate of the sample mean, and the more diverse set of samples contributes to a more general model.

5.3. Comparison of adaptation methods

Figure 3 contrasts MLLR and mixed-style adaptation performance for L1 and L2 adaptation material and 15 adaptation speakers. Both show similar trends, with mixed-style slightly outperforming MLLR.

We see clearly from all the experiments shown here that using L1 acoustic material for adaptation to low-proficiency non-native speech without re-evaluation of the polyphone set results in a degradation of recognizer performance, while adaptation with accented L2 data boosts performance.

6. RETRAINING WITH ACCENTED DATA

It was shown in Section 5 that while using accented data for adaptation improves recognition performance, adapting with L1 data results in a performance degradation. In speaker adaptation, the model inventory is kept the same, but the expectation of what a model sounds like is shifted towards what has been seen in the limited set of adaptation speech. The L1 data does not have the chance to make its maximal contribution, as the model inventory is based on the polyphones found in native speech; two allophones that are quite different in L1 may be used to update the same model if the two contexts do not trigger variation in English. By rebuilding the system based on the contexts that are meaningful in L1, we may be able to use the L1 data to its full advantage.

In this section, we compare systems trained with L1 data with systems trained with accented L2 data. Both full rebuilding of the system (rebuilding from scratch) and repetition of the final step of training (additional forward-backward iterations) with the new data are examined.

6.1. Rebuilding from scratch

In this experiment, two new systems were built, using L1 and accented L2 data. In both cases, initial labels were written using the baseline acoustic models, and a context-dependent system was trained along the specifications given in Section 2.3. Because the adaptation data available was sparse for fully training a recognizer, it was pooled with native English data in these experiments. The large amount of native data contributes to the robustness of the model, while the smaller amount of L1 or accented L2 data ensures that L1-specific phone sequences and phone realizations are seen during clustering and training. Training data consisted of 3 hours of L1 or accented L2 acoustic data pooled with the original native training data.

6.2. Additional forward-backward iterations

In this experiment, the new system was not retrained from scratch; rather, two additional forward-backward iterations are run on the fully trained baseline models using the accented L2 acoustic data.[1] In Section 5, we saw how recognition improves with adaptation to the non-native condition when accented data is used. By training with the accented data, we are essentially extending this approach, updating not only the mixture means but also the mixture weights and covariances. We also benefit from the second re-estimation. The effect of additional forward-backward iterations with the L1 data was not examined in this experiment.

6.3. Comparison of training methods

Figure 4 contrasts performance of fully-rebuilt and partially retrained systems. With the rebuilt systems, we see a small improvement when training with L1 data and a much larger improvement when training with accented L2 data.

The improvement from the additional training iterations is even larger. This may be because in retraining (described in Section 6.2), we are capitalizing on consistency in the data in the two phases of system building with native speech and retraining with non-native speech. When the two data sets are combined from the outset (as described in Section 6.1), we may incorporate a broader range of polyphones but be harmed by the mismatch between native and non-native speech. By simply retraining, we fix the identities of the acoustic models with native data, and then use the non-native data to adjust the expectation of how those models correspond to phonetic realization in non-native speech.

6.4. Model interpolation

Simply running additional forward-backward iterations with the three hours of accented data resulted in a 24% relative improvement over the baseline error rate. In this new model, however, the parameters were trained on a small amount of data. This introduces a danger of

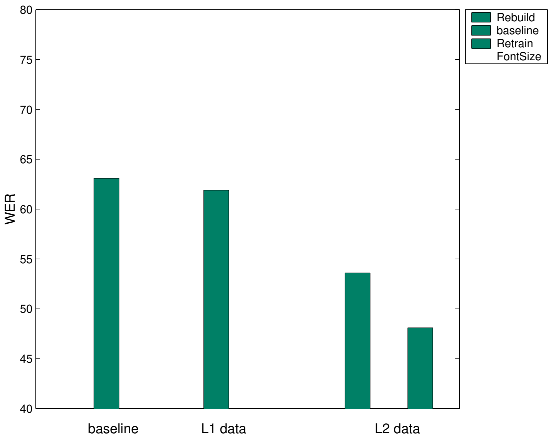

Fig. 4. Comparing rebuilding from scratch with L1 and L2 data and partial retraining with L2 data

overfitting, a problem which can addressed by smoothing the models via interpolation with a more robust model [7]. A direct parameter interpolation technique has also been shown to be successful in creating context-independent non-native models from source and target language model sets [14]. In model interpolation experiments, it was our goal to move the retrained distribution back towards the native distribution to the point of maximum robustness.

In the interpolation method that we used, corresponding codebook weight, mean, and covariance matrix elements are linearly interpolated for each baseline system / retrained system acoustic model pair. This results in a covariance space that covers an area between the two original covariances, rather than the union of the two. We are able to interpolate the individual models in this way because there is a clear one-to-one mapping between models; the decision of which models to interpolate would be much more difficult if we were working with the re*built* system of Section 6.1 instead of the re*trained* system of Section 6.2. Our method is described in detail in [10]. Performance of the interpolated system is 29% above that of the baseline system, a significant improvement over the retraining alone. The effect of the interpolation weight on recognition accuracy is shown in Figure 5; optimal performance is found when the retrained models are weighted at .72.

7. SUMMARY

In this paper, we have examined how application of acoustic model training and adaptation techniques affects recognition accuracy on non-native speech. A summary of the individual contributions of each method is shown in Figure 6.

Generally speaking, adaptation to the non-native condition (and by adaptation we refer to both the speaker adaptation techniques of MAP and MLLR and retraining techniques) using L1 data does not improve performance, and in some cases causes a large degradation. Accented L2 data, on the other hand, contributes positively to the acoustic model. The largest gains are seen when using the full 3 hours of accented data to run ad-

[1] For this experiment, the baseline models were also trained an additional two iterations to ensure that the comparison was fair. We did not observe any significant change between the original 7-iteration training and the 9-iteration training with the native data, however.

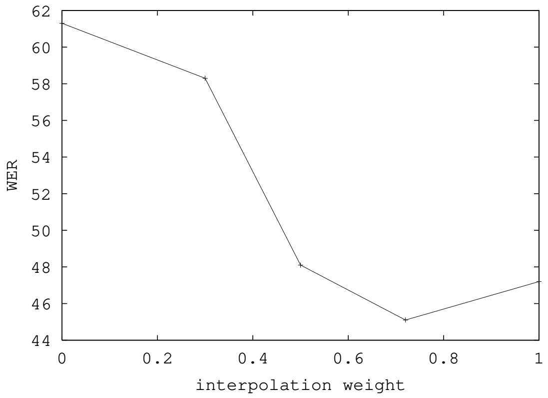

Fig. 5. Results for interpolation with different interpolation weights. A weight of 0 represents performance with the original acoustic models. A weight of 1 represents performance with the new models.

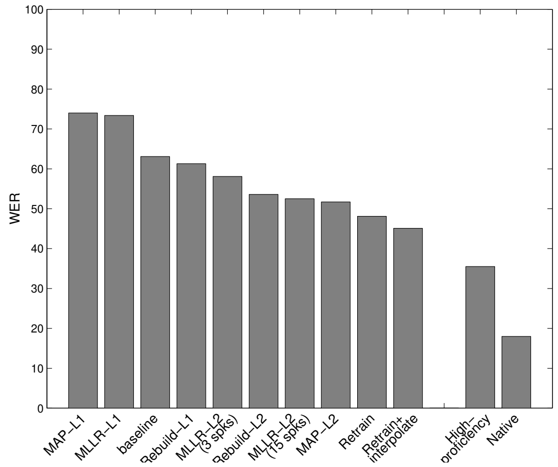

Fig. 6. Summary of adaptation results. Bars labeled "MAP" refer to mixed-style adaptation, which is a simplified form of MAP adaptation.

ditional forward-backward training iterations and then interpolating the retrained model back with the more robust baseline models. Significant gains are also seen with MAP and MLLR adaptation, where performance of the system improves proportionally to the amount of accented adaptation speech.

In the best case, word error rate for the lower-proficiency speakers is lowered from 63.1% to 45.1%, which represents a 29% relative reduction in error. This approaches, but does not match, performance on the higher-proficiency speakers.[2] With an absolute reduction in error of 18%, we have closed half of the gap in recognizer performance on native and low-proficiency non-native speech; how close this brings us to the upper limit, however, remains to be seen.

[2] We see the same trends when applying adaptation techniques to proficient non-native speech, although the effect is far less dramatic [10].

8. REFERENCES

[1] Eugene Brière. An investigation of phonological interference. *Language*, 42(4):768–796, 1966.

[2] William Byrne, Eva Knodt, Sanjeev Khudanpur, and Jared Bernstein. 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] Robert Eklund and Elizabeth Shriberg. Crosslinguistic Disfluency Modeling: A Comparitive Analysis of Swedish and American English Human-Human and Human-Machine Dialogs,. In *Proc. ICSLP*, 1998.

[4] Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries, and Torsten Zeppenfeld. The Janus-RTk Switchboard/Callhome 1997 Evaluation System. In *Proc. the LVCSR Hub5-e Workshop*, 1997.

[5] James Emil Flege. Production and perception of a novel, second-language phonetic contrast. *J. Acoust. Soc. Am.*, 93(3):1589–1608, March 1993.

[6] Robert Allen Fox and James Emil Flege. The perception of English and Spanish vowels by native English and Spanish listeners. *J. Acoust. Soc. Am.*, 97(4):2540–2551, April 1995.

[7] 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.

[8] Wai Kat Liu and Pascale Fung. MLLR-based accent model adaptation without accented data. In *Proc. ICSLP*, 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. *Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in Speech Recognition.* PhD thesis, Carnegie Mellon University, 2001.

[11] Tanja Schultz and Alex Waibel. Language independent and language adaptive large vocabulary speech recognition. In *Proc. ICSLP*, Sydney, 1998.

[12] Richard Schwartz, Hubert Jin, Francis Kubala, and Spyros Matsoukas. Modeling Those F-Conditions - Or Not. In *Proc. the 1997 DARPA Speech Recognition Workshop*, 1997.

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

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