COMPARISON OF ACOUSTIC MODEL ADAPTATION TECHNIQUES ON NON-NATIVE SPEECH

Zhirong Wang, Tanja Schultz, Alex Waibel

Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, PA, 15213
Email: {zhirong, tanja, ahw}@cs.cmu.edu

ABSTRACT

The performance of speech recognition systems is consistently poor on non-native speech. The challenge for non-native speech recognition is to maximize the recognition performance with small amount of non-native data available. In this paper we report on the acoustic modeling adaptation for the recognition of non-native speech. Using non-native data from German speakers, we investigate how bilingual models, speaker adaptation, acoustic model interpolation and Polyphone Decision Tree Specialization methods can help to improve the recognizer performance. Results obtained from the experiments demonstrate the feasibility of these methods.

1 INTRODUCTION

With maturing speech technology, the recognition of speech as uttered by non-native speakers of the language is becoming a topic of interest. Any deployed speech recognizer must be able to handle all of the input speech, which includes the speech from non-native speakers. Despite the large progress in fields like large vocabulary continuous speech recognition or noise robustness, recognition accuracy has been observed to be drastically lower for non-native speakers of the target language than for the native ones. One reason is because the non-native speakers' pronunciation differs from those native speakers' pronunciation observed during system training.

A number of methods for handling non-native speech in speech recognition have been proposed. The most straightforward approach is to use the non-native speech from the target language spoken by the group of non-native speakers in question for recognizer training [6], however the problem of this method is that the non-native speech data is only rarely available. Another approach is to apply general speaker adaptation techniques such as MLLR and MAP on speaker-independent models to fit the characteristics of a foreign accent [5]. Some researchers are also working on using multilingual HMM for non-native speech [4], or applied recognizer combination methods and multilingual acoustical models on the non-native digit recognition task [2].

For this study, we implement a number of acoustic modeling techniques to compare their performance on non-native speech recognition. Here we restrict our study to non-native English spoken by native speakers of German. In more detail, we explore how the acoustic models can be adapted to better handle the non-native speech. First we try to use a multilingual recognizer to do the decoding on non-native speech. Unlike [2], we are not testing on a small domain task like digit recognition, but on a conversational speech task. Furthermore we use the traditional MLLR and MAP to do the speaker adaptation experiments, with a different test setup to see how the variety among speakers will contribute to the recognition performance. Interpolation is useful in building language models for speech systems; here we explore this idea on acoustic models for non-native speech recognition. Additionally the Polyphone Decision Tree Specialization (PDTS)[1] method which was originally applied to port a decision tree to a new language in a multilingual environment; we adopt this approach for our task to see whether it can also help to improve the performance on non-native speech recognition.

This paper is structured as follows: The database is presented in section 2. In section 3, we describe the baseline system of our experiments and in section 4 we document how bilingual models, speaker adaptation, acoustic model interpolation and PDTS can help to improve the recognizer performance. Section 5 gives a brief conclusion of this paper.

2 DATABASE DESCRIPTION

Our study has been confined to sentences from German-accented speakers. We use German-accented in-house data set that has been recorded with close-head microphone. The recording scenario is based on spontaneous face-to-face dialogues in the domain of appointment scheduling. Table 1 shows the corpus and the partition for training and testing data set in this study.

Data	Partition	SPKs	UTTs	Minutes
Non-native Data	Training/adaptation	64	452	52
	Cross-validation	20	100	24
	Testing	40	260	36
Native Data	Training	2118	17000	2040
	Testing	40	312	52

Table 1 Database overview

Using the same 3-gram language model and vocabulary, the perplexity of the non-native test data is 211.27 and the OOV rate is 1.29%, the perplexity of the native test data is 323.41 and the OOV rate is 1.59%. The perplexity of native data is bigger than that of non-native data; this may come from the fact that the non-native speakers restrict themselves to smaller but well-known vocabulary and phrases in spontaneous spoken scenario.

3 BASELINE SYSTEMS

3.1 Baseline native system

All recognition experiments described in this paper use the Janus recognition Toolkit JRTK [7].

The baseline system for native English speech use acoustic models trained on 34 hours ESST data. ESST data was collected for the Verbmobil project, a long-term research project aimed at automatic speech-to-speech translation between English, German and Japanese. Here we use the first phrase of Verbmobil (VM-I) English data to do the training, the domain is limited and the speaking style is cooperative spontaneous speech, the scenario is the same as the non-native data. The baseline recognition engine is a 3-state quintphone HMM system with 48 Gaussians per state, 2000 codebooks sharing 4000 distributions. Vocal tract length normalization and cepstral mean subtraction is applied at the spectral level. Linear discriminate analysis (LDA) is used to find the most discriminated MFCC, and power features and reduces the dimension of the feature vector to 40. The WER of the baseline system on native test data is 16.2%.

3.2 Baseline non-native system

The non-native acoustic models are trained on non-native training set described in table 1. The measurement, label classes and training procedure are kept the same as those used to train the baseline native models. The non-native speech engine is a 3-state quintphone HMM system with 48 Gaussians per state, 1000 codebooks sharing 1800 distributions. Table 2 shows the performance on the non-native test set when using the native, non-native models.

Models	Native Models	Non-Native models
WER	49.3%	43.5%

Table 2 Performance on non-native test data

As expected, the non-native models perform better on the non-native set than do the native models. This result provides some assurance that the non-native models are adequately trained. We also tried a non-native system with the same number of parameters as the native system, the result is worse.

4 EXPERIMENTS

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

4.1 Pooled models

Although non-native training data is better than native training data for recognition of non-native speech, including native English data in the training set may be helpful. So our first experiment is to pool the ESST native training data and the non-native training data (see table 1) together to build so-called “pooled” English acoustic models. In this case we have much more native English data than that of non-native English data. When testing on non-native test data, the pooled models get a WER of 42.7%, performing slightly better than that of the baseline non-native system.

4.2 Bilingual models

The usefulness of multilingual acoustic models has been demonstrated before in non-native digit recognition task [2], here we extend our investigation on the recognition of non-native conversational speech.

Since we are studying the non-native data from German speakers, we use bilingual acoustic models trained earlier [3] that share training data from English part of Verbmobil (ESST) and German part of Verbmobil (GSST), to improve the robustness of the recognizer against accent of non-native speakers. We investigated the knowledge-based (IPA) approach and the data-driven approach to define a common phone set for English and German bilingual acoustic models. We achieved the best WER of 48.7% by using IPA-based bilingual models on non-native speech, only a 0.6% absolute reduction from the baseline native system.

Adding German native speech to the acoustic models without using any non-native speech seems not working well on improving the performance of non-native speech. With limited non-native data available, doing adaptation on native models is a promising way to improve the recognition performance on non-native data. So we investigate speaker adaptation, acoustic model interpolation and PDTS and compare their performance on non-native speech. We tried the adaptation experiments on both the baseline native system and the bilingual system. Since we did not see significant difference between them, we present the results from the experiments on the baseline native system.

4.3 Speaker adaptation

In speaker adaptation, acoustic models that have been trained for general speech are adjusted so that they better model the speech characteristic of a specific condition. Those adaptation techniques do not have to be limited to speaker adaptation; general models can be specialized to

compensate for differences in acoustic environment or the characteristic of a group of speakers.

Most widely used adaptation techniques include maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation. MLLR is an example of what is called transformation based adaptation, here one single transformation operation is applied to all models in a transformation class; the transformation function is estimated from a small amount of held-out data. In MAP adaptation, the model parameters are re-estimated individually, using held-out adaptation data. Sample mean values are calculated. An updated mean is then formed by shifting the original value toward the sample value. If there was insufficient adaptation data for a phone to reliably estimate a sample mean, no adaptation is performed.

We use various amount of adaptation data for our two adaptation experiments. In the first experiment, both the number of adaptation speakers and the amount of speech data are varied; the range of the speaker number is from 0 to 64. In the second experiment, the number of speakers is fixed at the maximum of 64, and only the amount of speech data is varied. For both experiments, the performances are calculated at 52, 48, 42, 37, 35, 32, 28, 22, 17, 13, and 7 minutes of adaptation data.

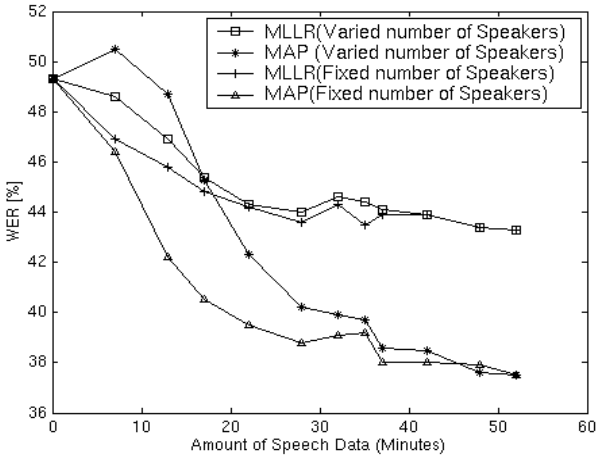

Figure 1 MLLR & MAP adaptation

Figure 1 shows the results of applying MLLR and MAP on baseline native system with non-native data. Adaptation with non-native data leads to improved performance, and as expected MAP adaptation would be the better choice as long as there is enough adaptation data. Also with the same amount of speech data, adaptation from more speakers' data is better than from that of fewer speakers, suggesting that the variety among speakers contributes more to the gain.

4.4 Acoustic model interpolation

The improvement we have seen from pooling the native and non-native training data indicated that the recognition of non-native speech could benefit from native training data (see table 2). However, the pooled training set gives very little weight to the non-native training utterances, while there are overwhelmingly more native training utterances. One way to achieve the desired weighting is by interpolating the native and non-native models.

Interpolating of acoustic models refers to the weighted averaging of the PDFs of several models to produce a single output. In this case we are combining the native and non-native models, the native models are better trained and the non-native models are more appropriate for the test data.

For our case, there are only two different models, so the interpolated model can be defined as:

$$P_I(O) = W_{Native} P_{Native}(O) + W_{Non-Native} P_{Non-Native}(O)$$

Where $W_{Native} + W_{Non-Native} = 1$; O: observed vector of acoustic features; $P(O)$: acoustic models; W: Interpolation weights.

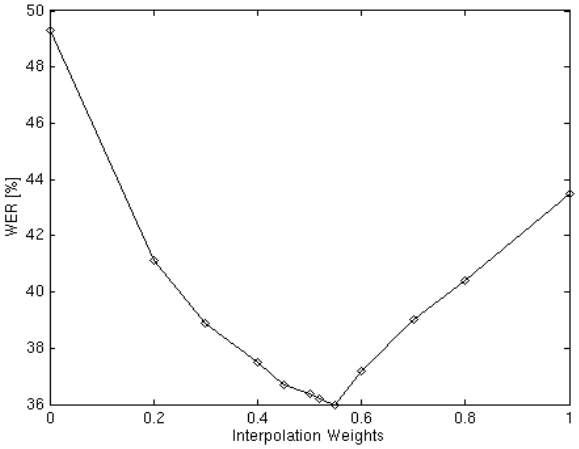

Figure 2 Performance with various interpolation weights

We create a series of interpolated models by varying the related weights assigned to the baseline native and baseline non-native models. Figure 2 shows the results of interpolated models with different interpolation weights. A weight of 0 represents the performance with the native acoustic models, and a weight of 1 represents performance with the non-native acoustic models. The optimal weighting factor was found to be 0.56 achieving 36.0% word error rate.

4.5 PDTS

Non-native speakers are known to have difficulties acquiring context-conditioned phonetic contrasts when the English phoneme is perceived as corresponding to one of their native language's phoneme that is not subject to the same variation. There is a big mismatch of the context between the speech of native speakers and that of non-native speakers. However when we do the decoding we are using the context decision tree that was built from

native speech to model the context of non-native speech. This decision tree does not represent the context of the non-native speech very accurately.

By building the tree from scratch with a sufficient amount of non-native data, one would expect to capture important patterns of allophonic distribution in accented English. The problem here is we need enough non-native training data to build the tree. In order to include questions relevant to non-native speech in the decision tree without building it from scratch, we adopt the Polyphone Decision Tree Specialization (PDTS)[1] method for porting a decision tree to a new language. This method was originally designed to overcome the problems of the observed mismatch between represented context in the multilingual polyphone decision tree and the observed polyphones in the new target language. In this approach, the clustered multilingual polyphone decision tree is adapted to the target language by restarting the decision tree growing process according to the limited adaptation data available in the target language. Each time a new language is added, it brings with its phonemes and polyphones that have not yet seen by the system. PDTS allows question to be asked about these new polyphones in the decision tree and new model mixture weights to be trained for them without discarding the questions about the polyphones that the new language share with the old one.

For the non-native speech, the recognizer selects the best acoustic match for each word during alignment, generating a list of new polyphones. The new polyphones are then integrated into the decision tree, with branches pruned back to the point where the new polyphone data could be inserted, and re-grow with new specialization where the new data show sufficient internal diversity or divergence from the native data.

After applying PDTS, the adapted decision tree represents contexts of the non-native speech data. After doing MAP adaptation using this new decision tree, the system is expected to improve the recognition results for the non-native speech. The best result we got so far from PDTS approach was 35.5%. However, further investigation need to be done using PDTS on non-native speech such as we did not change the pronunciation dictionary that limits the occurrence of new polyphones.

4.6 Summary

In this section, we show how applications of acoustic model adaptation techniques contribute to increase the recognition accuracy on non-native speech. Figure 3 compares the results for each approach. For the training data of each system, the baseline native system is trained on ESST data, the bilingual system is trained on ESST and GSST data, the baseline non-native system is trained on non-native training data, and the pooled system is trained on ESST and non-native training data. The MAP, interpolation and PDTS systems are all using the non-native adaptation data adapting on the baseline native system. The testing data is the non-native test set (see table 1). From our experiments, while using 52 minutes of adaptation data, the PDTS approach works best reducing the word error rate from 49.3% to 35.5%, a 27.9% relative reduction in error rate over the baseline native system.

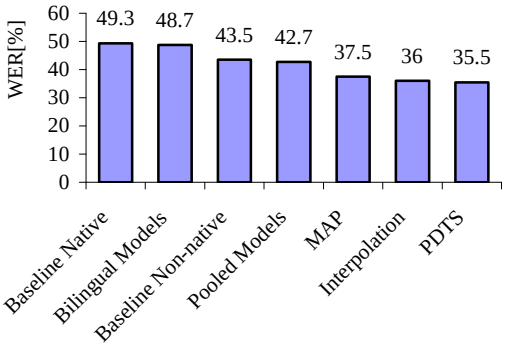

Figure 3 Best results of various systems

5 CONCLUSION

In this paper, we explore how the acoustic models can be adapted to better recognize the non-native speech. The results present in this paper show that while there are many elements of non-native speech such as the non-native pronunciation patterns that remain to be investigated, a small amount of non-native data can be used effectively in improving the recognition performance on non-native speech.

REFERENCES

[1] T. Schultz, A. Waibel. *Polyphone Decision Tree Specialization for Language Adaptation* Proc. ICASSP, Istanbul, Turkey, June 2000.

[2] V. Fischer, E. Janke, S. Kunzmann, *Likelihood Combination and Recognition Output Voting for the Decoding of Non-native Speech with Multilingual HMMs*, Proc. ICSLP, 2002.

[3] Z. Wang, U. Topkara, T. Schultz, A. Waibel, *Towards Universal Speech Recognition*, Proc. ICMI 2002.

[4] L. Mayfield Tomokiyo. *Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in Speech Recognition*. Ph.D. thesis, Carnegie Mellon University, 2001.

[5] G. Zavaliagkos, R. Schwartz, J. Makhoul, *Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition*, Proc. ICASSP, 1995.

[6] U. Uebler, M. Boros, *Recognition of Non-native German Speech with Multilingual Recognizers*, Proc. Eurospeech, Volume 2, pages 911-914, Budapest, 1999.

[7] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal, *the Karlsruhe-verbmobil Speech Recognition Engine*, ICASSP, Munich, 1997.

[8] H. Soltau, T. Schaaf, F. Metze, A. Waibel. *The ISL Evaluation System for Verbmobil II*. ICASSP 2001, Salt Lake City, May 2001.