Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization

Zhirong Wang, Tanja Schultz

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

ABSTRACT

With more and more non-native speakers speaking in English, the fast and efficient adaptation to non-native English speech becomes a practical concern. 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 effectiveness of using polyphone decision tree specialization method for non-native speech adaptation and recognition. Several recognition results are presented by using non-native speech from German speakers. Results obtained from the experiments demonstrate the feasibility of this method.

1 INTRODUCTION

With more and more non-native speakers speaking in English, the fast and efficient adaptation to non-native English speech becomes a practical concern. 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 one of the major time and costs factor for developing such a system is the need of large amount of training data, a formidable task especially for collecting data from non-native speakers. Another approach is to apply general acoustic models adaptation techniques such as MLLR or MAP on speaker-independent models to fit the characteristics of a foreign accent [5].

For this study, we adopt a polyphone decision tree specialization method for non-native speech adaptation and recognition. Here we restrict our study to non-native English spoken by native speakers of German. The polyphone decision tree specialization (PDTS)[1] method was originally designed 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 of non-native speech recognition. For the acoustic models adaptation, we use the traditional MAP algorithm, with the polyphone decision tree before PDTS and with the polyphone decision tree after PDTS respectively.

The 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 show the phonetic context mismatch between native speech and non-native speech. In section 5, the PDTS method is adopted for non-native speech, and section 6 shows the experimental results.

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	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 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 consists of a fully continuous 3-state HMM system with 2000 triphone models. Each HMM-state is modeled by a codebook containing a mixture of 48 Gaussians. The preprocessing is based on 13 Mel-scale cepstral coefficients with first and second order derivatives, power and zero crossing rates. 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 word error rate (WER) of the baseline system on native test data is 16.2%. And the WER of this system on our non-native test data set is 49.3%.

4 PHONETIC CONTEXT MISMATCH

In our system, each phone in the data is associated with a polyphone comprising that phone and the two preceding (one if the phone is word initial) and two following (one if the phone is word final) phones. In case of utterance-initial and utterance-final phones, no preceding or following phones are included in the polyphone sequence. Totally we have 375K polyphone types with different context sizes in our training corpus.

When creating the recognition system by JRTK, we use a divisive clustering algorithm that builds context querying decision trees. As selection measure for dividing a cluster into two sub-clusters, we use the maximum entropy gain on the mixture weight distributions. The recognition system uses these decision trees to find the optimal groupings and classify input speech samples during decoding.

The 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. That means there are big mismatches of the phonetic contexts between the native speech and the non-native speech. Two of the phonetic context mismatches are studied in this work, first the different pattern of polyphone usage and second the different polyphone coverage.

Figure 1 shows the pattern of polyphone usage by different group of speakers.

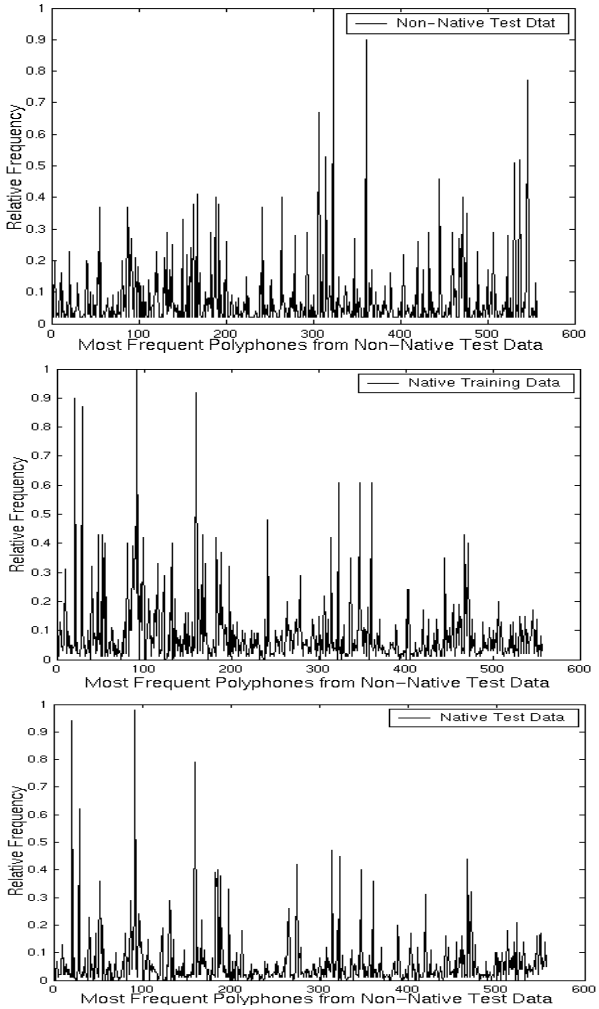

Figure 1 the pattern of polyphone usage by different group of speakers (using the same polyphone set)

For all the graphs in figure 1, x-axes are the polyphone set and y-axes are the frequency of these polyphones in different data set, here the frequency is normalized to 1. All the three graphs use the same polyphone set, which is extracted from the non-native test data set by frequency. The first graph shows the pattern of polyphone usage in non-native test data set, the second one in native training data set, and third one in native test data set. From these graphs we can see that the pattern of polyphone usage in the native test data set is very similar to that in native training data set, while the pattern of polyphone usage in non-native test data set is not.

The second phonetic context mismatch between native speech and non-native speech is the difference of

polyphone coverage. We use the following equation to calculate the polyphone coverage for each test data set:

$$PolyphoneCoverage \quad P_m / P_n * 100\%$$

P_m: The number of polyphone types in the test data that also appeared in the training corpus.

P_n: The total number of polyphone types in the test data.

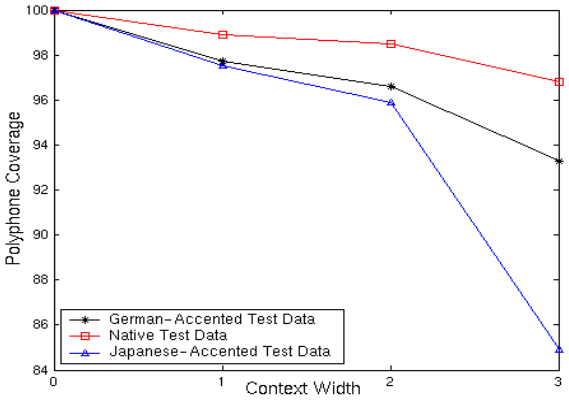

Figure 2 polyphone coverage for different test data set

Figure 2 shows the polyphone coverage of our German-accented non-native test data. For the purpose of comparison, we also calculate the polyphone coverage of the native test data set and a Japanese-accented non-native test data set. This Japanese-accented data set was recorded in our lab. From this graph we can see that the polyphone coverage is much lower for non-native test data than for the native test data.

Because of the phonetic context mismatch, the polyphone decision tree will not describe the non-native speech as well as they did for the native speech. 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.

5 POLYPHONE DECISION TREE SPECIALIZATION

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 contexts relevant to non-native speech in the decision tree without building it from scratch, we adopt the polyphone decision tree specialization (PDTS)[1] method which was originally designed for porting a decision tree to a new language in multilingual environment. The goal of this method is 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.

Because there are phonetic context mismatches between native speech and non-native speech, we hope that the PDTS method will also help to improve the recognition performance on non-native speech. Here is how the PDTS is adopted to the non-native data: 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.

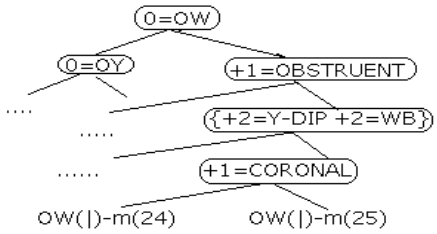

Figure 3 Part of polyphone decision tree for middle state of phone OW before PDTS

Figure 3 illustrates one part of the polyphone decision tree for the middle state of phone OW before PDTS; Figure 4 shows the polyphone decision tree after adaptation. The adapted polyphone decision tree now represents valid contexts of non-native speech data for the phone OW and is expected to improve the recognition on non-native speech.

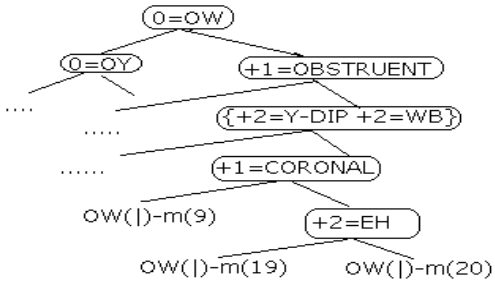

Figure 4 Part of polyphone decision tree for middle state of phone OW after PDTS

6 EXPERIMENTS

Although the adapted polyphone decision tree represents contexts of the non-native speech data after applying PDTS, the acoustic models that have been trained for general native speech are still need to be adjusted so that they better model the speech characteristic of non-native speakers. The acoustic model 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 acoustic model 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.

The goal of the experiments is to see the effectiveness of the PDTS method on improving the performance of non-native speech recognition. From previous experiments, we know that MAP adaptation would be the better choice as long as there is enough adaptation data. Also with the same amount of speech data, the variety among speakers contributes more to the gain. So we decide to use MAP as our acoustic model adaptation method, to do adaptation experiments with the original polyphone decision tree (the polyphone decision tree before PDTS) and with the polyphone decision tree after PDTS.

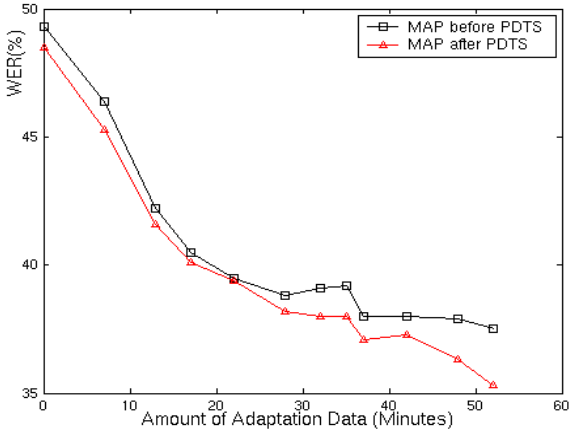

Figure 5 MAP adaptations with and without PDTS

Figure 5 shows the results of the MAP adaptations with and without the PDTS. The x-axes shows the amount of adaptation data for each experiment. The y-axes shows the word error rate. Although the amount of the adaptation data is varied during the experiments, the number of speakers is fixed at the maximum of 64. The performances are calculated at 52, 48, 42, 37, 35, 32, 28, 22, 17, 13, 7 and 0 minutes of adaptation data. For 0 minute of adaptation data, it simply means there is no MAP adaptation at all.

From figure 5 we can see that the MAP after PDTS method outperforms the MAP before PDTS method with any amount of adaptation data. While using 52 minutes of adaptation data, the MAP after PDTS method works best reducing the word error rate from 37.5% to 35.3%, a 2.2% absolute reduction in error rate over the MAP before PDTS method.

7 CONCLUSION

In this paper, we explored the phonetic context mismatch between native speech and non-native speech and showed how the polyphone decision tree specialization method could be adopted for the non-native speech. The results presented in this paper demonstrate that PDTS method could 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.