Improving ASR Performance On Non-native Speech Using Multilingual and Crosslingual Information

Ngoc Thang Vu, Yuanfan Wang, Marten Klose, Zlatka Mihaylova, Tanja Schultz

Karlsruhe Institute of Technology, Germany

thang.vu@kit.edu, tanja.schultz@kit.edu

10.21437/Interspeech.2014-3

Abstract

This paper presents our latest investigation of automatic speech recognition (ASR) on non-native speech. We first report on a non-native speech corpus - an extension of the GlobalPhone database - which contains English with Bulgarian, Chinese, German and Indian accent and German with Chinese accent. In this case, English is the spoken language (*L2*) and Bulgarian, Chinese, German and Indian are the mother tongues (*L1*) of the speakers. Afterwards, we investigate the effect of multilingual acoustic modeling on non-native speech. Our results reveal that a bilingual L1-L2 acoustic model significantly improves the ASR performance on non-native speech. For the case that L1 is unknown or L1 data is not available, a multilingual ASR system trained without L1 speech data consistently outperforms the monolingual L2 ASR system. Finally, we propose a method called *crosslingual accent adaptation*, which allows using English with Chinese accent to improve the German ASR on German with Chinese accent and vice versa. Without using any intra lingual adaptation data, we achieve 15.8% relative improvement in average over the baseline system.

Index Terms: Multilingual ASR, Non-native speech

1. Introduction

Non-native speech recognition is a very challenging task. There are many reasons why an automatic speech recognition (ASR) system which performs quite well on native speech is challenged by non-native speech. Two of them are the characteristics of non-native speech itself and the lack of training data. Some of the speaker-related factors that have negative impact on the ASR performance on non-native speech are 1) high intra- and inter-speaker inconsistency of phonetic realizations, 2) different second language acquisition methods and backgrounds, and, thus different acoustic or grammatical realizations and proficiency levels, 3) the speakers' perception of the non-native phones, 4) reading errors in read speech, and 5) slower reading with more pauses in read speech [2, 3].

There are many previous studies on handling non-native speech in speech recognition. In [3, 4], the acoustic model was adapted to each test speaker individually or to a class of non-native speakers. The adaptation was based on the direct use of MAP or MLLR. In [5, 6], the authors applied multilingual weighted acoustic models to improve recognition accuracy for non-native speech recognition. Bouselmi et. al [7] showed improvements by modifying the acoustic model using phonetic confusion rules which are extracted from a non-native speech database for a given L1 and L2 based on L1's and L2's ASR systems. Their results indicate that multilingual information might be useful to improve ASR performance on non-native speech.

In this paper, we explore the use of multilingual and crosslingual information in different ways. We investigate the effect of bilingual L1-L2 acoustic models on non-native speech. If L1 is unknown or the data of L1 is not available, a multilingual acoustic model trained without L1 training data is examined on non-native speech. For scenarios, in which the adaptation data is not available, we propose a new method called *crosslingual accent adaptation* which allows using English with Chinese accent to improve the German ASR on German with Chinese accent and vice versa. To conduct all the experiments, this paper also presents a non-native speech corpus - an extension of the GlobalPhone database [1] - which contains English (L2) with Bulgarian, Chinese, German and Indian (L1) accent and German with Chinese accent.

2. Data resources

2.1. GlobalPhone data corpus

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages [1]. It contains more than 400 hours of speech spoken by more than 1900 adult native speakers. For this work, we selected Bulgarian, German, Mandarin, and Tamil from the GlobalPhone corpus to train the bilingual L1-L2 and the multilingual acoustic models. In addition, we used English speech data from WSJ0 to build the English baseline system.

2.2. Extension 1: English with non-native accents

To conduct the experiments, a non-native speech corpus which contains 63 non-native speakers of English (approximately 10 hours) was recorded [8]. Since there are many differences between the accents of people with various language backgrounds, this research is focused on four major groups of speakers: Native speakers of Bulgarian, Chinese (Mandarin or Cantonese), German and some of the languages spoken in India (Hindi, Marathi, Bengali, Telugu, Tamil). The choice of these speaker groups was based on the availability of subjects as well as on the fact that these languages belong to different language families. Bulgarian is from the Slavic language family, Mandarin and Cantonese are members of the Sino-Tibetan language family, German is a Germanic language and the Indian languages belong to several language families, such as the Indo-European or the Dravidian language family. Each subject was asked to read English sentences which are excerpts from Wall Street Journal (WSJ). The majority of the topics are economy related news. In total, each subject within an accent read 30 unique sentences and 6 sentences that are the same for everyone. Depending on the speaker's self confidence and experience with English, the recording of the sentences took between 30 minutes and an hour. Table 1 shows some corpus statistics.

The speech data for the Chinese and Indian databases were

Table 1: *Corpus of English with the non-native accents Bulgarian (BG), Chinese (CH), German (GE) and Indian (IN)*

	Total	BG	CH	GE	IN
#speakers	63	16	17	15	15
male/female	42/21	9/7	11/6	10/5	12/3
audio length [min]	490	125	149	107	109
time/speaker [min]	7.47	7.46	8.42	7.8	7.14
#tokens	57.4k	14.3k	15.8k	13.6k	13.9k
#tokens/speaker	911	890	927	904	924
#utterances	2,368	583	640	565	580
#utts/speaker	37	36	37	37	38

collected in the USA, while the recordings of the Bulgarian and German speech data were performed in Bulgaria and Germany respectively. Although the residence time is not a sufficient indicator of language proficiency, we also collected this information for further analysis. The speakers from India have spent two years in average as residents in the USA, the Chinese speakers approximately 2.5 years. The numbers for the German and the Bulgarian databases are 4.5 months and less than a month, respectively. All speakers were at an age between 21 and 30: BG (21 - 29), CH (22 - 30), GER (22 - 30), IND (21 - 29). All recordings were made in a quiet room.
The division of the speakers that is used for the experiments is as follows: For each accent, 5 speakers were used for the test set, 5 speakers for the development set and additional 5 speakers for the acoustic model adaptation experiments.

2.3. Extension 2: German with non-native accent

To conduct crosslingual accent adaptation experiments, we collected about three hours of speech with German with Chinese accent. Chinese students at Karlsruhe Institute of Technology were asked to read about 150 German sentences selected from the German GlobalPhone database in a relatively quiet room. The recordings took between 30min and 70min per person. In total, the corpus contains 21 speakers whose ages are between 19 and 32. Their native language is Mandarin. Table 2 presents some statistical information about the adaptation, development and testing data.

Table 2: *Corpus of German speech with Chinese accent*

	Total	Adaptation	Dev	Test
#speakers	21	9	6	6
male/female	12/9	5/4	3/3	4/2
audio length [min]	186	75	52	59
time/speaker [min]	8.86	8.30	8.73	9.83
#utterances	1,057	454	301	302

3. Baseline system

The English and German baseline recognizers can be described as follows: Each system used Bottle-Neck front-end features with a multilingual initialization scheme as proposed in [9, 10]. In this approach, a multilingual multilayer perceptron (ML-MLP) was trained using training data from 12 languages (Bulgarian, Chinese Mandarin, English, French, German, Croatian, Japanese, Korean, Polish, Russian, Spanish, and Thai). To initialize the MLP training for the English and German system, we selected the output from the ML-MLP based on the IPA phone set and used it as a starting point for MLP training.
To rapidly bootstrap the system, the phone models were seeded by the closest matches of the multilingual phone inventory MM7 [11] derived from an IPA-based phone mapping. The acoustic model used a fully-continuous 3-state left-to-right Hidden-Markov-Model. The emission probabilities were modeled by Gaussian Mixtures with diagonal covariances. For context-dependent acoustic models, we trained a quintphone system and stopped the decision tree splitting process at 2,500 leaves. After context clustering, a merge-and-split training was applied, which selects the number of Gaussians according to the amount of data. For all models, we used one global semi-tied covariance (STC) matrix after a Linear Discriminant Analysis (LDA). The language model was built with a large amount of text data crawled with the Rapid Language Adaptation Toolkit [12]. The English and German ASR obtained a word error rate (WER) of 9.4% and 14.3% on the native data set, respectively. On the non-native speech data, our baseline ASR performance varies between 60.0% WER on English data with Bulgarian accent, 57.6% with Chinese accent, 62.2% with German accent, 67.5 % with Indian accent and 59.6% on German data with Chinese accent. Since the acoustic conditions of the native and non-native corpus are quite similar, we assume that the highly drop of WER from the native to non-native speech data is due to a phonetic mismatch between non-native and native speech.
We applied MAP and MLLR to our baseline system to improve the ASR accuracy for each accent. Table 3 provides an overview of our baseline system on English with different non-native accents with and without adaptation. The results show that, using MAP adaptation we gained a lot of improvement over the baseline system and much more than using MLLR. The combination of MLLR and MAP leads to the best performance on English with Bulgarian and Indian accent. Furthermore, the best WER after adapting the German ASR to Chinese accent is 43.2%.

Table 3: *Word error rates (WER) on English with non-native accents using a monolingual acoustic model*

Accents	BG	CH	GE	IN
English ASR (1)	60.0	57.6	62.2	67.5
(1) + MAP	43.1	**38.4**	**43.1**	36.1
(1) + MLLR	49.6	46.2	51.7	48.7
(1) + MAP + MLLR	**43.0**	40.2	43.6	**33.1**

4. Improving ASR performance on non-native speech using multilingual information

4.1. Bilingual L1-L2 acoustic model

Many previous studies ([7, 13, 14, 15, 16]) showed that the native language L1 has an impact on the pronunciation of L2. Therefore, it is reasonable to use not only L2 but also L1 audio data to train the acoustic model which covers the L1 and L2 phonetic space and, therefore, improves the ASR performance. We train a bilingual acoustic model for each accent using English data from WSJ0 and data from the native language from the GlobalPhone database. We merge all the phones which share the same symbol in the IPA table and apply the same training procedure as the training of the baseline system. To model more contexts, we increase the number of leaves of the decision tree to 3,000 quintphones. Table 4 shows the WER of the bilingual models on non-native test data. The results show im-

provements up to 27% relative for all accents. On top of the bilingual acoustic models, we applied MAP, MLLR and their combination for adaptation. Similar to the experiments of the baseline system, using MAP gained much more improvement than MLLR. However, in contrast to the baseline system, the combination of MLLR and MAP gives consistently some improvements in terms of word error rate for all accents.

Table 4: *Word error rates (WER) on English with non-native accent using bilingual acoustic models*

Accents	BG	CH	GE	IN
English ASR	60.0	57.6	62.2	67.5
Bilingual L1-L2 ASR (2)	53.2	52.2	45.3	60.2
(2) + MAP	38.4	34.3	36.8	34.0
(2) + MLLR	43.3	41.1	41.7	45.3
(2) + MAP + MLLR	**37.6**	**34.1**	**36.5**	**31.8**

4.2. Multilingual acoustic model

In many cases, the information about L1 is not available or the L1 data is not available. The question here is whether multilingual information still helps or not. Hence, we train four different multilingual AMs for each accent in which we omit the L1 speech data, i.e. for English with German accent, a multilingual AM is trained on English, Mandarin, Bulgarian, and Indian speech data. Table 5 summarizes the WER on the test set of four different accents. Compared to the monolingual system, we observe improvements in all cases. Except for the case of Indian accent, the WER is worse than using the bilingual L1-L2 acoustic model even if the number of parameters of the multilingual acoustic model is higher than the corresponding bilingual L1-L2 acoustic model. It indicates that L1 has a strong effect on L2 and therefore we could improve the ASR performance by using L1 speech data. However, we achieved the best WER on English with Indian accent with 29.6% by using multilingual acoustic model trained with Bulgarian, Chinese, German and English data. It corresponds to a relative improvement of about 7% compared to the bilingual L1-L2 AM. The reason could lie in the fact that the multilingual acoustic model trained with four different languages might cover more variations in the phonetic space than the monolingual and also the bilingual English-Tamil acoustic model. Since English with Indian accent has a lot of variations, it might benefit more than other accents by using this multilingual model.

Table 5: *Word error rates (WER) on English with non-native accent using multilingual acoustic models*

Accents	BG	CH	GE	IN
English ASR	60.0	57.6	62.2	67.5
Bilingual L1-L2 ASR	53.2	52.2	45.3	60.2
Multilingual ASR (3)	54.0	49.4	51.1	50.8
(3) + MAP	42.0	37.4	39.7	32.3
(3) + MAP + MLLR	**41.6**	**36.2**	**39.5**	**29.6**

5. Crosslingual accent adaptation

The approaches described in the previous sections rely on the availability of L2 speech data to adapt the background model. In this section, we describe a method called *crosslingual accent adaption* which can be applied when no such data is available.

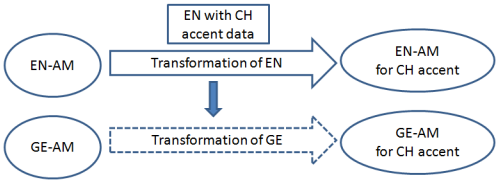

Figure 1: *Crosslingual accent adaptation* approach

5.1. Proposed Method

Typically, when an acoustic model in an HMM/GMM system is adapted to an accent, the mean and variances of all the Gaussians are modified in different ways so that the acoustic model fits better to the accent. This kind of modification is referred to as "transformation" in this paper. The idea of this proposed method is to use the transformation which was learned to adapt the native language to the non-native one across languages assuming that the accent stays the same. Figure 1 illustrates this proposed approach for the scenario in which an English and a German acoustic model should be adapted to English and German with Chinese accent. In this example, the English with Chinese accent adaptation data is available but no German with Chinese accent data is provided, i.e. 1) the transformation T which is used to adapt the English model to English with Chinese accent can be estimated using the provided adaptation data but 2) there is no chance to estimate the transformation to adapt the German model to German with Chinese accent. The key point is that the accent is the same, i.e. L1 stays the same and the effect of L1 on different L2 might share some common characteristics. Therefore, using T to adapt German models might improve the ASR performance on German with Chinese accent. This research idea allows borrowing transformations across languages for accent adaptation.

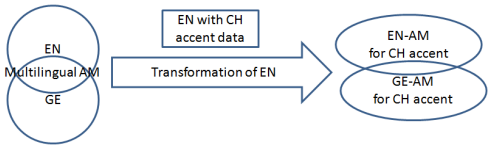

Figure 2: *Crosslingual accent adaptation* with multilingual AM

The main challenge of this proposed approach is to determine the context dependent HMM states e.g. of the German AM which should be adapted with the borrowed transformation e.g from English. Similar states between languages are a reasonable solution. To decide which states are similar, there are several possibilities. For example, the distance between Gaussian Mixtures such as Kullback-Leibler distance [17] can be used. Afterwards, those states should be adapted using the same transformation in the phonetic space. In this paper, we propose to train a multilingual model in which the states are shared between languages (see Figure 2). The phone set should be merged between languages if they share the same symbols in the IPA table. By doing that, the context dependent HMM states are merged together if they are similar during building the context decision tree of the multilingual acoustic model. Therefore, they are implicitly transformed by adapting the multilingual acoustic model to the accent. The main advantages of

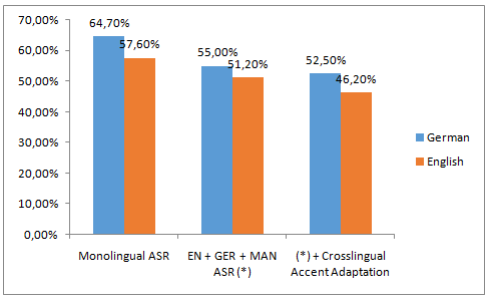

Figure 3: WER on German and English with Chinese accent

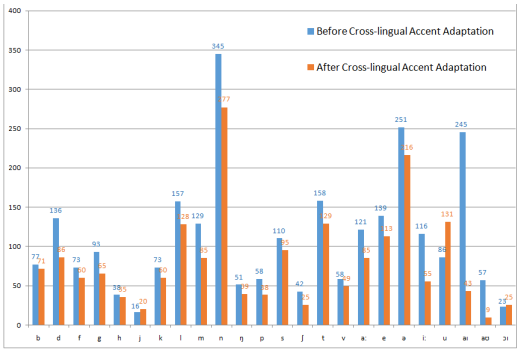

Figure 4: Substitution errors of shared phones before and after using *crosslingual accent adaptation* for German

this approach are 1) that the similarity of the context dependent HMM states across languages is determined implicitly during the training and 2) that the adaptation can be performed automatically for all the languages. Furthermore, we propose to perform only MAP adaptation since in contrast to MLLR the Gaussian mixtures of each HMM state are independently adapted. This allows us to better understand the crosslingual effect in which the performance of each shared phone can be analyzed before and after applying the proposed approach.

5.2. Experiments and results

To verify the *crosslingual accent adaptation* approach, we conducted two experiments: The first one assumed that English with Chinese accent was not available. Therefore, we used German with Chinese accent to improve the background acoustic model. In the second experiment, German with Chinese accent was not available and therefore, English with Chinese accent was utilized for adaptation. Based on the results of the experiments in Section 4.1, we used not only English and German but also Mandarin data to train the multilingual model which served as the background model in both experiments. This multilingual acoustic model has 5,000 quintphones. In our case, there are 24 phones which are shared between English and German. They correspond to 1,606 context dependent states which represent 32.12% of all the states. When English quintphone states are adapted to English with Chinese accent, all the German quintphone states which are shared with English quintphone states are also adapted implicitly and vice versa. In the first experiment, when we adapted the background model on German data with Chinese accent, 2,075 states were adapted in total. Of those, 1,367 states were shared between English and German. Compared to the first experiment, less states were adapted in the second experiment. More specifically, 1,662 states were adapted using English data with Chinese accent. 1,195 of them were shared between English and German. The reason lies within the fact that the amount of German data with Chinese accent is greater than the English one. Figure 3 summarizes the WER on English and German with Chinese accent. The results show that we achieved in total about 19.8% relative improvement on English with Chinese accent and 11.9% on German with Chinese accent without using any adaptation data of the target language compared to the monolingual baseline system. In the case of testing on English with Chinese accent, the multilingual acoustic model was adapted with German data and, therefore, more states were adapted than in the case of testing on German with Chinese accent. Therefore, it can be explained why the improvement on the English test set with Chinese accent is larger than on the German data with Chinese accent.

5.3. Result analysis

The results indicate that we can share data across L2 languages with the same accent to improve the ASR system on non-native speech. This can be applied to the case that we do not have any training or adaptation data of the target L2 language and the target accent. To obtain a better understanding of the ASR improvement, we performed an error analysis on phone level in which we compared the ASR errors of German and English with Chinese accent before and after applying *crosslingual accent adaptation*. Figure 4 shows all 24 shared phones and how often they were misrecognized for the German case. In total, we observed consistent improvements for these shared phones after applying the *crosslingual accent adaptation* approach on the German non-native test set except for the case of the phoneme /u/. The same effect was seen on the English with Chinese accent test set. It seems to be that /u/ might sound differently between English and German in the context of non-native speech even that they share the same symbol in the IPA. These results indicate that the L1 language has the same effect on different L2 languages, i.e. L1 native speakers may not be able to pronounce or wrongly pronounce the same phones in the L2 languages. Based on the experimental results and the error analysis, we can conclude that the improvement in our experiment is predictable. Since L1 native speakers may pronounce the same phones of L2 in the same way according to their accent, the accent transformation can be shared among different L2 languages.

6. Conclusions

This paper presented our latest investigations of using multilingual and crosslingual information to improve the ASR performance on non-native speech. To conduct all the experiments, a new speech corpus was presented as an extension of the GlobalPhone database containing about 8 hours of English speech data for four different accents: Bulgarian, Chinese, German, and Indian as well as about three hours of German speech with Chinese accent. We showed that bilingual L1-L2 acoustic models can improve ASR performance on non-native speech. If L1 is unknown or L1 data is not available, multilingual ASR trained without L1 speech data outperforms monolingual ASR on non-native speech. For the case that no adaptation data for the target accent is available, *crosslingual accent adaptation* provided 15.8% relative improvement in average over the baseline system. In the future, we plan to verify our approach *crosslingual accent adaptation* on language pairs which do not belong to the same language family.

7. References

[1] T. Schultz, N.T. Vu, T. Schlippe. GlobalPhone: A Multilingual Text & Speech Database in 20 Languages. In Proc. ICASSP, Canada, 2013.

[2] K. Livescu. Analysis and Modeling of Non-native Speech for Automatic Speech Recognition. Masters thesis, MIT, 1999.

[3] L. M. Tomokiyo, A. Waibel. Adaptation methods for non-native speech. In Multilinguality in Spoken Language Processing, 2001. SR: A preliminary study. In InSTIL, 2000.

[4] Z. Wang, T. Schultz and A. Waibel. Comparison of acoustic model adaptation techniques on non-native speech. In Proc. Interspeech, 2003.

[5] M. Raab, R. Gruhn, and E. Noth. Multilingual Weighted Codebooks for Non-Native Speech Recognition. In Proc. TSD, pp. 485 - 492, 2008.

[6] T.P. Tan, L. Besacier. Acoustic Model Interpolation for Non-Native Speech Recognition. In Proc. ICASSP, 2007.

[7] G. Bouselmi, D. Fohr, I. Illina, and J.P. Haton. Multilingual Non-Native Speech Recognition using Phonetic Confusion-Based Acoustic Model Modification and Graphemic Constraints, In Proc. of ICSLP, 2006.

[8] Zlatka Mihaylova. Lexical and Acoustic Adaptation for Multiple Non-Native English Accents. Diplomarbeit in Karlsruhe Institute of Technology (KIT), 2011.

[9] N.T. Vu, F. Metze, T. Schultz. Multilingual Bottle-Neck feature and Its Application on New Languages. In Proc. of SLTU, 2012.

[10] N.T. Vu, W. Breiter, F. Metze, T. Schultz. An Investigation on Initialization Schemes for Multilayer Perceptron Training Using Multilingual Data and Their Effect on ASR Performance. In Proc. of Interspeech, 2012.

[11] T. Schultz and A. Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. In Speech Communication August 2001., Volume 35, Issue 1-2, pp 31-51.

[12] N.T. Vu, T. Schlippe, F. Kraus, T. Schultz. Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit. In Proc. Interspeech, Japan, 2010.

[13] J.E. Flege. Phonetic approximation in second language acquisition. Language Learning 30, 117 - 134, 1980.

[14] J.E. Flege. The production of new and similar phones in a foreign language: evidence for the effect of Equivalence Classification. Journal of Phonetics 15, 47 - 65 , 1987.

[15] M. Gonzlez-Bueno. The effects of formal instruction on the acquisition of Spanish stop consonants. In Contemporary Perspectives on the Acquisition of Spanish. Volume 2: Production, Processing, and Comprehension, 57 - 75, 1997.

[16] J.E. Flege, E.M. Frieda and T. Nozawa. Amount of native-language (L1) use affects the pronunciation of an L2, Journal of Phonetics 25 , 169 - 86, 1997.

[17] S. Kullback. Letter to the Editor: The KullbackLeibler distance. The American Statistician 41 (4): 340341, 1987.