interspeech 2008 incorporating SST 08

Multi-Accent and Accent-Independent Non-Native Speech Recognition

Ghazi Bouselmi, Dominique Fohr, Irina Illina

Speech Group, LORIA-CNRS & INRIA
BP 239, 54600 Vandoeuvre-les-Nancy, France
{bousselm, fohr, illina} @loria.fr

Abstract

In this article we present a study of a multi-accent and accent-independent non-native speech recognition. We propose several approaches based on phonetic confusion and acoustic adaptation. The goal of this article is to investigate the feasibility of multi-accent non-native speech recognition without detecting the origin of the speaker. Tests on the *HIWIRE* corpus show that multi-accent pronunciation modeling and acoustic adaptation reduce the WER by up to 76% compared to results of canonical models of the target language. We also investigate accent-independent approaches in order to assess the robustness of the proposed methods to unseen foreign accents. Experiments show that our approaches correctly handle unseen accents and give up to 55% WER reduction, compared to the models of the target language. Finally, the proposed pronunciation modeling approach maintains the recognition accuracy on canonical native speech as assessed by our experiments on the *TIMIT* corpus.

Index Terms: non-native speech recognition, acoustic adaptation, multi-accent

10.21437/Interspeech.2008-670

1. Introduction

Nowadays, the use of automatic speech recognition (ASR) is spreading more and more as the performance of ASR systems increases. Technological advances allow ASR to be integrated into mobile applications, like mobile phones, despite their relatively limited resources. A large number of public services like call centers, ticketing or booking over the phone have already adopted ASR. Unfortunately, such applications could suffer from a flaw inherent to their nature: these wide public service applications are likely to deal with non-native speakers. ASR systems are well known to be less accurate with non-native speakers than with native ones.

Numerous approaches have already been developed to target the drastic performance drop of ASR against non-native speech. These methods can be classified into three main categories.

The *first class* deals with the acoustic adaptation approaches. The aim of such an approach is to grasp the accent deviation -produced by non-native speakers- at the acoustic level. Classical methods of acoustic adaptation, such as *Maximum Likelihood Linear Regression* (MLLR) and *Maximum a Posteriori* (MAP), are the most widely used. A small foreign-accented speech corpus is used to adapt the acoustic models of the target language (TL) ASR system to the foreign accent. A recent work presented in [1] proposed to take into account non-native accents in the very process of acoustic models training. The approach used by [1] proposed to merge HMM model states according to a phonetic confusion matrix based on the errors produced by non-native speakers.

The *second category* of methods that deal with non-native accents is the pronunciation modeling. The goal here is to detect the pronunciation errors that non-native speakers produce and take them into account in the ASR system. This pronunciation modeling consists in associating each phoneme of the TL with one (or several) alternate pronunciation(s). These alternate pronunciations can be composed to TL phones or speaker native language (NL) phones [1, 2]. A straight-forward method to integrate these alternate pronunciations is to modify the lexicon: for each word in the vocabulary, new entries are inserted, taking into account all possible pronunciations. Another approach is to modify the TL acoustic models. This could be done through merging the GMM models at the state level of both TL models and the models of their corresponding alternate pronunciations [3]. Another way of achieving this goal is to add to each TL acoustic model new HMM paths corresponding to the non-native pronunciations [2].

Finally, the *third category* consists in methods of language model adaptation to non-native speech. These methods aim at taking into account the errors produced by non-native speakers at the grammatical level. Non-native speakers might produce incorrect grammatical structures, wrongly conjugate verbs, omit syntactic connectors, or even use inexistent words.

Most approaches for non-native speech recognition rely on the knowledge of the origin of the non-native speakers and need a preliminary step for detecting that origin. In this article, we porpose a new approach that do not need this step: non-native speech recognition based on approaches that are an extension of our previous work [2]. We focus on “multi-accent” and “accent-independent” methods. By “multi-accent”, we mean an approach that is able to handle several non-native accents. By “accent-independent”, we mean an approach that is able to handle several non-native accents that are seen neither in the training nor in the adaptation. Experimental conditions and the *HIWIRE* corpus are presented in section 2. In section 3, we describe two “multi-accent” non-native speech adaptation methods. The first method is based on acoustic adaptation, while the second consists in a pronunciation modeling approach. In addition, the accents that are fed to the ASR system for recognition are among the accents that were used to adapt the system. In section 4, we describe an “accent-independent” non-native speech recognition approach. In other words, we have evaluated the robustness of both acoustic and pronunciation modeling against unseen foreign accents. We conclude this article with a brief discussion and conclusion.

Accepted after peer review of full paper

2. Experimental setup

Experiments were carried out on the *HIWIRE*[1] non-native speech corpus. It is composed of 31 French, 20 Greek, 20 Italian and 10 Spanish speakers, each has uttered 100 English sentences. The sentences follow the CPDLC[2] grammar which is a strict command language used in the communications between pilots and air traffic controllers. The vocabulary is composed of 134 application words. We have chosen an MFCC parameterization with 13 coefficients and their first and second derivatives. 40 3-state HMM mono-phone models were trained on a *TIMIT* corpus, with 128 Gaussians per state. In all the experiments, the cross-validation scheme is used and the tested speaker's data is never encountered in the foreign accent adaptation. That is to say, when testing one speaker, the entire *HIWIRE* corpus (except the data of the tested speaker) is used as adaptation/training material. In the context of this project, for speaker adaptation, the first 50 sentences are used to adapt the acoustic models and the last 50 sentences are used for testing. For all experiments, speaker adaptation is performed after non-native accent adaptation. In the tables the best results in each column are in bold face. For *HIWIRE* corpus, the 95% confidence interval is $\pm 0.4\%$ for "word error rate" (WER). For *TIMIT* corpus, the 95% confidence interval is $\pm 0.5\%$ for WER.

3. Multi-accent non-native speech recognition

It is well known that foreign accents decrease the recognition accuracy of ASR systems. When processing non-native speech, ASR systems are confronted with an inter-speaker variability that is wider and less predictable than the speaker variability of canonical TL speech. Indeed, the fluency level of non-native speakers in the TL may vary widely. This can be seen in figure 1 where the histogram of non-native speakers versus the word error rate (WER) is shown, using canonical English models. For all origins, the WER of non-native speakers varies from nearly 0% to 20%, suggesting a highly variable proficiency. We can also note that the middle fluency levels, where the WER is between 3% and 7%, include the major part of the speakers.

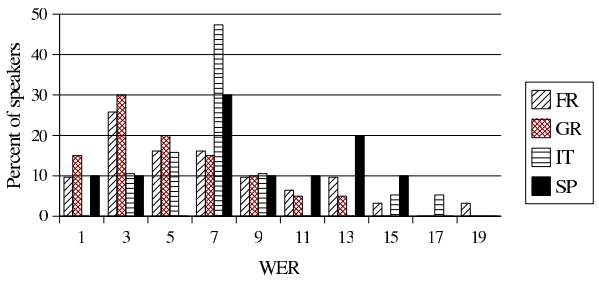

Figure 1: *Percent of speakers versus "word error rate" given by a canonical English models, HIWIRE corpus.*

Generally, non-native speech recognition approaches are intended for pairs of TL and NL [2, 3, 4]. The resulting software are based on the prior knowledge of the origins of the speakers to be processed. For an application that is likely to encounter several non-native accents, either an explicit user choice or an accent detection approach must be implemented [5]. Unfortunately in some cases, the reply delay should be short and the transaction duration is limited. Implementing an accent detection might prove to be difficult to achieve, as it requires a certain number of utterances to be captured and processed.

In [1, 6], a method for multi-accent foreign speech recognition is proposed that is based on the integration of several sets of models. The authors combine the HMM model of each TL phoneme with several models that were each adapted to one foreign-accent. The modified models are able to deal with the canonical accent and the foreign-accents under consideration. However, the method presented in [1, 6] implies important growth in the size of the models, as several models are combined for each phoneme.

In our current work, we target multi-accent non-native speech recognition in such a way that no origin detection would be needed and no excessive model growth implied. In the next two subsections, we will detail two approaches for multi-accent non-native speech recognition based on acoustic adaptation and pronunciation modeling.

3.1. Multi-accent acoustic adaptation

For non-native speech recognition, the ideal case would be to record a large foreign accented speech corpus to train accent-specific models. Unfortunately, recording such large corpora for each pair of TL and NL is very difficult to achieve. Instead, acoustic adaptation has been used to adapt TL phone models to non-native accents using a small corpus of foreign accented speech.

In classical foreign accent adaptation, acoustic models are adapted to the accent of a unique foreign origin. In our work, we propose to use acoustic adaptation for multi-accent non-native speech recognition. We use a corpus of foreign accented TL speech in order to adapt the TL acoustic models. In other words, we propose to acoustically adapt TL phone models to several foreign accents simultaneously. Through this procedure, the adapted phonetic models are expected to be closer than the original ones to each of the considered accents.

We propose the use of MAP adaptation and model retraining for multi-accent acoustic adaptation. The canonical TL HMM models are adapted simultaneously on the all foreign accents speech. In our experiments the entire *HIWIRE* corpus is used to adapt the English acoustic models to the foreign-accents (except for the data of the test speaker). MAP adaptation and model retraining are carried out following the same procedure: the labelled utterances are used to re-estimate the models' parameters. During adaptation, only means are modified, but during retraining all HMM parameters are updated.

Table 1 outlines the average results for all accents of this multi-accent adaptation. We can note that the multi-accent acoustic adaptation reduces the WER by 70%-76% (relative) compared to the canonical TL models (without speaker adaptation). One can also see that speaker adaptation benefits more to the canonical English models than to the accent adapted ones. This can be explained by the fact that the canonical models are much farther from the speaker accent than the adapted models.

[1] Human Input that Works In Real Environments, http://www.hiwire.org

[2] Controller Pilot Data Link Communications

Table 1: *Multi-accent non-native acoustic adaptation, HIWIRE corpus.*

	No Speaker Adaptation		*MLLR Speaker Adaptation*		*MAP Speaker Adaptation*	
System	*WER*	*SER*	*WER*	*SER*	*WER*	*SER*
T	7.2	14.6	4.8	10.6	2.6	6.3
T_{MAP}	2.1	5.1	1.9	4.5	**1.4**	3.4
T_{Retr}	**1.7**	**4.4**	**1.6**	**4.0**	1.5	**3.3**

- T : canonical English models (TIMIT models).
- T_{MAP} : TIMIT models adapted to multiple non-native accents through MAP.
- T_{Retr} : TIMIT models adapted to multiple non-native accents through model retraining.

3.2. Multi-accent pronunciation modeling

In [2], we have presented an approach for non-native speech recognition based on pronunciation modeling and acoustic adaptation. We propose to use this method in our current work. In the next we recall the method of [2].

The first step of the method consists in detecting the pronunciation variants using a non-native speech corpus. For that, we have used two sets of acoustic models:

- **the first set of models** represents the canonical pronunciation. We have investigated the use of several models for this set, such as canonical TL models or TL models that have been acoustically adapted to the foreign accent.
- **the second set of models** represents the foreign accent or what the speakers actually pronounced. We have also tested different sets of models for this second set, such as TL/NL models or acoustically adapted TL/NL models to the foreign accent.

The first set of models is used in a forced alignment procedure on the non-native speech corpus, in order to find the time interval where each phoneme was pronounced. Whereas, the second set of models is used in a phonetic recognition procedure in order to determine which phonemes were uttered and in which time intervals. For each sentence of the non-native corpus, the transcriptions given by the phonetic alignment and phonetic recognition are compared (time-aligned). This comparison gives the associations between each phone of the first set of models with the sequence of phones of the second set that were pronounced in the same time interval. Only the most frequent phone associations are then taken into account to form what we call "phonetic confusion rules", i.e. the most frequent pronunciation variants for each phone of the TL.

The next step in our method modifies the phones of the first set of models according to the confusion rules. To each phone model of the first set, we add new HMM paths corresponding to the concatenation of the models of the confused phones from the second set. The modified HMMs contain one path for the canonical phone from the first set along with several paths each corresponding to an alternate pronunciation. The resulting phone model is expected to represent the canonical pronunciation along with its foreign accented variants. An example of the structure of a modified HMM model is presented in figure 2 for the phoneme $[aI]_1$ of the first set of models. In this example, the phoneme $[aI]_1$ is associated with the sequences of phonemes from the second set of models: $\{[a]_2, [i]_2\}$ with a probability of 0.6, and $\{[a]_2, [e]_2\}$ with a probability of 0.4. 0.4 and 0.6 correspond to the occuring frequence of the rules in the training corpus. Here "β" is a weight that we have set to 0.5: experiments have shown that it has no effect on the ASR accuracy.

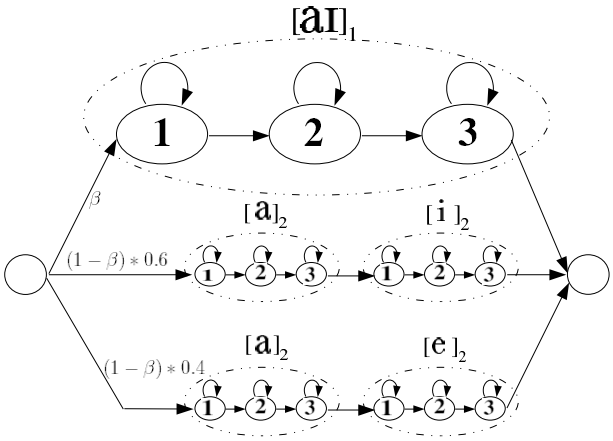

Figure 2: *Structure of a modified HMM model through pronunciation modeling.*

In [2], the pronunciation modeling is performed for each couple of languages TL/NL, using only the part of the non-native speech corpus uttered by speakers of one origin. During the recognition, the *a priori* knowledge of the origin of the tested speaker is required in order to select the suitable modified acoustic models.

We propose to extend this approach by taking into account several foreign accents in the pronunciation modeling. Instead of detecting the pronunciation variants for only one foreign accent, we propose to detect these variants simultaneously for several non-native accents. In other words, we suggest using the entire non-native speech corpus in order to perform a multi-accent pronunciation modeling. For this, we chose to use two model sets:

1. "T:T_{MAP}": using canonical TL models as the first set of models, and TL models that have been adapted through MAP to multiple foreign accents (see 3.1) as the second set.
2. "T:T_{Retr}": using canonical TL models as the first set of models, and TL models that have been retrained to multiple foreign accents (see 3.1) as the second set.

Table 2 outlines the results of the above two methods on the *HIWIRE* corpus. We can note that, as in table 1, both methods improve the ASR accuracy compared to the canonical TL models (by around 63% relative). We can also note that speaker adaptation improves the accuracy of the canonical TL models to a larger extent, for the same reasons explained in the previous paragraph. Besides, we observe that using the retrained models ("T_{Retr}") in the pronunciation modeling gives a larger error reduction than the use of the models adapted through MAP ("T_{MAP}").

3.3. Robustness to TL native speakers

In this section, we evaluate the multi-accent non-native ASR approaches on TL speech. The goal of this experiment is to

Table 2: *Multi-accent non-native pronunciation modeling, HIWIRE corpus.*

	No Speaker Adaptation		*MLLR Speaker Adaptation*		*MAP Speaker Adaptation*	
System	*WER*	*SER*	*WER*	*SER*	*WER*	*SER*
T	7.2	14.6	4.8	10.6	2.6	6.3
T:T_{MAP}	2.7	6.0	2.4	5.2	1.7	3.7
T:T_{Retr}	**2.4**	**5.4**	**2.2**	**5.0**	**1.4**	**3.4**

assess the performance of the presented methods with canonical TL speech. We evaluate the methods used in sections 3.1 and 3.2 on the test part of *TIMIT*. As we can see in the Table 3, the acoustic adaptation to non-native accents ("T_{MAP}" and "T_{Retr}") induces a large degradation of the accuracy of the models for native English speakers. On the other hand, the pronunciation modeling approach introduces only a slight degradation of the accuracy for native English speech. Indeed, this approach integrates the canonical TL models with alternate accented models to form the modified HMM models. Thus, the ASR system is able to select either the canonical or the non-native pronunciations of each phoneme. For native TL speech, the canonical part of the modified acoustic models is close to the phonemes' pronunciations and is more likely to be selected.

Table 3: *Multi-accent non-native acoustic adaptation and pronunciation modeling, TIMIT corpus.*

Approach	*System*	*WER*	*SER*
Canonical models	T	**11.3**	**34.2**
Acoustic adaptation to non-native accents	T_{MAP}	16.2	43.8
	T_{Retr}	38.1	66.9
Pronunciation modeling	T:T_{MAP}	11.6	36.4
	T:T_{Retr}	11.6	35.9

4. Accent-independent non-native ASR

In order to assess the robustness of the methods we described in section 3, we have evaluated them against unseen foreign accents. For instance, in order to test the French-accented English speech, we use Greek, Italian and Spanish accented English speech in the procedure of multi-accent ASR described in section 3. This way, the approaches are evaluated on a foreign accent that is never seen neither in the training nor in the adaptation.

The results of the robustness tests are shown in table 4 for the accent-independent acoustic adaptation and pronunciation modeling. We can see that all the accent-independent adaptation methods perform better than the canonical TL models and achieve about 55% WER reduction without speaker adaptation. Nonetheless, we observe that omitting the tested accent from the non-native adaptation corpus decreases the accuracy of the resulting models (see tables 1 and 2). The most sensitive approach to unseen accents is the acoustic adaptation through model retraining "T_{Retr}". Indeed, excluding the accent of speakers' origin from the adaptation data decreases the performance of the latter method by about 45% (WER increases from 1.7 to 3.0). All in all, the latter observations suggest that multi-accent acoustic adaptation and pronunciation modeling are quite robust in the face of unseen accents.

Table 4: *Accent-independent non-native acoustic adaptation and pronunciation modeling, HIWIRE corpus.*

	No Speaker Adaptation		*MLLR Speaker Adaptation*		*MAP Speaker Adaptation*	
System	*WER*	*SER*	*WER*	*SER*	*WER*	*SER*
T	7.2	14.6	4.8	10.6	2.6	6.3
T_{MAP}	**2.6**	**6.1**	**2.3**	**5.2**	**1.6**	**3.7**
T_{Retr}	3.0	6.4	2.5	5.3	2.2	4.3
$T : T_{MAP}$	3.2	7.0	2.9	6.0	1.8	4.2
$T : T_{Retr}$	2.8	6.5	2.4	5.5	1.7	4.0

5. Conclusion

In this article we have presented two approaches for non-native "multi-accent" and "accent-independent" speech recognition, based on acoustic adaptation and pronunciation modeling. The proposed approaches do not need detection of the origin of the speaker. We have used a non-native speech corpus containing several foreign accents in order to adapt the acoustic models. For the pronunciation modeling, the modified models are expected to handle several non-native accents with minimal growth in their size. A satisfying error rate reduction of 55% to 63% is achieved by our method. We have shown that multi-accent pronunciation modeling and MAP adaptation are robust to unseen accents. Finally, the main advantage of our pronunciation modeling approach is that recognition performance is dramatically increased for non native speakers without significantly decreasing the performance for TL speakers.

It would be interesting to assess the robustness of our methods on a larger number of foreign accents with a wider pronunciation variability.

6. References

[1] K. Bartkova and D. Jouvet, "Using Multilingual Units for Improved Modeling of Pronunciation Variants", In Proc. ICASSP, pp. 1037-1040, Toulouse, France, May, 2006.

[2] G. Bouselmi, D. Fohr and I. Illina "Combined Acoustic And Pronunciation Modelling for Non-Native Speech Recognition", In Proc. of Interspeech'07, pp. 1449-1452, Antwerp, Belgium, August 2007.

[3] J. Morgan, "Making a Speech Recognizer Tolerate Non-Native Speech Through Gaussian Mixture Merging". In Proc. InSTIL/ICALL 2004, pp. 213-216, Italy, 2004.

[4] Y. R. Oh, J. S. Yoon and H. K. Kim, "Acoustic Model Adaptation based on Pronunciation Variability Analysis for Non-Native Speech Recognition". In Proc. ICASSP, pp. 137-140, Toulouse, France, May 2006.

[5] G. Bouselmi, D. Fohr, I. Illina and J.-P. Haton, "Discriminative Phoneme Sequences Extraction for Non-Native Speaker's Origin Classification", In Proc. of the 9th International Symposium on Signal Processing and it's Applications ISSPA'07, Sharjah, UAE, 2007.

[6] K. Bartkova and D. Jouvet, "Multiple Models for Improved Speech Recognition for Non-Native Speakers", Proceedings SPECOM'2004, 9-th International Conference on Speech and Computer, St Petersburg, Russia, September 2004.