Adaptation in the Pronunciation Space for Non-Native Speech Recognition

Stefan Steidl[1], *Georg Stemmer*[2], *Christian Hacker*[1], *Elmar Nöth*[1]

[1] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
[2] ITC-irst – Centro per la Ricerca Scientifica e Tecnologica, Povo di Trento, Italy
steidl@informatik.uni-erlangen.de, stemmer@itc.it

Abstract

We introduce a new technique to improve the recognition of non-native speech. The underlying assumption is that for each non-native pronunciation of a speech sound, there is at least one sound in the target language that has a similar native pronunciation. The adaptation is performed by HMM interpolation between adequate native acoustic models. The interpolation partners are determined automatically in a data-driven manner. Our experiments show that this technique is suitable for both the off-line adaptation to a whole group of speakers as well as for the unsupervised online adaptation to a single speaker. Results are given both for spontaneous non-native English speech as well as for a set of read non-native German utterances.

1. Introduction

1.1. Motivation

A foreign accent makes automatic speech recognition even more difficult than it already is. This has been measured for instance by Glass and Hazen [1] for utterances that have been collected with the spoken dialogue system *Jupiter*: the word error rate (WER) for non-native speakers is more than twice as high than the WER for native speakers. Similar results have been obtained for users of the *Evar* spoken dialogue system in [2], where non-natives correspond to an increase in WER of 80 % relative. Teixeira et al. report in [3] a large drop in word recognition performance depending on the nationality of the speaker if a recognizer for native British English is applied to non-native speakers. While native speakers achieved more than 95 % recognition score, the recognition scores for non-natives ranged from only 15 % for German and 35 % for Italian speakers to up to 70 % for Dutch speakers. Van Compernolle discusses the main difficulties in speech recognition for non-natives [4]. The traditional approaches to increase performance in automatic speech recognition are to collect more data and to employ detailed acoustic models with sharp distributions. However, non-native speakers are no homogeneous group and their individual way to speak depends highly on their mother tongue and their speaking proficiency. Therefore it is usually impossible to collect enough data for each language pair and each level of proficiency. At the same time, recognition rates cannot be increased by detailed acoustic models as non-native speech has typically a higher variability than native speech.

1.2. Approach

In this work, we propose a new adaptation method that improves the recognition of non-native speech. The underlying assumption is that acoustic models of native speech are sufficient to adapt the speech recognizer to the way how non-native speakers pronounce the sounds of the target language. The HMM states of the native acoustic models are interpolated with each other in order to approximate the non-native pronunciation. Two different scenarios are employed for evaluation:

In the first scenario, a native English speech recognizer is adapted to non-native speech of German speakers. As all speakers share the same accent, the recognizer is adapted off-line to the whole group of speakers using a small training set of utterances. As the speakers of the training and the test set are disjoint, this experiment shows that the proposed method has the ability to adapt the acoustic models to the general properties of the German accent and not only to an individual speaker.

The second scenario is the recognition of non-native German speech. The speakers come from many different countries. Because of the large number of different accents, an off-line adaptation on a training set is impossible. Therefore, the recognizer is adapted to each speaker in an unsupervised manner. This experiment shows that the proposed adaptation method can also be applied when the accent of the test speaker is not known in advance. This is usually the case in spoken dialogue systems.

1.3. Related work

There are several ways to adapt a speech recognizer to non-native speech. One possibility is to include pronunciation variants to the lexicon of the recognizer. These variants can be created with data-driven or knowledge-based methods. For a literature survey, refer to [5]. It can be difficult to apply pronunciation adaptation when the accent of the non-native speaker is not known in advance. Furthermore, adding too many pronunciation variants to the lexicon can increase the confusability between the words. Our approach to interpolate HMM states with each other can be interpreted as a smoothed version of pronunciation adaptation: instead of substituting HMM states of the model, which is done when the phonetic transcription of a word is adapted, we only interpolate the states.

Another group of possibilities adapts the output densities of the acoustic models. Well known are for instance Maximum Likelihood Linear Regression (MLLR) or Maximum A Posteriori (MAP) adaptation [6]. For the recognition of non-native speech, acoustic model interpolation has already been used in [7, 8]. In these works, each model of a native-speech recognizer is interpolated with the same model from a second recognizer which depends on the speaker's accent. All models share one common interpolation weight. This technique has been extended in [9] so that a model can be interpolated with an arbitrary number of partners and each partner had its own interpolation weight. For the approach introduced in the next sections, no second recognizer is required. Therefore, it can also be applied when the speaker's accent is not known in advance.

A third possibility to improve the recognition of non-native speech is the adaptation of the language model. A recent

overview on language model adaptation can be found for instance in [10]. Below we will give some results for spontaneous non-native speech where an adapted language model is used, however, as this is beyond the scope of this paper, we will not give any details on this approach. Instead, the reader is referred to [2].

1.4. Overview

Our idea of interpolating acoustic models of the native speech to recognize non-native speech is explained in Sec. 2. Sec. 3 is dedicated to the details of the interpolation of semi-continuous HMMs. The corpora are described in Sec. 4, the baseline systems in Sec. 5. A description of our experiments and our results is given in Sec. 6. At the end, we draw a short conclusion and state our plans for our future work.

2. Adaptation in the pronunciation space

The goal of this work is to improve the recognition for non-native speakers. We expect that the changes of the model parameters during the adaptation process should be mainly related to changes in the pronunciation of the phones. The underlying assumption is that for each non-native pronunciation of a speech sound, there is at least one sound in the target language that has a similar native pronunciation. For instance, the pronunciation of the English /T/[1] sound by a German speaker may be located somewhere between /T/ and /s/. In our opinion, this assumption is justifiable for most language pairs, as many languages have very similar phone inventories [11]. Fig. 1 illustrates the situation by an example. The *pronunciation space* is spanned

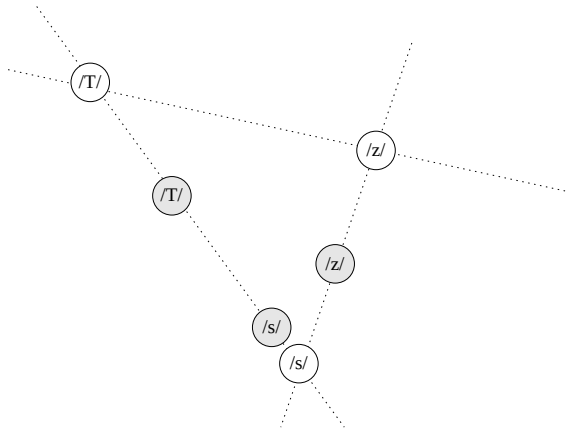

Figure 1: Pronunciation space. Acoustic models are illustrated by circles. Shaded circles represent non-native, white circles native pronunciation. The non-native models are located in lower-dimensional subspaces between several native models.

by the acoustic models trained on native speech. Points in the pronunciation space are HMMs. The optimized models for non-native speech are located between the native models. For instance in Fig. 1, the best representation for the non-native /T/ is located between the native /T/ and the native /s/. In Fig. 1 the non-native model is always in a one-dimensional subspace between two native models; of course, this can easily be generalized to higher-dimensional subspaces where the non-native model is located between several native models. In order to perform acoustic model adaptation, one has to locate a point in the pronunciation space that is optimal for the new speaker or speaker group. This is done by HMM interpolation as described in the next section.

[1] All phone transcriptions are given in the computer readable phonetic alphabet SAMPA (http://www.phon.ucl.ac.uk/home/sampa/)

3. HMM interpolation

In the following, we deal with semi-continuous HMMs. The output probabilities of semi-continuous HMMs are a mixture of M Gaussian densities. The mixture weights of an HMM state s_i are denoted by $c_{i,m}$. Assuming that the HMMs have the same number of states, we interpolate one HMM with $K-1$ partners by firstly interpolating the mixture weights of state $s_i = s_{i_1}$ with the ones of the corresponding partner states $s_{i_2}, \ldots, s_{i_K}$:

$$\forall\, m: \quad \hat{c}_{i,m} = \rho_1 \cdot c_{i_1,m} + \ldots + \rho_K \cdot c_{i_K,m} \qquad (1)$$

The interpolation weights ρ_k sum up to 1. This interpolation problem can be interpreted as an HMM as it is depicted in Fig. 2. The interpolation weights match the transition probabilities from state s_i, which is to be interpolated, to the partner states s_{i_k} and can then be estimated using the EM algorithm. Therefor, a small training set is necessary. If the interpolation weights are known, the corresponding transition probabilities can be interpolated in a similar way as in Eq. 1 using these weights. Note that it is not necessary to alter the Gaussian densities as in a semi-continuous recognizer all HMMs share the same set of densities. For details on the EM formula see [9, 2].

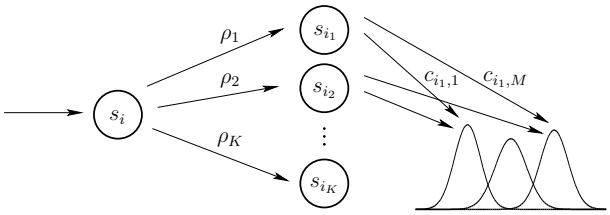

Figure 2: The linear interpolation problem (Eq. 1) can be represented by a semi-continuous HMM.

Once HMMs can be interpolated, the next problem is to select adequate interpolation partners. The number of interpolation partners determines the number of free parameters that have to be estimated from the data. For the data sets used in this work, using a single interpolation partner for each HMM ($K = 2$) leads already to good results. However, additional reductions in WER can be achieved with up to 50 partners. Good partners for an HMM can be found by interpolating this HMM with all possible candidates individually as the interpolation with only one partner is quite fast. Evaluating the benefit of each candidate leads to an n-best list. Taking the first $K-1$ entries of this list is only suboptimal: polyphones for the same phone and with similar context yield almost the same improvement but the combination of both is not better than the interpolation with only one of them. In [9], we introduced an algorithm which rejects those polyphones that are too similar to the ones already taken.

4. Corpora

4.1. Native speech corpora

Both baseline systems for native speech are trained on data from the *Verbmobil* project. *Verbmobil* was a bi-directional translation project for spontaneous speech. The German system is

trained on the *Verbmobil German* corpus which consists of 27.7 hours of spontaneous German speech of 578 different speakers. For the English baseline system, the *Verbmobil English* corpus with 22 hours of spontaneous American English speech of 260 different speakers is used.

4.2. Non-native read German speech

We use the BAS Strange Corpus 1 ("Accents")[2] for experiments with non-native German speech. The corpus consists of 1.25 hours of read speech. The 72 non-native speakers (26 female, 46 male) come from 50 countries and have 55 different mother tongues. Additionally, 16 native German speakers (7 female and 9 male) are available as a reference. All speakers read the same text (the German version of *The Northwind and the Sun*). The data is downsampled to 16 kHz.

4.3. Non-native spontaneous English speech

For experiments with non-native English speech, we use the *Verbmobil Denglish* corpus which comprises 1.5 hours of spontaneous English speech from 44 German speakers (20 female, 24 male). The signals are recorded with 16 kHz. We partitioned this corpus into a training and a test set. The training set is used for the selection of the interpolation partners and the estimation of the interpolation weights. The test set is used for evaluation. Both set are disjoint w. r. t. the speakers; each set contains 22 speakers (12 male, 10 female).

5. Baseline system

The baseline system for non-native German speech is our recognizer for spontaneous speech trained on the *Verbmobil German* corpus. It is based on semi-continuous HMMs with 500 shared full-covariance Gaussian densities. With 6825 words in the lexicon, the WER on the *Verbmobil German* test set is 20.7 %. A detailed description can be found in [2]. For our experiments, the lexicon was replaced with the 71 words of the text *The Northwind and the Sun*. Decoding is done using a unigram language model trained on this text. When applied to the German reference speakers (*BAS native*, Tab. 1), the WER is 18.5 %. If applied to the non-native speakers (*BAS non-native*), the WER increases to 34.0 %.

For the baseline system for non-native English speech, we trained our recognizer on the *Verbmobil English* corpus. For details see [2]. The lexicon was extended by the words of the *Verbmobil Denglish* corpus; the total size of the vocabulary is then 4228 words. The language models trained on the *Verbmobil English* corpus are used for decoding (a bigram for the beam search to build a word graph, a 4-gram for the A^* algorithm to rescore the graph). The WER on the *Verbmobil English* test set (*VM-English*, Tab. 1) is 35.0 %. When applied to the *Verbmobil Denglish* test set (*VM-Denglish*), the WER increases to 65.6 %.

6. Experiments and results

6.1. Recognizer for the *BAS non-native* corpus

The goal is to adapt the native speech recognizer to reduce the WER when it is applied to non-native German speech. No offline adaptation can be performed as each test speaker has a different accent. The following procedure is applied to adapt the recognizer to the current non-native speaker: Firstly, a preliminary transcription of the utterance is generated with the baseline recognizer. Only those acoustic models can be adapted which occur in this preliminary transcription. Therefore, each polyphone in the lexicon of the recognizer is replaced by the corresponding monophone. This prevents that only highly specialized polyphones are adapted. Next, based on the automatic transcription, the optimal interpolation partners for each HMM are selected and the interpolation is carried out. Finally, a second decoding pass using the interpolated acoustic models generates the final recognition result. Using only one interpolation partner, the WER can be reduced from 34.0 % to 28.3 % (Tab. 1). With 50 partners, the WER can be lowered to only 26.4 %. Both improvements are significant at a level of 0.001[3]. Fig. 3 illustrates the improvement of the WER for each speaker. The largest gains are reached for those speakers with a high WER. For speakers with a low WER, it is possible that in some cases the results are even worse than the baseline. This is due to the fact that the interpolated system is based on monophones while the baseline system uses polyphones. Using a confidence measure could help to apply our adaptation technique only to those speakers whose WER is above a given threshold.

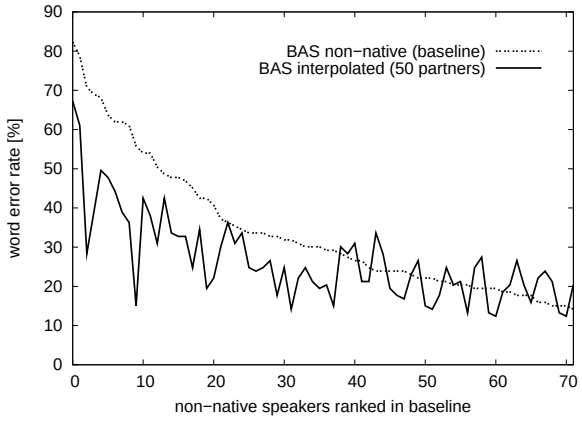

Figure 3: Results on the BAS corpus: WER for each non-native speaker for the baseline and the interpolated system (50 partners). The speakers are sorted w.r.t. their baseline performance.

experiment	WER	rel. impr.
BAS native	18.5 %	–
BAS non-native	34.0 %	0.0 %
BAS interpolated (1 partner)	28.3 %	16.8 %
BAS interpolated (50 partners)	26.4 %	22.4 %
VM-English (native)	35.0 %	–
VM-Denglish (non-native)	65.6 %	0.0 %
VM-Denglish interpolated (1 partner)	61.6 %	6.1 %
VM-Denglish adapted	57.2 %	12.8 %

Table 1: Results of the HMM interpolation on the BAS corpus and the *Verbmobil Denglish* corpus: WER and relative improvement with respect to the non-native baseline WER.

6.2. Recognizer for the *Verbmobil Denglish* corpus

The goal of the approach described here is to adapt the recognizer that has been trained on native English speech in order to reduce the WER when it is applied to non-native English

[2] Available at http://www.phonetik.uni-muenchen.de/Bas/

[3] We applied the NIST implementation of the MAPSSWE test available at http://www.nist.gov/speech/tests/sigtests/mapsswe.htm

monophone	core phone of partner
/3:/	/{/
/I:/	/i:/
/U/	/@U/
/i:/	/I/
/V/	/e/
/{/	/A:/
/S/	/z/
/T/	/s/
/D/	/d/
/dZ/	/t/
/v/	/f/
/z/	/s/

Table 2: Core phones of the HMM interpolation partners for *Verbmobil Denglish*.

speech. As the accent of the non-native speakers is the same, the HMM interpolation can be performed off-line using a training set which contains non-native speech. Experiments are done with one interpolation partner. The WER decreases from 65.6 % to 61.6 % (s. Tab. 1). In Tab. 2, the interpolation partners for a selection of monophones are given. For simplicity, only the core phones of the partners are shown.

It has been observed that non-native speakers tend to use a sentence structure different from native speakers [2, 7]. Therefore, we adapted the language model to non-native speech. From the transliteration of the *Verbmobil Denglish* training corpus, a small language model is estimated. This language model for non-native speech is combined with the one for native speech by interpolation [12, 2]. The resulting adapted recognizer for non-native speech has a WER of 57.2 % (*VM-Denglish adapted*, Tab. 1). When compared to the baseline system which has a WER of 65.6 % a relative improvement of 12.8 % has been achieved (Tab. 1). This is significant at a level of 0.001.

6.3. Discussion

Our new adaptation technique for non-native speech is based on the interpolation of the acoustic models. The results of our experiments justify the assumption that native acoustic models are sufficient to adapt to non-native speech. As *Verbmobil English* has been recorded in Pittsburgh, U.S.A., while *Verbmobil Denglish* has been recorded in Bonn, Germany, we have to consider the possibility that the acoustic models are adapted only to channel characteristics and not to the non-native accent. In our opinion, this is disproven by the list of interpolation partners shown in Tab. 2. Remember that the algorithm selects the partners in a data-driven manner. For instance, the English monophone /T/ is interpolated with a polyphone that has an /s/ as core phone. This fits to our expectation as the /T/ sound does not exist in the German language and is therefore often replaced by a similar sound like /s/. The same holds for other sounds like /D/ and /dZ/ which are hard to pronounce for German speakers.

7. Conclusion and future work

We have already mentioned in the literature review in Sec. 1.3 that often conventional speaker adaptation methods like MAP and MLLR are applied to improve the recognition rates for non-native speakers (e. g. in [6]). It would be interesting to examine to what extent these approaches can be combined with the HMM interpolation method that has been introduced in Sec. 3. Note that the interpolation method does not alter the densities in the codebook while techniques like MLLR exclusively adapt the parameters of the Gaussian densities. Thus, we can hope that the combination of both methods can lead to further improvements.

8. Acknowledgments

This work was partially funded by the European Commission (IST programme) in the framework of the PF-STAR project under Grant IST-2001-37599. The responsibility for the content lies with the authors.

9. References

[1] J. Glass and T. Hazen, “Telephone-Based Conversational Speech Recognition in the Jupiter Domain,” in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1998.

[2] G. Stemmer, “Modeling Variability in Speech Recognition,” Ph.D. dissertation, Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Germany, 2004.

[3] C. Teixeira, I. Trancoso, and A. Serralheiro, “Recognition of Non-Native Accents,” in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, vol. 5, 1997, pp. 2375–2378.

[4] D. Van Compernolle, “Speech Recognition by Goats, Wolves, Sheep and ... Non-Natives,” in *Proc. ESCA-NATO Workshop on Multi-lingual Interoperability in Speech Technology*, 1999, pp. 3–9.

[5] H. Strik and C. Cucchiarini, “Modeling Pronunciation Variation for ASR: A Survey of the Literature,” *Speech Communication*, vol. 29, pp. 225–246, 1999.

[6] G. Zavagliakos, R. Schwartz, and J. McDonough, “Maximum A Posteriori Adaptation for Large Scale HMM Recognizers,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 725–728.

[7] L. Mayfield Tomokiyo, “Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in LVCSR,” Ph.D. dissertation, Carnegie Mellon University, 2001.

[8] K. Livescu and J. Glass, “Lexical Modelling of Non-Native Speech for Automatic Speech Recognition,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2000.

[9] S. Steidl, G. Stemmer, C. Hacker, E. Nöth, and H. Niemann, “Improving Children’s Speech Recognition by HMM Interpolation with an Adults’ Speech Recognizer,” in *Pattern Recognition, Proc. of the 25th DAGM Symposium*. Berlin: Springer, 2003, pp. 600–607.

[10] J. Bellegarda, “An Overview of Statistical Language Model Adaptation,” in *Proc. Workshop Adaptation Methods for Speech Recognition*, 2001, pp. 165–174.

[11] I. Maddieson, *Patterns of Sounds*. Cambridge University Press, 1984.

[12] G. Stemmer, E. Nöth, and H. Niemann, “The Utility of Semantic-Pragmatic Information and Dialogue-State for Speech Recognition in Spoken Dialogue Systems,” in *Proc. of the Third Workshop on Text, Speech, Dialogue - TSD 2000*. Berlin: Springer, 2000, pp. 439–444.