AUTOMATIC PRONUNCIATION MODELLING FOR MULTIPLE NON-NATIVE ACCENTS

Silke Goronzy, Kathrin Eisele

Sony International (Europe) GmbH
Sony Corporate Labs Europe - Advanced Software Lab
Hedelfinger Str. 61, D - 70327 Stuttgart
{goronzy, keisele}@sony.de

ABSTRACT

This paper describes an automatic method for generating non-native pronunciations and its combination with speaker adaptation to solve the problem of a performance decrease if state-of-the-art speech recognisers are faced with non-native speech. Although being a data-driven approach it overcomes the problem of gathering accented speech data for deriving the non-native variants. It rather uses solely native speech of both languages - the language the speaker is speaking as well as his/her mother tongue. Our experiments showed that using the generated accented variants to enhance the standard pronunciation dictionary in combination with weighted MLLR speaker adaptation outperforms the baseline system by up to 20% and speaker adaptation alone by up to 3%. The approach was tested on English-accented German and on German-accented French and results were consistent throughout both languages. Since our approach relies on native speech data only, it can easily be extended to various accents for different languages.

1. INTRODUCTION

While automatic speech recognition (ASR) systems that recognise native speech have reached a certain level of maturity, the recognition of non-native speech is still a problem. Often, speech recognition performance degrades drastically if an ASR system is exposed to non-native speech. However, non-native speech is often encountered and its handling should therefore not be neglected. We develop recognisers for many different languages, each of them being frequently faced with non-native speech in real applications.

The reasons for the decrease in performance are manifold. One of the main causes is that the acoustic models (AMs) are often trained exclusively on native speech and the pronunciation dictionary only reflects native canonical pronunciations or native pronunciation variants. To a certain extent pronunciation variation can be captured by high complexity Hidden Markov models (HMMs), such as triphones with many mixtures, as was shown e.g. in [1]. Jurafsky et al. [2] however showed that while this is true for some kind of variation, such as vowel reduction and phoneme substitution, other variation, such as syllable deletion, cannot be captured in this way. Another important result of their research is, that modelling pronunciation variation in this way requires an increased amount of training data reflecting the considered variation, which is in case of non-native speech particularly difficult to obtain.

Furthermore, non-native speakers tend to replace certain constraints of the language they are trying to speak — the *target language* — by constraints they know from their native language — the *source language*, cf. [3]. That means that syllable deletions and insertions and other perturbation of the phonological structure are an important issue that should not be neglected and that are inappropriately modelled just by increasing the complexity of the AMs.
Using standard acoustic speaker adaptation algorithms is another attempt to improve performance for non-native speakers. While taking into account that certain phonemes are pronounced differently, the fact that non-native speakers might use different phoneme sequences than expected is totally neglected. To reflect this the pronunciation lexicon needs to be extended to also account for such non-native pronunciations. This extended dictionary should then be combined with the adaptation of the AMs to account for both — phonemes pronounced in a non-native way as well as non-native phoneme sequences. In [4] and [5] the necessity to combine both techniques is also underlined.

The main difficulty however is, how to derive the non-native pronunciation variants. Traditional approaches either require phonetic knowledge or accented databases to derive pronunciations or pronunciation rules. Since we are interested in flexibly modelling various different accents, we introduced a method in [6] that works automatically. It is thus very flexible concerning the kind of accent we want to model. It does neither need linguistic knowledge, nor accented databases but exclusively requires native databases of the source and the target language. These are used to automatically derive pronunciation variations that reflect the target language spoken with the accent of the source language. It is thus very flexible and can easily be applied to many languages.

We further developed this approach and tested the performance on various accented corpora and found that extending the dictionary in the proposed way can improve the recognition results for non-native speech. We then combined the extended dictionary with weighted Maximum Likelihood Linear Regression (MLLR) speaker adaptation which further improved the results.

The next section describes the different possible approaches to model pronunciations in general, i.e. native and non-native pronunciations, before our approach of automatically generating non-native pronunciations is explained in detail in Section 3. In Section 4 the experimental set-up that was used for our recognition experiments is described. Section 5 then discusses the results before a summary of this paper is given in Section 6.

2. PRONUNCIATION MODELLING

The research work done in the past concerning the modelling of pronunciation variation mainly concentrated on native speech. In [7] an extensive overview over the different techniques is given. There the source of pronunciation variation information is used to divide the different approaches into knowledge-based and data-driven ones.

2.1. Knowledge-based approaches

These approaches make use of existing knowledge that was derived by experts. This can be dictionaries or results from linguistic studies on pronunciation variation. The gathered information is often used to derive rules that are able to generate typical pronunciation variants from canonical pronunciations or from the orthography of a word, cf. [8, 9]. The pronunciation variants that are generated by applying the rules can then be added to the dictionary.

The advantage of this approach is that it is completely task-independent, since it uses general linguistic and phonetic rules and can thus be used across corpora and especially for new words that are introduced to the system.

The drawback however is that the rules are often very general and thus many variants are generated, some of which might not be observed very often. On the other hand the existing knowledge might not cover all aspects that are needed for the current task and thus not enough variants might be generated. Furthermore no information on how often the generated variants appear in the data under consideration is given.
For the task under consideration in this paper using a knowledge-based approach would require to generate a separate set of rules for each pair of source and target languages that is to be captured. Our goal is to be able to generate pronunciation variants for a variety of source and target language combinations, so it is very likely that not for all of these the required knowledge is available. As a consequence for this particular task the rules would have to be generated manually which would require expert knowledge about both the source and the target language and would be a very costly and time consuming task.

2.2. Data-driven approaches

Data-driven approaches try to derive the pronunciation variants directly from the speech signal. This can help avoiding over-generation since only variants that really occur in the data are used. It furthermore allows the computation of application likelihoods. On the other hand this method is very much database-dependent and variants that occur frequently in one speech corpus do not necessarily occur frequently in other corpora.

Besides the manual phonetic annotation of the corpus with the actual spoken phoneme sequence the variants can e.g. be derived by using a phoneme recogniser, cf. [1, 10]. In this case one is often faced with high phoneme error rates, which requires further processing of the generated variants. The highest achievable phoneme recognition results without using any further constraints for the recogniser have been between 50 and 70%, cf. [13]. Further restrictions could be made by using phoneme bi- or trigrams, but since the goal is to find unknown variants, the freedom of the recogniser should be as high as possible.

In order to filter out erroneous pronunciations and to obtain a certain generalisation capability to new words the generated variants are often not used as they are but rules can be generated therefrom, cf. [11]. Similar to the method proposed in this paper, [1] and [12] grow decision trees that are supposed to learn a phoneme-to-phoneme mapping from a canonical pronunciation to pronunciation variants, with the variants for decision tree training derived from a corpus. In [15] decision trees are used for pruning the variants but also confidence measures could be used to select only the reliably recognised variants, cf [10].

Amdal et al. [14], who also consider non-native speech, use statistics on co-occurrences of phonemes to remove erroneous variants.

In our case, for finding non-native pronunciation variants, the traditional data-driven method would require the collection of accented databases for various combinations of source and target languages which is again an extremely costly and time-consuming task.

A general problem for both knowledge-based and data-driven methods is that adding the found variants to the baseline pronunciation dictionary can increase confusability which can potentially lead to an increase in word error rate (WER).

3. AUTOMATICALLY GENERATING NON-NATIVE PRONUNCIATIONS

Since none of the approaches seems to be appropriate for applying it to various source and target language combinations as is our intention, we propose a fully automatic method for generating non-native pronunciations. It is a data-driven approach but in contrast to traditional approaches does not need any accented speech data, but solely native data for the source and the target language. In [6] we introduced this approach that tries to simulate some observations that were made when investigating the acquisition of a new language. Here we extend this method to various languages.

3.1. Simulating the reproduction of foreign words

Before we describe the new approach it needs to be mentioned that the area of language learning, comprising the acquisition of a second, third or more languages, is a research area of its own (see e.g. [16]). Learning a language other than the mother tongue is influenced by a lot of factors. People can learn a language by listening to other people speaking the target language, they can read texts written in the target language and study the pronunciation of the words and the grammar of the language. And there are many more factors that influence the learning process.

Our approach clearly does not attempt to model all of these different, interfering processes. It rather attempts to simulate one of the many — namely to listen to other (preferable native) speakers of the target language and to try to repeat what was just heard. Simulating solely this we try to find out whether this is sufficient to generate accented variants that can improve the performance of our speech recognition system for non-native speakers.

When trying to reproduce what was just heard in a foreign language, we consider that several things might happen: If the speaker does not know the target language at all he might use many

- phonemes he knows from his mother tongue and
- phonotactic constraints inherent to his mother tongue, resulting in ’wrong’ phoneme sequences

In reality, depending on the proficiency level, the speaker might not exclusively use source language phonemes but rather a mixture between source and target language phonemes, i.e. target language phonemes with a different realisation. Also only certain phonotactic constraints will be applied and where possible those of the target language will be used.

In [16] it was shown that experienced non-native speakers of English (who were native speakers of French) tend to approximate phones that have a direct mapping to their native language to the native pronunciation, while this is not the case for phones that do not have a direct counterpart in the native language. However, if they do not have much experience with the non-native language, the pronunciation deviates a lot, also for phones without a direct counterpart in the native language. Our approach attempts to simulate exactly this case of non-experienced speakers.

One example for the phonotactic constraints that we think need to be accounted for could be Italian speakers speaking English. Due to the constraints of the Italian language they often insert vowels at word endings to transform closed syllables to open ones, e.g. "team /t i: m/ → /t i: m @/" [1], cf. [17]. Also the native pronunciation of the English word "linked" is "/l I N k t/" but for Italian speakers a vowel insertion as in "/l I n k @ t/" can be observed quite often. If any kind of speaker adaptation was applied without taking this into account, the actually inserted /@/ would either be assigned to the /k/ or /t/ or both in the recogniser segmentation. The speaker adaptation would thus adapt /k/ and /t/ with partly wrong speech data. This should explicitly be avoided by taking the accented pronunciation into account.

Because of these two above mentioned processes we think that both adaptation of the AMs and adaptation of the dictionary should be applied, with the speaker adaptation accounting for a different realisation of certain phonemes, the latter one accounting for the 'wrong' phoneme sequences.

In our attempt to simulate the described listening and reproducing process, we use the following set-up: Let's assume the target language is German and it is supposed to be spoken by native English speakers. To simulate the English-accented reproduction of German words, we use a phoneme recogniser based on HMMs that were trained on British English. The English phoneme recogniser is used to decode German speech. This is also depicted in the first part of Figure 1. There the word 'Alarm' is spoken by a native German speaker (indicated by using the German canonical pronunciation[2] as input to the phoneme recogniser). If all German words in the database are processed like this, we will get English transcriptions for the German words. In our case we recognised several repetitions of German command words uttered by several native German speakers. The 'English realisation' of the German phonemes is accounted for by using HMMs trained on British English. The English phonotactic restrictions are applied to the phoneme recogniser by using an English phoneme n-gram model. Although we described that in approaches of (native!) pronunciation modelling using a phoneme recogniser the freedom of the recogniser should be as high as possible to find unknown variants, here we are particularly interested in those phoneme sequences that are influenced or caused by the source language.

The resulting English-accented phoneme transcriptions are then used to train a decision tree that learns a mapping from the German canonical pronunciation to an English-accented variant which is also indicated in the first part of Figure 1. The second part of the Figure then shows the final word recognition experiments that use a native recogniser with an augmented dictionary and recognises non-native speech.

The experiments that were described exemplarily for English-accented German were additionally conducted for German-accented French.

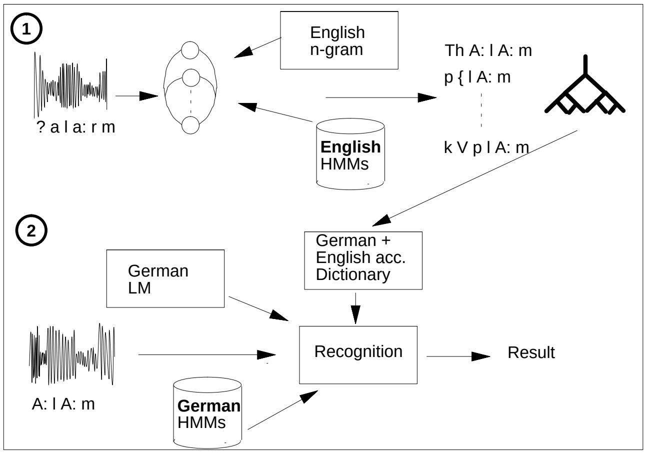

Fig. 1. 1: Procedure for generating non-native pronunciations 2: Using the augmented dictionary to recognise non-native speech

4. EXPERIMENTS

In this section we will first describe the general set-up and preprocessing of the speech data that we use in all of the experiments described in this paper. Then we will explain how we conducted the phoneme recognition experiments that generated the accented pronunciation variants. Section 4.2 outlines how decision trees are applied to learn a mapping from the pronunciation of the target language to an accented pronunciation of the target language with an accent of the source language. Finally in Section 4.3 we describe those experiments that expose our 'native' recognisers for German and French to non-native data and how the results can be improved by enhancing the corresponding dictionaries with the previously generated variants.

The speech data that was used in our experiments was recorded in our noise free studio and mainly consists of isolated words and short phrases. The data we recorded were isolated commands or short phrases (that would be suitable for any kind of device control) in many different European languages. The vocabulary size in all languages was 375 commands. All speech data were sampled at 16 kHz and coded into 25ms frames with a frame shift of 10ms. Each speech frame was represented by a 38-component vector consisting of 12 MFCC coefficients and their corresponding first and second order time derivatives. The energy was not used, but its first and second order time derivatives.

4.1. Phoneme recognition experiments

In order to generate the accented variants we built two different phoneme recognisers:

- To derive English-accented German: English phoneme recogniser (to recognise German speech)
- To derive German-accented French: German phoneme recogniser (to recognise French speech)

Please note that for each of these experiments we used only native speech data: The source language to train the HMMs for the phoneme recogniser and the target language to be recognised by the phoneme recogniser. Both phoneme recognisers used 3-state monophone speaker independent HMMs with one Gaussian

[1] Using the British English SAMPA notation

[2] Using the German SAMPA notation

density. We chose monophone rather than triphone models on purpose since we assume that for speakers that are not used to speak the foreign language, the level of 'native' co-articulation will be neglectable, especially because we are considering isolated speech. Furthermore we wanted to explicitly exclude effects of higher complexity models modelling pronunciation (although this might improve results) to be sure that potential improvements result from our approach rather than from sophisticated AMs. Also it was shown by He [18] that for non-native speakers of English a not very sophisticated AM produced better results than models of higher complexity. The detailed statistics for HMM training are summarised in Table 1. Phoneme label transcriptions were used to train a phoneme bi-gram for each recogniser.

language	# speakers	# utterances
English	152	60.720
German	129	76.790

Table 1. Phoneme recogniser HMM training data statistics

4.2. Generalisation using decision trees

In order to be able to deal with high phoneme error rates we trained decision trees to generalise from the data generated by the phoneme recognisers. For each source - target language pair one decision tree was trained using the native phoneme sequences (as included in the native pronunciation dictionary) as input and the accented variants (the output from the phoneme recognisers) as the desired output to learn a mapping between these two phoneme sequences, for example a mapping from the native French pronunciation to the German-accented French pronunciation. Table 2 shows the exact number of training phoneme sequences (utterances) used for the two decision trees.

language	# speakers	# utterances
German (English phone rec)	81	18.630
French (German phone rec)	19	9.100

Table 2. Training data statistics for the two decision trees

For decision tree training we used C5.0 [19]. We used a context of three phonemes to the left and to the right for the phoneme under consideration. Also we used the last predicted phoneme. We applied 'boosting', which is a special method to improve the classification rates of the decision tree classifier by training several classifiers from the same data set. The first classifier is trained in the usual way, yielding a certain classification rate. The second classifier is then trained using the same data set but especially focusing on those cases that were erroneously classified by the first classifier. A third classifier is trained and so on, until a predefined number of classifiers, in our case ten, is reached. In the end all classifiers are combined to make the final prediction.

After the described training procedure, the final decision tree classifier was applied to the native canonical pronunciations in the dictionary of the native German and French recogniser. Thus for each native pronunciation, one accented variant is generated and added to the dictionary (yielding two variants per word that means doubling the number of dictionary entries). This resulted in two different dictionaries: A German one that now included English-accented variants and a French one including German-accented variants.

To be able to use the standard HMMs for each recogniser a mapping of the non-native phonemes to the native phoneme set was necessary. This mapping was applied to the generated accented variants that were generated by the decision tree. The closest German phonemes for English were those used in the SMARTKOM project, cf. [20] and determined by a phonetician for French.

It should be mentioned at this point, that adding more than one accented variant to the dictionary should be investigated in the future since for some words more than one variant might be necessary. However, this would require a method to select the variants depending on the speaker to avoid an increase in confusability.

4.3. Recognising non-native speech

Since the overall goal is to be able to handle as many accents as possible, we wanted to change as little as possible in our standard speech recognisers, to keep the systems manageable also if various accents are accounted for simultaneously. Thus, adding the accented variants to the native dictionaries of each recogniser is the only modification that is done to our recognition systems.

The HMMs and the grammar used remain the same as if only native speech was recognised i.e. they are trained on German or French speech, respectively. However, the test set now comprises accented speech of 10 English speakers speaking German and 16 German speakers speaking French. All speakers spoke the target language with a moderate to strong accent. The strength of the accent was determined manually by a phonetician, who listened to the speech data and categorised the speakers. Again the test set comprised approximately 234 commands per speaker (approximately two words per utterance that were treated in the grammar as one entry). The dictionary itself covered a set of 375 different commands.

It should be mentioned that the accented speech data was not recorded for the purpose of these experiments but was already available. We decided to use our speech corpora since it is very difficult to obtain publicly available databases that contain accented speech from a lot of speakers of the same mother tongue. Those databases that include accented speech (e.g. ISLE, WSJ Spoke3), often recorded a variety of non-native speakers, but all having different mother tongues. Also often continuous speech was recorded. We considered isolated speech to be more suited here, since we wanted to exclude any co-articulation effects at this stage of testing the validity of the approach. Furthermore by using our databases that contain exactly the same kinds of commands for all involved languages and that were recorded under the same conditions, we tried to eliminate effects on recognition rates that might arise from different database settings as much as possible.

For training the German HMMs, 179 speakers (76.790 short utterances) and for the French HMMs 112 speakers (53.598 utterances) were used.

For each target language two sets of experiments were conducted. First the native baseline dictionary was compared to the dictionary extended with one accented variant per word. Then weighted maximum likelihood linear regression (MLLR) speaker adaptation was applied to both settings in online, incremental mode as described in more detail in [6]. In contrast to standard MLLR [21] the weighted approach is capable of successfully adapting the AMs on very small amounts of data, such as a few (in our case every four) isolated words. In the standard approach, several sentences

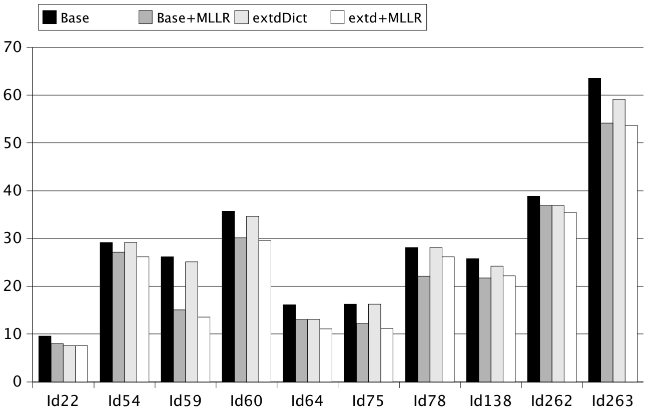

Fig. 2. WERs for English speakers speaking German

of adaptation data are required for adaptation, otherwise performance will decrease. The results we obtained are shown in the following section.

5. RESULTS AND DISCUSSION

Figure 2 shows the final recognition results in WER for all English speakers speaking German after the dictionary was enhanced with the predicted variants. We call this dictionary 'extended dictionary'. In the same figure the results for using these dictionaries in conjunction with weighted MLLR adaptation are shown.

First of all, it can be seen that extending the dictionary with the accented variants (third bar 'extdDict') is better compared to the baseline dictionary ('Base') for seven out of ten speakers and for three the results remain unchanged. It is not surprising that the improvements are not greater since with this experiment we accounted for the 'accented' phoneme sequences only, not paying any attention to the deviating phoneme realisations. These can be taken into account by using the weighted MLLR speaker adaptation, as shown in the second and fourth ('Base+MLLR' and 'extd+MLLR') bar of Figure 2, respectively, where 'Base+MLLR' means the baseline dictionary combined with online, incremental MLLR adaptation and 'extd+MLLR' means the extended dictionary combined with online, incremental MLLR adaptation.

For all speakers MLLR adaptation can improve the results as compared to the German baseline system even on these small amounts of data. This is what we expected since this adaptation improves the AMs and therefore accounts for the 'accented' phoneme realisations. When combining the extended dictionary with MLLR, the results can be further improved for eight out of ten speakers. Only for speakers Id78 and Id138 MLLR alone performs better, although the combination of the extended dictionary and MLLR is still better than the baseline for both. For all other speakers the combination yields the best results. This indicates that even though the extended dictionary alone could not improve for three speakers, in combination with MLLR it yields the biggest overall improvements.

Figure 3 show the results for 16 German speakers speaking French. Please note that even though partly the same speaker IDs are used, for different languages they also mean different speakers. The results show the same tendency as before: for all speakers except speaker Id36, MLLR adaptation outperforms the baseline dictionary. The extended dictionary alone is worse than MLLR and for nine speakers even than the baseline dictionary, but interestingly in combination with MLLR achieves the best results for 12 out of 16 speakers. Only for speakers Id24, Id25, Id33 and Id34 the 'extd+MLLR' results are worse than MLLR alone.

To get a clearer picture of the overall performance of our method, Table 3 again summarises the results averaged over all speakers. It shows that the extended dictionary alone is better than the baseline but not than MLLR alone for English-accented German and worse than the baseline and MLLR alone for German-accented French, but in combination with MLLR the best results can be achieved for both settings.

Language	Base	Base +MLLR	extdDict	extd +MLLR
Eng-acc German	28.91	24.06	27.42	23.66
Ger-acc French	19.63	16.17	20.56	15.69

Table 3. Overall recognition results averaged over all speakers

The extended dictionary was also tested on our native reference test sets (not shown in the Figure), in order to see whether for native speakers performance degrades if we simply double the number of pronunciations in the dictionary. The baseline WER averaged over all 15 speakers in the German reference test set of 11% increased to 11.5%. For the French recogniser, the baseline WER of 12.3% on a test set of 15 native speakers even decreased to 11.9%. Both is quite acceptable and means that the confusability was not increased too much in both cases. If we compare the baseline WERs of 11% and 12.3% to the baseline WERs for non-native speech of 28.9% and 19.6% in Table 3, this underlines how much the performance degrades if the native recognisers are faced with non-native speech.

Summarising the results we can state that the method we developed for generating non-native pronunciation variants is able to generate variants that can improve the recognition results for non-native speech. However, just adding these variants to the dictionary does not help for all speakers. This is what we expected since when we add non-native phoneme sequences we do so using 'native' phonemes. We also have to account for the differing realisations of phones by applying weighted MLLR speaker adaptation. Then we can improve the results for the majority of the speakers and the results are better than MLLR alone. The validity of the approach was demonstrated on two different language pairs. Results could be improved by up to 20% for some speakers compared to the baseline and by up to 3% compared to MLLR alone. We are currently experimenting with Spanish-accented German and preliminary results show the same tendency.

6. SUMMARY

In this paper we proposed a new approach to automatically generate non-native pronunciation variants and its combination with weighted MLLR speaker adaptation to improve the performance of speech recognisers for non-native speech. The automatic method we propose is a data-driven approach that uses only native speech data for the two languages under consideration and tries to simulate a speaker learning a foreign language by listening to native speakers and reproducing what was just heard. A phoneme recogniser trained with data from the source language is used to

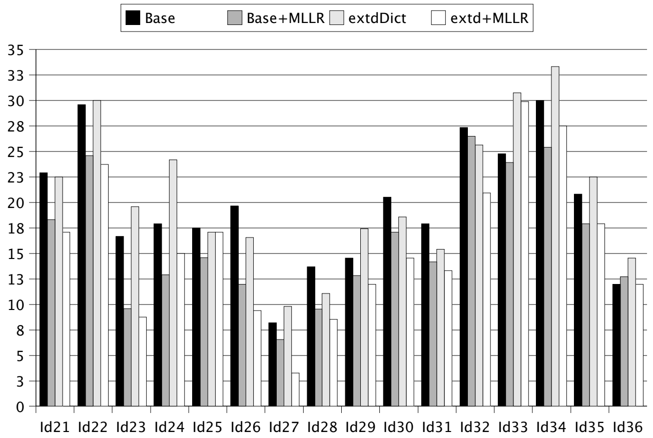

Fig. 3. WERs for German speakers speaking French

decode speech data of the target language to generate pronunciations for the target language spoken with the accent of the source language. A decision tree is then used for generalisation. In the final recognition experiments where recognisers trained on native speech were faced with non-native speech, we enhanced the dictionary with the generated accented variants and combined this with weighted MLLR adaptation. We tested our method on English-accented German and on German-accented French. For all experiments the WERs could be reduced by around 20% as compared to the baseline dictionary and by up to 3% compared to MLLR alone.

This shows that for non-native recognition tasks it is necessary to account for the different phonetic realisation (accounted for by the speaker adaptation) as well as for potentially different phonotactic structures in non-native speech (accounted for by adding accented pronunciation variants to the dictionary). The results also revealed that for a few speakers results did not improve which indicates that not all generated variants are appropriate for all speakers and a careful (and automatic) selection of the variants for each speaker would be beneficial. However, the proposed method, that in contrast to traditional methods does not need any accented speech, which is difficult to obtain, is very flexible and can be applied to generate variants for any kind of accent.

7. ACKNOWLEDGEMENTS

This research was conducted within the SMARTKOM project and partly funded by the German Federal Ministry of Education and Research under grant 01IL905I7.

8. REFERENCES

[1] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolie, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic Pronunciation Modelling from hand-labelled phonetic Corpora" *Speech Communication*, vol. 29, pp. 209–224, 1999.

[2] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What Kind of Pronunciation Variation is Hard for Triphones to Model?" *ICASSP 2001*, Salt Lake City, USA, vol. 1, pp. 577–580.

[3] S. M. Witt and S. J. Young, "Off-line Acoustic Modeling of Non-native Accents" *Eurospeech99*, Budapest, Hungary, pp. 1367–1370.

[4] P. C. Woodland, "Speaker Adaptation: Techniques and Challenges" *ASRU*, Colorado, USA, IEEE, 1999, vol. 1, pp. 85–90.

[5] C. Huang, E. Chang, J. Zhou, and K.-F. Lee, "Accent Modeling Based on Pronunciation Dictionary Adaptation for Large Vocabulary Mandarin Speech Recognition" *ICSLP2000*, Beijing, China, pp. 818–821.

[6] Silke Goronzy, *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition*, Number 2560 in Lecture Notes on Artificial Intelligence. Springer Verlag, 2002.

[7] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature" *Speech Communication*, vol. 29, pp. 225–246, 1999.

[8] S. Downey and R. Wiseman, "Dynamic and Static Improvements to Lexical Baseforms" *ESCA Workshop on Modeling Pronunciation Variation*, Rolduc, 1998, pp. 157–162.

[9] A. Kipp, M.-A. Wesenick, and F. Schiel, "Automatic Detection and Segmentation of Pronun. Variants in German Speech Corpora" *ICSLP96*, Philadelphia, USA, pp. 106–109.

[10] G. Williams and S. Renals, "Confidence Measures for Evaluating Pronunciation Models" *ESCA WS on Modeling Pronunciation Variation*, Rolduc, 1998, pp. 151–155.

[11] N. Cremelie and J.-P. Martens, "Automatic Rule-Based Generation Of Word Pronunciation Networks" *Eurospeech97*, Rhodes, Greece, pp. 2459–2462.

[12] J. J. Humphries, P. C. Woodland, and D. Pearce, "Using Accent-Specific Pronunciation Modelling for Robust Speech Recognition" *ICSLP96*, Philadelphia, USA, pp. 2324–2327.

[13] Renato di Mori, Ed., *Spoken Dialogues with Computers*, Academic Press, 1998.

[14] I. Amadal, F. Korkmazskiy, and A. C. Suredan, "Data-driven Pronunciation Modelling for Non-Native Speakers using Association Strength between Phones" in *ISCA WS Automatic Speech Recognition and Understanding*, Sophia-Antipolis, France, 2000, vol. 1, pp. 85–90.

[15] E. Fosler-Lussier, "Multi-Level Decision Trees for Static and Dynamic Pronunciation Models" *Eurospeech99*, Budapest, Hungary, pp. 463–466.

[16] J.E. Flege, "The production of "New" and "Similar" phones in a foreign language: evidence for the effect of equivalence classification" *Journal of phonetics*, vol. 15 1, pp. 47–65, 1987.

[17] S. Goronzy and M. Sahakyan and W. Wokurek, "Is Non-Native Pronunciation Modelling Necessary?", *Eurospeech 2001*, Aalborg, Denmark, vol. 1, pp. 309–312.

[18] X. He and Y. Zhao, "Model Complexity Optimization for Nonnative English Speakers", *Eurospeech 2001*, Aalborg, Denmark vol. 2, pp. 1461–1463.

[19] "http://www.rulequest.com" .

[20] "http://www.smartkom.org" .

[21] C. J. Leggetter and P. C. Woodland , " Speaker Adaptation of HMMs using linear regression, *Technical Report*, Cambridge Univ., 1995.