ISCA Archive
http://www.isca-speech.org/archive

ITRW on Adaptation Methods
for Speech Recognitionn
Sophia Antipolis, France
August 29-30, 2001

Generating Non-Native Pronunciation Variants for Lexicon Adaptation

Silke Goronzy, Ralf Kompe, Stefan Rapp

Sony International (Europe) GmbH
Advanced Technology Center Stuttgart, Man Machine Interface Lab
Heinrich-Hertz-Strasse 1, D-70327 Stuttgart, Germany
{goronzy,kompe,rapp}@sony.de

Abstract

Traditional approaches to model pronunciation variations either require expert knowledge or extensive speech databases. In the cases where non-native speech is considered they are too costly, especially if a flexible modelling of various accents is desired. We propose to exclusively use native speech databases to derive non-native pronunciation variants. We use a phoneme recognizer to generate English pronunciations for German words and use these to train decision trees that are able to predict the respective English-accented variant from the German canonical transcription. In first experiments we achieved promising results using the enhanced dictionary for decoding accented-data.

1. Introduction

It is a known problem for nowadays speech recognition systems to handle non-native speech. Usually, recognition rates decrease drastically. The reasons for this are manyfold. One of the main causes is, that the HMM models are trained exclusively on native speech in the majority of the cases. Even if non-native speech is included for HMM training, only a limited number of accents can be covered this way. Using too many different accents would result in too diffuse models that would not be able to handle native speech adequately anymore. Standard acoustic speaker adaptation algorithms can be used to improve performance for non-native speakers. While taking into account that certain phonemes are pronounced differently, they totally neglect the fact that, non-native speakers often use different phoneme sequences than expected. They 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. [1]. To reflect this, the pronunciation lexicon needs to be modified to also account for such non-native pronunciations.

The research work done in the past concerning the modelling of pronunciation variation mainly concentrated on native speech. It can be divided into 3 different categories:

Rule-based approaches try to find rules that can generate typical pronunciation variants from canonical pronunciations, cf. [3, 4]. This involves expert knowledge and would have to be done separately for each pair of source and target languages under consideration. Furthermore the rule sets are often too general and generate too many variants.

Data-driven approaches try to derive the pronunciation variants directly from the speech corpora. This avoids over-generation, but on the other hand is very much data base-dependent. The variants can be derived by using a phoneme recogniser. In order to deal with the high phoneme error rates, Amdal, cf. [5], who also considers non-native speech, uses statistics on co-occurrences of phonemes to remove erroneous ones. In [6] decision trees are used for this purpose. For finding non-native pronunciation variants however, the data-driven method would require the collection of accented databases for various combinations of source and target languages, which is an extremely costly and time-consuming task.

Combined approaches also use speech corpora but instead of directly deriving the variants, rules are derived from the corpus to ensure the generalisation capability to new words, cf. [7].

It is a general problem for all approaches, 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).

None of the approaches seems to be appropriate to be applied to various source and target language combinations. Kat and Fung recently used only native data of two languages for adapting the HMM models to a foreign accent, cf. [2]. In order to be as flexible as possible w.r.t. the source and target language, we also propose to use solely native databases for the derivation of non-native pronunciation variants. Using native corpora of the source and target language, we generate pronunciations that represent the target language spoken with the accent of the source language.

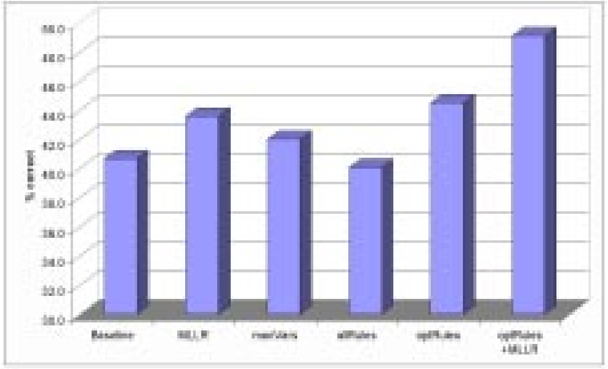

Figure 1: Results on the ISLE corpus

2. Feasibility Study

In order to investigate the influence of pronunciation adaptation, we examined the Interactive Spoken Language Education corpus (ISLE), cf. [8]. It contains speech from German and Italian second language learners of English.

For parts of the ISLE data, manual transcriptions of the actually spoken phoneme sequences were available in addition to the canonical transcriptions. Since we are interested in evaluating the effect of adding relevant non-native variants to the dictionary, we used the manual and canonical transcriptions to derive pronunciation rules by inspection. These were applied to automatically generate the corresponding German- or Italian-accented variant from the native British English pronunciation. A detailed description of the rule set can be found in [9].

Twelve speakers with a perceivable accent were chosen for testing. Figure 1 shows the results averaged over all twelve speakers. The first bar is the baseline recognition rate using our British English recogniser and the baseline dictionary that included the English canonical pronunciations only.

The rather poor baseline performance can be explained by the fact that the only data available for language model training were the transcriptions of the read sentences. Since the goal of the ISLE project was to build a language learning tool, the sentences focused on special problematic cases and were thus very different in structure. Our interest however, is focused on the acoustic-phonetic side, so we did not investigate this issue any further. In a second experiment online, unsupervised MLLR speaker adaptation, cf. [10], was used together with the baseline dictionary, which could improve the results. The third bar ('man Vars') shows the results, if the dictionary was enhanced with the manually found variants directly. It is important to note, that only for two thirds of the utterances, these manual variants are available, i.e. only for these words variants were included (yielding on the average 1.5 pronunciations per word).

The baseline recognition rate can be improved, however, not as much as with MLLR adaptation. The fourth bar ('allRules') shows the result if all pronunciation rules were applied to the baseline dictionary, automatically generating non-native pronunciations for all words (on the average 3.2 pronunciations per word). In this case the results are worse than the baseline, underlining that adding too many variants increases confusability and thus the WERs. Since not all rules were equally effective for all speakers, an optimal set of rules was determined for each speaker by testing each rule separately. Only those, that had a positive effect on the WER were used to construct a speaker-specific dictionary. The results of this experiment are shown in the fifth bar ('optRules'). It can be seen that the improvements are bigger than using the manually found variants and than MLLR speaker adaptation. Finally, the speaker-optimised dictionaries were combined with MLLR speaker adaptation ('optRules+MLLR'). Note that the combination of MLLR with the speaker-optimised dictionaries is clearly better than using MLLR or the optimised dictionaries separately.

As stated previously this rule-based approach is not suited if we want to treat a variety of source and target languages but the experiments clearly demonstrate that this kind of pronunciation adaptation is beneficial. Furthermore the results indicate that a speaker-wise selection of the rules is necessary and that the improvements of MLLR adaptation and speaker-optimised dictionaries are at least partly additive and that these methods should therefore be combined.

3. Decision Tree-Based Generation of Non-Native Pronunciation Variants

Since the methods described in the previous section do not offer the desired flexibility, we follow a completely different approach. The basic assumption behind it is that people speak as they hear. If speakers of the source language try to learn the target language, they will, after having heard a word spoken by a native speaker, try to reproduce it. They will tend to use the phonemes and phonotactics known from the source language, somehow assimilating them to what they heard. This time the used language pair is the same than before but the languages are combined the other way round. That means we investigate how English native speakers speak German. We simulate this process by a two-step procedure that is depicted in Figure 2. First, a phoneme recogniser based on HMM models that were trained on the source language (British English) is used to decode speech in the target language (German). This results in English transcriptions for the German words. In our case we recognized several repetitions of German command words uttered by several native speakers of German. Phonotactic restrictions are applied to the phoneme recogniser by an English phoneme n-gram model.

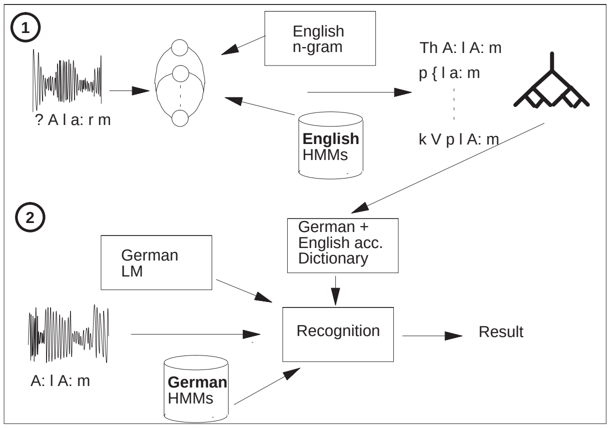

Figure 2: 1: Method to generate non-native pronunciations. 2: Final recognition system with enhanced dictionary

The transcriptions found are used to train a decision tree, that maps the standard German pronunciation to the English variants. The trained tree is then used to predict English-accented variants from the German canonical ones, which are added to the baseline dictionary of the German recognition system. To be able to do so, a mapping of the English phonemes to the German phoneme set is necessary. This mapping was applied to the generated variants after decision tree training. In a second step, the English-accented German speech is then recognised with the standard German system, using the enhanced dictionary.

Some of the pronunciations, that were generated by the English phoneme recogniser for the German word ’Aktuelles’ (engl. current topics) are given below. Please note that the phoneme sets used are the German and the British English one, respectively, in SAMPA notation.

German word	German canonical	English variant
Aktuelles	? a k t u: ? E l @ s	{ k t w e l @U s
	? a k t u: ? E l @ s	{ k t w 3: m @ z
	? a k t u: ? E l @ s	k t w 3: l @ s
	? a k t u: ? E l @ s	{ k t w e l @U z
	? a k t u: ? E l @ s	{ t k w 3: r @ s
	? a k t u: ? E l @ s	@ k w e@ r e s

Though being quite different, it can be seen, that among the English-accented variants there are some consistencies that are likely to be captured by appropriately trained decision trees.

4. Experimental setup

The phoneme recogniser employed to derive the English variants for decision tree training used models that were trained on the British English Wall Street Journal (WSJ). For both, the German and the English recognisers we used speaker-independent, 3-state, one-mixture monophone models with one Gaussian per state. We chose monophone rather than triphone models for both recognisers on purpose, since we assumed that for non-native speakers there will not be a high degree of co-articulation. For German, 50 speakers (21434 short utterances) and for English, 44 speakers (7861 utterances) were used for the training.

The German database was recorded in our noise free studio and mainly consists of isolated words and short phrases. Both, the English and the German speech 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 time derivatives. The energy was not used, but its first and second time derivatives.

The test set consisted of eight native English speakers, speaking German with a more or less severe accent. A set of commands for all kinds of consumer devices was used for testing. It comprises on the average 234 utterances per speaker (approximately 2 words per utterance).

5. Results

Due to the usage of a phoneme recogniser to generate the English-accented pronunciations, we are faced with high phoneme error rates. When testing the phoneme recogniser on the British English WSJ data the recognition rate was roughly 50%. As a consequence we have to try to eliminate the erroneous phonemes from the generated transcriptions. We generally assume that, if a phoneme is ’correctly’ recognised, it will appear more often in the different generated transcriptions for a word than a phoneme that was erroneously recognised. So part of the erroneous phonemes will be automatically ’removed’ through the generalisation that is done during the training of the decision tree. However, by looking at the generated transcriptions we also found a lot of words, for which there was not much consistency in the generated English transcriptions. In order to be able to make reasonable predictions also for these cases, we added several correct German transcriptions to the English decision tree training set, hoping that these would ’win’ for those parts for which no consistent English transcriptions were found. The best results were so far achieved by adding two German pronunciations to approximately 100 English pronunciations.

Figure 3 shows the final recognition results in WER after the dictionary was enhanced with the predicted variants (yielding two variants per word, that means doubling the number of dictionary entries). Also the results for combining these dictionaries with our weighted speaker adaptation method, cf. [10] are shown.

First of all, it can be seen that, for all speakers MLLR adaptation can improve the results. Using the dictionary

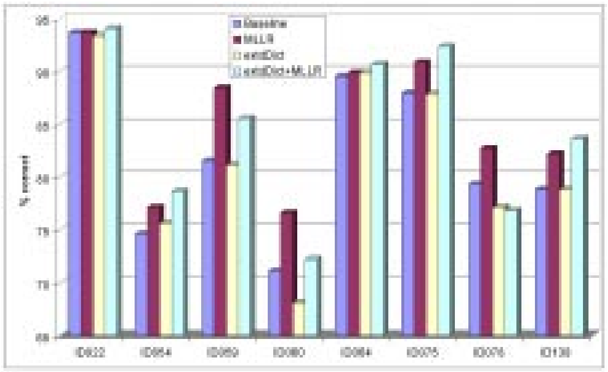

Figure 3: Results using the extended dictionary

enhanced with the generated non-native variants ('extdDict') is worse than the baseline for most of the speakers. Note that the number of entries in the dictionary was doubled in this case. When combining the extended dictionary with MLLR ('extdDict+MLLR'), the results are better than using the baseline dictionary with MLLR for five out of eight speakers. It is clear to see that some kind of speaker adaptation is necessary, since we use German models for decoding and even if we include the relevant 'accented phoneme sequences' there will definitely be a mismatch between the models and the phonemes used by the English speakers.

The results demonstrate that our approach is capable of automatically generating appropriate non-native pronounciations, without using any non-native speech data. Unfortunately for three speakers, performance decreases. Please note that the strength of accent is very different in our test set, which is reflected by the very different baseline results. We do not account for these differences when we construct the extended dictionary.

As we have learned form our experience with the ISLE database, a speaker-wise selection of rules is clearly superior to using all rules, in order to keep confusability as low as possible. As a consequence instead of directly generating variants by applying the decision tree to the baseline dictionary, rules need to be generated that allow a speaker-wise selection.

6. Summary and Outlook

In this paper a new approach for pronunciation adaptation for non-native speakers was presented. It requires no non-native speech data to generate non-native pronunciation variants.

English-accented variants for German are generated by decoding German speech with an English phoneme recognizer and training a decision tree on the found variants. Recognition experiments with native English speakers speaking German were conducted. It was shown that the German dictionary, that was enhanced with the automatically derived English-accented pronunciations and combined with MLLR speaker adaptation, achieved better results than the baseline dictionary combined with MLLR adaptation for the majority of the speakers. The proposed approach is very flexible w.r.t. the combination of various source and target languages without the need of acquiring accented data. Traditional approaches are much more limited in this respect.

In accompanying experiments it was shown, that a speaker-wise selection of pronunciation rules is beneficial. We therefore expect to further improve the results by generating rules, that can be specifically chosen for each speaker. A method that allows the online selection of pronunciation rules for each speaker, depending on the strength of his accent, is currently under investigation.

7. Acknowledgments

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

8. References

[1] Witt, Young. Off-line Acoustic Modeling of Non-native Accents. In Eurospeech99, Budapest, pages 1367–1370.

[2] Kat, Fung. MLLR-based accent model adaptation without accented data. In ICSLP2000, Beijing, pages 738–741.

[3] Downey, Wiseman. Dynamic and Static Improvements to Lexical Baseforms. In WS on Modeling Pronunciation Variation, pages 157–162, Rolduc 1998.

[4] Kipp, Wesenick, Schiel. Pronunciation Modeling Applied To Automatic Segmentation Of Spontaneous Speech. In Eurospeech97, Rhodes, pages 1023–1026.

[5] Amadal, Korkmazskiy, Suredan. Data-driven Pronunciation Modelling for Non-Native Speakers using Association Strength between Phones. In *ASRU2000*, volume 1, pages 85–90. ISCA, 2000.

[6] Fosler-Lussier. Multi-Level Decision Trees for Static and Dynamic Pronunciation Models. In Eurospeech99, Budapest, pages 463-466.

[7] Cremelie, Martens. Automatic Rule-Based Generation Of Word Pronunciation Networks. In Eurospeech97, Rhodes, pages 2459–2462.

[8] http://nats-www.informatik.uni-hamburg.de/˜isle/.

[9] Sahakyan. Variantenlexikon ital. und dt. Lerner des Englischen für die autom. Spracherkennung. Master's thesis, IMS, University of Stuttgart, May 2001.

[10] Goronzy, Kompe. A MAP-like weighting scheme for MLLR speaker adaptation. In Eurospeech99, Budapest, pages 5–8.