Is Non-Native Pronunciation Modelling Necessary?

Silke Goronzy, Marina Sahakyan
Sony International (Europe) GmbH
Advanced Technology Center Stuttgart, MMI Lab
Heinrich-Hertz-Str. 1, 70327 Stuttgart, Germany
{goronzy,sahakyan}@sony.de

Wolfgang Wokurek
IMS, University of Stuttgart
Azenbergstr.12
70174 Stuttgart, Germany
wokurek@ims.uni-stuttgart.de

Abstract
It is difficult to recognize accented or non-native speech with speech recognition systems that are trained using native speech. While standard acoustic speaker adaptation techniques are often applied in these cases, they can only reduce the recognition errors that are due to mispronunciations on the phoneme level. They are not able to handle severe deviations from the expected pronunciation. Also, there has been a lot of interest in native pronunciation modelling recently. However results often were not as good as expected. This paper concentrates on non-native speakers and examines, if a special treatment of these speakers is necessary. The effect of adding special non-native pronunciation variants to the pronunciation lexicon is investigated. In contrast to native pronunciation modelling the results show that for the non-native case the enhanced dictionary is really necessary to obtain acceptable recognition rates. Recognition rates on the Interactive Spoken Language Education corpus (ISLE) were improved by up to 10% for German and even up to 28% for Italian learners of English. When combining this with Maximum Likelihood Linear Regression (MLLR) adaptation, these results can be further improved.

1. Introduction
It is a well-known problem that state-of-the-art speech recognition systems encounter drastic increases in word error rate (WER), when they are faced with non-native speech. However, the ability to deal with non-native speech becomes more and more important. A speech recognition or dialogue system is almost always limited to a certain number of languages. As a consequence, the probability that the system will be used by a non-native speaker of one of the supported languages will be high, especially if public systems, like information systems, are concerned. One reason for the high error rates is that most of the speech recognition systems use HMM models that are trained using native speech, but speakers speaking with a foreign accent tend to use the phonemes they know from their native language, cf. [Wit99], thus introducing a mismatch between the speech and the HMM models. One solution could be to include non-native speech during HMM training, cf. [UB99], or to use standard speaker adaptation techniques, cf. [Dig97]. While part of the problem can be remedied by these techniques, both approaches do not take into account that non-native speakers also tend to sometimes use completely different phoneme sequences than expected.

The major characteristic of foreign accented speech could be described as a deviating realization of corresponding phonological units that can often be traced back to specific features of the respective speakers’ native language. German speakers of English are likely to have problems with pronouncing the dental fricatives. These segmental aspects of foreign accent may be accompanied by non-segmental aspects, provided there is a substantial difference between the native and the foreign language. A strong Italian foreign accent, for example, is often noticed to add /@/ to closed syllables, creating additional open syllables. This particular phenomenon concerning consonant-vowel (CV) structure can thus be classified as affecting phonotactic and rhythmic aspects of speech.

From this point of view, our approach is to improve the speech recognizer by a modified phonetic lexicon. We generate such modified lexica for the ISLE corpus, cf. [ISL]]. This corpus contains non-native speech of German and Italian learners of English. By detailed analysis of the corpus a set of non-native pronunciation rules was generated and applied to the baseline dictionary. The derived pronunciation lexica were evaluated and showed that the performance can be improved by up to 28% for individual speakers. A problem that remains is to limit the automatically generated variants to a reasonable number. However, this is necessary to keep the confusability in the lexicon as low as possible. Hence, the most relevant phonotactic rules are identified by testing their individual contribution (improvement or deterioration) for each speaker.

The results show that a speaker specific selection of rules is necessary depending on the problems the individual speaker has while trying to speak the new language. In contrast to recent results on native pronunciation modelling, where the positive effect that was achieved by including variants was often eaten up by the increased confusability in the lexicon, our experiments show that for non-native speech the results are much better. The combination of these speaker adapted pronunciation lexica with online, unsupervised MLLR speaker adaptation further improved the results.

2. The ISLE corpus
The speech corpus that was used in this study was recorded for the ISLE project. The goal of this project was to develop a tool that helps learners of English to improve their pronunciation and thus their intelligibility. Therefore 23 German and 23 Italian speakers were recorded reading English sentences. The sentences to be recorded were carefully chosen so as to cover those cases, that are known to be problematic for German and Italian speakers, e.g. vowel reduction, vowel lengthening and others. For parts of the corpus manually corrected phoneme labels are available, that reflect the phoneme sequences that were actually spoken by the speakers. On this basis a ‘manual word recognition rate’ was computed, that counted the words that contained no phoneme errors. Figure 1 shows these recognition results can be further improved.
rates together with the results, that were obtained by our baseline automatic speech recognition (ASR) system. Speakers are numbered and their native language is indicated by appending either ‘G’ or ‘I’ for German and Italian, respectively. It was expected that the fewer errors were made by the speakers, the closer their level of speaking English will be to a native speaker. As a consequence the recognition rates of the ASR system are expected to be higher, than for speakers, who make a lot of errors. But the results show that nearly every speaker of this set is judged better by the human listener. This indicates that often it is a bigger problem for ASR systems to deal with non-native speech, than it is for humans.

It can be further observed, that in general the Italian speakers seem to have more difficulties in pronouncing English phonemes than the German speakers. A second thing is, that the pronunciation quality of a speaker is only to a certain extend related to the recognition rate of the ASR system. However, a clear tendency of its influence on the ASR system can be observed.

A careful inspection of the speech files of the test speakers revealed a lot of other factors occurring in this corpus, that might be problematic for ASR systems. Sometimes the speech of certain speakers was rated quite high in terms of word correct rate by the transcribers, but loud background noises resulted in bad recognition rates.

3. Derivation of pronunciation rules

Our goal was to find from the data pronunciation rules, that are able to predict non-native pronunciations from the canonical, native pronunciation. Therefore, a set of non-native pronunciation rules has been derived manually by examining the manually corrected phoneme labels and comparing them to the expected canonical English pronunciations. The most important rules, from the phonological point of view, are described in the following. How often the phoneme obeying the rule was involved in errors is reported as ‘occurrence’. We use SAMPA notation.

For German speakers

- Word final devoicing: German speakers have difficulties in pronouncing final voiced obstruents in English (in German all final obstruents are voiceless).
  Occurrence: 14%

- Vowel Reduction: This takes place in English in unstressed syllables. The vowel that is mostly affected is /ɜ/. In German unstressed vowels are, with few exceptions, pronounced with their full quality.
  Occurrence: 21.7%

- Vowel /ɪ/: German speakers replace /ɪ/ by the nearest vowel /i:/ in words like “shall”, “pan”. These mispronunciations are typical examples of intra-lingual mistakes.
  Occurrence: 3.6%

- /w/: German speakers have difficulties in pronouncing /w/, as this sound does not exists in modern German. They often replace /w/ by /v/ in words like “workforce”. Some German speakers also replace /w/ by /wl/. This is termed hypercorrection, e.g. /w / i: w E l/ for “very well”.
  Occurrence: /wl/ to /v/: 2.9%, /v/ to /wl/: 1.0%

- /Ng/-problem: This consonant combination does not exist in German. Therefore the second consonant /ɡ/ is very often deleted at morpheme boundaries, e.g. “finger”.
  Occurrence: 0.2%

- /D/-problem: German (also Italian) speakers have great articulatory difficulty in pronouncing the dental fricative /D/, especially in combination with other fricatives. They replace it by the plosive consonant /d/ in words like “then” /d /n/.
  Occurrence: 0.8% for German, 5% for Italian speakers

For Italian speakers

- Vowel Lengthening: Since Italian speakers have difficulties in perceiving the distinction between the two sounds, often /l/ is replaced by /l/.
  Occurrence: 10%

- /@/-problem: Most Italian speakers append a /@/ at word final consonants.
  Occurrence: 13.5%

- /l/-problem: Italian speakers have difficulties in pronouncing /l/ in word initial position and thus it is often deleted.
  Occurrence: 1.7%

The vowel reduction turned out to be a very important rule, however its application to the baseline dictionary was not as straightforward as for the other rules. While for the other rules a mapping on the phoneme level was directly possible, adding the variants that were due to the vowel reduction required a back off to the orthography. In English many vowels when being reduced are replaced by /ə/ in unstressed syllables. As stated above, especially German but also Italian speakers have difficulties in realizing this /ə/ and tend to pronounce it in its full quality. This realization depends on the grapheme and its context that was found in the orthography.

4. Experimental Setup

Twelve speakers (6 German and 6 Italian speakers) were chosen for testing. They were selected carefully, since we are especially interested in the ‘problematic’ speakers. Bad recordings needed to be excluded, to ensure that the bad performance was mainly caused by a strong accent rather than some background noises, etc. The speech was 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 mel-frequency cepstral (MFCC) coefficients and their corresponding first and second time derivatives. Energy was not used, but its first and second time derivatives. The data base used for training of the one-mixture monophone models was the British English.
Figure 2: Using variants from transcriptions, German speakers

Figure 3: Using variants from transcriptions, Italian speakers

Figure 4: Speaker-wise best rule set, German speakers

Wall Street Journal (WSJ). We used very simple models on purpose, since we assumed that for non-native speakers there will not be a high degree of co-articulation. The language model (LM) was trained solely on the sentences available from the transcriptions. The corpus is separated into so-called blocks, that correspond to the exercise types of the language teaching tool. Thus the structure of the sentences was very different, each time focusing on the particular problem faced in the exercise. Since extremely long as well as extremely short sentences were included in the corpus, it is clear that the used LM was not optimal and that a more detailed design could clearly improve the results. However, since the main focus of the presented research was the pronunciation lexicon, this was not investigated any further. The perplexity of the LM was 5.91. The baseline dictionary contained native, canonic pronunciations only and consisted of 810 entries.

5. Experiments and Results

5.1. Adding manually derived variants

The first experiments tried to investigate the effect of adding relevant variants, that means variants that really occurred in the corpus. So the variants from the manually corrected label files were directly added to the lexicon. Figures 2 and 3 show the results for the baseline ASR system and after all German and Italian variants, respectively, were added. The number of variants per speaker was 1.2 and 1.8 if the German and Italian rules were applied, respectively.

The baseline recognition rates are relatively poor, especially for Italian speakers. One main reason is that the only data available for training the LM was the data of the transcriptions, as mentioned in the previous section, and the fact, that we used very simple HMM models. Unfortunately the ISLE corpus does not contain any native English speakers for reference, that would help to optimize the baseline system. However, the results from the manual transcriptions indicate, that the speakers selected for testing are really ‘problematic’ and thus such error rates for the ASR system are reasonable.

It can be seen that the recognition rates can be improved for almost all speakers in the test set if the lexicon is enhanced with the corresponding variants of that respective language. Interestingly also for some German speakers the Italian variants improved the results and vice versa, indicating that the necessary variants depend not only on the mother tongue of a speaker but also on the speaker himself. Please note, that only for parts of the data, manual phonetic transcriptions were available, that means that only for words, that occurred in these parts pronunciation variants were added to the lexicon.

5.2. Adding automatically generated variants

In a next step, the variants were automatically generated by applying the pronunciation rules that were generated as described in Section 3. As can be seen from the third bar in Figures 4 and 5, simply applying all rules often increases the WER compared to the baseline dictionary. In a second step, all rules were separately applied to the baseline dictionary and tested. It turned out that some rules only improved the results for very few speakers but caused an increase in WER for all other speakers. Thus all speakers were evaluated separately, by finding out which of the rules had a positive effect. Thus an optimal set of rules for each speaker was determined. This is shown in the second bar in Figures 4 and 5. The average number of variants per speaker was 1.75 for the optimal rule set and 2.4 if all variants were added.

These detailed results clearly show, that a speaker-wise selection of the rules is superior to adding all rules, that are typical for a speaker of that mother tongue. This was already observed previously, where Italian rules improved the performance for some German speakers and vice versa. This indicates that using rules, that do not reflect the speaker’s articulatory habits can indeed lower the recognition rates. There, too many variants are added, which increases the confusability in the lexicon, thus lowering the recognition rates.

A further result is that applying the rules outperforms the experiments using the manually found variants. One reason might be the already mentioned fact, that only for parts of the
data manual phonetic transcriptions were available, so that only part of the realized variants that really occur are present in the dictionary. When using the pronunciation rules to generate the variants, of course variants for more entries in the dictionary will be generated. Still we can deduce from the above results that the rule set we generated is capable of generating reasonable variants, in the sense of variants that do really occur in practice.

6. Combination with MLLR Speaker Adaptation

Apart from the fact that simple phoneme substitutions occur or that phoneme sequences that are different from the canonic, native pronunciation, are used, also speakers tend to pronounce the phonemes the way they know them from their native language. That means even if we are capable of predicting the phoneme sequence that is used by non-native speakers, there will still be this model mismatch. To reduce this, MLLR speaker adaptation was used. We used one global regression class and adaptation was conducted online and unsupervised, adapting the HMM models after each four utterances. The results are shown in Table 1, listing the baseline dictionary with and without MLLR adaptation and the optimal set of rules for each speaker with and without adaptation.

As can be seen, MLLR can already improve the results using the baseline dictionary. Using the extended dictionaries together with speaker adaptation, we can achieve further improvements. This again proves, that speaker adaptation alone is not sufficient in the case of non-native speech. For two speakers, the unsupervised adaptation decreased recognition rates with the baseline dictionary, but using the variants only improved the results. Only for one speaker, the combination of the baseline dictionary with adaptation was superior to using the enhanced lexicon. Unsupervised adaptation is often problematic if the initial recognition rates are relatively low, as it is the case in this study. Here the approach to enhance the lexicon with variants seems to be more robust.

7. Conclusion and Future Work

In this paper the necessity of non-native pronunciation modelling was demonstrated. Recently, there has been a lot interest in modelling native pronunciation variation. Even though there seems to be an agreement that this a necessary thing to do, the results often were beyond expectations.

8. References


