Towards a Localised German Automatic Speech Recognition

Michael Stadtschnitzer, Christoph Schmidt, Daniel Stein
Fraunhofer Institute IAIS, 53754 Sankt Augustin, Germany
Email: {michael.stadtschnitzer,christoph.andreas.schmidt,daniel.stein}@iais.fraunhofer.de
Web: www.iais.fraunhofer.de

Abstract

Spoken languages are often rich in regional accents and dialects. These local variations often pose challenges to automatic speech recognition. In this study, we analyse the influence of German regional accents on the performance of a large vocabulary continuous speech recogniser trained on standard German data. The experiments show a large variation in the error rate over different regions. We investigate the influence of phoneme-level variations on recognition errors and use our findings to build a dialect recognition system.

1 Introduction

In contrast to written language which is usually standardised with respect to orthography, spoken languages often show a rich variety in regional accents and dialects even within one country. These accents and dialects affect the pronunciation of words in several ways. First of all, the pronunciation of certain phonemes of the language might differ from region to region. Moreover, especially in colloquial conversations the pronunciation of words considerably differs from their standard form and is often characterised by omissions of some sounds. This variation in pronunciation often poses severe problems for automatic methods such as speech recognition, which usually rely on standard phonetic dictionaries. In this study, we want to examine the effect of German regional accents on the performance of our ASR system which is trained on broadcast news spoken in standard German. We investigate whether recognition errors on the word level are due to different pronunciations of the words in dialectical speech. As a first step to an adaptation of our speech recognition system, we try to detect the dialect of a given utterance. We investigate two different approaches, one based on a speaker identification system, and one based on a localised phoneme language model.

2 Related Work

Only few studies on the influence of regional accents of German on speech recognition exist. For other languages, studies are more prevalent, especially for English and Arabic, some studies even covering both languages ([1], [2], [3], [4] and [5]).

[6] examines the influence of non-native accents on speech recognition performance and the adaptation of acoustic models and lexica in the special case of Turkish pronunciation of German. The study takes the phonological differences between the speakers’ native language Turkish and the foreign language German into account. [4] automatically adapt a German pronunciation dictionary to Austrian German, and [8] describes the adaptation of a German LVCSR system to Austrian speakers.

More closer to our work, [9] examine regional variants in the German VERBMOBIL corpus and use selection of localized dictionary entries, but report only average error rates over all regions and do not achieve significant improvements by their selection methods.

In the field of phonetics, [10] use the Levenshtein distance on phoneme strings to define a distance metric between dialects.

3 Speech Recognition Experiments

We use the open-source speech recognition toolkit Kaldi [11] to train a large vocabulary continuous speech recognition (LVCSR) system. For acoustic modelling, we apply tri-phone Hidden Markov Models (HMM) based on linear discriminant analysis, maximum likelihood linear transformation and speaker adaptive training. The deep neural network (DNN) based models, which are trained upon, use an architecture with four hidden layers, each consisting of 1,024 neurons. The acoustic models are trained on 636 hours of German broadcast data taken from the GER-TV1000h corpus [12]. For word decoding, we use a 5-gram language model with a pronunciation lexicon consisting of 350,000 words with subsequent recurrent neural network (RNN) language model rescoring (using [13]) of the 20-best hypothesis list. Both n-gram and RNN language model are trained on a German text corpus consisting of 71.8 million words.

For phoneme decoding we use an open-loop phoneme recogniser. It consists of the acoustic models and a grammar, where any phoneme sequence can be produced without additional cost. Hence, the phoneme hypothesis corresponds to the likeliest sequence with respect to the acoustic models.

To examine the effect of regional accents and dialects on ASR results, we test our system on the RVG1 (Regional Variants of German) corpus [14]. The corpus contains utterances recorded by German speakers from different di- rectional regions. While the corpus consists of different kinds of utterances such as read digits or spontaneous speech of approximately one minute length, we restrict ourselves to phonetically balanced sentences which speakers from different regions of Germany were to read. We only use the data recorded with the high-quality microphone to avoid mismatching recording conditions in the training data. Figure 1 shows the word error rate (WER) obtained on speakers from the different regions. The results show a strong variance in WER between the regions, ranging from 24.0 % for Niederfränkisch to 35.8 % for Ostfränkisch. The former result is not surprising, since the pronunciation of Standard German which is used in the training data featuring broadcast news is similar to the formal German spoken in the region around Hannover.

In the following, we want to examine which pronunciation variants are mostly responsible for this divergence in error rates.
4 Pronunciation Variants

We use an open-loop phoneme recogniser, i.e. a phoneme recogniser that uses no language model, to obtain a transcription of the spoken words on the phoneme level. While the sentences in the RVG1 corpus have been automatically annotated on the phoneme level using the Munich Automatic Segmentation System (MAUS) [15], this annotation turned out to be unsuitable, because the MAUS system restricts deviations from the standard pronunciation provided by the dictionary using certain microrules. These rules however were not designed to include pronunciation deviations created by dialectal speech. An open-loop phoneme recogniser makes no such assumptions and is therefore more suitable for the task at hand. To examine the impact of accents and dialects on speech recognition errors, we compare phoneme error rate, i.e. the distance between the dialectal speech and the standard pronunciation in our dictionary, to the word error rate of the speech recognition system. The assumption is that a strong dialect, i.e. a strong deviation in pronunciation, leads to an increase in the recognition word error rate.

Table 1 shows the average word error and phoneme error rate for each dialectal region. There is a strong correlation between the total phoneme error and the resulting recognition error with a correlation coefficient (Pearson’s r) of 0.92 (cf. Fig. 2).

Analysing the result even further, we examine the impact of errors in different phoneme groups on the recognition error rate. The motivation for this analysis is that some phoneme errors such as changes in the length of vowels do not necessarily lead to recognition errors, because the ASR system can compensate such errors in its hidden Markov model states.

In Table 2, we grouped the phonemes into classes and measured the correlation between phoneme errors in these classes and the resulting recognition errors. The phoneme class with the strongest correlation to recognition errors are the plosives, especially confusions of the pairs p⇔b and t⇔d are common. The group with the weakest correlation are the diphongs. Here, it turned out that the phoneme recogniser often confused the diphongs with the constituting vowels, e.g. aI⇒a I, leading to phoneme errors but to correctly recognised words.

5 Dialect detection

To adapt our system to the different dialectal regions, we first want to detect the dialect region to which a given speaker belongs. We try two different approaches: first, we train a speaker identification system on speakers for each regions. Secondly, we use dialectal peculiarities on the phoneme level to detect the region to which a speaker belongs.

In our first approach, we apply a system usually used to detect the identity of individual speakers, i.e. a speaker identification system. We use the open source toolkit ALIZE 3.0 [16], which features state of the art speaker identification technologies such as the i-vector approach. Af-
We investigated the effect of dialectal speech on a speech recognition system trained on standard German. Pronunciation variants can be partly detected at the phoneme level, which means that a phoneme lexicon adapted to a particular dialectal region might improve system results. To detect the dialect of a specific utterance, we used an approach based on speaker identification and one on localised phoneme language models. The latter performed significantly better.

In the near future, we want to continue our efforts to use the i-vector approach of the speaker identification system to detect dialectal properties of speech. We also want to investigate the adaptation of our standard German lexicon to German dialects.

Acknowledgements

This work has been partly funded by the European Community’s Seventh Framework Programme (FP7-ICT) under grant agreement No 287911 LinkedTV and No 269980 AXES.

References


<table>
<thead>
<tr>
<th>Order</th>
<th>Precision [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.3</td>
</tr>
<tr>
<td>2</td>
<td>14.3</td>
</tr>
<tr>
<td>3</td>
<td>16.5</td>
</tr>
<tr>
<td>4</td>
<td>19.2</td>
</tr>
<tr>
<td>5</td>
<td>18.2</td>
</tr>
<tr>
<td>6</td>
<td>18.0</td>
</tr>
<tr>
<td>7</td>
<td>17.2</td>
</tr>
</tbody>
</table>

Table 3: Dialect detection precision using language models of different orders.