GERMAN REGIONAL VARIANTS - A PROBLEM FOR AUTOMATIC SPEECH RECOGNITION?

N. Beringer[1], *F. Schiel*[1], *P. Regel-Brietzmann*[2]

[1] Institut für Phonetik und Sprachliche Kommunikation, Schellingstr. 3, 80799 München, Germany
[2] Daimler-Benz AG, Wilhelm-Runge-Straße 11, 89081 Ulm, Germany

ABSTRACT

A well known problem in automatic speech recognition (ASR) is robustness against the variability of speech between speakers. There are several ways to normalise different speakers; one of them is to deal with the problem of regional variation. In this paper we discuss the problem of whether moderate regional variants of German influence the automatic speech recognition process and whether there is a way to improve performance through knowledge of the regional origin of the unknown speaker.
The basic idea in our experiment is to cluster test speakers into distinct dialectal regions and derive observations about the typical pronunciation within these regions from a classified training set. In a cheating experiment where the origin of the test speakers is known we verify whether the use of the dialect-specific pronunciation forms will improve the overall performance of the recognizer.
It turns out that simply using dialect-specific pronunciation does not significantly improve word accuracy on the VERBMOBIL 1996 task.

1. INTRODUCTION

One way of improving an automatic speech recognizer is to exploit language-specific phenomena like dialectal influences. Real dialects are easy to spot automatically and can be treated as own languages (new training, new dictionary, new rules etc.) Therefore in this investigation we are looking for the influences of regional variants in standard German ('High German').
Given a non-prompted German speech corpus like VERBMOBIL - that includes such variants - the following question can be asked:
Is it possible to improve on automatic speech recognition by generating and including weak regional variants during the recognition process?
It is known from previous experiments (e.g. [10]) that simply extending the pronunciation dictionary of an HMM based recognizer to multiple pronunciations will not improve performance; in most cases the word accuracy even degrades because of the extended search space. Our question here is whether the knowledge about the regional origin of the unknown speaker and subsequently the use of a specific pronunciation dictionary for that region may help to reduce the ambiguity of the search space and simultaneously yield better word modelling in ASR.
If no success can be obtained, this will lead to the conclusion that weak regional variants can be neglected because of an already existing good robustness of the recognizer.
The following two sections briefly describe the used data and how the variants were obtained from the training set. In conjunction with that we briefly describe the training process of the used recognizer. Section 4 deals with some of the experiments we conducted in this investigation:

- the baseline test (reference test)
- the 'naive approach' of this experiment (embedding all variants of a defined region)
- statistical constraints
- lexical constraints and
- constraints by using a forced-alignment

Finally, results and future work are discussed in the last section.

2. DATABASE

Most currently available speech corpora contain read or prompted speech which is not suitable for our experiment, because the speech of these corpora contains only very little dialectal variation (although there is evidence that even in prompted digit strings dialectal influences may be found, see for instance [8]). The only currently available German spoken database with non-prompted speech is the German portion of the VERBMOBIL corpus ([6]).

Concerning the situations in which automatic speech recognition is to be used, it is evident that people don't speak like "pronunciation dictionaries" but rather use a controlled version of their everyday speech. Of course, more official situations like in the VERBMOBIL task (asking for departures or arranging an appointment with a business partner) require weaker regional variants than personal affairs because people tend to hyper-articulate in these situations.

The German VERBMOBIL database contains sufficient speech data for training (12 000 turns[1]) and testing (about 1800 turns) the Daimler Benz VERBMOBIL recognizer [1] and contains moderate regional variants which can be automatically selected from the provided transliteration files. ([7])

All observed variants can be derived out of their citation forms using a fixed set of phonological rules ([2],[3]). It

[1] One turn in the VERBMOBIL database has about 22.8 words in average

is also important that the database-scenario deals with scheduling appointments which are real-life-situations with currently used speech. For this reason our vocabulary is not too dialectal and the regional variants which are needed to build a dictionary do not contain any dialectal vocabulary. The average number of variants per word is 1.8.

3. DICTIONARY

A set of dictionaries containing regionally clustered variants was built by using the dialect-specific transcribed pronunciation forms (done by the Munich VERBMOBIL group ([5]); the forms were corrected or re-transcribed using a set of phonological rules ([2],[3]).

The transcribed variants were clustered into 10 broad dialectal regions of Germany + 2 not clearly identifiable classes (north and south). The used dialectal regions can be seen in figure 1.

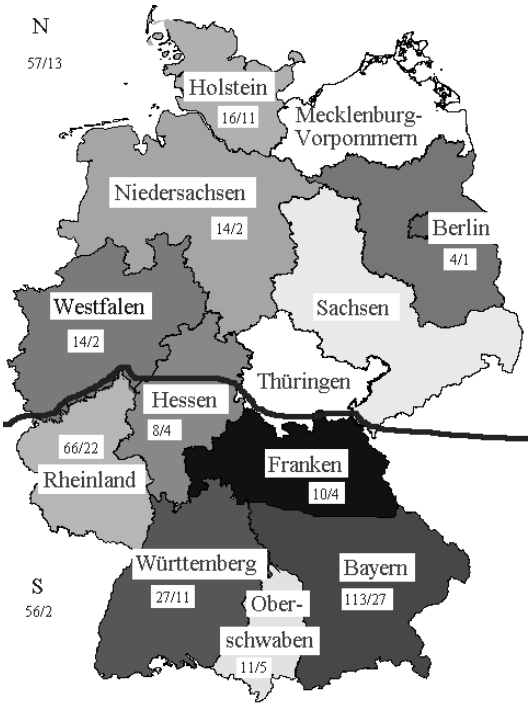

Figure 1: Dialectal regions used in the experiments: note that for Thüringen, Mecklenburg-Vorpommern and Sachsen no test data were available; numbers represent speakers in the training and test set respectively

The individual dictionaries were filled up with the citation form for those words of the standard test dictionary (5362 words) which were not observed in the corresponding training subset resulting in 12 regionally specific sets of pronunciation. Each cluster contained at least 19 turns of the training set.

The size of the dictionaries for different experiments can be seen in table 1. The baseline dictionary had 5362 distinct words. The acoustic models of the Daimler Benz recognizer

dialectal region	absolute pruning	function words	forced align.	naive approach
hessen	5362	5562	5455	5752
berlin	5363	5583	5550	5897
ober-schwaben	5363	5591	5362	6014
franken	5363	5592	5626	6208
south	5369	5609	5441	6340
westfalen	5369	5621	5441	6485
wuerttem-berg	5382	5648	5716	6827
north	5385	5662	6037	6996
nieder-sachsen	5385	5662	6026	7104
holstein	5404	5681	6294	7638
bayern	5459	5706	6357	7992
rheinland	5420	5692	6511	8225

Table 1: Number of distinct words in the dictionaries

([1]) were trained using a dictionary including all observed regional variants; therefore no regionally specific acoustic models were used in this experiment.

4. EXPERIMENTS

With this setup we performed the following cheating experiments: The 104 speakers of the test set were clustered into the same 12 dialectal regions and their speech data was tested using only the corresponding dictionary. We conducted 19 recognition experiments to use regional variant dictionaries in different flavours.

4.1 Baseline Test

We defined a baseline test by using only the canonical form in all 12 speaker groups resulting in an average word error rate of 30.91%[2]

The following tests - divided into 4 groups as seen below - were expected to show how the average word error rate could be improved.

4.2 Naive Approach

Apart from the average word error rate which we obtained from our baseline test we also need the average word error rate from a test involving all found variants of our dialects. Therefore we used the pronunciation variants of each of the 12 regions in the corresponding test speaker groups. Here, we actually expected a higher word error rate because embedding all found variants of a region will also include some unusual ones. So it was no surprise that our test led to a significant degradation of 33.48% average word error rate.

[2]Word error rate throughout this paper is calculated with the standard formula
(total-substitutions-deletions-insertions) / total

It should be noticed that smaller dictionaries caused lower word error rates (e.g. Franken 25%), which is to be expected because there the search space for an utterance is reduced.

A possible explanation is that simply increasing the dictionary also increases the lexical ambiguity within each region. All the following experiments aim to improve this by applying certain constraints to the pronunciation subsets.

4.3 Statistical Constraints

Next we repeated the naive approach using an absolute statistical pruning and a statistical a-posteriori pruning of the variants.

Absolute Pruning
In this test the minimum number of dialectal variants in the corresponding regional training set was set to 50. All remaining variants were discarded. This threshold was determined by a set of heuristic experiments. The average word error rate was 30.40% which is a significant improvement to our naive approach, but not significantly better than the baseline test.

A Posteriori Pruning
In these experiments only variants were allowed that had a better a posteriori probability $P(V|W)$ for a variant V given the lexical word W than a fixed threshold. Once again, in a separate heuristic experiment we determined the optimal value for $P(V|W)$ to be over 15%.

The average word error rate amounted to 30.62% in this experiment.

Both methods of statistical prunings show that our naive-approach word error rate can be significantly improved but this is not true compared to the baseline test. It is interesting to note that the amount of lexical entries used in the above experiments was within the range of that used in the baseline test (see also table).

4.4 Linguistical Constraints

Instead of using statistical constraints there might be a benefit of including linguistic knowledge to reduce the ambiguity of the individual search spaces.

In another series of experiments only function words of the training corpus with high word probabilities were kept in the regionally specific dictionaries. Function words are defined as the closed word class of articles, pronouns, prepositions, auxiliaries and conjunctions. The elements are often found in an enclitic or proclitic form which can - again - be dialectal. Function words are rather limited but contain the most frequently used words of a language which have to occur in every dialect. Therefore, this word class seems to be a good candidate for dialectal pronunciation modelling.

But although the lexical ambiguity of the dictionary (see table) was reduced by this method we encountered an increased average word error rate of 31.59% compared to our baseline system.

This result confirms similar findings by Silvia Moosmueller in [4]: As function words always form a closed word group of often used (mostly) one-syllable-words, their variants don't differ much between the dialectal groups and the canonical form.

4.5 Forced Alignment Variants

In the last series of experiments we restricted the regional dictionaries to the variants found by a forced alignment over the training data. This resulted in a reduction of the used variants in all regional dialects of about 30%.

No different results were found using monophone- or triphone-based forced alignment. Therefore only results with triphones are reported here. Based on this new corpus we repeated the naive approach and the absolute statistical pruning experiment.

This naive approach showed 0.5% improvement compared to the previous experiment with all variants. But it is still significantly worse than the baseline result.

Using the absolute statistical pruning yielded almost the same results as the baseline system (30.66%), which is not surprising since the resulting dictionaries included almost the same pronunciations as the baseline tests. The reason for this might be that the forced alignment in most cases preferred the citation form and the few deviations were pruned by the statistical threshold.

5. CONCLUSION

Table 2 summarises the average word error rates over the 12 dialectal regions for all described experiments in order of ascending word error rate. As can be seen, only the use of absolute pruning improves the average word error rate. This improvement is not significant.

test	values in percent
statistical - absolute	30,40
baseline	**30,91**
statistical - a-posteriori	30,65
forced-alignment (statistical)	30,66
linguistic(function)	31,59
forced-alignment	32,98
naive approach	33,48

Table 2: Mean average word error rates for each experiment in percent

It has been reported that improvement within the VERBMOBIL task has been achieved using a statistical model of pronunciation in conjunction with properly trained acoustic models ([10]). One possible reason why our experiment failed might be that we used the same acoustic models for

all dialect-specific regions. Another reason might be that the pronunciation has to be modelled statistically by the use of well-estimated a posteriori probabilities for each individual variant. However, since the training set had to be splitted into 12 sub-sets for this work the available data was not sufficient for such an approach.

Another possible explanation might be the fact that in a task like the VERBMOBIL scenario only very weak dialectal variation can be found, which is already captured by the robustness of the recognizer.

Future work remains in testing and eventually training stronger regional variants in an ASR system. Such data are now beginning to emerge from several ongoing investigations (e.g. SpeechDat or RVG corpora ([9], [8]).

References

[1] Class, Kaltenmeier, Regel-Brietzmann : 'Optimisation of an HMM-based continuous speech recognizer', Daimler-Benz AG, Research Institute, Ulm, in Proceedings of the EuroSpeech 93.

[2] Rudolph Merkle : 'Bairische Grammatik', Heimeran Verlag Muenchen, 1. Auflage 1975, p. 11 - 39.

[3] Beringer Nicole : 'Untersuchungen zur dialektalen Faerbung des Deutschen', Studienarbeit am IMS, Stuttgart, December 96.

[4] Moosmueller : 'Hochsprache und Dialekt in Oesterreich: soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck', chap. 2, Wien, Koeln, Weimar, Boehlau-Verlag 1991

[5] German VERBMOBIL Corpus : www.phonetik.uni-muenchen.de/Bas

[6] S. Burger : 'Transliterationslexikon', VERBMOBIL TechDok 36-95, University of Munich, 1995.

[7] Burger, Kachelrieß: 'Aussprachevarianten in der VERBMOBIL-Transliteration - Regeln zur konsistenteren Verschriftung' VERBMOBIL MEMO 111, University of Munich,August 1996.

[8] S. Burger, F. Schiel (1998): RVG 1 - A Database for Regional Variants of Contemporary German; Proceedings of the First International Conference on Language Resources and Evaluation 1998, Granada, Spain.

[9] S. Burger, Chr. Draxler (1998): Identifying Dialects of German from Digit Strings; Proceedings of the First International Conference on Language Resources and Evaluation 1998, Granada, Spain.

[10] F. Schiel, A. Kipp, H.G. Tillmann (1998): Statistical Modelling of Pronunciation: It's not the Model, it's the data; Proceedings of the ESCA Tutorial and Research Workshop on 'Modelling Pronunciation Variation for Automatic Speech Recognition', May 1998, Kerkrade/Netherlands