PROSODY-BASED RECOGNITION OF SPOKEN GERMAN VARIETIES

V. Dizdarevic, M. Hagmüller, G. Kubin, F. Pernkopf

Graz University of Technology
Institute of Communications and Wave Propagation
Inffeldgasse 12, 8010 Graz, Austria

*Micha Baum**

SPEX/CLST
University of Nijmegen
Netherlands

ABSTRACT

An approach to the recognition of regional language varieties is presented. The algorithm is tested on utterances of 3 to 6 seconds duration taken from large speech databases (SpeechDat) of Austrian and German German. The features are based only on the prosody of the speech and include parameters derived from the Fujisaki model and statistics of the fundamental frequency. Classification is performed using a multi layer perceptron and yielded a rate of 64% correct identification of the regional variety.

Those results are then further evaluated for the use of a regional variety recognizer as a front-end of an automatic speech recognizer for different regional varieties. In case there is no a priori information of the distribution of the regional varieties spoken by the users, this approach yields a considerable improvement in the robustness of the speech recognition rates.

1. INTRODUCTION

German is a pluricentric language, which means, that it is realized in more than one country. The realization of such a language in one of these countries is called national variety, for German the major varieties are German, Swiss and Austrian German. The differences between the German spoken in Austria and Germany can be compared to the differences between British English and American English. It is stated [1, 2] that every national variety must have the status of an official administrative and public language and hence evolves a certain linguistic and communicative independence. Since the work is also applicable to regional varieties of a language, we will use that term in the rest of the paper.

The variety differences cause difficulties for Automatic Speech Recognition (ASR) systems and result in higher word error rates (WER) if the variety of the training is not equal with the variety of the test set. A front end, prior to the phone recognizer, can recognize the regional variety of the speaker, cf. fig. 1. Using this information for each German variant, a different acoustic model (AMO) can be used for recognition. Alternative approaches would either be to include all varieties for the training of the AMO or to do parallel recognition with AMOs, trained for each variety, and pick the most likely result [3].

First results of recognition of German varieties using prosodic features with a small database were very promising. Recognition rates of German and Austrian German were up to 72% [4]. Considering recent language identification (LID) literature [5] this is comparable to results for identification of completely different languages. Later a larger database became available so new experiments were performed to verify those experiments and put them on a statistically reliable foundation [6]. The paper is organized as follows. After a short description of the speech data in chapter 2, the features used for the regional variety recognition (RVR) are introduced in chapter 3. Chapter 4 describes the feature selection, classification methods and RVR results. Then chapter 5 discusses the relevance of those results for ASR of pluricentric languages. The paper finishes with a conclusion in Chapter 6.

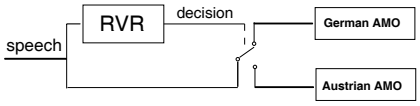

Fig. 1. *Block diagram for regional variety recognition (RVR).*

2. THE SPEECH DATA

Training and test material is taken from the German SpeechDat(II) FDB 4000 [7] and the Austrian SpeechDat-AT FDB-1000 telephone speech databases [8]. The format of the speech files is A-Law, 8-bit, 8 kHz. In the following we describe the subsets of the databases used for both the training and testing of AMOs for speech recognition and for the recognition of regional varieties.

2.1. Speech data for the training of the acoustic models

Both subsets of the databases contain 980 speakers. 90% are training data, test data are taken from the remaining 10%. The splitting of the corpus was carried out in three different ways so that there are three different training sets and test sets. The results are average values over the three splits. The total number of speech files in each database is 42268. The databases are subdivided into the following tasks: yes/no, typical words used in command recognition tasks, connected digits, natural numbers, money amounts, dates, times, directory assistence, spelling, phonetically rich words, phonetically rich sentences. For each of these tasks, either a bigram based language model was trained (e.g., for phonetically rich sentences) or a regular grammar was written (e.g., for digits). The vocubulary size for the Austrian database is 17965 and 19664 for the German database. The demographic distribution of the subsets of the databases corresponds to the SpeechDat criteria.

*The author performed the work while at ftw—Vienna Telecommunication Research Center. He gratefully acknowledges Siemens for making the SpeechDat-DE data available for the experiments.

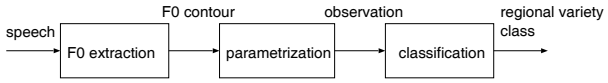

Fig. 2. *Block diagram for feature extraction.*

2.2. Speech data for the training of variety recognition

As the RVR algorithm only deals with prosody we needed utterances with a certain minimum amount of time in order to capture the suprasegmental features. Therefore, only phonetically rich sentences were used for both training and testing. We selected 1200 sentences from the Austrian database and 1200 sentences from the German database. The length of the sentences was between 3 and 6 seconds only, thus reflecting the requirement for a fast reaction time of the RVR front-end.

Note that SpeechDat databases primarily contain read speech. Therefore, we did not expect as significant differences between the two national varieties as we might obtain from spontaneous or colloquial speech.

3. FEATURE EXTRACTION

Three steps lead to the decision on the language variety of the given speech sample. As shown in figure 2, these are the extraction of the fundamental frequency contour, feature extraction and classification. The fundamental frequency (F0) of the utterance is the only prosodic parameter used for the classification of the regional variety of the language. F0 is calculated using YIN, a time-domain algorithm based on the evaluation of the average magnitude difference function. YIN was proposed by de Cheveigné [9] and it was adapted for the automatic regional variant recognizer. The parameters used by the algorithm are calibrated to minimize the number of outliers of the F0 contour for the data from SpeechDat. For the parameterization steps the overall form of the contour is more important than the exact values of F0. With regard to this a post-processing stage to YIN is implemented to remove all the outliers by filtering the contour as well as by an octave error correction mechanism.

The calculated contour is then parameterized and represented as a set of 26 features for an utterance. Three main categories of features are calculated:

- Fujisaki features
- Mean log intervals
- Percentiles and range

The Fujisaki model described in figure 3 allows a quantitative analysis of certain linguistic features [10, 11]. Originally developed for the generation of F0 contours for synthesized speech. In this work it is used to analyze the utterance, the calculated Fujisaki parameters are then used for classification. Fujisaki models a fundamental frequency contour as a linear superposition of an offset frequency, a local accent component and a global phrase component thus providing linguistically meaningful features for further processing. Parameters of the second order linear system representing duration and amplitude of the components are used to calculate 8 Fujisaki features, 3 based on the phrase and 5 based on the accent components. These are:

- Mean time difference between phrase control impulses

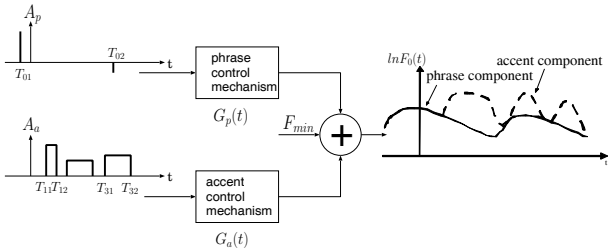

Fig. 3. *Fujisaki Pitch Contour Parameterization. The parameterized pitch contour is a superposition of a minimum frequency, a local accent compontent and a global phrase component.*

- Mean amplitude of phrase components
- Mean difference amplitude between consecutive phrase components
- Mean amplitude of accent components
- Mean difference amplitude of accent components
- Mean difference between two consecutive accent starts
- Mean duration of accent
- Mean times between two consecutive accents

The second block of six features is the representation of the mean log intervals. The logarithm of a continuously approximated F0 contour and the intervals between adjacent local minima and maxima is calculated. This information is mapped in the form of a histogram thus representing the frequency of occurrence of quantitatively different frequency jumps within the utterance. The whole F0 range within an utterance is divided in six equal parts. The frequency jumps falling within a single class are normalized by the overall number of jumps within an utterance.

The last twelve features are the representation of elementary statistical properties of the calculated F0 values. Ten features represent the percentiles of the contour and two range features are introduced. P10, P25, P75 and P90 of the F0 contour are calculated. To avoid the impact of the generally decreasing value of F0 towards the end of the sentence the same percentiles are calculated for the contour with a linear trend removed from it. Additionally, P75 and P90 for the differential value of the contour are calculated. Last two features are the range of F0 (F0max - F0min) and the minimal value of F0 observed in the entire utterance, relative to the median value of the fundamental frequency contour.

4. MACHINE LEARNING APPROACH

The principal component analysis of the parametrized data (features) shows the difficulty of this classification task since the classes are strongly overlapping.

We used two different classification approaches: Feature selection combined with *k*-NN and Neural Networks

4.1. Feature Selection with *k*-NN classifier

In our classification problem the relevant features are unknown a priori. Thus, many features are derived and those which do not contribute or even degrade the classification performance are removed from the set of extracted features during classification. The

sequential forward floating search (SFFS) algorithm [12, 13] is employed since a good tradeoff between computational demands and obtained classification rate is achieved compared to optimal feature selection methods. The SFFS method includes a feature to the current feature subset which maximize the performance of the 5-NN classifier. Afterwards, conditional exclusions of the previously updated subset take place. If no feature can be excluded anymore, the algorithm proceeds again with adding features. This floating behavior allows to correct wrong decisions made in previous steps of the search. The experiments are based on 2400 samples uniformly distributed between both classes, German and Austrian. A five-fold cross-validation classification performance of 61.25% is obtained for a feature subset of size 6.

4.2. Neural Network Classification

Different neural network architectures are trained for 16 different subclasses of features. The full set of 26 features does not guarantee the best classification results. Experiments showed that the full set of the statistical parameters group combined with the Fujisaki parameters provides the best basis for the classification. This combination of 16 parameters was tested on a number of different neural networks varying the number of neurons in the hidden layer of a multilayer feed-forward network between between 30 and 100.

The data set consisting of 1200 German and 1200 Austrian speech files (120 speakers each) is chosen to determine the average recognition rate. Using 80% for training and 20% for testing, the final recognition rate is obtained averaging the results of three cross validation steps. A recognition rate of 64% is achieved using a three layered NN with 40 neurons in the hidden layer. The neural networks output is trained to reach the values 0 and 1 for the German and Austrian variant, respectively. The optimal decision boundary in this case was 0.51, based on experimental results.

5. EVALUATION

The recognition rate of 64% of the RVR algorithm is relatively low. In order to judge its efficiency for certain scenarios, ASR recognition rates (respectively word error rates (WER)) have to be compared using RVR with two AMOs on the one hand and a single AMO without pre-processing variety recognition on the other hand. The recognition rates for single AMOs for speakers from Germany and Austria were investigated in [14]. Training and testing was carried out with data introduced in section 2.1. All AMOs were trained and tested with the Hidden Markov Toolkit (HTK) [15]. The results are laid down in table 1.

	German AMO	**Austrian AMO**
German test speakers	12.09%	21.78%
Austrian test speakers	16.32%	9.89%

Table 1. *Word error rates for German and Austrian test speakers using AMOs trained with speakers from Germany and Austria*

5.1. Using a single AMO

If a single AMO is used by different groups of speakers the word recognition rate (WRR) can be estimated by averaging the WRRs for each group of test speakers using a common AMO. Each of the WRRs has to be weighted with the probability that a speaker is from a certain area. This probability can be considered the relative frequency of the speakers of that region. The WER then amounts to $WER = 1 - WRR$. With equation (1) the estimates for WER using any desired number of AMOs and speaker groups can be computed. $p(spk_i)$ stands for the probability that a speaker of group i uses the speech recognizer and $WRR(AMO, spk_i)$ means the word recognition rate that was scored for a particular AMO for a corresponding group of speakers.

$$WER(ONE\_AMO) = 1 - \sum_{i=1}^{n} p(spk_i) \cdot WRR(AMO, spk_i) \quad (1)$$

Figure 4 illustrates the case for the two groups of speakers (Germans, Austrians) using the speech recognizer trained with the German SpeechDat database.

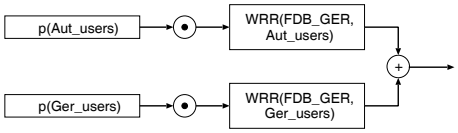

Fig. 4. *Block diagram for estimation of the percentage of correctly recognized words for two groups of test speakers (Austrians and Germans) using the German AMO.*

5.2. Using the RVR algorithm

For the RVR algorithm the WRRs for the different groups of speakers are weighted too. Then, however, for each group of speakers the probability for assigning the speakers to a certain AMO is multiplied with the WRR for this group of speakers using the AMO the speaker was assigned to. With equation (2) the WER can be estimated, $p(RVR(AMO_j, spk_i))$ is the probability that a speaker from group i is assigned to AMO_j. Note that the probability to recognize the speaker's origin correctly is 0.64 and hence an incorrect assignment is expected to occur in 36% of the cases. There is no distinction between variety recognition rates for German speakers and Austrian speakers. Figure 5 shows a block diagram for estimating the WRR for the RVR algorithm and two AMOs (trained with German and Austrian speakers, respectively) and two groups of users (Germans and Austrians).

$$WER(RVR) = 1 - \sum_{i=1}^{n} \sum_{j=1}^{m} p(spk_i) \cdot p(RVR(AMO_j, spk_i)) \cdot$$

$$\cdot WRR(AMO_j, spk_i) \quad (2)$$

5.3. Results and discussions

The best method to decide whether to choose a single AMO or the RVR algorithm is to consider the WERs as a function of the distribution of speakers. These can be determined by solving (1) and (2) for $p(spk_i)$ assuming that $i = 2$ and $p(spk_1) = p(spk_{ger})$ whereas $p(spk_2) = p(spk_{aut}) = 1 - p(spk_{ger})$. The values for $p(RVR(AMO_j, spk_i))$ and $WRR(AMO_j, spk_i)$ are known. The three functions are illustrated in figure 6. It can be seen clearly that the RVR algorithm has the flattest slope, i.e. the WERs vary

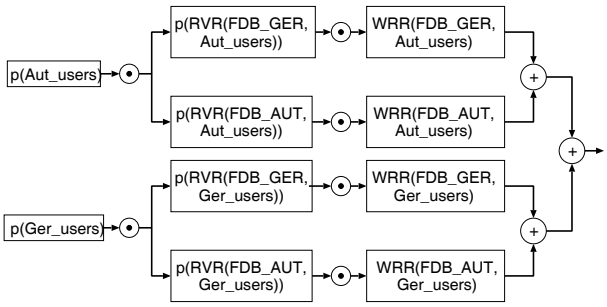

Fig. 5. *Block diagram for the estimation of the percentage of correctly recognized words for two groups of test speakers (Austrians and Germans) and two AMOs (trained with Germans and Austrians respectively) using the RVR algorithm.*

less than for single AMOs. The WERs for the use of the RVR algorithm vary between 12.20% (100% Austrian speakers) and 15.54% (100% German speakers). For single AMOs, the maxima of the WERs are known, namely 12.09% and 16.32% using the German AMO and 9.89% and 21.78% for using the Austrian AMO. Naturally the RVR algorithm which is the method corresponding to the flattest slope is the most appropriate one. In particular, the maximal WER for the RVR algorithm is less than the maximal WER of either the German or the Austrian AMO.

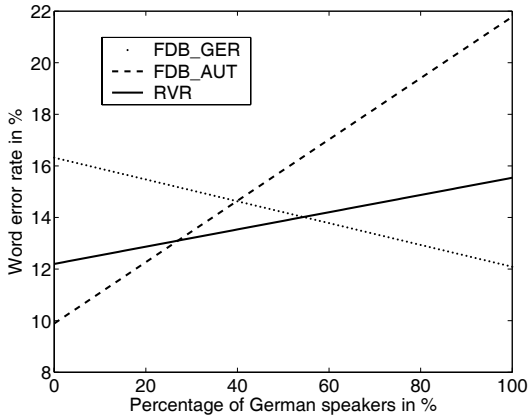

Fig. 6. *Word error rates for the RVR algorithm (solid line), the German AMO (dotted, $\cdots$), and the Austrian AMO (dashed, - - -) as a function of the percentage of German German speakers.*

6. CONCLUSION

We presented a machine learning approach for regional variety recognition. The work was carried out using SpeechDat-AT and SpeechDat-DE as a source. Though the variety recognition rate of 64% is relatively low, it is still comparable to other results for the identification of German versus English [5]. Even better results could be expected for utterances with durations longer than 6 seconds. It has been shown that for speech recognition tasks where the regional variety of the speaker is not known a priori, even this result improves the robustness of the overall ASR system, considerably.

Due to the large database, the current result can be seen as a solid baseline, which shows, that there are significant differences in the prosody of German and Austrian. Further research will be necessary to improve the recognition results and to extend these methods to other pluricentric languages.

7. REFERENCES

[1] Rudolf Muhr, "Die plurizentischen Sprachen Europas. Ein Überblick," in *Vielsprachiges Europa. Zur Situation der regionalen Sprachen von der Iberischen Halbinsel bis zum Kaukasus*, pp. 191–232. Guggenberger, Frankfurt am Main, Germany, 2003.

[2] Michael Clyne, "German as a pluricentric language," in *Pluricentric languages: Different norms in different nations*, pp. 117–143. Nouton de Gruyter, Berlin, Germany, 1992.

[3] R. Chengalvarayan, "Accent-independent universal HMM-based speech recognizers for American, Australian and British english," in *Proc. Eurospeech*, Aarlborg, Sept. 2001, pp. 2733–2736.

[4] Martin Hagmüller, "Recognition of regional variants of German using prosodic features," Diploma thesis, Graz University of Technology, 2001.

[5] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. Andre-Obrecht, "Modeling prosody for language identification on read and spontaneous speech," in *Proc IEEE ICASSP*, Apr. 2003, vol. 1, pp. 40–43.

[6] Vedran Dizdarevic, "A machine learning approach to recognition of spoken German variants," Diploma thesis, Graz University of Technology, 2003.

[7] "http://www.elda.fr/catalogue/speech/s0063.html," downloaded in October 2003.

[8] Micha Baum, Gregor Erbach, and Gernot Kubin, "Speechdat-AT: A telephone speech database for German," in *Proc. LREC 2000 workshop "Very large telephone speech databases (XL-DB)"*, Athens, Greece, 2000, pp. 51–56.

[9] Alain de Cheveigne and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 1917–1930, 2002.

[10] Hiroya Fujisaki, "Dynamic characteristics of voiced fundamental frequency in speech and singing," in *The production of speech*, P. F. Mac-Meilage, Ed. Springer, Berlin, 1983.

[11] Hansjörg Mixdorff, "Fujisaki analysis software," downloaded in March 2003 from http://www.tfh-berlin.de/~mixdorff/fujisaki_analysis.htm.

[12] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119,1125, 1994.

[13] Franz Pernkopf, *Automatic visual inspection of metallic surfaces*, Ph.D. thesis, University of Leoben, 2002.

[14] Micha Baum, *Improving speech recognition for pluricentric languages exemplified on varieties of German*, Ph.D. thesis, Graz University of Technology, 2003.

[15] S. Young, D. Kershaw, J. Odell, D. Ollason, and V. Valtchev, *The HTK book (for HTK Version 3.0)*, Microsoft Corporation, Cambidge, UK, 2000.