Joint Standard German and Bavarian Subdialect Identification of Broadcast Speech

Michael Stadtschnitzer[1], Christoph Schmidt[2]

[1] *Fraunhofer IAIS, Schloß Birlinghoven, 53757 Sankt Augustin, Deutschland, michael.stadtschnitzer@iais.fraunhofer.de*

[2] *Fraunhofer IAIS, Schloß Birlinghoven, 53757 Sankt Augustin, Deutschland, christoph.andreas.schmidt@iais.fraunhofer.de*

Abstract

Automatic speech recognition is a very important technique for numerous applications like automatic subtitling, dialogue systems and information retrieval systems. Speech recognition systems usually perform very well in clean and controlled environments. However they still often fail when mismatches between the trained models and the testing data are present, e.g. due to noise, reverberation, or dialects. A method to cope with dialects is to identify the dialect in advance and then use specialized dialectal speech recognition models for the decoding. Also, dialect identification systems have recently been used for targeted advertising, service customization, forensics tasks and for text-to-speech synthesis of regional speech. In this work, we annotate a large quantity of dialectal and standard German speech from a German broadcaster and exploit the data to train and evaluate a joint standard German and Bavarian subdialect identification system that is able to distinguish between standard German and three Bavarian subdialects, namely Bavarian, Swabian and Franconian, with promising performance.

Introduction

Automatic speech recognition systems nowadays perform particularly well in controlled environments, i.e. when the training data has the same characteristics as the testing data. However, if a mismatch is present, e.g. due to noise, reverberation or the presence of dialectal speakers, the performance of these systems usually degrades. In the case of dialectal speakers, it can be beneficial to infer the dialect in advance to choose the optimal speech recognition model to decode the data. For this task, usually a dialect identification system is employed. Dialect identification is also used e.g. in audio forensics, to retrieve the dialect and hence the origin of a unknown speaker, or in audio mining applications, where as much as possible information about the speech and audio data shall be retrieved.

Convolutional neural networks are becoming increasingly popular for dialect identification in the research community [4]. In this work, we focus on the creation of a dialect identification system that is able to distinguish between standard German and south German dialectal speech in the broadcast domain. The dialects that are investigated are Bavarian, Franconian and Swabian. To the best of our knowledge there is no database available that covers broadcast domain data of standard German speakers and speakers of the mentioned dialects. Therefore we create a new database with the generous support of the *Bayerischer Rundfunk* (BR), who provided the requested data to perform this research. In the following sections we describe the dialectal database that we created, and we describe the training and evaluation of a dialect identification system based on convolutional neural networks (CNN).

Dialectal database

The Bayerischer Rundfunk (BR) provided us a set a 302 broadcast media files with a total size of 146 hours and a average media file length of 29.0 minutes. The data contains mostly regional broadcasts from Bavaria, covering a large number of dialectal and standard German speakers. The data has already been clustered into the dialects Bavarian, Swabian and Franconian in advance by the BR. However, a fine annotation of the speakers and the segments was necessary, which was performed manually using the annotation tool ELAN [1]. Also the gender (male/female) of the speakers was annotated. The annotation of the speaker name, if available, is crucial to avoid that the same speaker is present across the training, validation and testing subsets. 1,732 speech segments with a total of 398 speakers have been annotated so far with a total size of 6.7 hours. An average of 60.1 seconds and an average of 4.4 segments have been annotated per speaker in this dataset. 75 standard German speakers, 149 Bavarian dialect speakers, 89 Swabian dialect speakers and 85 Franconian speakers have been annotated. We convert the audio data into RIFF/WAVE format (16 kHz sampling rate, 16 bits per sample, 1 channel).

Dialect identification

In this section we make use of the dialectal database to train a dialect identification system. First, we split the data into a training, validation and a test set. First of all we exclude segments that are smaller than one second from the data. In a second step we only keep a maximum of two minutes speech data per speaker. To have a balanced dataset we balance the number of speakers per dialect for each subset. Also, speakers are disjunct across the subsets, so that a speaker can only occur in one of the subsets. The selection which speaker goes to which dataset was performed in a random fashion. The training set covers the segments of 35 speakers per dialect with a total of 635 segments (2.6 hours). The validation set covers 20 speakers per dialect with a total of 341 segments (1.3 hours) and the test set also covers 20 speakers per dialect with a total of 321 segments (1.3 hours).

For the task of dialect identification we train a convolutional neural network using the Keras Toolkit [2] with

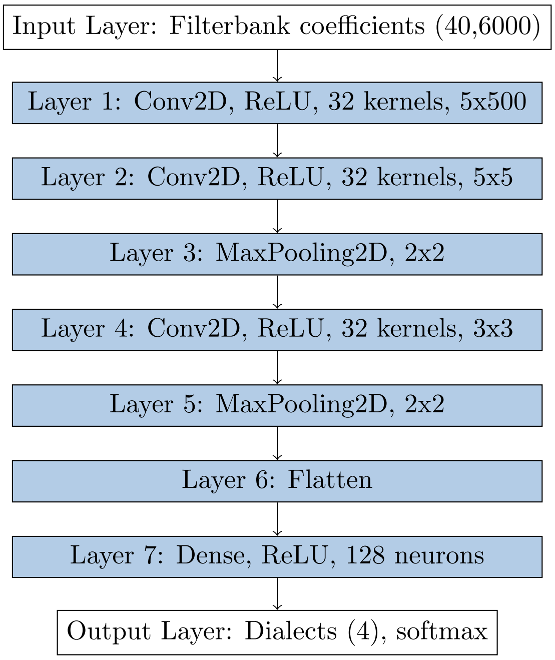

Input Layer: Filterbank coefficients (40,6000)
Layer 1: Conv2D, ReLU, 32 kernels, 5x500
Layer 2: Conv2D, ReLU, 32 kernels, 5x5
Layer 3: MaxPooling2D, 2x2
Layer 4: Conv2D, ReLU, 32 kernels, 3x3
Layer 5: MaxPooling2D, 2x2
Layer 6: Flatten
Layer 7: Dense, ReLU, 128 neurons
Output Layer: Dialects (4), softmax

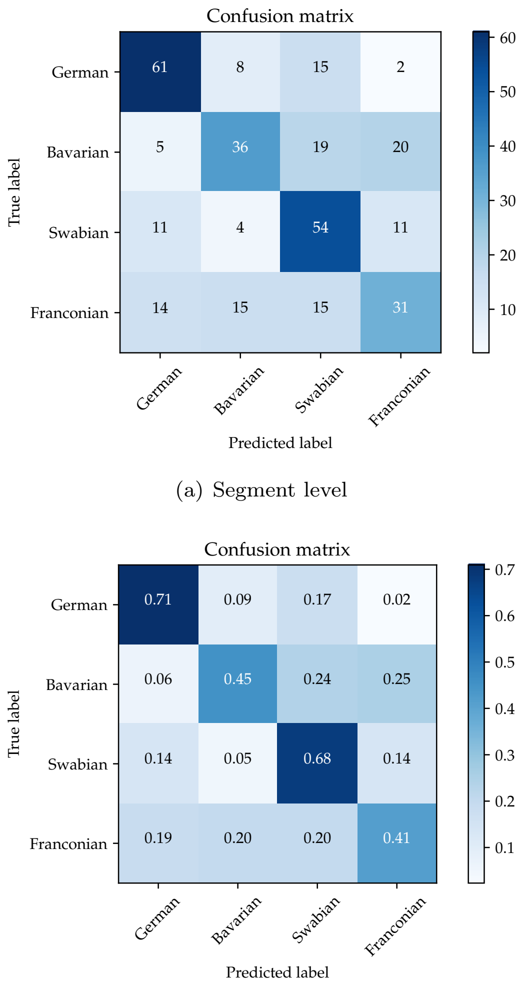
Figure 1: CNN architecture

Tensorflow [3] backend. First, the audio signal is filtered by a first order IIR preemphasis filter ($a = 0.97$). A set of 40 mel-spaced filterbank coefficients are extracted for windows of length 25 ms with a hopsize of 10 ms. The filters cover the whole range from 0-8 kHz. The filterbank coefficients of the whole segment (zero-padded, 1 minute maximum) are then fed into a convolutional neural network (CNN), whose output is a probability function over the investigated dialects. The employed CNN architecture is depicted in Figure 1.

The results of this approach on the test set in terms of segment level and normalized confusion matrices are depicted in Figure 2. The accuracy of the system on the test set is 56.7%. This is a promising result, considering the number of classes (4), the small amount of training data (2.6 hours) and also the fact that this data is collected from a heterogeneous set of broadcast media files.

Conclusion

In this work we described the creation and exploitation of a dialect identification database for the broadcast domain. The dialects that are investigated are Bavarian, Franconian, Swabian and also standard German is included. 1,732 Utterances from a total of 398 speakers have been collected. The size of the annotated data is 6.7 hours. This data has been used to train and evaluate a dialectal identification system based on CNNs. Therefore the data has been split into a training, validation and a test set, with a balanced amount of speakers per dialect. The speakers are disjunct across the data subsets. The results, as described in the evaluation section of this paper, show promising results and we are eager to increase the amount of training data in the near future to further enhance the performance.



(a) Segment level

(b) Normalized

Figure 2: Confusion Matrix

Acknowledgements

The authors want to thank *Bayerischer Rundfunk* (BR) for providing the data and assisting the research. Without their help this research would not have been possible to undertake.

References

[1] H. Brugman and A. Russel. Annotating Multimedia/Multi-modal resources with ELAN. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 2004.

[2] François Chollet et al. Keras: The Python Deep Learning library. `https://github.com/fchollet/keras`, 2015.

[3] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems. `https://www.tensorflow.org`, 2015.

[4] Sameer Khurana, Maryam Najafian, et al. QMDIS: QCRI-MIT Advanced Dialect Identification System. In *Proc. of INTERSPEECH*, pages 2591–2595, 2017.