Spoken Language Identification in the Image Domain

Alva Liu, Mikaela Åstrand
KTH Royal Institute of Technology
DT2119 Speech and Speaker Recognition

Abstract

Spoken language identification (LID) systems allow for automatic language detection. Among the many methods that can be applied to this classification task, modern machine learning approaches have been reported as effective. A previous study approached LID in the image domain by transforming speech samples to spectrograms[1]. In this study, we attempt the same approach with data from the SpeechDat database, and aim to investigate how well two types of convolutional neural networks can generalize on speech samples from another source then SpeechDat. The results indicate that even though the models can achieve over 80 % in test accuracy on SpeechDat data, they cannot perform well on speech samples that do not originate from the SpeechDat database.

Background

- The application of voice when interacting with modern technology is rapidly increasing.
- A limitation to many voice-controlled products is that they have to be explicitly told which language is being used. LID systems address the problem of automatically determining the language being spoken.
- Neural Networks and Deep Learning have in recent years been proven to be effective at different tasks concerning speech recognition, and can be applied to language identification.
- Since speech is generally difficult to work with in the raw audio format, previous studies have instead attempted to use frequency representations of speech as input. Bartz et al.[1] constructed a LID system that approached the problem of language identification in the image domain by using spectrograms of speech samples, and was the main inspiration for this project. Their model was based on a convolutional recurrent neural network (CRNN) architecture for image-based sequence recognition originally proposed by Shi et al.[2] in 2017.

Data

- The SpeechDat database contains speech samples from numerous European languages and variants recorded either over the fixed or the mobile network[3].
- We used in total ~100,000 SpeechDat speech samples recorded over the fixed network in Swedish, English, German, Spanish and French: 400-600 unique speakers per language with 40-48 audio files per speaker.
- We also recorded 2-3 native speakers of each language at the same occasion for approximately 2 min each while they were reading a newspaper article in their language.
- The recording was performed with sampling rate 16 kHz in 16 bits linear format using a headset and WaveSurfer.

Preprocessing

- The SoX software[4] was used to transform the audio files into grayscale spectrograms in the .png format.
- The audio files were cut into 5 s fragments (all shorter files were disregarded) which resulted in spectrograms of size 129x250 px.
- The SpeechDat spectrograms were split into three sets, with approximately 70 % of the speakers of each language for training, 20 % for validation, and 10 % for test. Each dataset was then balanced to have the same number of spectrograms from each language, randomly picked from all speakers.
- This gave us 30,360 training samples, 9,010 validation samples, and 4,405 test samples.
- Our own recordings were converted to 8 bits a-law format, transformed to spectrograms in the same way, and then used for testing.

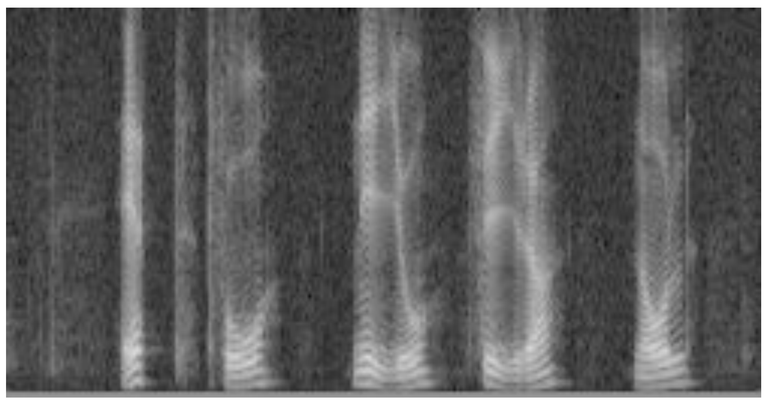

Figure 1: A spectrogram example.

Network Topologies

Two deep convolutional neural network architectures were compared in this study:

CNN: The first architecture consists of five convolutional layers with max pooling and batch normalization after each layer, and a final dense layer with dropout.

CRNN: The second architecture was originally proposed by Shi et al. in 2017[2]. It consists of the same convolutional layers as the first architecture, followed by a Bidirectional Long Short-Term Memory (BLSTM) layer.

We trained two models of each type: first using only half of the training and validation data and then using all of it. Each model was trained for 50 epochs with batch size 128 and the Adam-optimizer with learning rate 0.001. Validation accuracy during training was used to choose the four models to be evaluated on the test data.

Results

The models' performances were evaluated on the test data set. Table 1 shows the test accuracy of each model. Figure 2 shows confusion matrix and evaluation metrics for the model with highest test accuracy (the CRNN trained on all data).

Model	Training data	Test accuracy [%]
CNN	Half / all	82.9 / 85.2
CRNN	Half / all	81.6 / 85.6

Table 1: Test accuracy of each model.

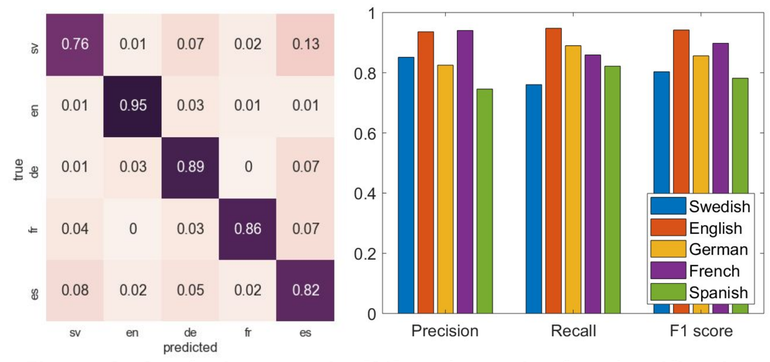

Figure 2: Confusion matrix [%] and metrics for classification on test data for CRNN trained on all data.

The CRNN model with a test accuracy of 85.6 % still performed poorly on our own recordings with an accuracy of only 18.2 %. A closer look at the confusion matrix in Figure 3 shows that the model classified almost all samples as German, regardless of the true language.

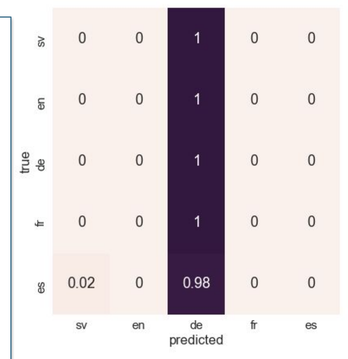

Figure 3: Confusion matrix for classification of our own recordings [%].

Two things we will look into that might improve this is image normalization and filtering based on the long term average spectrum from the two sources.

Summary

- Doubling the amount of training data slightly improved the models' performances on data from SpeechDat.
- The CRNN and CNN performed almost equally well on the test data.
- The models could not generalize well on our own recordings, likely due to mismatches in the recording settings.

Contact

Alva Liu
alvaliu@kth.se

Mikaela Åstrand
miastra@kth.se

References

1. C. Bartz, T. Herold, H. Yang, and C. Meinel, "Language Identification Using Deep Convolutional Recurrent Neural Networks," in International Conference on Neural Information Processing. Springer, 2017, pp. 880–889.
2. B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 11, pp. 2298–2304, 2017.
3. K. Elenius and J. Lindberg, Documentation: FIXED1SV / FDB5000 - A 5000 Speaker Swedish Database for the Fixed Telephone Network, Department of Speech, Music and Hearing, KTH, Mar 1999.
4. (2015, Feb) SoX - Sound eXchange. [Online]. Available: http://sox.sourceforge.net/