A Study on Spoken Language Identification using Deep Neural Networks

Alexandra Draghici
Semantic Music Technologies Group
Fraunhofer IDMT
Ilmenau, Germany

Jakob Abeßer
Semantic Music Technologies Group
Fraunhofer IDMT
Ilmenau, Germany
jakob.abesser@idmt.fraunhofer.de

Hanna Lukashevich
Semantic Music Technologies Group
Fraunhofer IDMT
Ilmenau, Germany

ABSTRACT

In this paper, we investigate a previously proposed algorithm for spoken language identification based on convolutional neural networks and convolutional recurrent neural networks. We improve the algorithm by modifying the training strategy to ensure equal class distribution and efficient memory usage. We successfully replicate previous experimental findings using a modified set of languages. Our findings confirm that both a convolutional neural network as well as convolutional recurrent neural networks are capable to learn language-specific patterns in mel spectrogram representations of speech recordings.

CCS CONCEPTS

• **Information systems → Speech / audio search**.

KEYWORDS

spoken language identification, speech recognition, convolutional neural networks, convolutional recurrent neural networks

ACM Reference Format:
Alexandra Draghici, Jakob Abeßer, and Hanna Lukashevich. 2020. A Study on Spoken Language Identification using Deep Neural Networks. In *Proceedings of the 15th International Audio Mostly Conference (AM'20), September 15–17, 2020, Graz, Austria.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3411109.3411123

1 INTRODUCTION

Spoken language identification deals with classifying the language associated to a speech recording. Possible application scenarios involve user interfaces for information retrieval systems as well as systems for automatic speech recognition, digital law enforcement, multilingual translation systems, emergency call routing, and spoken document retrieval. The main challenge of spoken language identification is find meaningful audio feature representations which are robust to individual variations in pronunciation as well as to similarities of languages within the same language families.

AM'20, September 15–17, 2020, Graz, Austria

ACM ISBN 978-1-4503-7563-4/20/09…$15.00
https://doi.org/10.1145/3411109.3411123

In this paper, we aim to reproduce a system for spoken language identification originally proposed by Bartz et al. in [1], which processes spectrogram features using convolutional neural networks (CNN) and convolutional recurrent neural networks (CRNN). We keep the original six-language classification scenario, but restrict the choice of languages to the Indo-European family of languages. We chose languages with similar phonetics from the Germanic languages (*English* and *German*), Roman languages (*French*, *Spanish*, and *Italian*), as well as *Greek* as an independent branch of the Indo-European family of languages. The original paper investigated the languages Chinese and Russian instead of Italian and Greek, in the six-language classification scenario. As a second contribution, we propose a training strategy, which allows for efficient memory usage and equal class distribution when training with imbalanced datasets and long audio recordings.

2 RELATED WORK

Several approaches have been proposed for language identification (LID) in speech signals before. Early approaches focus on handcrafted audio features. As an example, Torres-Carrasquillo et al. proposed a phone recognition followed by language modelling (PRLM) system using a single phone tokenizer based on Gaussian Mixture Models (GMM) in [6]. As audio features, mel-warped cepstral coefficients as well as delta coefficients are computed on a 10 ms time resolution. After being trained, the GMM tokenizer outputs the index of the Gaussian component with the highest score for a given feature vector. This system provides performance that is competitive with state-of-the art phone tokenization system at lower computational cost, without requiring prior transcribed speech material.

After the general transition towards data-driven methods, recently proposed LID systems mostly rely on deep neural networks. Montavon uses mel-spectrograms with 39 mel bands as audio features and a time-delay neural network (TDNN) with two-dimensional convolutional layers for temporal feature modeling and language classification [4]. In the evaluation, the system achieved an accuracy of 0.83 in a three-language classification scenario based on 5 second long speech recordings.

Heracleous et al. presented a comparative study for deep learning based LID in [2]. The authors compared fully-connected deep neural networks (DNN) and CNN architectures and found a similar performance while the CNN model required significantly fewer parameters. In contrast to spectrogram-based input features, the authors processed sequences of i-vectors using the networks. 300 training i-vectors and 100 test i-vectors are used for each of the 50 in-set languages involved. Results show that for the identification

of the 50 in-set languages, EERs (equal error rates) of 3.6% and 3.5% were obtained using DNN and CNN, respectively. When CNN was fused with SVM, an EER of 2.79% was achieved. When DNN was fused with SVM, an EER of 2.84% was obtained.

Kotsakis et al. [3] collected broadcast radio content in four different languages, namely native Greek, English, French, and German. As a pre-processing step, audio segments were classified as "speakers", "music", or "phone calls" before language classification was applied to the non-music parts. The authors propose to jointly apply multiple supervised and unsupervised classification approaches based on different analysis window lengths and then to fuse the classification results. The highest classification accuracy of 99.58 % was achieved using artificial neural networks as classifiers and an analysis window length of 1s.

Revay and Teschke [5] interpret the LID task as image classification task on mel-spectrogram "images". They use a pre-trained Resnet50 network, which combines multiple convolutional layers with intermediate skip connections to avoid vanishing gradients. In a comparable six-language scenario including English, Spanish, German, French, Russian, and Italian speech recordings, the system achieves an accuracy of 0.89.

Zhang and Hansen proposed in [7] to add two types of unsupervised deep learning approaches to the i-Vector based feature extraction part of a language and dialect classification. The first approach is an unsupervised bottleneck feature extraction, which is trained with estimated phonetic labels requiring no secondary transcribed data. The basic concept is similar to traditional bottleneck feature extraction, but without the requirement of an extra transcribed English corpus. A universal background model (UBM) is trained with all enrollment data based on Mel-Frequency Cepstral Coefficient (MFCC) with Shifted Delta Cepstral (SDC) features. Subsequently, frame level phonetic labels are estimated according to posterior probabilities. A four-layer fully-connected neural network is used. Furthermore, variational and adversarial autoencoders were included into the speech feature processing. The proposed improvements were evaluated using multiple dialect identification datasets and have been shown to outperform a baseline system based on traditional MFCC features.

Another deep learning based LID approach was presented by Bartz et al. in [1], which we use as reference for this paper. The system processes mel-spectrograms of speech signals using convolutional neural networks (CNN) and convolutional recurrent neural networks (CRNN). More details on the feature extraction and model architecture be discussed in Sections 3.1 and 3.2, respectively. For evaluation, the three datasets EU Speech Repository, Youtube News and Voxforge are used, which will be detailed further in Section 4.1.

3 SYSTEM OVERVIEW

3.1 Audio Pre-Processing

All audio files are single-channel (mono) recordings and were downsampled to a sample rate of 22.05 kHz since most relevant frequency components of speech signals are below 11 kHz. In [1], further refereed as "original paper", the type of time-frequency transformation was not explicitly mentioned. We applied the mel spectrogram as a more compact time-frequency representation. We used a window size of 512 samples, a hop-size of 441 samples (20 ms), and a frequency axis containing 129 mel bands. Finally, we compensated for different recording levels in the datasets by normalizing file-level spectrograms to a maximum of 1. All files with a duration below 10 s as well as files with silence were discarded. The normalization of spectrogram patches will be detailed in Section 3.3.

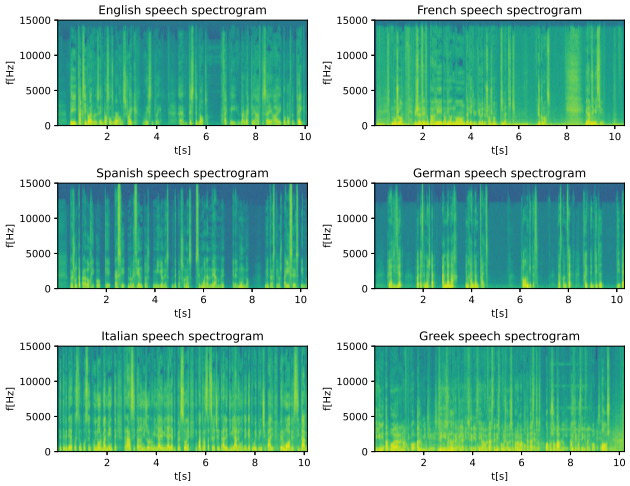

Figure 1: Examples of speech mel-spectrograms, taken as an input of neural network for LID, for six languages in this study: English, Spanish, Italian, French, German, and Greek

3.2 Neural Network Architectures

Table 1: Layer-wise model architecture for the CNN and CRNN model. A "ConvBlock" consists of a 2D convolutional layer (with the given number of 3x3 filters and ReLU activation function), a batch normalization layer, as well as a max pooling layer with pool size and stride size of 2x2. The final layer uses a softmax activation function for N classes.

Layer	CNN	CRNN
ConvBlock(64)	✓	✓
ConvBlock(128)	✓	✓
ConvBlock(256)	✓	✓
ConvBlock(256)	✓	✓
ConvBlock(512)	✓	✓
ConvBlock(512)	✓	✓
ConvBlock(512)	✓	✓
GlobalMaxPooling	✓	✓
Dense (1024)	✓	
Bidirectional LSTM (256)		✓
Dense (N)	✓	✓

We re-implemented the CNN model and the CRNN model based on a freely available implementation[1] in which some of the model

[1] https://github.com/HPI-DeepLearning/crnn-lid

parameters such as number of layers and filters differ from the original paper [1].

Table 1 summarizes the layer-wise model architectures. Both models share a feature learning front-end of seven convolutional blocks each consisting of 2D convolutional layers, batch normalization, and max pooling for downsampling along time and frequency. The learning of more abstract features in higher layers is favoured by increasing the number of kernels from 64 to 256 from the first to the last convolutional layer. While the CNN model uses two final dense layers for classification, the CRNN model applies a bidirectional long short-term memory (LSTM) recurrent layer for temporal modeling prior to the final classification layer. Both models were trained using the categorical cross-entropy loss function. In our experiments, we found that adding a weight decay of 0.01 as regularization to all convolutional layers and the penultimate dense layer in the CNN model reduces the model's tendency to overfit to the training data.

3.3 Training Process

Due to the large size of the applied datasets (compare Section 4.1), we implemented a generator-based approach for training the neural networks discussed in Section 3.2 using the Keras API in TensorFlow[2]. This way, spectrogram data can be loaded efficiently from multiple files at the same time and fed to the model training process. During each epoch, 10 files are randomly selected for each language class and within each file, a random segment (patch) of 10 s (200 frames) duration is extracted. This procedure ensures an equal distribution of training items per class. Z-score normalization is applied to each patch to speed up the convergence of the learning algorithm. Finally, the training data for each epoch is stored in a tensor $X \in \mathbb{R}^{40\times200\times129\times1}$ with 40 mel spectrogram patches each having 200 frames and 129 frequency bins. The Adam optimizer with a learning rate of 10^{-5} is used.

4 EVALUATION

4.1 Datasets

In our experiments, we investigate three datasets with speech recordings. The *European Speech Repository*[3] is a collection of videos recorded at the European Parliament, as well as press conferences, interviews and dedicated training materials from EU interpreters. The *YouTube News Collection* is a collection of videos from various Youtube news channels. We gathered data from channels like BBC news, France24, DW News, and Noticias Telemundo.[4] The third dataset *Voxforge*[5] is an public-domain speech dataset that was set up to collect transcribed speech recordings to be used with free and open-source speech recognition systems.

The EU Speech and Youtube News datasets share a very high audio quality as well as a high diversity of speakers. As a potential drawback which might confuse a language classifier, all datasets also include occasional music jingles, silent segments as well as transitions between different news reports. The EU Speech dataset includes files with variable lengths ranging from two minutes up to 30 minutes. Similarly, the Voxforge dataset contains recordings between a few seconds to a few minutes. Only the Youtube News dataset includes audio files with a fixed length of 10 seconds.

[2]https://www.tensorflow.org
[3]https://webgate.ec.europa.eu/sr/
[4]The dataset is published at https://zenodo.org/record/3968292.
[5]http://www.voxforge.org

Table 2: Experiment configurations with applied network architecture, number of languages to be classified, training and test set, as well as validation and test accuracy A_V and A_T, respectively.

Network Architecture	# Languages	Training Set	A_V	Test Set	A_T
CNN	4	Youtube News	0.99	Voxforge	0.32
CNN	4	Youtube News (80%)	0.97	Youtube News (20%)	0.86
CNN	4	EU Repo (80%)	0.97	EU Repo (20%)	0.82
CRNN	4	EU Repo (80%)	0.98	EU Repo (20%)	0.83
CNN	6	EU Repo (80%)	0.87	EU Repo (20%)	0.71

4.2 Experiments

As shown in Table 2, we performed a number of experiments to evaluate the effectiveness of the two model architectures for the language identification task. The first column lists the applied network architecture. The second and third column lists the training and test set. In all configurations, we randomly selected 20% of the training set files as validation set. Finally, the last two columns show the validation set accuracy A_V and test set accuracy A_T. We performed language identification with either four classes (English, German, French, Spanish) or six classes (English, German, French, Spanish, Italian, Greek).

First, we trained the CNN model on the Youtube News dataset and observed a rather poor test accuracy of $A_T = 0.32$ on the Voxforge dataset. Here, no weight decay regularization was performed. Presumably, the number of model parameters is too high such that it overfits to the training set characteristics, which is confirmed by a very high validation accuracy of $A_V = 0.99$. While the Youtube News dataset contains professionally recorded broadcast data, the Voxforge data was self-recorded by non-experienced users. At the same time, this model fails to generalize to a different dataset. As a result, we introduced weigh decay regularization as explained in Section 3.2.

In a second set of experiments, we focused on train/test splits within particular datasets to evaluate different model architectures in different language scenarios. We used the same CNN model to train and test on 80%/20% file-wise splits of the European News Repository dataset given the four-language classification task. The CNN model achieves very good classification rates of $A_V = 0.97$ and $A_T = 0.82$. By adding the recurrent bidirectional LSTM layer in the CRNN model, only a small improvement by 0.01 for both accuracy measures can be observed, which indicates that there is no actual benefit of CRNN models for this task. Finally, we evaluated the CNN model on the European News repository for the six-language scenario. Despite the increasing complexity of the task, the model achieves good accuracy values of $A_V = 0.87$ and $A_T = 0.71$, which confirms that the CNN models are appropriate for the task at hand.

4.3 Conclusion

In this paper, we performed a study to evaluate a previously proposed method for language identification in speech recordings. Our results show both for scenarios with four and six languages that convolutional neural networks combined with mel spectrogram representations of speech signals are an adequate processing pipeline for the given task.

Languages are not easily treated as discrete, identifiable units with precise boundaries between them. This is especially valid for languages coming from the same language family such as Roman, Slavic, Germanic, or Baltic languages. Moreover, every language is characterized by variations between the communities that use it, which often leads to a variety of local accents and dialects.

Given a potential application scenario of a larger number of languages to be identified, a hierarchical classification approach could be beneficial. Here, the language family should be classified first. Then, specific classification models, which are optimized towards languages of certain families, will likely perform better than a general-purpose language classifier.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786993 and has been supported by the German Research Foundation (AB 675/2-1).

REFERENCES

[1] Christian Bartz, Tom Herold, Haojin Yang, and Christoph Meinel. 2017. Language identification using deep convolutional recurrent neural networks. In *International Conference on Neural Information Processing (ICONIP)*. Springer, Guangzhou, China, 880–889.

[2] Panikos Heracleous, Kohichi Takai, Keiji Yasuda, Yasser Mohammad, and Akio Yoneyama. 2018. Comparative Study on Spoken Language Identification Based on Deep Learning. In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*. Rome, Italy, 2265–2269.

[3] Rigas Kotsakis, Maria Matsiola, George Kalliris, and Charalampos Dimoulas. 2020. Investigation of Spoken-Language Detection and Classification in Broadcasted Audio Con-tent. *Information* 11, 4 (2020), 211.

[4] Gregoire Montavon. 2009. Deep learning for spoken language identification. In *NIPS Workshop on deep learning for speech recognition and related applications*. Vancouver, BC, Canada, 1–4.

[5] Shauna Revay and Matthew Teschke. 2019. Multiclass language identification using deep learning on spectral images of audio signals. *CoRR abs/1905.04348* (2019).

[6] Pedro A. Torres-Carrasquillo, Douglas A. Reynolds, and John R. Deller. 2002. Language identification using Gaussian mixture model tokenization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Orlando, FL, USA, I–757–I–760.

[7] Qian Zhang and John HL Hansen. 2018. Language/dialect recognition based on unsupervised deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 5 (2018), 873–882.