LangSonic

Fast Multilingual Speech Classification with CNNs

Henry Lefebvre [1], **Basel Ahsan** [1], and **Abrar Mohammad Fuad** [1] [1] McGill University

Abstract

LangSonic is a simple Convolutional Neural Network designed for swift and accurate spoken-language classification. Our system analyzes audio spectrograms to identify languages with precision. This final report illustrates the project's culmination, where we present our final training results and the integration of LangSonic into a web application.

Keywords CNN, Speech Classification, Machine Learning

1. Introduction

This paper presents the culmination of the LangSonic project, a Convolutional Neural Network designed for the rapid and precise classification of spoken languages. Focused on analyzing audio spectrograms, LangSonic follows a straightforward yet effective approach to language identification using Convolutional Neural Networks. The process detailed in this final report encapsulates the evolution of the project, highlighting the considerable strides made in enhancing the model's accuracy and performance since its preliminary stages.

Drawing on a robust subset of Mozilla Common Voice dataset (Ardila et al., 2020), the final training of LangSonic involved over 90,000 spectrograms for each language, ultimately achieving a commendable validation accuracy of 76%. This accomplishment positions LangSonic favorably, aligning with the 79% accuracy achieved by a comparable CNN developed by Sergey Vilov for the same task (Vilov, 2023).

2. Methods

The dataset used to train the LangSonic model was a subset of Mozilla Common Voice dataset comprising roughly 90,000 spectrograms for each language. Due to limited or low-quality data for others, we restricted ourselves to just 5 major languages with enough different speakers and validated data to train on.

2.a. Data Processing

The Mozilla Common Voice dataset is provided in the form of MP3 files with speaker and sentence-related metadata (Ardila et al., 2020). Except in selecting which languages to train on, we ignored the metadata and just used the audio itself. To get the audio to a state interpretable by our model, we first padded and truncated the dataset to a consistent length, cutting out any samples we deemed too short. We further trimmed 5% from the start and tail of each recording, as the data often began and ended with silence and clicking noises from volunteers beginning their recordings (Pietz, 2017).

Next, we converted these equal-length audio signals into their log-mel spectrogram representations. Mel spectrograms are a way of representing audio designed to mimic how humans process it, and have been used regularly in speech-processing and deep learning contexts, including state-of-the-art models like OpenAI's Whisper (Radford et al., 2022). We

Speech Classification Model

Published Nov 26, 2023

Key Points
- 76% Accuracy on 5 languages
- Fast prediction times
- Low training times

Correspondence to
Henry Lefebvre
henry.lefebvre@mail.mcgill.ca

Data Availability
Code is available on GitHub.
Try the live demo

chose a feature size of 13 mels because it is both large enough to be effective in language identifaction problems and small enough to be fast (Vilov, 2023). Sticking to a smaller mel count considerably reduces the model's size, allowing for faster iteration cycles, lower training time, a smaller memory footprint, and quicker prediction times. Due to the relatively quick training times for the model, we found that the primary rate-limiting step for test iterations was data processing. If we had more time or computational power, we would likely have done further experimentation with our data procesing pipeline before settling.

2.b. Model Architecture

We chose to use a Convolutional Neural Network (CNN) as our architecture due primarily to its simplicity, low training cost, and fast iteration time. Other models, such as Time Delay Neural Networks (Desplanques et al., 2020) or transformer-based architectures (Radford et al., 2022) have been shown to produce more accurate results when applied to speech processing problems, however they are often slow and computationally expensive to train and run.

The final LangSonic model architecture comprises several layers, including Conv2D layers with increasing filter sizes of 16, 32, 64, and 128, each followed by max-pooling layers for spatial downsampling. These layers extract hierarchical features from the input audio spectrograms. The flattened representation is then fed into a dense layer with 512 neurons, facilitating high-level feature learning. To prevent overfitting, a dropout layer is incorporated before the final dense layer, which produces the output classification. The model contains a total of 4.1M parameters.

3. Results

3.a. Final Model Performance

The final version of the LangSonic model achieved a final validation accuracy of 76%. This is comparable to the 79% achieved with a similar CNN made by Sergey Vilov and tested on the same dataset, albeit on a different set of languages (Vilov, 2023). We present several relevant metrics demonstrating the model's efficacy below.

3.b. Data & Metrics

As indicated in the confusion matrix, the classifier most often confuses Italian and Spanish, and English and German. Some of this can be explained as English and German sharing many auditory features due to both being Germanic languages. When allowing randomness in the training process, there is some run-to-run variance, and the most and least accurate models can change.

We present graphs depicting the validation accuracy and loss during the training process, offering insights into the model's learning over time. After approximately 10 epochs, further training offerred no increase in performance, while suffering from the usual drawbacks of overfitting. In order to mitigate overfitting, an early-stopping mechanism was used to save the best model during training and restore its weights once the model's validation loss stopped improving. On an M1 Pro chip, training on the entire dataset of 450,000 spectrograms for 10 epochs took approximately 20 minutes.

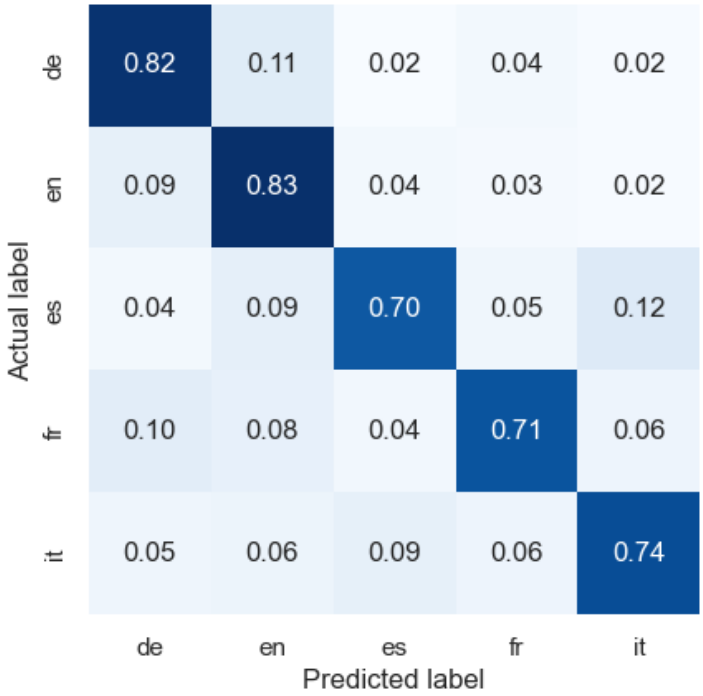
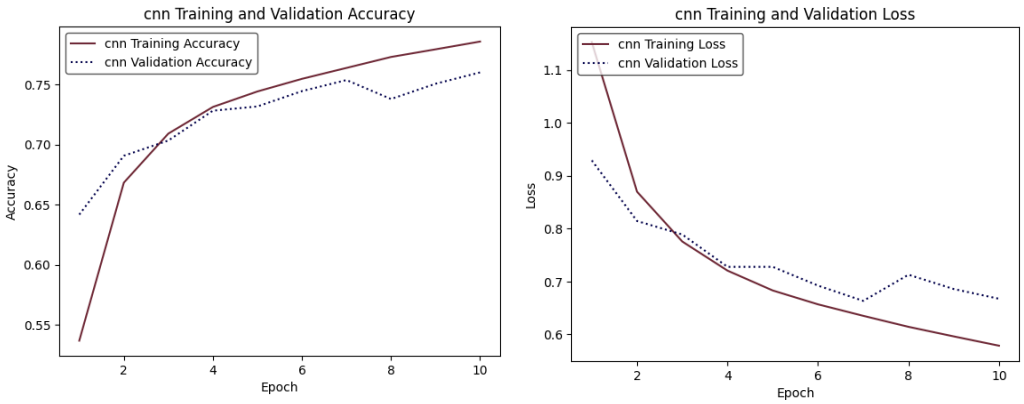


Figure 1: Confusion matrix

Language	Accuracy
English	83%
German	82%
Italian	74%
French	71%
Spanish	70%

Figure 2: Accuracy by language



(a) Validation Accuracy (b) Validation Loss

Figure 3: Validation Accuracy and loss during training

4. Demonstration

4.a. Website

The final product is a user-friendly web application, accessible at this URL. The web application displays a landing page where users can record audio and receive instant language classification predictions.

4.b. Tech Stack

Our team chose Flask for the web framework and Heroku for hosting. Flask's compatibility with Python allowed us to leverage existing audio processing code, ensuring consistently processed data. Our team's familiarity with Flask reduced the learning curve and streamlined development. Heroku was chosen for ease of deployment, although its 500MB slug size limit prevented us from utilizing several useful dependencies due to their size.

5. Conclusion

LangSonic's final training and integration have proven the model's potential in delivering swift and reliable language classification. Further enhancements to the model could lead to increased accuracy, and with careful training and data selection, the range of supported languages be expanded upon.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). *Common Voice: A Massively-Multilingual Speech Corpus*. 4211–4215.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020, October). *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*. https://doi.org/10.21437/interspeech.2020-2650

Pietz. (2017). *Pietz's Language Classifier*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv. https://doi.org/10.48550/ARXIV.2212.04356

Vilov, S. (2023). Spoken Language Recognition on Mozilla Common Voice. *Toward Data Science*.