Deep Learning Shared Task Group 39 – Assigning speaker labels to spoken utterances

Anne Fleur Donker (2081566), Balázs Gönczy (2078135), Hans Kuijpers (2068103), Muriël Valckx (2063049)

Feature engineering

We started with exploratory data analysis (EDA), e.g., identifying the number of speakers to classify. Through EDA, we noticed differences in the length of samples. Also, we noticed that many samples contained silence and noise. Therefore, we applied a **voice extraction** algorithm (Silero-VAD, 2022) to identify parts of the spoken audio sound samples. On the other hand, the original samples were passed through if the algorithm could not identify voice sound.
Furthermore, we also applied **noise removal** to retrieve audio samples containing spoken utterances with this method. Subsequently, we standardised the data by shortening the WAV-file samples to 11 seconds per unit and 8000 Hz.
Our initial feature engineering and transformation approach considered converting the WAV-file samples into spectrograms that could be used as 2D image input in the deep learning architecture. However, we opted for a numerical feature extraction approach due to storage issues. For this workflow, we used tools from the Librosa library to **extract the following features**, among others: a constant-q chromatogram, CENS features, Mel-spectrogram, spectral contrast, spectral flatness, the Fourier tempogram of the sample, the Roloff and multiple Mel frequency cepstral coefficients (MFCC). In addition, MFCCs can capture the phonetic characteristics of speech, which are essential in speaker identification (Hasan et al., 2004). Subsequently, we used descriptive statistics such as mean, standard deviation, and minimum and maximum (Maël Fabien, 2019). Next, we prepared for modelling by converting to tensors. After that, applied standard scaling to make the data more suitable for MLP. Finally, we split the data into train and test CSV files.

Considered Deep Learning Architectures and Training steps

A study from 2021 by the University of Chinese Academy of Sciences demonstrated the implementation of a CNN-GRU workflow to conduct speaker identification with > 97% accuracy (Ye & Yang, 2021). The CNN-GRU approach gave the right impression because it captured both temporal and spatial information in the input data. However, only the CNN part of the implementation was successful. Unfortunately, this did not yield satisfactory results with an accuracy of less than 0.1. Therefore, we considered a more straightforward MLP approach covering different feature engineering. Still, the model turned out to be overtrained, which led to an unreliable output of the validation set. Therefore, we retrained the data, which returned an accuracy of less than 0.3. Eventually, we used **Attentive Interpretable Tabular Learning Algorithm** from TabNet. TabNet is a fine-tuned, ready-to-use model optimised for classification, regression, and multi-task learning purposes (Arik & Pfister, 2020). Subsequently, we tried to eliminate backend errors (source) and then we ran the algorithm (Optimox, 2021).

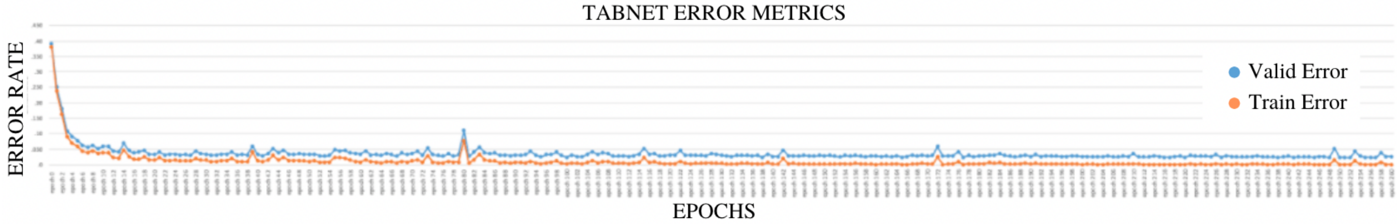
Hyperparameters and Optimisation

The combination industry-grade voice extractor algorithm and the TabNet algorithms with MLP architecture yielded an accuracy of more than 0.97. We tried fine-tuning the model with simple iterative replacing hyperparameters compared to the baseline, but this did not increase accuracy. Therefore, we reached out to the authors of the TABNET and the voice extractor algorithm to substantiate our findings after fine-tuning.

Discussion of Solution Performance

The graph below shows the learning rate of our training data and a plot of the training, validation and test set. Again, there was **minor** (less than 2.5% on average) **overfitting** due to the early stopping regularisation of the TABNET algorithm.

Figure 1
Tabnet Error Metrics



Our first submission returned an error rate of 0.0355. This version contained 25 epochs. The second submission had 357 epochs and produced an error rate of 0.0209. With the use of Cuda, we were able to run our first model within 30 minutes and the second within 2 hours 18 minutes. Also, the voice extractor and noise reducer algorithm could recognise speech in 99.75% of the cases within our 2000 subsample set (EDA). These results show that our final approach is highly efficient and precise, even in noisy environments.

CodaLab account

AnBaHaMu

Contribution per Group Member

Anne Fleur Donker: I contributed to the literature review for best performing deep learning workflows, emphasising feature extraction. I identified relevant feature extraction functions from the Librosa library and created a combinatory workflow with a voice extractor algorithm to create a CSV file containing all relevant extracted features. Furthermore, I contributed to the finalisation of the report.

Balázs Gönczy: I contributed with the following: Literature review for best-performing algorithms, conducting EDA on WAV files, preprocessing data for CNN analysis, applying voice extractor algorithm, creating basic CNN workflow and tuning, Adding more features to tabular learning, debugging and fine-tuning MLP algorithm, implementing and tuning TABNET algorithm, code refactoring for submission, outlining the logical framework of report.

Hans Kuijpers: I also contributed with the literary review for algorithms and online research on the dataset. Experimented with the CNN-GRU structure together with Balázs and set up the MLP algorithm. I helped with debugging where I could and testing the workflows to whether they were flawless. Also contributed to finalising the report.

Muriël Valckx: I helped find ways to convert the WAV-files into JSON or CSV format. Then I researched the use of spectrograms and the application of MLP. Next, I examined how we could apply feature extraction, frame the data, and look at converting time duration into a frequency. Finally, I contributed to writing the report, and ensured the contents were coherent and clear.

Literature

Arık, S. O., & Pfister, T. (2020). *TabNet: Attentive Interpretable Tabular Learning.* Sunnyville, CA: Google Cloud AI. Retrieved from https://arxiv.org/pdf/1908.07442.pdf.

Hasan, M. R., Jamil, M., & Rahman, M. G. R. M. S. (2004). Speaker identification using mel frequency cepstral coefficients. *variations*, *1*(4), 565-568.

Librosa. (n.d.). *Feature extraction*. Retrieved from www.librosa.org: https://librosa.org/doc/main/feature.html.

Maël Fabien. (2019, 7 december). Sound Feature Extraction. Consulted on 29 maart 2022, from: https://maelfabien.github.io/machinelearning/Speech9/#1-statistical-features.

Optimox. (2021, February 2). *TabNet : Attentive Interpretable Tabular Learning*. Retrieved from www.github.com: https://github.com/dreamquark-ai/tabnet.

Silero-VAD. (2022, February 22). *Silero VAD: pre-trained enterprise-grade Voice Activity Detector, Language Classifier and Spoken Number Detector.* Retrieved from www.github.com: https://github.com/snakers4/silero-vad.

Ye, F., & Yang, J. (2021). *A Deep Neural Network Model for Speaker Identification.* Beijing, China: Institute of Microelectronics of the Chinese Academy of Sciences.