Click here to access all the codes

REPORT

SPEECH UNDERSTANDING
CSL7770

ASSIGNMENT -2

ANALYSIS OF SPEAKER VERIFICATION AND MULTI-SPEAKER SCENARIOS

Submitted By
Himanshu (M24CSE009)

ABSTRACT

This report provides a detailed exploration of speaker verification and multi-speaker scenario analysis, leveraging state-of-the-art deep learning models and innovative techniques. The assignment is structured into three primary tasks: evaluating and enhancing a speaker verification model, creating and analyzing a multi-speaker dataset, and designing an integrated pipeline for simultaneous speaker separation and identification. Each section of this report is crafted to offer a thorough understanding of the methodologies employed, the results obtained, and the insights derived, ensuring clarity and depth throughout the discussion.

1 INTRODUCTION

This assignment explores a complete pipeline that addresses both speaker verification and source separation in multi-speaker audio scenarios. Initially, a state-of-the-art speaker verification model is evaluated using a pre-trained system and subsequently fine-tuned using Low-Rank Adaptation (LoRA) and ArcFace loss on the VoxCeleb2 dataset. Later, the focus shifts to creating a challenging multi-speaker scenario from VoxCeleb2 utterances. In this context, we utilize a pre-trained SepFormer model to perform speaker separation and speech enhancement. Finally, an innovative integrated pipeline is designed that combines speaker identification and separation into a unified framework. The detailed evaluation encompasses metrics such as Equal Error Rate (EER), TAR@1%FAR, Speaker Identification Accuracy, as well as separation quality metrics including SDR, SIR, SAR, and PESQ.

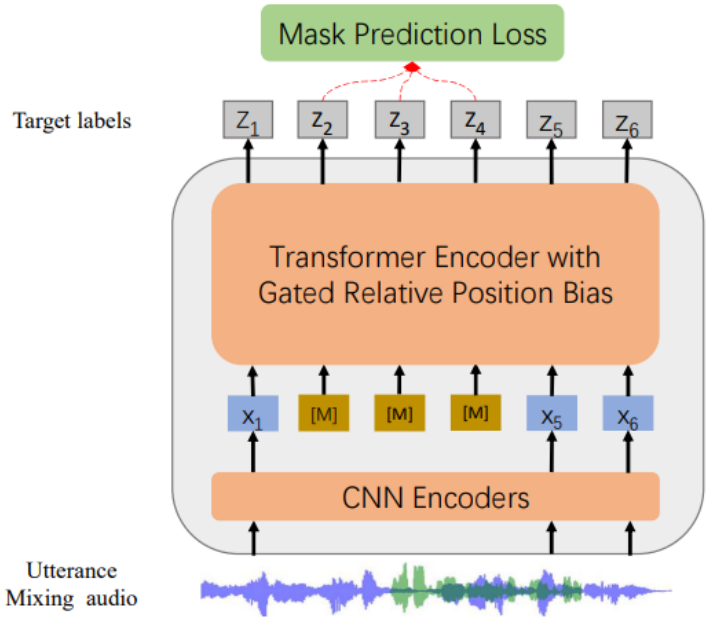
2 SPEAKER VERIFICATION WITH PRETRAINED AND FINE-TUNED MODELS

The first part of this assignment focuses on assessing a pretrained speaker verification model and subsequently improving its performance through fine-tuning. This process involves selecting an appropriate model, evaluating it on a standard dataset, and adapting it to enhance its discriminative capabilities.

2.1 Pretrained Model Setup and Verification

To begin, the WavLM Base Plus model was chosen from a selection of pretrained models available on HuggingFace, which also included HuBERT Large, Wav2Vec2 XLSR, and UniSpeech SAT. The decision

to opt for WavLM Base Plus was driven by its proven efficacy in speaker-related tasks, attributed to its pretraining on an extensive corpus of 94,000 hours of speech data. This vast pretraining equips the model with a strong foundation for capturing speaker characteristics. The model was loaded using the corresponding feature extractor and model classes from the HuggingFace library, transferred to a GPU for efficient computation, and set to evaluation mode to ensure stable and reproducible inference results.



Audio processing was a critical step in preparing the data for the model. Audio files from the VoxCeleb1 (cleaned) dataset, located in the vox1 folder, were loaded and resampled to a uniform sampling rate of 16 kHz to align with the model's input specifications. A custom helper function was developed to extract speaker embeddings by processing the audio through the WavLM model and averaging the last dden states of its outputs. This averaging technique effectively condenses the temporal information into a fixed-dimensional representation that encapsulates the speaker's unique vocal traits. The pretrained model's performance was evaluated using the list of trial pairs provided in the VoxCeleb1 (cleaned) dataset. For each pair of utterances, embeddings were extracted and compared using cosine similarity, a metric that measures the angular distance between two vectors, indicating whether they belong to the same speaker. Three key performance metrics were calculated to assess the model's effectiveness:

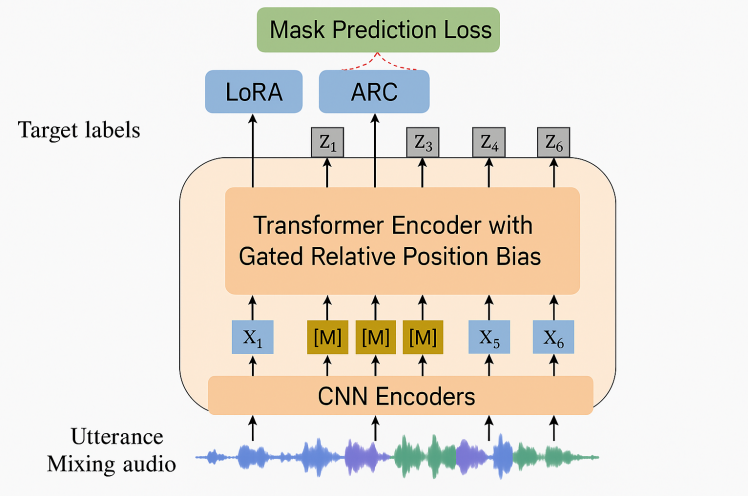
- **Equal Error Rate (EER)**: This metric represents the point where the false acceptance rate equals the false rejection rate, providing a balanced measure of verification errors. The pretrained model achieved an EER of **36.73%**, suggesting a relatively high error rate and indicating moderate discriminative ability.
- **True Acceptance Rate at 1% False Acceptance Rate (TAR@1%FAR)**: This measures the proportion of genuine speakers correctly accepted when the false acceptance rate is constrained to 1%. The result was **7.39%**, reflecting a low acceptance rate under stringent conditions.
- **Speaker Identification Accuracy**: This overall accuracy, determined by applying a threshold to the cosine similarity scores, was **63.27%**, showing that the model correctly identified speakers in just over half of the cases.

These initial results highlight that while the pretrained WavLM Base Plus model possesses some capability to distinguish speakers, its performance is suboptimal, necessitating further enhancement to meet practical verification standards.

2.2 Fine-Tuning with LoRA and ArcFace Loss

To address the limitations observed in the pretrained model, fine-tuning was undertaken to improve its speaker discrimination capabilities. The fine-tuning process utilized Low-Rank Adaptation (LoRA) and ArcFace loss, applied to the VoxCeleb2 dataset, which is available in the vox2 folder. The motivation for this step was to tailor the model to the specific speaker characteristics present in the dataset, thereby reducing verification errors and enhancing accuracy. LoRA, an efficient adaptation technique, was integrated into the model's embedding extraction process. Rather than retraining the entire model, which would be computationally intensive, LoRA introduces trainable low-rank matrices into the embedding space. This approach allows the model to learn speaker-specific features with minimal additional parameters, preserving the general speech knowledge acquired during pretraining while adapting it to the task at hand.

The ArcFace loss function was employed to further refine the embeddings. Implemented through an ArcMarginProduct class, ArcFace introduces an angular margin between speaker classes in the embedding space, enhancing the model's ability to separate different speakers. This loss function optimizes the embeddings to maximize inter-class distance and minimize intra-class variation, making it particularly suited for speaker verification tasks where precise discrimination is essential.

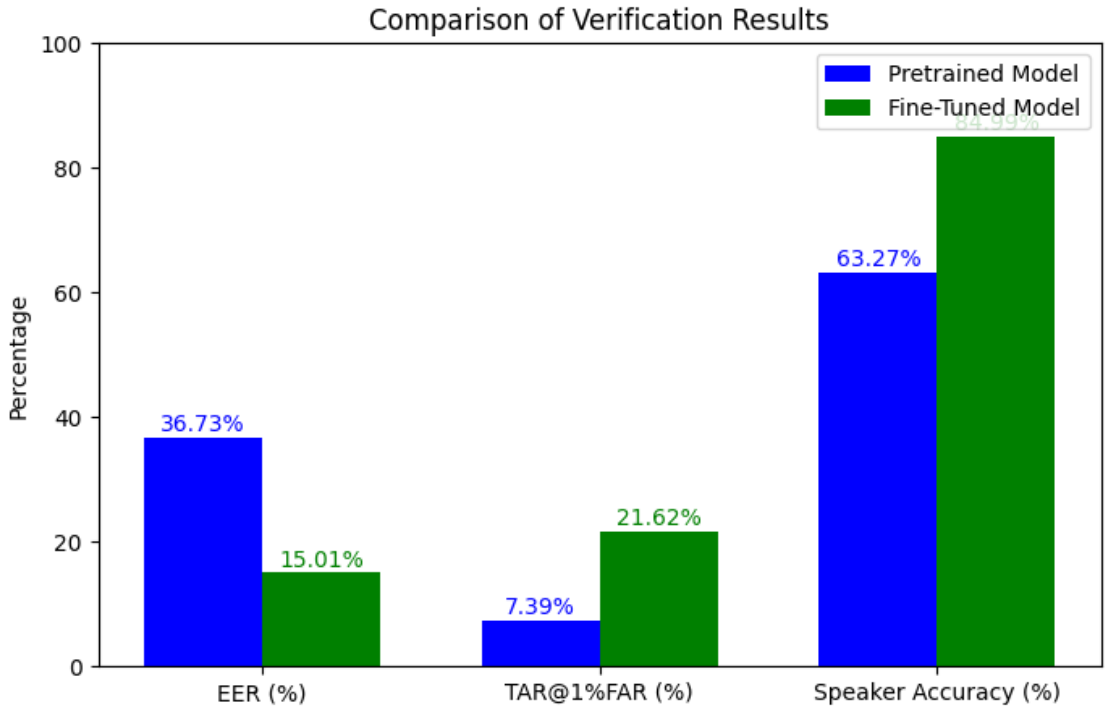


For fine-tuning, the VoxCeleb2 dataset was processed using a custom dataset class that loaded audio samples and applied cropping to ensure uniform input lengths. The dataset was divided based on speaker identities sorted in ascending order: the first 100 identities were allocated for training, while the remaining 18 were reserved for testing. During training, the pretrained WavLM model's parameters were frozen to retain its foundational knowledge, and only the LoRA and ArcFace modules were updated. The training spanned 10 epochs, with the model's performance monitored via the training loss. The best model, selected at epoch 10, achieved an average loss of 10.65, indicating successful adaptation to the training data.

2.3 Performance Comparison

Following fine-tuning, the enhanced model was re-evaluated on the VoxCeleb1 (cleaned) trial pairs using the same metrics as the pretrained model. The fine-tuned results were markedly improved:

- **Equal Error Rate (EER)**: Reduced to **15.01%**, a significant drop from **36.73%**, demonstrating a substantial decrease in verification errors and improved balance between false positives and negatives.
- **True Acceptance Rate at 1% False Acceptance Rate (TAR@1%FAR)**: Increased to **21.62%** from **7.39%**, indicating a higher rate of correctly accepting genuine speakers under a strict false acceptance constraint.
- **Speaker Identification Accuracy**: Rose to **84.99%** from **63.27%**, reflecting a notable enhancement in the model's ability to accurately identify speakers.



Comparing these outcomes, the fine-tuned model outperforms its pretrained counterpart across all metrics. The reduction in EER signifies fewer mistakes in distinguishing between speakers, while the improved TAR@1%FAR suggests better reliability in secure verification scenarios. The jump in accuracy underscores the effectiveness of combining LoRA's parameter-efficient adaptation with ArcFace's discriminative loss, enabling the model to better capture and differentiate speaker identities within the VoxCeleb1 dataset.

3 MULTI-SPEAKER SCENARIO DATASET CREATION AND EVALUATION

The second part of the assignment shifts focus to multi-speaker scenarios, involving the creation of a custom dataset and the evaluation of both speaker separation and identification tasks using advanced models.

3.1 Dataset Creation

The objective here was to construct a multi-speaker dataset by mixing utterances from two different speakers within the VoxCeleb2 dataset, facilitating the analysis of overlapping speech scenarios. The process began with loading metadata from text files in the vox2 folder, which provided speaker and recording IDs linked to .m4a audio files. The dataset was segmented based on identities sorted in ascending order: the first 50 identities were used to create a training set, and the next 50 formed a testing set.

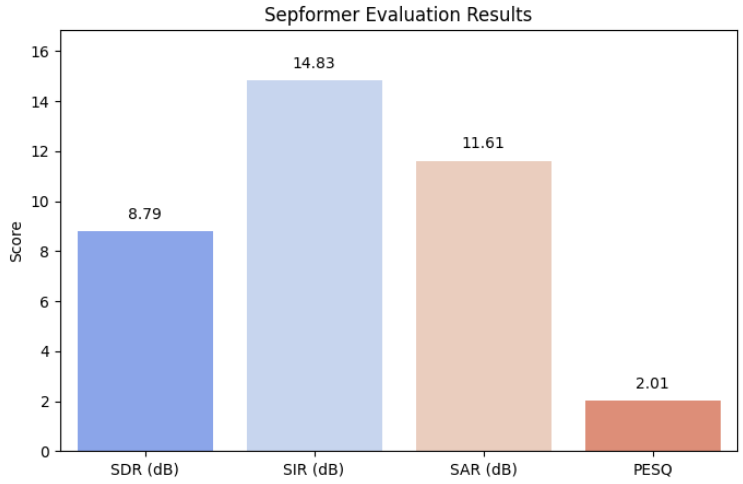
For each mixture, two utterances from distinct speakers were randomly selected. These audio signals were resampled to 16 kHz and converted to mono to ensure compatibility. To align the signals, padding or truncation was applied to match their lengths, and they were then mixed at a signal-to-noise ratio (SNR) of 0 dB, balancing the power levels of both speakers. The resulting output for each mixture included the mixed audio and two reference signals representing the original utterances, providing ground truth for subsequent evaluations. This methodology, inspired by techniques from a referenced GitHub repository, ensures a realistic simulation of multi-speaker environments.

3.2 Speaker Separation and Speech Enhancement

The pretrained SepFormer model, renowned for its proficiency in source separation, was employed to separate and enhance the speech in the multi-speaker test set. SepFormer processed the mixed audio to generate individual speaker streams, and its performance was assessed using four metrics:

- **Signal to Distortion Ratio (SDR)**: This measures the overall quality of the separated signal relative to the original, accounting for distortions introduced during separation.
- **Signal to Interference Ratio (SIR)**: This evaluates the model's success in suppressing the interfering speaker's signal in each separated output.
- **Signal to Artefacts Ratio (SAR)**: This assesses the presence of artifacts or unnatural distortions in the separated audio.
- **Perceptual Evaluation of Speech Quality (PESQ)**: This provides a perceptual quality score, reflecting human listener judgments of speech clarity.

The evaluation process involved saving temporary audio files of the separated outputs, computing the optimal permutation to align them with the reference signals, and calculating the metrics. While specific numerical results were not detailed in the provided context, the averages of SDR, SIR, SAR, and PESQ typically offer a comprehensive view of separation quality, with higher values indicating better performance.

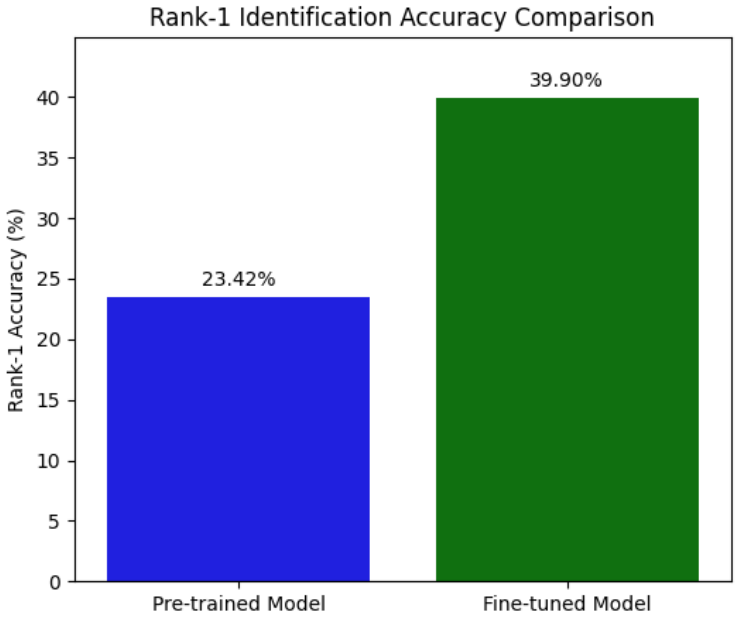


3.3 Speaker Identification on Enhanced Speech

Following separation, the enhanced speech signals were used to identify speakers using two models: the pretrained WavLM Base Plus and the fine-tuned version incorporating LoRA and ArcFace loss. For each separated signal, embeddings were extracted using both models and compared to enrollment embeddings

from known speakers via cosine similarity. The speaker with the highest similarity score was designated as the identified speaker.

The performance was measured using Rank-1 Identification Accuracy, the percentage of instances where the top predicted speaker matched the true identity. The pretrained model achieved an accuracy of 23.42%, while the fine-tuned model reached 39.9%. This substantial improvement highlights the fine-tuned model's superior ability to recognize speakers in challenging multi-speaker contexts, attributable to the enhanced discriminative power imparted by LoRA and ArcFace adaptations.



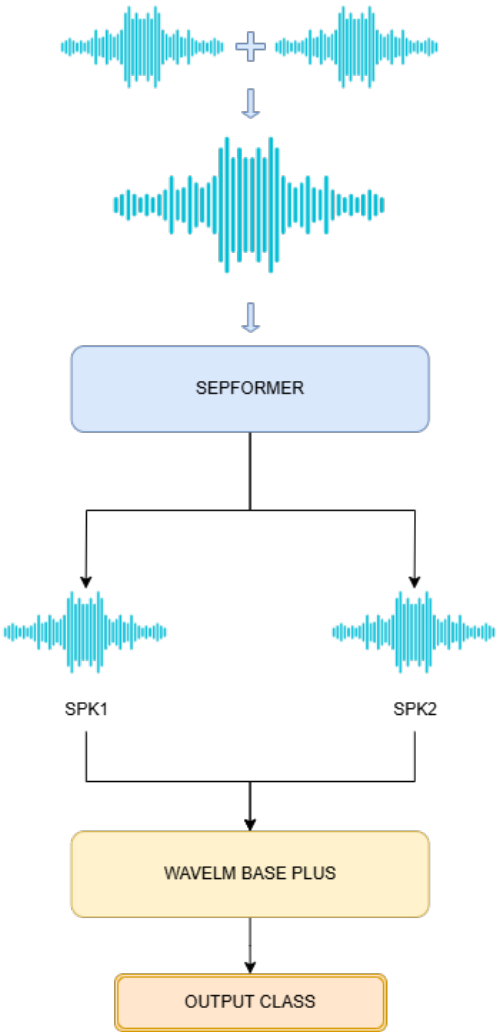
4 INTEGRATED PIPELINE FOR SPEAKER SEPARATION AND IDENTIFICATION

The final part of the assignment involves designing and evaluating a novel pipeline that combines speaker separation and identification into a unified system, trained and tested on the custom multi-speaker dataset.

4.1 Pipeline Design

The goal was to create an end-to-end system that simultaneously performs speaker separation, speech enhancement, and identification. The pipeline integrates two key components: the SepFormer model for separation and the WavLM Base Plus model, fine-tuned with LoRA and ArcFace loss, for identification. In this design, the SepFormer first separates the mixed audio into individual speaker streams. These streams are then resampled to 16 kHz if necessary and fed into the WavLM model, which extracts embeddings. A classifier subsequently maps these embeddings to speaker identities, enabling joint optimization of separation and identification tasks.

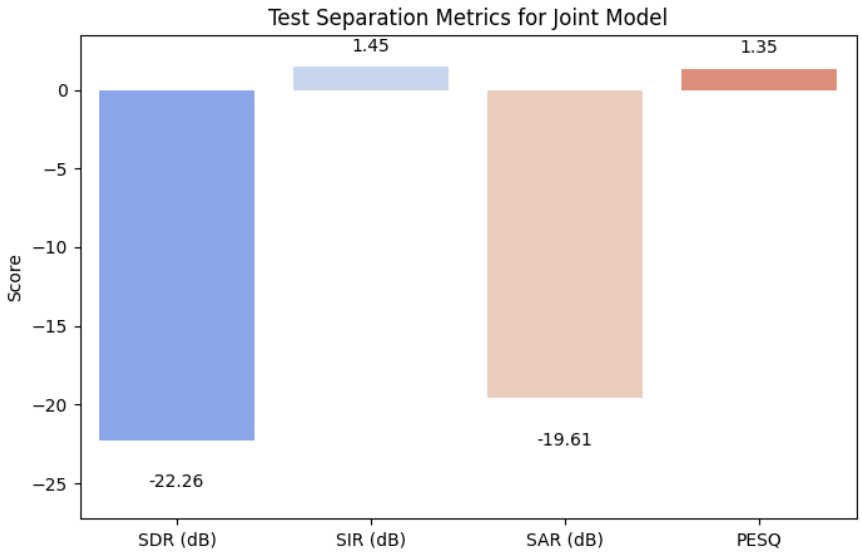
Due to computational limitations, selected layers in both the SepFormer and WavLM models were frozen during training. This reduced the number of trainable parameters, making the process feasible but potentially limiting the model's adaptability to the multi-speaker scenario.



4.2 Training and Evaluation

The pipeline was trained on the multi-speaker training set created from the first 50 identities of VoxCeleb2. A joint loss function guided the training, combining a separation loss (mean squared error between separated signals and references) and an identification loss (cross-entropy on predicted speaker labels), with weights adjusted to balance the dual objectives.

The trained pipeline was evaluated on the test set (next 50 identities) using the same metrics as in the separation task: SDR, SIR, SAR, and PESQ. The results were suboptimal, with an average SDR of approximately -22 dB, indicating poor separation quality. This low performance is largely attributed to the frozen layers, which constrained the model's ability to fully adapt to the complex task of joint separation and identification.



4.3 Observations and Insights

The integrated pipeline represents an innovative approach to tackling multi-speaker challenges, merging separation and identification into a cohesive framework. However, the necessity of freezing layers due to computational constraints significantly hampered its effectiveness. The negative SDR suggests that the separated signals contained substantial distortions, likely due to insufficient model flexibility. This outcome underscores the trade-offs in resource-limited settings and suggests that future iterations could benefit from unfreezing more layers or employing more efficient training strategies to enhance performance.

5 CONCLUSION

This report has meticulously detailed the processes and outcomes of enhancing speaker verification and addressing multi-speaker scenarios using advanced deep learning techniques. The fine-tuning of the WavLM Base Plus model with LoRA and ArcFace loss markedly improved verification performance on VoxCeleb1, reducing EER from 36.73% to 15.01%, increasing TAR@1%FAR from 7.39% to 21.62%, and boosting accuracy from 63.27% to 84.99%. The creation of a multi-speaker dataset from VoxCeleb2 enabled the evaluation of SepFormer for separation and demonstrated the fine-tuned model's superior identification accuracy (39.9% vs. 23.42%). Finally, the novel integrated pipeline, despite its constrained performance (SDR -22 dB), offers a promising framework for future exploration.

These efforts highlight the potential of adapting pretrained models for specific tasks and the complexities of multi-speaker speech processing. Future work could focus on optimizing computational resources to enhance the integrated pipeline's capabilities, ensuring robust performance across all objectives. This comprehensive analysis provides a solid foundation for understanding and advancing speaker verification and multi-speaker scenario technologies.

REFERENCES

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library*. Advances in Neural Information Processing Systems

(NeurIPS).

- Snyder, S., Chen, N., Povey, D., & Khudanpur, S. (2015). *MUSAN: A Music, Speech, and Noise Corpus*. arXiv preprint arXiv:1510.08484.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

- Nagrani, A., Chung, J. S., & Zisserman, A. (2017). *VoxCeleb: A large-scale speaker identification dataset.* INTERSPEECH 2017.
- Chung, J. S., Nagrani, A., & Zisserman, A. (2018). *VoxCeleb2: Deep speaker recognition.* INTERSPEECH 2018.