Evaluating the Effectiveness of Transformer Layers in Wav2Vec 2.0, XLS-R, and Whisper for Speaker Identification Tasks

Michael Saxer, Linus Stuhlmann
ZHAW - School of Engineering
{saxermi1, stuhllin}@students.zhaw.ch
https://github.com/slinusc/speaker_identification_evaluation

Abstract

This study evaluates the performance of three advanced speech encoder models—Wav2Vec 2.0, XLS-R, and Whisper—in speaker identification tasks. By fine-tuning these models and analyzing their layer-wise representations using SVCCA, k-means clustering, and t-SNE visualizations, we found that Wav2Vec 2.0 and XLS-R capture speaker-specific features effectively in their early layers, with fine-tuning improving stability and performance. Whisper showed better performance in deeper layers. Additionally, we determined the optimal number of transformer layers for each model when fine-tuned for speaker identification tasks.

1 Introduction

Speaker recognition is a critical component of Natural Language Processing (NLP) and audio processing. Advanced encoders like XLS-R (Conneau et al., 2020), Wav2Vec2 (Baevski et al., 2020), and Whisper (Radford et al., 2022) use multiple transformer layers to extract detailed acoustic and phonetic features from audio files.

Our previous research indicated that speech encoders like XLS-R capture the most relevant speaker information in the early layers. However, this conclusion was based on a single model and a limited dataset of 25 speakers, which may not be sufficient for a robust validation. In this study, we aim to extend our research by using a larger and more diverse dataset and employing more sophisticated methods to evaluate the contribution of specific layers to speaker identification tasks. We will also compare multiple state-of-the-art models, namely Wav2Vec2, XLS-R, and Whisper's encoder. By fine-tuning these models and analyzing the performance of different layers within each model, we aim to identify the optimal configurations for maximizing effectiveness and efficiency in speaker recognition tasks.

2 Related Work

Pasad et al. examined the contributions of different layers to overall performance in various speech related tasks by using the Canonical Correlation Analysis (CCA). Lukic et al. demonstrated the effectiveness of CNNs in learning speaker-specific features from spectrograms for speaker identification and clustering. Fan et al. utilized t-SNE for visual analysis, finding that early layers of Wav2Vec 2.0 are more distinctive for speaker verification and language identification. Building on similar approaches, our study aims to extend the analysis across multiple models to gain a deeper understanding of speech encoders.

3 Experiment

The experimental setup is designed to evaluate speech encoder models for speaker identification tasks through a detailed multi-stage process.

1. In the initial stage, we fine-tune the models on a speaker identification task.
2. The second stage involves extracting the hidden states from each layer of three models: XLS-R, Wav2Vec, and Whisper and their respective finetuned versions.
3. In the next stage, we apply Singular Vector Canonical Correlation Analysis (SVCCA) to the extracted hidden states.
4. We then conduct k-means clustering on the hidden states, evaluating layer-wise clustering results.
5. To visualize the data in 2D, we apply t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction.
6. Finally, we use Optuna's hyperparameter optimization framework to determine the optimal number of transformer layers for speaker identification tasks.

3.1 Data

We curated a subset of the Mozilla Common Voice dataset (Ardila et al., 2019), encompassing a wide range of languages (English, German, French, Spanish, Chinese) and balanced in terms of gender, with an equal representation of female and male speakers, providing a comprehensive basis for evaluating the models' performance across different linguistic and gender groups. Each audio sample has an average duration of 4.9 seconds. For the feature extraction, we selected 50 speakers from this dataset, with each speaker contributing 50 mp3 samples. For the fine-tuning phase, we used 110 different speakers, each providing 100 samples.

3.2 Encoder Models

In this experiment, we use three encoder models: Wav2Vec 2.0, XLS-R 300M, and Whisper. For comparison purposes, we utilized versions of these models with the same number of transformer encoder layers. Detailed specifications of these models are provided in Table 1.

3.2.1 Wav2Vec 2.0

Developed by Facebook AI (Baevski et al., 2020), Wav2Vec 2.0 is a self-supervised speech representation model. It segments raw audio into fixed-length frames, transforms these into feature vectors using a convolutional neural network (CNN) encoder, and then processes the discretized features through 24 transformer encoder layers.

3.2.2 XLS-R 300M

XLS-R 300M, also created by Facebook AI (Conneau et al., 2020), is a multilingual extension of Wav2Vec 2.0. It processes audio by converting the waveform into frames, generating feature vectors through a CNN encoder, and then feeding the quantized features into 24 transformer encoder layers.

3.2.3 Whisper

Developed by OpenAI (Radford et al., 2022), Whisper is designed for robust speech recognition. For this project, we used the encoder of the large model version, which converts raw audio into a mel-spectrogram, transforms it into feature vectors, and processes these vectors through its 24 transformer encoder layers.

3.3 Fine-Tuning Process

During the fine-tuning phase, we trained the models on a speaker identification task, targeting an accuracy of approximately 90%. For all models, we utilized a linear classification head with a mean pooling layer, comprising a fully connected layer followed by a softmax activation function to output class probabilities. Further details of the fine-tuning process are summarized in Table 2.

3.4 Hidden State Extraction

We used the three previously introduced models and their fine-tuned counterparts, to extract hidden states from each encoder layer. Layer 0 represents the initial audio processing layer, which passes the tensor to the first transformer layer. After extracting the hidden states, we applied mean pooling along the temporal axis to standardize dimensions and reduce data volume.

3.5 SVCCA

The Singular Vector Canonical Correlation Analysis (SVCCA) proposed by Raghu et al., combines Singular Value Decomposition (SVD) (Andrews and Patterson, 1976) with Canonical Correlation Analysis (Hotelling, 1936) to analyze neural network representations. CCA examines the linear relationships between two multidimensional variables, while SVD decomposes a matrix into singular a vector. We used SVCCA to evaluate the effectiveness of different layers within an encoder model for speaker recognition by identifying the most significant layers. We compute the correlation between the hidden state embeddings X and speaker labels Y. Layers with high maximal canonical correlations tend to be informative and discriminative for speaker recognition.

3.6 K-Means Clustering

We applied K-Means clustering (MacQueen, 1967) to the hidden states of different encoder model layers to group speaker embeddings and evaluate their distinctiveness. Cluster quality was assessed using three metrics: Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), which measures the similarity between model-generated and true clusters; Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002), which quantifies the mutual dependence between predicted and true clusters; and Silhouette Score (Rousseeuw, 1987), which assesses how well an object matches its own cluster compared to other clusters.

3.7 t-SNE

The dimensionality of the data is further reduced using t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008), facilitating the visualization of the encoder model's hidden states in a two-dimensional space. This method allows for a clearer observation of how the speaker embeddings cluster and helps discern patterns and structures not visible through linear analysis. In this project, we visualized the 2D representation of 10 speakers at layers 0, 5, 10, 15, 20, and 24 for each model.

3.8 Optuna

We utilized Optuna, a hyperparameter optimization framework, to determine the optimal number of transformer layers for speaker identification tasks in the Wav2Vec, XLS-R, and Whisper models, using the number of layers as a hyperparameter. Optuna employs Bayesian optimization to maximize the performance of a model, dynamically adjusting trials based on prior results (Akiba et al., 2019).

4 Results

4.1 SVCCA

The original version of the Wav2Vec2 model, shown in Figure 1, exhibits a steep increase in maximal canonical correlation after layer 0, reaching its highest values between layers 1 to 5. Two significant declines are observed between layers 10 to 17 and 20 to 24, with the lowest correlation occurring between layers 22 to 24. The fine-tuned Wav2Vec2 model shows more stability across the layers but mirrors similar patterns, achieving the highest score at layer 7 and the lowest at layer 24.
The XLS-R models (Figure 2) show comparable patterns to Wav2Vec2 but with generally higher and more consistent correlations. Both versions decline between layers 7 and 21, with the highest scores between layers 1 and 5.
For the Whisper model (Figure 3), the correlation increases in the first 5 layers and stabilizes thereafter. The original model peaks at layer 13, while the fine-tuned model shows consistently lower performance from layer 5 onward.

4.2 k-Means Clustering

In Figure 4, the Wav2Vec2 model's ARI and NMI scores peak at layer 8 and decline steadily thereafter, while the Silhouette Score decreases significantly after layer 6. The fine-tuned model shows more stable and higher ARI and NMI scores throughout the layers, with the Silhouette Score remaining stable and reaching its highest value at layer 24.
For the XLS-R model in Figure 5, both the original and fine-tuned versions show high ARI and NMI scores from the beginning. The original model peaks around layers 3-4 and experiences a steep decline after layer 21. The fine-tuned model maintains high scores throughout, with a similar but less extreme decline after layer 21. The Silhouette Score remains relatively constant, with a valley between layers 10 and 18 in the original model.
In Figure 6, the original Whisper model's ARI and NMI scores increase up to layer 7 and then plateau, while the Silhouette Score remains constant. The fine-tuned Whisper model follows a similar pattern but with consistently lower ARI and NMI scores and slightly improved Silhouette Scores.

4.3 t-SNE Visualizations

The visualization 7 of the Wav2Vec2 model shows loosely packed clusters with clearly separated label categories in Layer 0. By Layer 5, these clusters become distinct. From Layer 10 onward, overlaps increase, making speaker separation difficult. By Layer 24, the data points are heavily intermixed. In contrast, the fine-tuned Wav2Vec2 model shown in Figure 8 consistently forms clear clusters across all layers. Starting from Layer 0, the model effectively distinguishes speakers. Layers 5 and 10 maintain well-defined separations, with slightly more overlap by Layer 20. By Layer 24, the clusters become less dense, with more overlap between the labels.
The original XLS-R model in Figure 9 forms clear clusters in Layer 0 with clearly separated label categories, becoming more distinct by Layer 5. However, from Layer 10 onward, overlaps increase, complicating separation. By Layer 24, data points are heavily intermixed. Conversely, the fine-tuned XLS-R model 10 shows improved differentiation at each layer. Layer 0 already has defined clusters, which remain well-separated through Layers 5, 10, and 15. By Layer 24, clusters start to blend but remain recognizable.

In Layer 0 of the original Whisper model 11, many overlaps are present. By Layer 5, clusters become better defined. Layer 10 clusters are denser, with clearer separation by Layer 15. In Layer 20, definitions improve slightly, but overlaps persist. By Layer 24, clusters are well-recognizable with some overlaps. The fine-tuned Whisper model 12 shows mixed data points in Layer 0 with hints of clusters. By Layer 5, clusters improve but still overlap. Layers 10, 15, and 20 show recognizable but not dense clusters. By Layer 24, clusters remain recognizable but not dense.

4.3.1 Layer Optimization

The optimal number of transformer layers for speaker identification tasks was determined for each model. The best configurations are: 7 layers for Wav2Vec2, 3 layers for XLS-R, and 16 layers for Whisper.

5 Discussion

The SVCCA results for the Wav2Vec2 model closely align with clustering performance metrics ARI, NMI, and the t-SNE visualizations, underscoring the importance of early layers in capturing speaker-specific features. High scores in the first eight layers highlight their significance, while declines in deeper layers suggest these layers encode more abstract information, consistent with (Pasad et al., 2023). Post fine-tuning, Wav2Vec2 exhibits higher and more stable performance, with increased Silhouette Scores but decreased ARI and NMI in the later layers, indicating improved overall cohesion but challenges in capturing complex, possibly non-linear structures.

XLS-R, an extension of Wav2Vec2 trained on a larger and more diverse dataset, shows higher overall performance, highlighting the benefit of its extensive training. XLS-R demonstrates more stable patterns and denser clusters, with further improvements after fine-tuning. The sharp decline after layer 21 suggests that later layers capture more complex information, which is harder to interpret. These findings align with our prior work, even when different methods were used.

Whisper initially exhibits lower SVCCA and clustering metrics, as evident in the t-SNE visualizations. This difference may stem from Whisper's different approach of processing raw audio files using Mel spectrograms instead of raw audio signals, as in Wav2Vec2 and XLS-R. However, deeper layers in Whisper show improved speaker differentiation, indicating its architecture becomes more effective with depth. The fine-tuning process for Whisper results in more consistent but overall poorer performance, possibly due to an insufficiently large fine-tuning dataset, especially given Whisper's considerably larger model size.

The layer optimization process identifies specific layers as optimal for speaker recognition tasks across different models: layer 7 for Wav2Vec2, layer 3 for XLS-R, and layer 16 for Whisper. These layers were determined using the Optuna optimizer, which aligns closely with those previously identified as having the highest information-capturing abilities. This alignment underscores the reliability of the optimization process. Furthermore, these findings suggest that by selecting fewer but more effective encoder layers, we can achieve improved performance in speaker recognition tasks while simultaneously reducing computational resources.

6 Limitations

The analysis techniques used in this study, namely Singular Vector Canonical Correlation Analysis and k-means clustering, are primarily suited for identifying linear relationships. Consequently, these methods may not adequately capture or address more complex, non-linear relationships.

For comparison purposes, we focused exclusively on model versions with 24 transformer layers. This uniformity might limit the exploration of potentially optimal configurations with different layer counts.

Additionally, the fine-tuning process for the Whisper model was conducted using a relatively small dataset due to limited computational resources. It is expected that employing a larger dataset could significantly enhance the performance of the fine-tuned models, particularly for Whisper.

7 Conclusion and Future Work

This study provides insights into the layer-specific performance of the Wav2Vec2, XLS-R, and Whisper models in speaker recognition tasks. Our analysis highlights the importance of early layers in capturing speaker-specific features, as evidenced by the strong alignment of SVCCA, t-SNE, and k-means clustering metrics, particularly in the initial layers. The identified optimal layers for fine-tuning these speech encoder models align with our observations.

Future work should focus on a detailed analysis to identify which speaker attributes most significantly contribute to recognizing speaker-specific characteristics. Additionally, a comparison of different model sizes should be conducted to further understand their impact on performance. Moreover, the differing patterns in layer performance between Whisper and the other models should be further investigated to uncover underlying factors contributing to these differences.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *Preprint*, arXiv:1907.10902.

H. Andrews and C. Patterson. 1976. Singular value decomposition (svd) image coding. *IEEE Transactions on Communications*, 24(4):425–432.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *Preprint*, arXiv:2006.13979.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2020. Exploring wav2vec 2.0 on speaker verification and language identification. *arXiv preprint arXiv:2012.06185*.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Yanick Lukic, Carlo Vogt, Oliver Dürr, and Thilo Stadelmann. 2016. Speaker identification and clustering using convolutional neural networks. In *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA. University of California Press.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *arXiv preprint arXiv:1706.05806*.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. In *Journal of Machine Learning Research*, pages 583–617. Citeseer.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

A Appendix

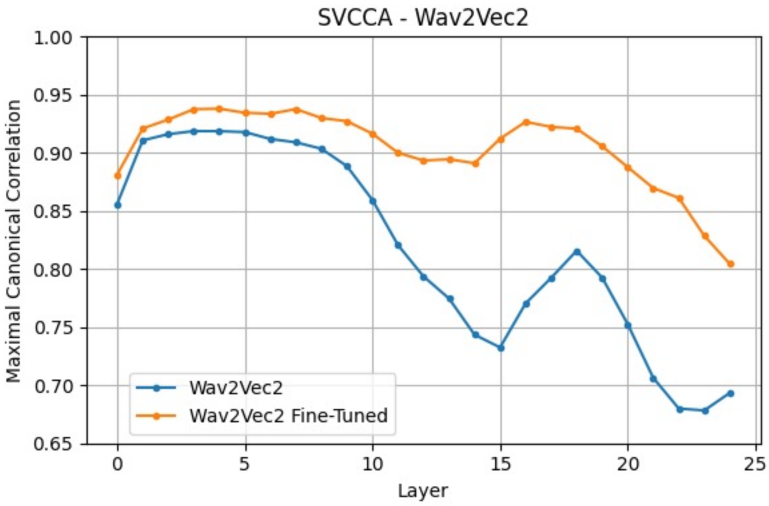

Figure 1: SVCCA- Wav2Vec2

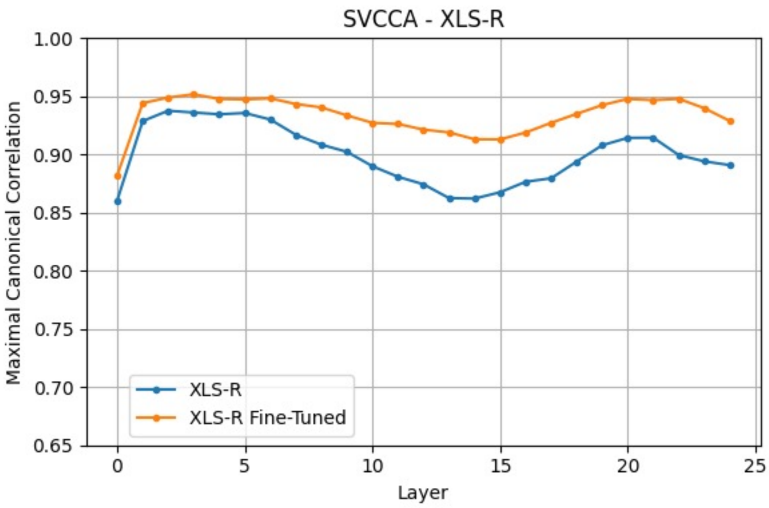

Figure 2: SVCCA- XLS-R

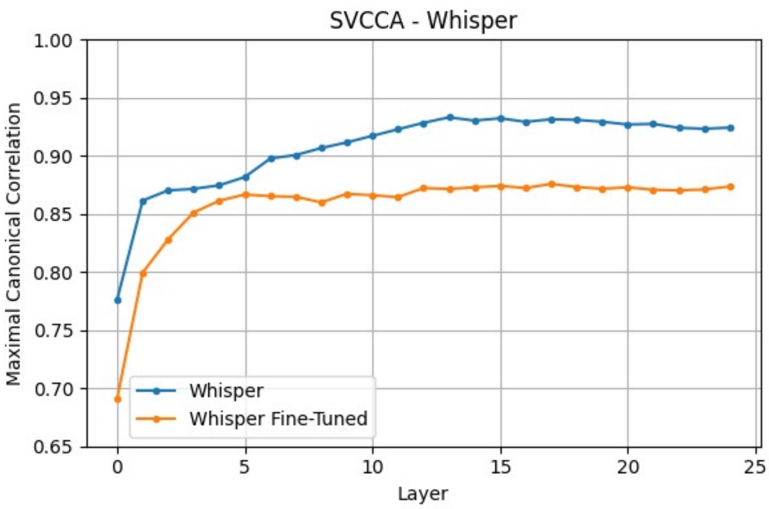

Figure 3: SVCCA - Whisper

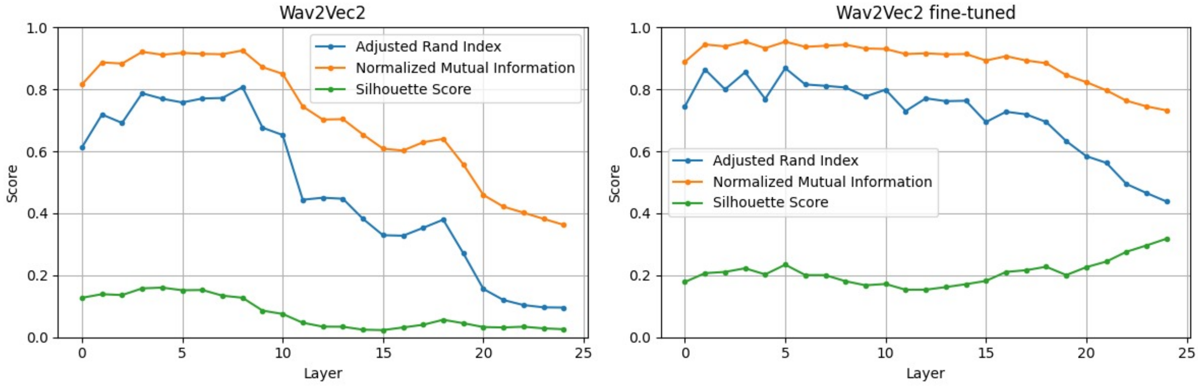

Figure 4: K-Means Clustering - Wav2Vec2

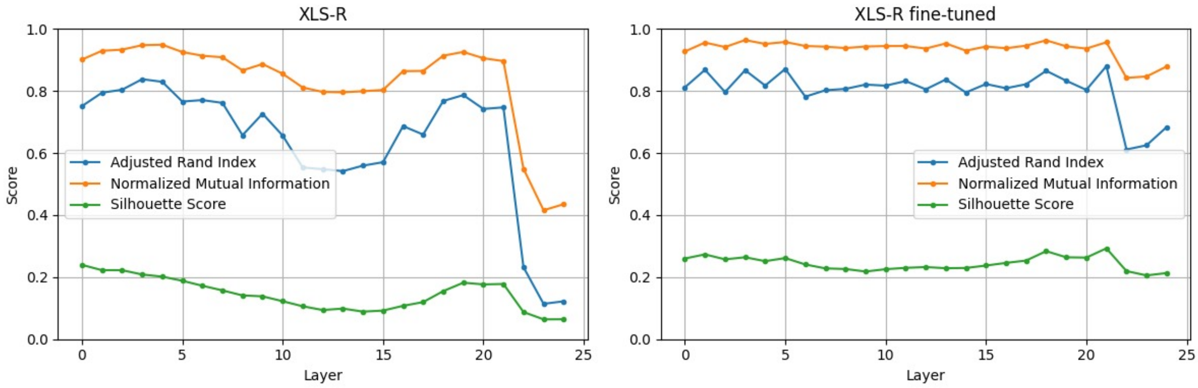

Figure 5: K-Means Clustering - XLS-R

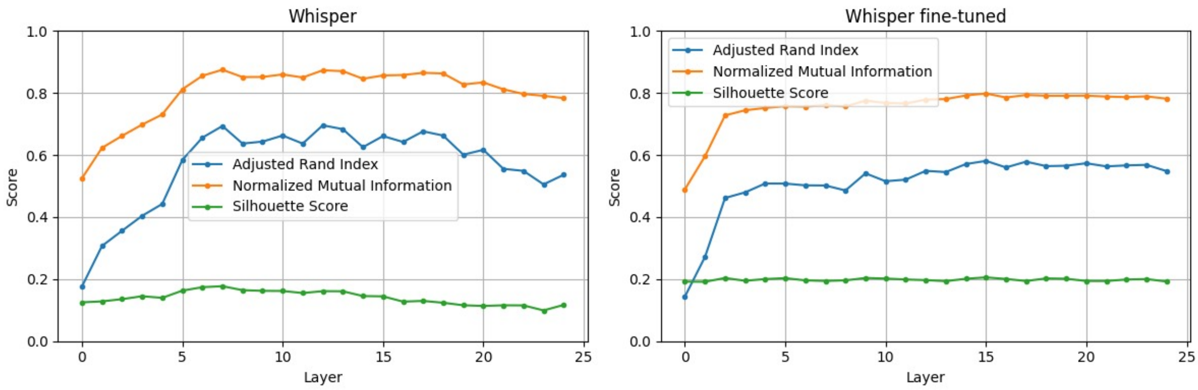

Figure 6: K-Means Clustering - Whisper

Figure 7: T-SNE - Wav2Vec2
The colors of the dots represent different speaker labels.

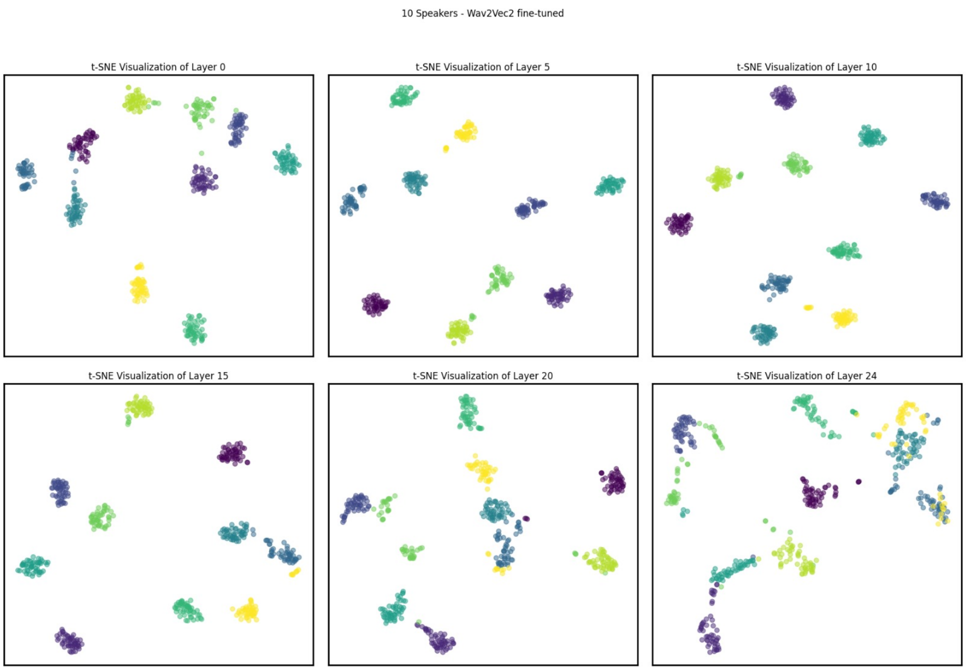

Figure 8: T-SNE - Wav2Vec2 Finetuned
The colors of the dots represent different speaker labels.

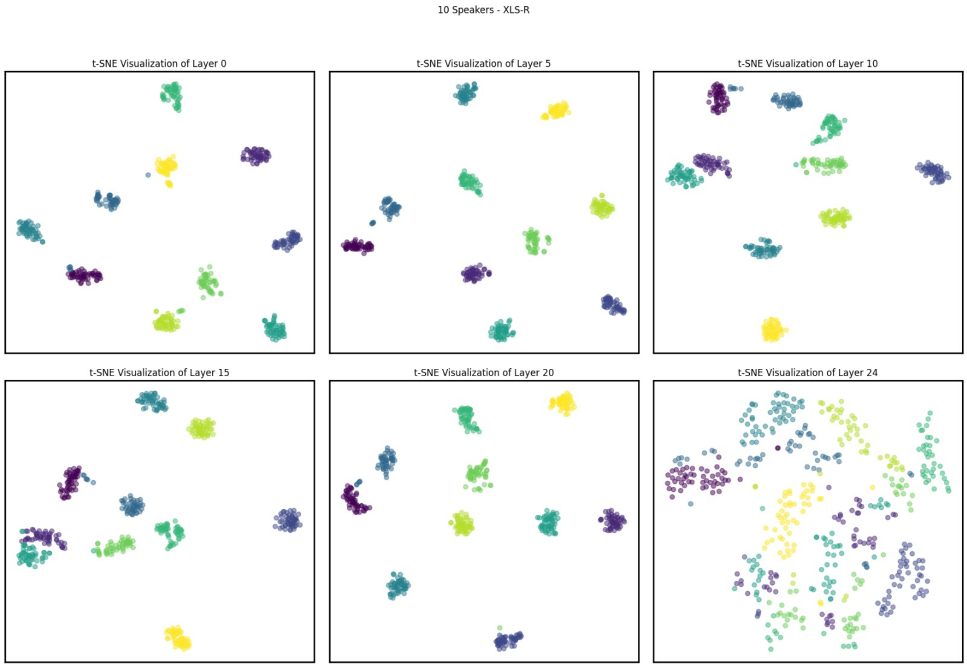

Figure 9: T-SNE - XLS-R
The colors of the dots represent different speaker labels.

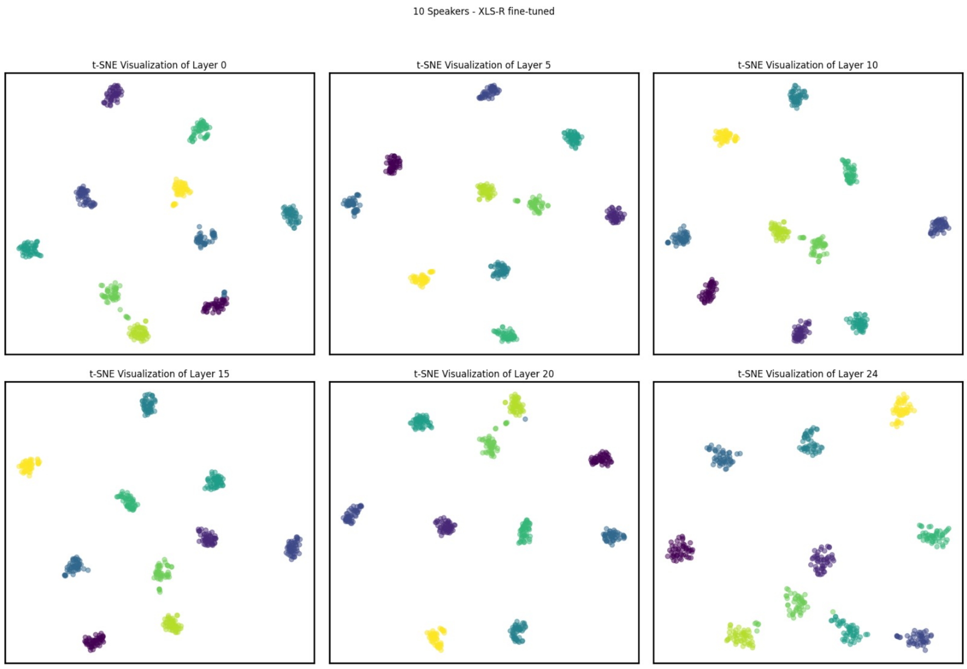

Figure 10: T-SNE - XLS-R Finetuned
The colors of the dots represent different speaker labels.

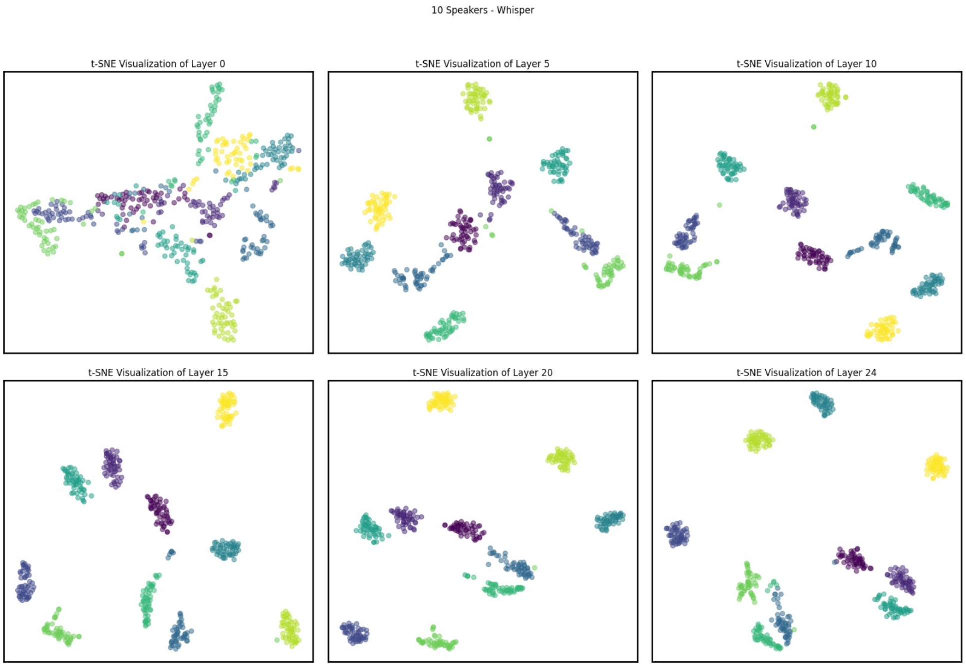

Figure 11: T-SNE - Whisper
The colors of the dots represent different speaker labels.

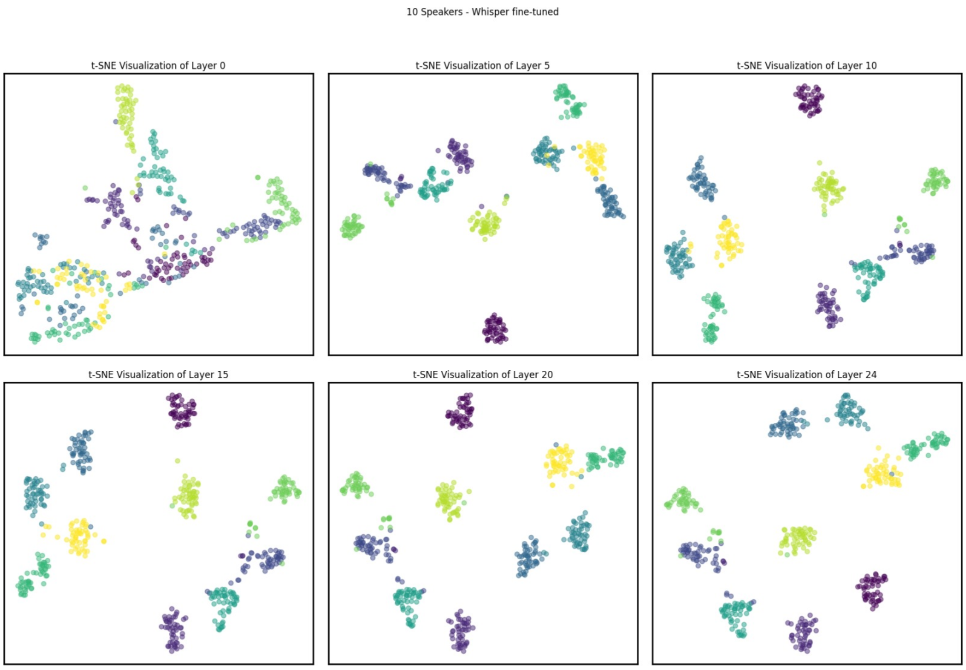

Figure 12: T-SNE - Whisper Finetuned

Model	**Wav2Vec2 Large 960h**	**XLS-R 0.3B**	**Whisper (encoder) Large**
Architecture	Transformer	Transformer	Transformer
Parameters	317 million	300 million	900 million
Hidden State Dimension	1024	1024	1280
Transformer Layers	24	24	24
Attention Heads	16	16	20
Audio Input Layer	7 CNN layers	7 CNN layers	Log-Mel Spectrograms
Training Data	LibriSpeech (960h)	436K hours, 128 languages	LibriSpeech (960h)
Training Objective	ASR, English SR	ASR, Multilingual SR	Robust speech recognition
Frame Length	25 ms	25 ms	20 ms
Hop Size	20 ms	10 ms	10 ms

Table 1: Comparison of Architectures for Wav2Vec2 Large 960h, XLS-R 0.3B, and Whisper (encoder) Large.

	Wav2Vec2	**XLS-R**	**Whisper**
Batch-Size	8	8	2
Early Stopping Patience	10	10	5
Learning Rate	5e-6	5e-6	1e-5
Epochs	100	100	30
Runtime	11 h	10.5 h	62 h
Optimizer	AdamW	AdamW	AdamW
Classification Head	Sequence Classification	Sequence Classification	Customized Classification Head
Used GPU	Nividia A30	Nividia A30	Nividia A30
Accuracy	0.911	0.897	0.870

Table 2: Overview of the Finetuning-process.