Investigating Layer-Specific Performance in Speaker Recognition with XLS-R Architecture

Anonymous ACL submission

Abstract

This study explores the impact of various layers of the XLS-R encoder on speaker recognition accuracy using CNNs and logistic regression. Findings indicate that earlier layers yield higher accuracy, highlighting their importance in feature capture. The study also reveals a significant gender disparity in accuracy. These results suggest the need for further investigation into model biases and optimizations.

1 Introduction

Speaker recognition is a key aspect of natural language processing (NLP) used in security authentication, personalized services, and forensic analysis. Different encoders, especially those like XLS-R (Conneau et al., 2020), utilize multiple transformation layers to extract features from audio files. Our research aims to investigate how these various layers influence feature extraction and representation specifically for speaker recognition tasks.

2 Related Research

(Pasad et al., 2023) used canonical correlation analysis (CCA) to study the impact of transformer-based audio encoder layers on speech tasks. Their analysis of intermediate representations from self-supervised speech models showed distinct patterns that often aligned with pre-training objectives. Remarkably, they found that performance using single layers could rival or exceed that of using all layers together, underscoring the efficiency of pre-trained models. In a similar vein, (Miwa and Kai) fine-tuned models using intermediate outputs from the XLS-R architecture. Adopting a similar approach, we extracted vector representations from various layers of the XLS-R model to assess their impact on speaker recognition accuracy.

2.1 XLS-R

The XLS-R model, developed by Meta AI, is a self-supervised speech processing model optimized for cross-lingual speech representation learning. It extends the "wav2vec 2.0" architecture, featuring a transformer-based neural network learning directly from raw audio data, without the necessity for annotated resources.

In our research, we utilized the "wav2vec2-XLS-R-300M" model, the most compact variant in the XLS-R series, selected due to our computational resource constraints. This version, with its 24 transformer layers and 300 million parameters, provides a robust framework for efficient and effective speech processing in resource-limited settings (Babu et al., 2021).

3 Experiment

The goal of the experiment is to explore how vector representations from different hidden states of the "wav2vec2-XLS-R-300M" model impact speaker recognition tasks. We extracted representations at several stages of the encoder and analyzed them by training and testing a convolutional neural network (CNN). To validate our results, we also employed a simpler logistic regression model as a comparative method.

3.1 Dataset

The dataset STT4SG-350, developed collaboratively by ZHAW, FHNW, and UZH, contains Swiss German MP3 audio files from various speakers (Plüss et al., 2023). We selected the audio files of 25 speakers from the test dataset, prioritizing those with high recording quality while maintaining an even gender and age distribution. These speakers encompass all ages, with a female-to-male ratio of 13:12.

3.2 Hidden State Extraction

We extracted vector representations from the "wav2vec2-XLS-R-300M" model at different stages of the encoder, specifically after the 1st, 5th, 10th, 15th, 20th, and 24th layers, for each audio sample.

3.3 Data Preprocessing

To accommodate the varying lengths of our audio files, we segmented the hidden state matrices into 1-second chunks without overlap. This method of segmentation was preferred over padding to ensure the integrity and quality of the feature representations (Pawar et al., 2018).

3.4 Models

For the speaker recognition task, we used a CNN model, see Table 1. The model's hyperparameters were optimized using vector representations from the final layer of the XLS-R model. After confirming the model's stability, we applied the same settings to evaluate the informational content of the fifferent hidden states for speaker recognition.
Additionally, a logistic regression model was trained on the same task, reinforcing our findings by reproducing the findings with a different and simpler model architecture.
The dataset was divided in an 80:20 train/test ratio to validate the model's performance. Model effectiveness was assessed using accuracy, precision, recall, and F1 score metrics, providing a comprehensive evaluation.

4 Results

Figures 1 and 2 reveal that both the CNN and logistic regression models exhibit similar trends in their speaker recognition accuracies across the various layers. The early layers, particularly layer 5, show the highest accuracy levels, with layer 1 also demonstrating strong performance. Despite a dip in accuracy across the middle layers, there is a pronounced increase at layer 20. The logistic regression performed better overall. The full details of the test set evaluations for both models are recorded in Tables 2 and 3.
The analysis reveals a gender-based disparity in accuracy across the model layers, with male speakers consistently achieving notably higher accuracy than female speakers. In the deeper layers, accuracy declines for both genders, with the reduction being stronger for female speakers.

5 Discussion

Our analysis showed consistent patterns in speaker recognition accuracy across various layers of both CNN and logistic regression models, indicating that early layers capture essential information for speaker recognition. This aligns with previous research (Pasad et al., 2023), suggesting that certain layers are more crucial for specific tasks. For example, the diminishing accuracy at deeper layers could reflect the complexity and abstraction of features at these stages, which might not be optimal for speaker recognition. A more detailed layer-wise analysis would be required to further explore this hypothesis.
A significant accuracy gap between male and female speakers was observed despite using balanced datasets for both models, hinting at a potential underrepresentaiton of female speakers in the encoder's training data (Feldman and Peake, 2021; Mehrabi et al., 2021). Further investigation using a broader and more diverse dataset is needed to conclusively determine this is the source of this bias.
Interestingly, logistic regression, despite its simplicity, outperformed the CNN, potentially due to the limited size of our dataset. Future experiments with larger datasets could clarify this unexpected result.

6 Limitations

With only 25 speakers and 5000 audio files, our dataset size may not fully represent the diverse range of speakers and speech characteristics present in larger datasets. Additionally, we only focused on six layers out of the 24 available in the XLS-R architecture. Therefore, our analysis cannot directly capture the influence of the not chosen layers.

7 Conclusion

Our study demonstrates the critical influence of various layers within the XLS-R architecture on speaker recognition, particularly highlighting that the initial layers typically yield higher accuracy. Additionally, our research exposes a notable gender discrepancy in accuracy, with male speaker recognition consistently surpassing that of female speakers, particularly in the deeper layers of the model. These findings emphasize the necessity for additional research into potential biases and the optimization of models to enhance gender equity in performance outcomes.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *Preprint*, arXiv:2111.09296.

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

Tal Feldman and Ashley Peake. 2021. End-to-end bias mitigation: Removing gender bias in deep learning. *Preprint*, arXiv:2104.02532.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Shogo Miwa and Atsuhiko Kai. Dialect speech recognition modeling using corpus of japanese dialects and self-supervised learning-based model XLSR. In *INTERSPEECH 2023*, pages 4928–4932. ISCA.

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.

R. V. Pawar, Rajesh M. Jalnekar, and J. S. Chitode. 2018. Segmental analysis of speech signal for robust speaker recognition system. In *ICACDS*.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Figures and Tables

Hyperparameter	Value
Number of Convolutional Layers	3
Number of Filters (Conv1)	32
Kernel Size (Conv1)	(5, 5)
Padding (Conv1)	(2, 2)
Pooling (MaxPool1)	(2, 2)
Number of Filters (Conv2)	64
Kernel Size (Conv2)	(3, 3)
Padding (Conv2)	(1, 1)
Number of Filters (Conv3)	128
Kernel Size (Conv3)	(3, 3)
Padding (Conv3)	(1, 1)
Dropout Probability	0.5
Fully Connected Layer 1 Output Size	128
Optimizer	AdamW
Loss	CrossEntropyLoss
Learning Rate	0.001
Epochs	30
Batch Size	64

Table 1: CNN hyperparameters and model architecture

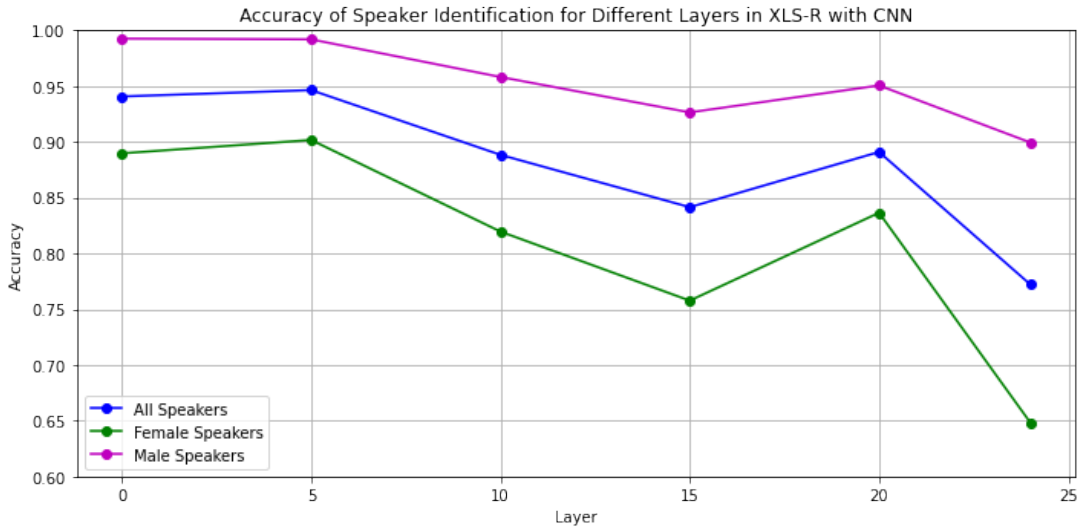

Figure 1: The depicted figure showcases the model's (CNN) accuracy across vector representations from various layers of XLS-R vector for the task speaker recognition. Accuracy scores are presented for the entire speaker dataset and are further disaggregated by gender into female and male subsets.

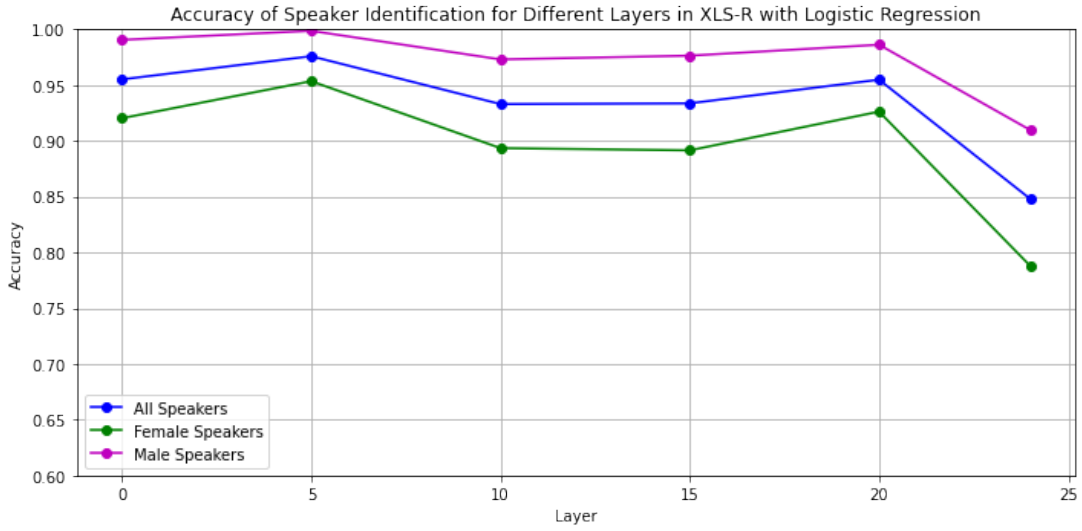

Figure 2: The depicted figure showcases the model's (Logistic regression) accuracy across vector representations from different layers of XLS-R vector for the task speaker recognition. Accuracy scores are presented for the entire speaker dataset and are further disaggregated by gender into female and male subsets.

Layer	Accuracy Total	Accuracy Female	Accuracy Male
1	0.94	0.89	0.99
5	0.95	0.90	0.99
10	0.89	0.82	0.96
15	0.84	0.76	0.93
20	0.89	0.84	0.95
24	0.77	0.65	0.90
Layer	Precision Total	Precision Female	Precision Male
1	0.95	0.91	0.99
5	0.95	0.91	1.00
10	0.89	0.84	0.97
15	0.86	0.80	0.95
20	0.90	0.86	0.96
24	0.76	0.62	0.94
Layer	Recall Total	Recall Female	Recall Male
1	0.94	0.89	0.99
5	0.95	0.90	0.99
10	0.89	0.82	0.96
15	0.84	0.76	0.93
20	0.89	0.84	0.95
24	0.77	0.65	0.90
Layer	F1-Score Total	F1-Score Female	F1-Score Male
1	0.94	0.89	0.99
5	0.94	0.90	0.99
10	0.88	0.82	0.96
15	0.83	0.74	0.94
20	0.89	0.84	0.95
24	0.73	0.58	0.90

Table 2: The accuracy, precision, recall and F1 score of the test set is show for all speaker and the gendered subgroups, for the trained CNN.

Layer	Accuracy Total	Accuracy Female	Accuracy Male
1	0.96	0.92	0.99
5	0.98	0.95	1.00
10	0.93	0.89	0.97
15	0.93	0.89	0.98
20	0.95	0.93	0.99
24	0.85	0.79	0.91
Layer	Precision Total	Precision Female	Precision Male
1	0.96	0.92	0.99
5	0.98	0.96	1.00
10	0.93	0.90	0.98
15	0.93	0.90	0.98
20	0.96	0.94	0.99
24	0.85	0.82	0.94
Layer	Recall Total	Recall Female	Recall Male
1	0.96	0.92	0.99
5	0.98	0.95	1.00
10	0.93	0.89	0.97
15	0.93	0.89	0.98
20	0.95	0.93	0.99
24	0.85	0.79	0.91
Layer	F1-Score Total	F1-Score Female	F1-Score Male
1	0.95	0.92	0.99
5	0.98	0.95	1.00
10	0.93	0.90	0.98
15	0.93	0.90	0.98
20	0.95	0.93	0.99
24	0.85	0.80	0.92

Table 3: The accuracy, precision, recall and F1 score of the test set is show for all speaker and the gendered subgroups, for logistic regression.