Applied Acoustics

Which layer of Wav2Vec2 is effective for age and gender classification from children's speech, and is it necessary to use all features?

--Manuscript Draft--

Manuscript Number:	APAC-D-24-02348
Article Type:	Research Paper
Section/Category:	Europe and Rest of the World
Keywords:	Children Speech; Self Supervised Learning (SSL); Wav2Vec2; CNN; Dimensionality Reduction
Corresponding Author:	Hemant Kumar Kathania, PhD National Institute of Technology Sikkim INDIA
First Author:	Abhijit Sinha, Master of Technology
Order of Authors:	Abhijit Sinha, Master of Technology
	Mohit Joshi, Bachelor of Technology
	Harishankar Kumar, Bachelor of Technology
	Hemant Kumar Kathania, PhD
	Sudarsana Reddy Kadiri, PhD
Abstract:	Speech patterns in children differ significantly from those in adults and evolve as children grow. Children's speech is characterized by higher pitch, variable speech rates, and inconsistent pronunciation due to their developing vocal anatomy and linguistic skills. These factors, combined with natural variability, makes distinguishing between age groups and genders particularly challenging. This study explores the layer wise effectiveness of Wav2Vec2 models for classifying age and gender in children's speech and evaluates the necessity of all extracted features. Features from four Wav2Vec2 models were analyzed using a Convolutional Neural Network (CNN) classifier on the PFSTAR and CMU Kids datasets. Our analysis reveals that the best results for both age and gender classification were achieved using features from the initial layers of these models. Applying dimensionality reduction using PCA further improved classification accuracy, highlighting that only a subset of features is crucial for accurate classification, optimizing both performance and efficiency. The Wav2Vec2-large-960h-lv60 model achieved the highest accuracy, with 97.14% and 98.20% for age and gender classification respectively, on the CMU Kids datasets. For the PFSTAR dataset, the base-100h and large-960h-lv60 models perform particularly well, achieving accuracies of 86.05% for age classification and 95% for gender classification. The findings highlight the potential of SSL models for optimizing children's speech processing tasks.
Suggested Reviewers:	Mittul Singh, PhD Researcher Associate, Aalto University mittul.singh@aalto.fi He is working in the area of machine learning for speech processing. So he suited person to review this article.
	Waquar Ahmed, PhD Assistant Professor, National Institute of Electronics and Information Technology - Calicut waquar@nitc.ac.in As his research interest in speech processing, so he is suited person to review this article.
	Manuel Rosa Zurera, PhD Professor, University of Alcala de Henares manuel.rosa@uah.es His work is related our paper, so he is good to review our article.

Opposed Reviewers:	

Which layer of Wav2Vec2 is effective for age and gender classification from children's speech, and is it necessary to use all features?

Abhijit Sinha[a], Mohit Joshi[a], Harishankar Kumar[a], Hemant Kumar Kathania[a], Sudarsana Reddy Kadiri[b]

[a]*Department of Electronics and Communication Engineering, National Institute of Technology Sikkim, Ravangla, India*
[b]*Speech Analysis and Interpretation Laboratory (SAIL), University of Southern California, Los Angeles, USA*

Abstract

Speech patterns in children differ significantly from those in adults and evolve as children grow. Children's speech is characterized by higher pitch, variable speech rates, and inconsistent pronunciation due to their developing vocal anatomy and linguistic skills. These factors, combined with natural variability, makes distinguishing between age groups and genders particularly challenging. This study explores the layer wise effectiveness of Wav2Vec2 models for classifying age and gender in children's speech and evaluates the necessity of all extracted features. Features from four Wav2Vec2 models were analyzed using a Convolutional Neural Network (CNN) classifier on the PFSTAR and CMU Kids datasets. Our analysis reveals that the best results for both age and gender classification were achieved using features from the initial layers of these models. Applying dimensionality reduction using PCA further improved classification accuracy, highlighting that only a subset of features is crucial for accurate classification, optimizing both performance and efficiency. The Wav2Vec2-large-960h-lv60 model achieved the highest accuracy, with 97.14% and 98.20% for age and gender classification respectively, on the CMU Kids datasets. For the PFSTAR dataset, the base-100h and large-960h-lv60 models perform particularly well, achieving accuracies of 86.05% for age classification and 95% for gender classification. The findings highlight the potential of SSL models for optimizing children's speech processing tasks.

Keywords:
Children Speech, Self Supervised Learning (SSL), Wav2Vec2, CNN, Dimentionality Reduction

1. Introduction

As demand for advanced speech-based applications grows, precise classification of age and gender becomes vital, especially for applications designed for children. These applications include content filtering, age-appropriate recommendations, and personalized interactions, all of which are essential for ensuring safe and engaging experiences for young users. Accurate

classification helps tailor content and services to the specific needs of children, enhancing both safety and relevance in digital environments.

However, automatic classification of age and gender from children's speech presents unique challenges compared to adult speech due to the distinct acoustic properties of children's voices [1, 2]. Children's speech is characterized by higher variability in pitch, formant frequencies [3], and pronunciation patterns, which differ significantly from those of adults [4, 5]. As children age, their speech patterns evolve: pitch generally lowers, formant frequencies become more stable, and pronunciation becomes more consistent with adult norms. These developmental changes add complexity to age and gender classification, as models must adapt to these dynamic acoustic variations throughout different stages of childhood. Furthermore, the scarcity of large annotated datasets for children's speech [6, 7] further complicates the development of robust models, as the limited availability of high-quality labeled data impedes the training of accurate and generalizable classification systems.

Previous research on classifying age and gender in children's speech has utilized methods like Gaussian mixture models (GMMs), deep neural networks (DNNs) [8, 9] and temporal convolutional neural networks (CNNs) [10] with features such as Mel-frequency cepstral coefficients (MFCCs) [11] and time-delay neural networks (TDNNs) with low-level descriptors [12]. Recent studies have explored the impact of acoustic and prosodic features on age classification [13, 14] and investigated the use of sincNet over the ERB scale for automatic speaker and age identification [15]. Additionally, work using TDNNs and LSTM networks [16] on raw waveforms has shown potential for improving classification accuracy, suggesting that further advancements with sincNet-based CNNs could offer even better results.

Recent advancements in self-supervised learning (SSL) models, such as Wav2Vec2 [17], have significantly advanced speech processing by utilizing large-scale unlabeled data. These SSL models have shown remarkable performance across various speech related tasks [18, 19, 20, 21, 22], including tasks like speaker verification, language identification, and emotion recognition [23, 24, 25, 26, 27]. However, their application in age and gender classification specifically for children's speech remains underexplored. Moreover, these studies often involve fine-tuning or combining features across multiple layers, offering limited insights into layer-specific behaviour for specific tasks.

This paper addresses these gaps by investigating Wav2Vec2 models with different pre-training and fine-tuning configurations to enhance the performance for age and gender classification from children's speech. We extracted features from four Wav2Vec2 variants base-100h, base-960h, large-960h-lv60, and large-960h-lv60-self and trained a Convolutional Neural Network (CNN) classifier on two benchmark datasets: PFSTAR [28] and CMU Kids [29]. This approach was employed to identify the best-performing layers in Wav2Vec2 models for optimizing age and gender classification performance. This layer wise analysis allows for targeted feature selection, enabling the identification of the most task relevant representations without fine tuning. By avoiding fine-tuning, we aim to assess how well pre-trained representations generalize to children's speech. In addition, we introduce balanced accuracy as a complementary metric to accuracy, providing a fairer evaluation of performance in potentially imbalanced datasets.

We also applied dimensionality reduction using Principal Component Analysis (PCA)

to determine whether all extracted features are necessary, further optimizing feature representation. Our results show that not all features from the Wav2Vec2 model are crucial for classification tasks; reducing dimensionality can enhance accuracy and streamline computational efficiency. This highlights the importance of optimizing feature representations for children's speech, as some features may be more relevant than others.

The main contributions of this study are as follows:

- Experimental evaluation on two children speech datasets, PFSTAR and CMU Kids, highlighting the challenges and opportunities in adapting SSL models to children speech.
- A comprehensive analysis of the performance of layer wise features in Wav2Vec2 models for age and gender classification in children's speech.
- An investigation of dimensionality reduction using PCA, demonstrating their role in improving accuracy and computational efficiency.

Table 1: Wav2Vec2 model details. M is short of a million, K for a thousand and h for hours. PT represents pretraining, FT represents finetuning and Ft. Dim. represents feature dimensions for each model.

Wav2Vec2	**Size**	**PT**	**FT**	**Layers**	**Ft. Dim.**
base-100h	95M	960 h	100 h	13	768
base-960h	95M	960 h	960 h	13	768
large-960-lv60	317M	60K h	960 h	25	1024
large-960-lv60-self	317M	60K h	960 h	25	1024

2. Proposed Framework

Figure 1 describes the proposed framework for classifying age and gender in children's speech using layer-wise features extracted from various Wav2Vec2 models. We experimented with four Wav2Vec2 models base-100h, base-960h, large-960h-lv60-self and large-960h-lv60-self. These models differ in size, number of layers and pre-training hours as described in table 1. Although pre-trained on adult speech, we used these models to extract layer-wise features without any fine-tuning, allowing us to explore how well the learned representations generalize to children's speech for age and gender classification. These features are then fed into a convolutional neural network (CNN) for classification. The CNN design was deliberately kept simple to directly assess the discriminative power of Wav2Vec2 features without introducing any additional complexity. Additionally, we applied dimensionality reduction using PCA to identify and retain the most critical features, thereby reducing computational complexity and enhancing interpretability.

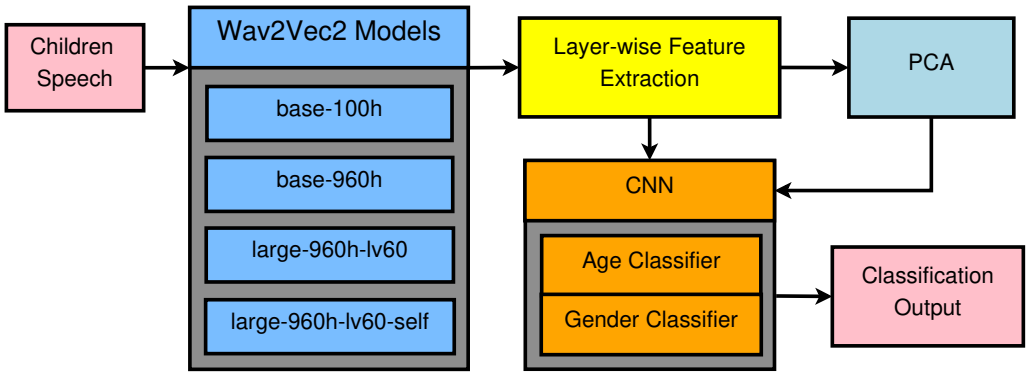

Figure 1: A block diagram illustrating the proposed framework for classifying age and gender in children using layer-wise features from various Wav2Vec2 models. Dimensionality reduction via PCA is applied to the most effective layers to enhance performance.

3. Database and Experimental Setup

This section describes the datasets used in the experiments and the experimental setup for evaluating the performance of self-supervised models on children's speech. It provides details on the database composition and the experimental configurations.

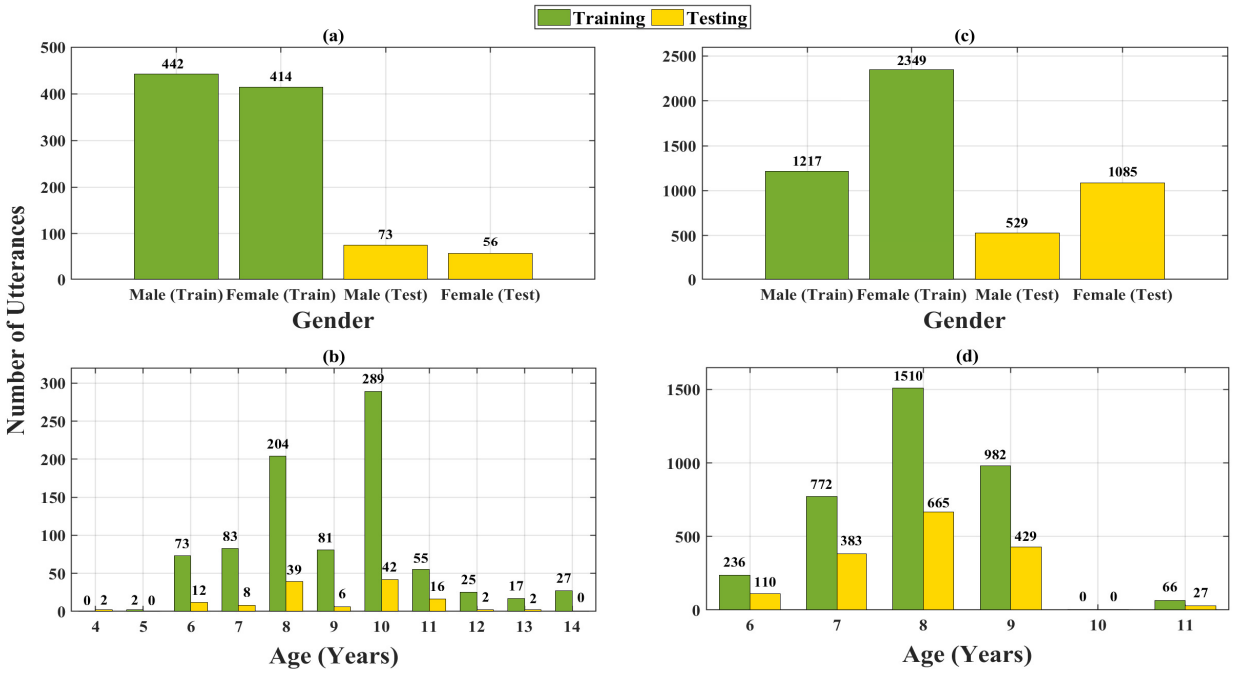

Figure 2: The figure illustrates the dataset distribution for age and gender classification in children using layer-wise features from various Wav2Vec2 models, specifically for the PFSTAR and CMU Kids datasets. It includes the training and testing splits, with the age-wise and gender distribution and the number of utterances. (a) PFSTAR gender distribution, (b) PFSTAR age distribution, (c) CMU Kids gender distribution, and (d) CMU Kids age distribution.

3.1. Database

This study employs two widely used children's speech datasets: PFSTAR [28] and CMU Kids [29].

The PFSTAR dataset [28] contains recordings of children's speech in British English, with an age range from 4 to 14 years. The training set of PFSTAR used for adaptation consists of 8.3 hours of speech from 122 speakers. The test set of PFSTAR consists of 1.1 hours of speech from 60 speakers.

The CMU Kids dataset [29] consists of recordings of children reading sentences aloud in American English. The corpus consists of recordings from 76 speakers, with an age range from 6 to 11 years, with a total of 5180 utterances.The training set utilized in this study accounts for exactly 70% of the overall data, totaling a duration of 6.3 hours. The test set comprises the remaining 30% of the data, totaling 2.83 hours.

Figure 2 provides a detailed overview of the dataset distribution used for age and gender classification tasks, utilizing layer-wise features extracted from various Wav2Vec2 models. For the PFSTAR dataset, the distribution includes training and testing splits, covering the age range of 4 to 14 years, with gender categories for male and female speakers. Similarly, the CMU Kids dataset is represented with its training and testing splits, covering the age range of 6 to 11 years, along with gender statistics. Additionally, the figure highlights the total number of utterances for each dataset. Unlike PFSTAR, the CMU Kids dataset features a balanced gender distribution, reducing the likelihood of bias in gender classification. However, the 70-30 split is not fully independent, potentially allowing overlaps in speaker characteristics.

3.2. Experimental Setup

For the experiments, we employed the framework illustrated in Figure 1. This approach involves extracting layer-wise features from each Wav2Vec2 model and subsequently processing these features through a 1D CNN with three convolutional layers, utilizing 64, 128 and 256 filters, respectively, with a kernel size of 5. ReLU activation functions are incorporated to introduce non-linearity, and batch normalization is applied to ensure stable training and optimize model performance.

To evaluate the effectiveness of Wav2Vec2 models for age and gender classification in children's speech, we conducted experiments with four distinct variants of the model, each with different pretraining and fine-tuning configurations. The four models used in our experiments include: base-100h, which is trained on 100 hours of LibriSpeech [30] data and with fewer parameters; base-960h, trained on 960 hours of LibriSpeech data, providing a more comprehensive range of speech features; large-960h-lv60, which is trained on the same 960 hours of LibriSpeech data and further fine-tuned with 60,000 hours of unlabeled data from Libri-Light [31] to capture a broader variety of speech patterns; and large-960h-lv60-self, which undergoes additional self-training on the 960-hour LibriSpeech dataset to enhance its learning capabilities.

We extracted layer-wise features from each Wav2Vec2 model. Table 1 describes the Wav2Vec2 model details used in this study. The base models consist of 13 hidden layers with a feature dimension of 768, while the large models have 25 hidden layers with a feature

dimension of 1024. Each model utilizes CNNs to transform raw audio into latent representations, effectively capturing local acoustic features from the waveform. These CNN-generated features are then processed by Transformer encoders, which incorporate multiple layers of self-attention and feed-forward networks to capture long-range dependencies and further refine the representations. The initial hidden layer reflects the CNN output, while the following layers are Transformer layers that enhance contextual understanding.

We extracted features from all the hidden layers of Wav2Vec2 models, including the CNN output (the first layer) and the Transformer layers. For the base models, this resulted in twelve feature sets, while for the large models, twenty-five feature sets were obtained. The features extracted from the base models formed a 768-dimensional feature matrix for each speech frame, maintaining frame-wise granularity. Similarly, the large models produced a 1024-dimensional feature matrix for each input speech frame.

Further, we applied dimensionality reduction using PCA on the best performimg layers of each Wav2Vec2 model to investigate the impact of reducing feature dimensions and to determine whether all features are critical for age and gender classification performance.

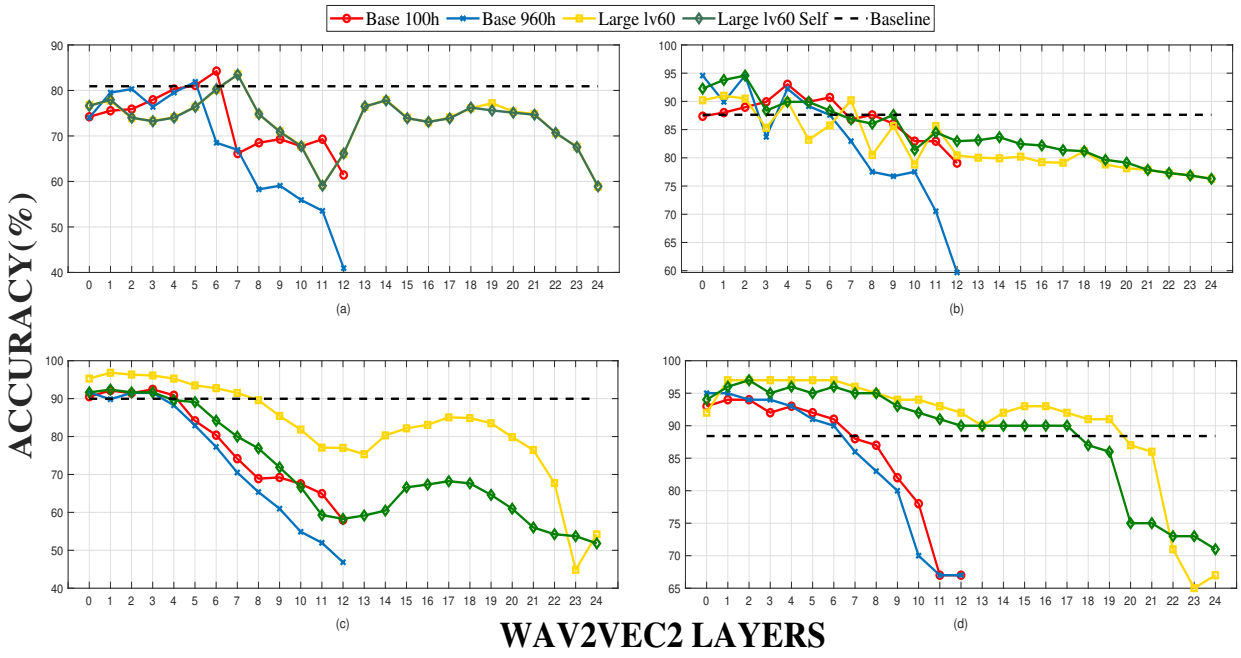

Figure 3: The figure illustrates layer-wise classification performance for age and gender across four Wav2Vec2 models: (a) age classification accuracies on the PFSTAR dataset, (b) gender classification accuracies on the PFSTAR dataset, (c) age classification accuracies on the CMU Kids dataset, and (d) gender classification accuracies on the CMU Kids dataset.

4. Results and Discussion

4.1. Baseline

We established a baseline by training a CNN classifier for children age and gender classification on 26-dimensional Mel-frequency cepstral coefficients (MFCC) features extracted from the PFSTAR and CMU Kids datasets. The baseline model achieved an accuracy (A) of 80.92% for age classification on PFSTAR, and an accuracy of 89.97% on CMU Kids, as represented in Table 2. For gender classification, the baseline model achieved an accuracy of 87.63% on PFSTAR, and an accuracy of 88.41% on CMU Kids, as shown in Table 3. These results underscore the limitations of traditional features and provide a reference for evaluating the enhancements achieved with Wav2Vec2-based features.

4.2. Layer wise Wav2Vec2 performance

We further evaluated the performance of a CNN classifier by extracting layer-wise features from four distinct Wav2Vec2 models, each with different configurations: base-100h, base-960h, large-960h-lv60, and large-960h-lv60-self. The base models are pretrained on 100 hours and 960 hours of speech data, respectively, and contain 13 hidden layers, each with a feature dimension of 768. In contrast, the large models are pretrained on 960 hours of speech data and consist of 25 hidden layers, each with a higher feature dimension of 1024. These features were aggregated over time to generate fixed-length representations suitable for classification tasks. This setup enabled us to compare the performance of different Wav2Vec2 model configurations, examining how variations in model size, training data, and pretraining objectives affect the quality of learned representations for classification.

Figure 3 presents a comprehensive comparison of layer-wise performance for age and gender classification across the various Wav2Vec2 models employed in this study. The figure visually illustrates how different layers of the models contribute to the classification task, providing insights into the relative effectiveness of early versus later layers for age and gender classification. This comparison highlights the model's ability to capture relevant features at different stages of its deep architecture, offering a better understanding of which layers are most critical for achieving optimal performance in this specific task.

Tables 2 and 3 further provide a comparison of age and gender classification performance, highlighting the best-performing layers for each model. For the PFSTAR dataset, the base-100h and large-960h-lv60 models stand out, achieving impressive accuracies of 84.25% for age classification and 94.57% for gender classification, respectively. These results demonstrate the model's ability to effectively capture age- and gender-specific acoustic features even with the relatively smaller 100-hour pretraining dataset. On the CMU Kids dataset, the large-960h-lv60 model achieves the highest performance, with an outstanding 96.84% accuracy for age classification and 96.68% for gender classification. This suggests that, when trained on larger, more diverse datasets, Wav2Vec2 models can leverage more robust representations that significantly improve classification accuracy.

These results indicate that the initial layers of Wav2Vec2, which capture low-level acoustic features, are crucial for age and gender classification. The later layers, on the other hand, appear to focus on more abstract, high-level representations that are less relevant for

Table 2: The table presents performance metrics- accuracy (A), precision (P), recall (R), and F1 score (F1) for the best performing layers across various Wav2Vec2 models in age and gender classification on the PFSTAR dataset.

Age					
Wav2Vec2 Model	**Best Layer**	**A**	**P**	**R**	**F1**
Baseline (MFCC Features)	26 Features	80.92	0.82	0.81	0.80
base-100h	6	84.25	0.86	0.84	0.83
base-960h	5	81.89	0.84	0.82	0.81
large-960h-lv60	7	83.59	0.84	0.83	0.83
large-960h-lv60-self	7	83.46	0.85	0.84	0.83
Gender					
Baseline (MFCC Features)	26 Features	87.63	0.90	0.88	0.88
base-100h	4	93.02	0.93	0.93	0.92
base-960h	2	94.57	0.96	0.95	0.95
large-960h-lv60	1	91.45	0.93	0.92	0.90
large-960h-lv60-self	2	94.57	0.95	0.94	0.94

Table 3: The table presents performance metrics- accuracy (A), precision (P), recall (R), and F1 score (F1) for the best performing layers across various Wav2Vec2 models in age and gender classification on the CMU Kids dataset.

Age					
Wav2Vec2 Model	**Best Layer**	**A**	**P**	**R**	**F1**
Baseline (MFCC Features)	26 Features	89.97	0.90	0.89	0.89
base-100h	1	92.13	0.92	0.92	0.92
base-960h	0	91.63	0.90	0.90	0.90
large-960h-lv60	1	96.84	0.97	0.97	0.97
large-960h-lv60-self	1	92.37	0.92	0.92	0.92
Gender					
Baseline (MFCC Features)	26 Features	88.41	0.89	0.88	0.88
base-100h	2	93.78	0.94	0.94	0.94
base-960h	1	94.96	0.95	0.95	0.95
large-960h-lv60	2	96.68	0.97	0.97	0.97
large-960h-lv60-self	2	96.53	0.97	0.97	0.97

this task, emphasizing the importance of selecting appropriate features based on the task's nature.

4.3. Dimensionality Reduction using PCA

To assess how dimensionality reduction affects classification performance, we applied PCA to the features extracted from the top-performing layer in each model for the classification task. We systematically reduced the feature dimensions in steps of 64, starting from 512 down to 64, with an additional evaluation at 32 dimensions.

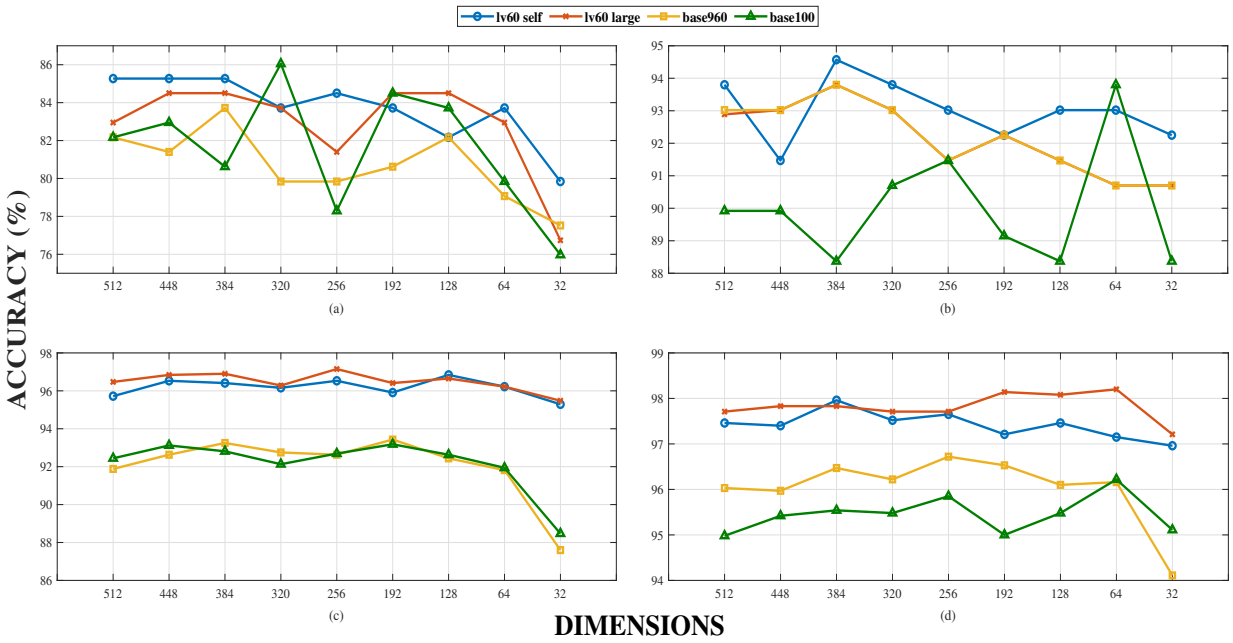

Figure 4: The figure illustrates classification performance across feature dimensions ranging from 512 to 64 in steps of 64 for age and gender after dimensionality reduction using PCA across four Wav2Vec2 models: (a) age classification accuracies on the PFSTAR dataset, (b) gender classification accuracies on the PFSTAR dataset, (c) age classification accuracies on the CMU Kids dataset, and (d) gender classification accuracies on the CMU Kids dataset.

Figure 4 illustrates the PCA results for age and gender classification on the PFSTAR and CMU Kids datasets, showing the performance across feature dimensions ranging from 512 to 64 in steps of 64. This figure demonstrates that even with reduced feature dimensions, the model maintains competitive classification accuracy.

Table 4: The table presents performance metrics—accuracy (A), precision (P), recall (R), and F1 score (F1)—for the best-performing layers of various Wav2Vec2 models in age and gender classification on the PFSTAR dataset, with dimensionality reduction applied using PCA.

Age					
Wav2Vec2 Model	**Reduced Feature Dimension**	**A**	**P**	**R**	**F1**
base-100h	320	86.05	0.87	0.86	0.86
base-960h	384	83.72	0.85	0.84	0.82
large-960h-lv60	256	84.8	0.85	0.84	0.83
large-960h-lv60-self	384	85.27	0.87	0.85	0.84
Gender					
base-100h	64	93.80	0.94	0.93	0.94
base-960h	384	93.80	0.95	0.94	0.94
large-960h-lv60	320	92.75	0.94	0.92	0.91
large-960h-lv60-self	384	95.00	0.95	0.95	0.95

Table 5: The table presents performance metrics—accuracy (A), precision (P), recall (R), and F1 score (F1)—for the best-performing layers of various Wav2Vec2 models in age and gender classification on the CMU Kids dataset, with dimensionality reduction applied using PCA.

Age					
Wav2Vec2 Model	**Reduced Feature Dimension**	**A**	**P**	**R**	**F1**
base-100h	192	93.18	0.93	0.93	0.93
base-960h	192	93.43	0.93	0.93	0.93
large-960h-lv60	256	97.14	0.97	0.97	0.97
large-960h-lv60-self	128	96.84	0.97	0.97	0.97
Gender					
base-100h	64	96.22	0.96	0.96	0.96
base-960h	256	96.71	0.97	0.97	0.97
large-960h-lv60	64	98.20	0.98	0.98	0.98
large-960h-lv60-self	384	97.95	0.98	0.98	0.98

Tables 4 and 5 summarizes the results for age and gender classification with dimensionality reduction for the best performing layers of each Wav2vec2 model. For the PFSTAR dataset, the base-100h model, with PCA applied to 320 dimensions, achieved the highest accuracy in age classification with a score of 86.05%. In gender classification, the large-960h-lv60-self model, with PCA applied to 384 dimensions, delivered the best performance with an accuracy of 95%. For the CMU Kids dataset, the large-960h-lv60 model, using PCA with 256 and 384 dimensions, outperformed the others, achieving accuracies of 97.14% for age classification and 98.20% for gender classification, respectively.

To gain deeper insight into the model's behavior, we also analyzed the confusion matrices corresponding to the MFCC features and the best-performing reduced feature set after PCA as illustrated in figure 5.

The confusion matrix for the original MFCC features revealed higher misclassification rates. In contrast, the confusion matrix for the best results after PCA indicated fewer misclassifications, emphasizing the robustness of SSL features and demonstrating that the reduced feature set focused more on the most informative aspects of the data. The confusion matrices further support that, after dimensionality reduction, the model's ability to correctly classify both age and gender categories was significantly improved, confirming the effectiveness of dimensionality reduction in optimizing classification for this task.

These findings highlight that not all features produced by the Wav2Vec2 model are critical for achieving optimal classification performance. Through dimensionality reduction, we observed that a reduced number of principal components can sufficiently capture the core information needed for accurate age and gender classification from children's speech. By retaining only the most relevant features, dimensionality reduction not only simplifies the model but also reduces computational complexity and improve the system's robustness. These results validate the effectiveness of dimensionality reduction as a strategy for optimizing feature sets, improving both classification accuracy and model efficiency, especially when working with high-dimensional data in tasks such as age and gender classification.

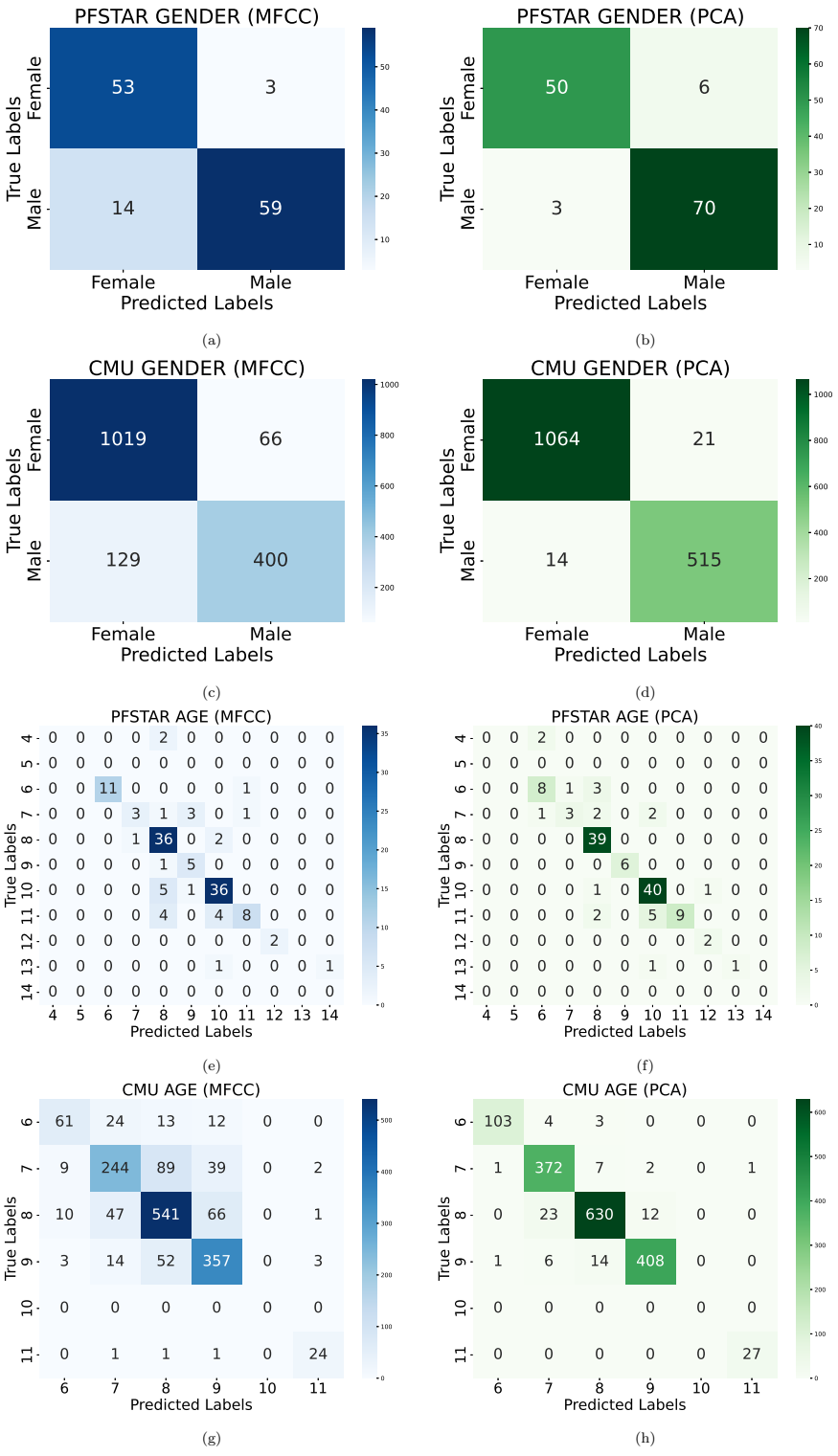

Figure 5: Comparison of confusion matrices for age and gender classification tasks on the PFSTAR and CMU Kids dataset. The left column represents the baseline system using MFCC features, while the right column shows results from the best-performing layer of the Wav2Vec2 model after dimensionality reduction with PCA.

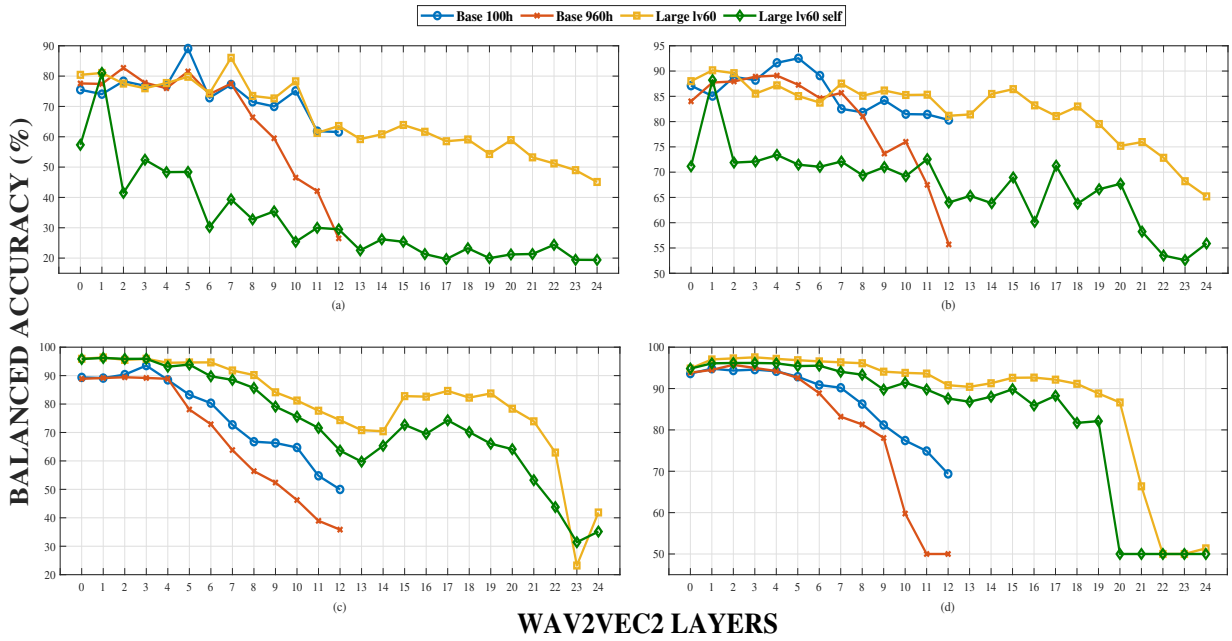

Figure 6: The figure illustrates layer-wise classification performance for age and gender using balanced accuraccy across four Wav2Vec2 models: (a) age classification accuracies on the PFSTAR dataset, (b) gender classification accuracies on the PFSTAR dataset, (c) age classification accuracies on the CMU Kids dataset, and (d) gender classification accuracies on the CMU Kids dataset.

4.4. Balanced Accuracy

To further evaluate the classification performance, we computed the balanced accuracy for age and gender classification across various Wav2Vec2 models. The balanced accuracy metric is particularly valuable for addressing class imbalance by ensuring that both minority and majority classes are equally represented in the evaluation.

Figure 6 presents a detailed comparison of layer-wise balanced accuracy for age and gender classification across various Wav2Vec2 models on the PFSTAR and CMU Kids datasets. For the PFSTAR dataset, the Wav2Vec2-base-100h model outperforms the other three models, achieving balanced accuracies of 89.01% and 92.53% at layer 5 for age and gender classification respectively. For the CMU Kids dataset the Wav2vec2-large-960h-lv60 performed the best with balanced accuracies of 96.39% at layer 1 and 97.55% at layer 3 for age and gender classification respectively. The trend observed in the balanced accuracy results, as shown in figure 6 is similar to that seen in traditional accuracy metrics. Notably, the performance is highest in the earlier layers of the models, with balanced accuracy gradually decreasing as the layer number increases. This pattern suggests that the initial layers of Wav2Vec2, which capture low-level acoustic features, are particularly effective for age and gender classification. As the model progresses to deeper layers, which focus more on abstract representations, the relevance of these features for age and gender classification diminishes, leading to a drop in balanced accuracy.

The consistency of trends across balanced and standard accuracy highlights the robust-

ness of early layer features, making them ideal for lightweight classification tasks.

This trend could also have implications for other speech related tasks such as automatic speech recognition (ASR) tasks, where the reduction in age and gender bias in the later layers might improve performance. As the model learns more generalized, higher-level features in the deeper layers, it may become better at handling diverse speech patterns, reducing the impact of demographic biases and potentially enhancing transcription accuracy in ASR tasks. The focus on abstract representation in the later layers explains there effectiveness in ASR tasks, where high level linguistic generalization is essential.

5. Conclusion

In this study, we evaluated the effectiveness of layer-wise features from various Wav2Vec2 models for classifying age and gender in children's speech. Our results showed that features from the initial layers of these models, especially from the larger variants, are the most effective for these classification tasks. This is likely because early layers capture fundamental acoustic properties, such as pitch and formants, which are crucial for distinguishing age and gender differences in children's speech. In contrast, later/last layers focus more on abstract linguistic information, which may be less relevant for these specific classification tasks.

Furthermore, applying PCA revealed that not all extracted features are necessary. Dimensionality reduction not only improved classification accuracy but also reduced computational complexity. This suggests that many features, particularly from the deeper layers, are either not crucial or contribute less to the classification process. By selectively retaining key features from the early layers, we achieved more efficient and accurate classification. These findings underscore the importance of optimizing feature selection for age and gender classification, ensuring that the model captures critical speech characteristics while avoiding unnecessary complexity.

Overall, this study demonstrates the effectiveness of using early-layer features from Wav2Vec2 models and the advantages of dimensionality reduction for optimizing age and gender classification in children's speech. These findings suggest that targeted feature selection and efficient representation can substantially enhance performance, paving the way for efficient feature selection strategies, enabling the development of lightweight speech systems for educational tools, personalized assistants, and content moderation applications for children.

References

[1] L. L. Koenig, J. C. Lucero, E. Perlman, Speech production variability in fricatives of children and adults: Results of functional data analysis, The Journal of the Acoustical Society of America 124 (5) (2008) 3158–3170.

[2] T. Tran, M. Tinkler, G. Yeung, A. Alwan, M. Ostendorf, Analysis of disfluency in children's speech, in: Interspeech 2020, 2020, pp. 4278–4282. doi:10.21437/Interspeech.2020-3037.

[3] H. Kumar Kathania, S. Reddy Kadiri, P. Alku, M. Kurimo, Study of formant modification for children asr, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7429–7433. doi:10.1109/ICASSP40776.2020.9053334.

[4] H. K. Vorperian, R. D. Kent, Vowel acoustic space development in children: a synthesis of acoustic and anatomic data., Journal of speech, language, and hearing research : JSLHR 50 6 (2007) 1510–45.
[5] S. Lee, A. Potamianos, S. Narayanan, Acoustics of children's speech: Developmental changes of temporal and spectral parameters, The Journal of the Acoustical Society of America 105 (3) (1999) 1455–1468.
[6] G. Yeung, A. Alwan, On the difficulties of automatic speech recognition for kindergarten-aged children, Interspeech (2018).
[7] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, R. Hoffmann, A survey about databases of children's speech., in: Interspeech, 2013, pp. 2410–2414.
[8] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, M. Rosa-Zurera, Age and gender recognition from speech using deep neural networks, in: Advances in Physical Agents II: Proceedings of the 21st International Workshop of Physical Agents (WAF 2020), November 19-20, 2020, Alcalá de Henares, Madrid, Spain, Springer, 2021, pp. 332–344.
[9] D. Kwasny, D. Hemmerling, Gender and age estimation methods based on speech using deep neural networks, Sensors 21 (14) (2021) 4785.
[10] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, M. Rosa-Zurera, Age group classification and gender recognition from speech with temporal convolutional neural networks, Multimedia Tools and Applications 81 (3) (2022) 3535–3552.
[11] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jančovič, M. Carey, Speaker recognition for children's speech, in: Interspeech, 2012, pp. 1836–1839. doi:10.21437/Interspeech.2012-401.
[12] N. Jia, C. Zheng, W. Sun, Children's speaker recognition method based on multi-dimensional features, in: Advanced Data Mining and Applications: 15th International Conference, ADMA 2019, Dalian, China, November 21–23, 2019, Proceedings 15, Springer, 2019, pp. 462–473.
[13] M. Li, K. J. Han, S. Narayanan, Automatic speaker age and gender recognition using acoustic and prosodic level information fusion, Computer Speech & Language 27 (1) (2013) 151–167.
[14] V. Kumari, A. Sinha, H. K. Kathania, Role of acoustics and prosodic features for children's age classification, in: 2024 International Conference on Signal Processing and Communications (SPCOM), IEEE, 2024, pp. 1–5.
[15] K. Radha, M. Bansal, R. B. Pachori, Automatic speaker and age identification of children from raw speech using sincnet over erb scale, Speech Communication 159 (2024) 103069.
[16] M. Sarma, K. K. Sarma, N. K. Goel, Children's age and gender recognition from raw speech waveform using dnn, in: Advances in Intelligent Computing and Communication: Proceedings of ICAC 2019, Springer, 2020, pp. 1–9.
[17] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.
[18] L. Pepino, P. Riera, L. Ferrer, Emotion recognition from speech using wav2vec 2.0 embeddings (2021) 3400–3404doi:10.21437/Interspeech.2021-703.
[19] A. Sinha, M. Singh, S. R. Kadiri, M. Kurimo, H. K. Kathania, Effect of speech modification on wav2vec2 models for children speech recognition, in: International Conference on Signal Processing and Communications (SPCOM), IEEE, 2024, pp. 1–5.
[20] Y. Getman, R. Al-Ghezi, K. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, S. Strömbergsson, wav2vec2-based speech rating system for children with speech sound disorder, in: Interspeech 2022, 2022, pp. 3618–3622. doi:10.21437/Interspeech.2022-10103.
[21] Y. Gao, C. Chu, T. Kawahara, Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining, in: Interspeech 2023, 2023, pp. 3637–3641. doi:10.21437/Interspeech.2023-756.
[22] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, M.-H. Wu, X. Fang, L.-R. Dai, A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 3174–3178. doi:10.1109/ICASSP43922.2022.9747379.
[23] Z. Fan, M. Li, S. Zhou, B. Xu, Exploring wav2vec 2.0 on speaker verification and language identification, 2021, pp. 1509–1513. doi:10.21437/Interspeech.2021-1280.

[24] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, H. Aronowitz, Speech emotion recognition using self-supervised features, 2022, pp. 6922–6926. doi:10.1109/ICASSP43922.2022.9747870.
[25] O. H. Anidjar, R. Marbel, R. Yozevitch, Harnessing the power of wav2vec2 and cnns for robust speaker identification on the voxceleb and librispeech datasets, Expert Systems with Applications 255 (2024) 124671. doi:https://doi.org/10.1016/j.eswa.2024.124671.
[26] B. Nasersharif, M. Namvarpour, Exploring the potential of wav2vec 2.0 for speech emotion recognition using classifier combination and attention-based feature fusion, J. Supercomput. 80 (16) (2024) 23667–23688. doi:10.1007/s11227-024-06158-x.
[27] S. Novoselov, G. Lavrentyeva, A. Avdeeva, V. Volokhov, N. Khmelev, A. Akulov, P. Leonteva, On the robustness of wav2vec 2.0 based speaker recognition systems, in: Interspeech 2023, 2023, pp. 3177–3181. doi:10.21437/Interspeech.2023-881.
[28] M. Russell, The pf-star british english childrens speech corpus, The Speech Ark Limited (2006).
[29] M. Eskenazi, J. Mostow, D. Graff, The cmu kids corpus, Linguistic Data Consortium 11 (1997).
[30] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: an asr corpus based on public domain audio books, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 5206–5210.
[31] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, et al., Libri-light: A benchmark for asr with limited or no supervision, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7669–7673.

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: