Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 10, No. 02, June 2024.

BEYOND WORDS: HARNESSING SPEECH SOUND FOR SPEAKER AGE AND GENDER DETECTION USING 1D CNN ARCHITECTURE WITH SELF-ATTENTION MECHANISM

Umniah Hameed Jaid[1] and Alia Karim Abdulhasan[2]

(Received: 22-Dec.-2023, Revised: 9-Mar.-2024, Accepted: 20-Mar.-2024)

ABSTRACT

Beyond the immediate content of speech, the voice can provide rich information about a speaker's demographics, including age and gender. Estimating a speaker's age and gender offers a wide range of applications, spanning from voice forensic analysis to personalized advertising, healthcare monitoring and human-computer interaction. However, pinpointing precise age remains intricate due to age ambiguity. Specifically, utterances from individuals at adjacent ages are frequently indistinguishable. Addressing this, we propose a novel, end-to-end approach that deploys Mozilla's Common Voice dataset to transform raw audio into high-quality feature representations using Wav2Vec2.0 embeddings. These are then channeled into our self-attention-based convolutional neural network (CNN) model. To address age ambiguity, we evaluate the effects of different loss functions such as focal loss and Kullback-Leibler (KL) divergence loss. Additionally, we evaluate the estimation accuracy at different speech durations. Experimental results from the Common Voice dataset underscore the efficacy of our approach, showcasing an accuracy of 87% for male speakers, 91% for female speakers and 89% overall accuracy, as well as an accuracy of 99.1% for gender prediction.

KEYWORDS

Speaker age, Speaker gender, Speaker profiling, Wav2vec embedding, Attention mechanism.

1. INTRODUCTION

Beyond mere verbal content, the sound of a person's speech offers profound insights into the speaker's identity, revealing hints about age, gender, ethnicity and emotional state [1]. The capability to infer demographic information from speech plays a pivotal role in numerous applications, from forensics [2] to personalized advertising [3]-[4], healthcare systems and human-robot interactions [4].

However, the accurate estimation of demographics from speech is a challenging task. The multifaceted nature of human speech, influenced by factors like emotions, health status, weight and context not only enriches the vocal expressions, but also makes them complex. A particular challenge lies in segregating the textual content from a speaker's physical attributes [5].

Traditionally, the process of speaker profiling has been structured in three stages: data accumulation and preprocessing, feature extraction and selection and finally, the estimation of physical attributes. Historically, voice-pattern analysis has largely relied on time-frequency representations, such as mel-frequency cepstral coefficients (MFCCs) [6], linear predictive coding (LPC) [7] and formant frequencies [8]. However, some studies have leaned towards statistical methods or Gaussian mixture models for speech modeling [9]-[11].

Recent developments in deep-learning (DL) techniques have emerged as powerful tools for identifying complex patterns in data. The multilayered architecture of DL models has demonstrated superior performance in speech processing and speaker-profiling tasks [12]. For instance, long short-term memory (LSTM) networks combined with features like MFCC have been employed for age estimation [13]. Additionally, research by Kalluri et al. [14] and Kaushik et al. [15] delved into the potential of deep neural networks (DNNs) for the estimation of various speaker attributes.

Traditional approaches have relied heavily on handcrafted feature-extraction techniques, such as MFCC and LPC, with classical machine-learning models, resulting in significant limitations in terms of accuracy, generalizability and efficiency. Other studies employed handcrafted features with DL models.

1. U. Jaid is with Department of Computer Science, College of Science, Uni. of Baghdad, Iraq. Email: umniah.h@sc.uobaghad.edu.iq
2. A. Abdulhassan is with Department of CS, University of Technology, Iraq. Email: Alia.K.AbdulHassan@uotechnology.edu.iq

These approaches, while being effective, introduce limitations in capturing the nuanced patterns within speech indicative of age and gender. Handcrafted features play a crucial role in the performance and accuracy of the recognition system; however, their implementation is challenging due to the complexity of feature engineering, as well as the significant time investment needed. Additionally, handcrafted feature extraction can underperform when the manually selected features aren't aligned with the task requirements.

To utilize the rich source of information available in the signal, including spatial cues, several studies adopted the utilization of raw input signal to DL models. Researches used raw signal directly as input to the DL model [16]-[17] or employed hybrid architectures to utilize both the spatial domain of the speech signal with handcrafted features [18]. Several researchers utilized pre-trained models, such as wave2vec and Titanet, to extract features from raw-speech signals directly [19]-[20].

The challenges of age-group prediction are further compounded by the intrinsic diversity of human speech, influenced by factors, such as emotion, health and accent, which can obscure critical demographic indicators. One major limitation to age-group prediction from speech is the ambiguity of the age, where speakers from adjacent age groups are often indistinguishable, due to the gradual change in speech characteristics with age. This problem is further emphasized with data imbalance with more samples in certain age groups than others. One approach to address this problem is the use of distribution learning, emphasizing the model's capability to output probability distributions that reflect the likelihood of each possible outcome, incorporating uncertainty into the predictions [21].

KL-divergence loss naturally accommodates this by comparing the predicted probability distribution against a target distribution that can represent soft labels, improving the model's ability to learn from nuanced differences in speech related to age. Instead of making hard predictions for a specific age group, using KL-divergence encourages the model to output a probability distribution over all possible age groups. This probabilistic approach is beneficial for capturing the uncertainty in age-group prediction, where speech features might not clearly distinguish between adjacent age groups.

Building on the strengths of using raw-speech signals with DL models and the strengths of KL-divergence loss, our proposed model addresses the aforementioned limitations and challenges, by introducing an end-to-end model that integrates Wav2Vec 2.0 embeddings with a self-attention-based CNN, utilizing Mozilla's Common Voice dataset. This methodology not only simplifies the feature extraction process, but also introduces a robust framework capable of discerning subtle age-related variations and gender characteristics in speech. By incorporating the principles of KL-divergence loss within a more comprehensive and advanced modeling approach, we address critical gaps in speaker profiling, including the challenges of age ambiguity and the need for robust, data-driven feature extraction.

In addition to employing KL-divergence loss and raw-speech signal with pre-trained feature extractors, the proposed model employs a self-attention mechanism. Attention mechanisms have recently revolutionized several fields, such as emotion recognition [22], natural-language processing [23] and speech recognition [24], enabling models to focus selectively on parts of the speech signal that are most relevant to the task at hand, by weighting different parts of the input differently, allowing the model to consider the context of the entire speech sequence when making predictions. A specific type of attention, self-attention allows models to capture dependencies and relationships between different parts of the speech signal, regardless of their distance within the sequence. This is particularly beneficial for understanding long-range dependencies in speech, where context from earlier parts of a sequence may influence the interpretation of later parts. This is particularly advantageous for age and gender prediction, where temporal dynamics across the entire speech sequence are analyzed, identifying patterns that are characteristic of different age groups and genders, allowing the model to dynamically focus on segments that are more informative for these predictions.

Additionally, this work provides an insight into the role that loss-function choice plays in the performance of the model, as we compare the performance of the model with several loss functions, such as regular corss-entropy loss, KL-divergence loss that is designed to handle age ambiguity and focal loss that is designed to handle age-group imbalance. A hybrid loss function is introduced in this work, focal-KL to introduce a balance between age-group imbalance and age ambiguity. Further, analyzing the relation between age-group sample size and the accuracy obtained for that age group,

showcases the effectiveness of the loss function in addressing the problem of data imbalance and age-group ambiguity.

The paper also demonstrates the robustness of the proposed model by conducting a thorough investigation into the impact of speech-segment duration on prediction accuracy with varying durations of speech ranging from 1 to 5 seconds of speech. This analysis informs our understanding of the balance between computational efficiency and the quality of our model's predictions.

Our proposed system outperforms existing DNN methods reliant on time-consuming, handcrafted feature extraction. Our work contributes to multi-task age group and gender detection from raw speech and introduces a novel combination of the self-attention mechanism with distributional learning.

Subsequent sections will explore our methodology in detail, present experiment results, compare our findings with existing literature, highlight potential applications and suggest future research directions in speaker profiling.

2. DATASET

To achieve our research objectives, it was crucial to select a dataset that is diverse and comprehensive. In light of this, we chose the Common Voice dataset by Mozilla [25]. This dataset is a crowdsourced, multi-language resource of spoken sentences. The dataset is rich in its demographic diversity, with data collected from speakers of various ages, genders and accents, making it an ideal resource for our research goal.

Each data entry consists of short spoken sentences, textual transcription and the demographic information of the speaker, including age group and gender. The age groups are categorized as 'Teens', 'Twenties', 'Thirties', 'Forties', 'Fifties', 'Sixties', 'Seventies' and 'Eighties and older'. Gender information as self-reported by the contributors is categorized as 'Male', 'Female' and 'Other'. The dataset is continually updated with new contributions; thus, the version used in this work is common_voice_11.

To maintain consistency and avoid ambiguity in the training data, only records marked as ‘Male’ and ‘Female’ were incorporated. Following the completion of data cleaning and removal of empty records, the dataset included 35,846 English-speaking samples from an array of global accents. This diverse collection includes accents from the USA, England, Australia, India, Canada, Malaysia, Scotland, Philippines, Singapore, Hong Kong and several other countries.

The dataset was divided as follows: 33,794 samples for training, 1,511 for validation and 577 for testing. From a gender-distribution perspective, it comprises 25,355 male samples and 8,439 female samples.

A detailed breakdown of the data distribution across various age and gender groups is provided in Table 1.

In the pre-processing phase, the audio data in the dataset originally stored in MP3 format, was converted into waveform samples for compatibility with the Wav2Vec model. For reasons of efficiency and memory management, longer utterances were cropped to 3 seconds, resulting in a maximum length of 48000 at a 16 kHz sampling rate.

Table 1. Description of the common voice dataset used in this work.

Age group	Training		Validation		Testing		Total
	Male	Female	Male	Female	Male	Female	
Teens	1,960	503	48	28	34	13	2,586
Twenties	8,601	1,830	389	87	149	39	11,095
Thirties	6,274	2,155	256	88	107	26	8,906
Forties	4,093	1,033	180	60	67	16	5,449
Fifties	2,307	2,055	116	87	42	32	4,639
Sixties	1,328	795	55	40	18	18	2,254
Seventies	691	58	36	1	14	-	800
Eighties	101	10	4	-	2	-	117
Total	25,355	8,439	1,084	391	433	144	35,846

3. PROPOSED METHOD

In this study, we propose an end-to-end methodology for speaker age and gender detection, leveraging the advanced capabilities of Wav2Vec2.0 for feature extraction from raw-audio signals. This approach eliminates the need for manual feature engineering, allowing the model to automatically learn the most informative aspects of the audio for our tasks.

The proposed methodology is comprised of three key aspects:

- Feature Extraction: The pre-trained Wav2Vec2.0 model is utilized to transform raw audio into high-quality feature representations. This unsupervised-learning technique captures complex speech characteristics essential for distinguishing speaker demographics.
- Self-attention-based CNN: The extracted features are processed through a self-attention-based convolutional neural network. This combination allows our model to dynamically focus on the most relevant parts of the audio signal for age and gender prediction.
- Loss Function Evaluation: To tackle the challenges of age ambiguity and class imbalance, various loss functions are explored, including focal loss and KL-divergence loss.
  A hybrid loss function combining focal loss and KL loss is introduced to offer a mixture for handling class imbalance and age ambiguity. This comparative analysis is crucial for optimizing our model's performance across diverse speech samples.

3.1 Network Architecture

In this study, we propose a novel architecture for audio-based gender and age classification. Our model employs the Wav2vec2.0 transformer-based architecture as an upstream model for feature extraction. Wav2Vec is an unsupervised-learning approach that transforms raw audio into rich, dense vector representations. These embeddings, also known as latent representations, capture significant information from the audio, such as speech content and speaker characteristics. Wav2vec2.0, pre-trained on a large corpus of unlabeled audio data, has demonstrated robustness in extracting meaningful representations from audio signals [26].

The extracted features are then passed through a series of three 1-dimensional convolutional layers, each followed by batch normalization. Each convolutional layer consists of filters of size 3, with the number of filters changing from 512 to 256 to 128 across the layers. The stride of 1 and padding of 1 are maintained in all convolutional layers.

Following feature extraction and convolutional processing, we employ adaptive average pooling with output size 64 to capture global temporal information. The output of the adaptive-pooling layer is then flattened before being passed to a self-attention mechanism. The self-attention mechanism assigns weights to features in the sequence based on their importance, thereby focusing the model's attention on the most informative parts of the audio signal. The attention mechanism consists of a linear transformation followed by a softmax activation function to generate attention scores.

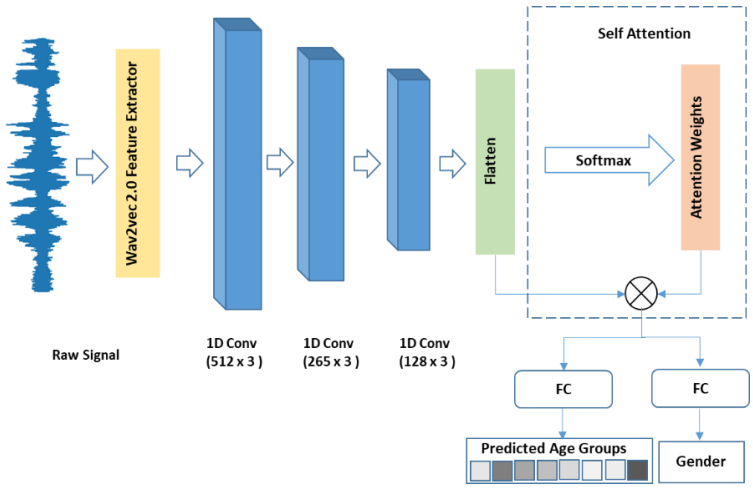

Figure 1. Overview of the proposed architecture.

Post the self-attention mechanism, a dropout layer is applied with a dropout rate of 0.5 to prevent overfitting. Subsequently, the processed features are passed to two separate fully-connected layers for the task of gender and age classification. The gender classifier consists of a linear layer with a single-output unit followed by a sigmoid activation function, classifying the audio clip into either of the two gender categories.

The age classifier, on the other hand, consists of a linear layer with an output size equal to the number of age categories. The proposed model effectively combines the strengths of transformer-based audio representation learning, convolutional processing, adaptive average pooling, self-attention mechanism and task-specific classification layers to perform the dual task of age and gender classification from raw-audio signals.

3.2 Loss Function

In our approach, the model is designed to simultaneously predict both age and gender. Thus, the composite loss function, as shown in Equation (1), merges the individual losses corresponding to age and gender predictions:

$$Loss = age_{loss} + gender_{loss} \tag{1}$$

For gender prediction, the loss is computed using the Mean Squared Error (MSE):

$$MSE_{gender} = \frac{1}{N}\sum_{i=1}^{N}(y_i - y\,pred_i)^2 \tag{2}$$

where N is the total number of predictions, y_i is the actual value of the i^{th} prediction and $y\,pred_i$ is the predicted value of the i^{th} prediction.

Regarding age prediction, we explore various loss functions including cross-entropy, focal loss and KL divergence. These will be detailed in the subsequent sub-sections.

3.2.1 Cross Entropy

Cross Entropy Loss is one of the most widely used loss functions for classification tasks. It measures the dissimilarity between the true label distribution and the predicted probabilities from the model.

Given a classification task with C classes, where each instance is assigned a label in the range [1, C], for a single data point, the predicted probabilities for each class can be determined using the softmax normalization function applied to the model's outputs. The predicted probability p_c of class c is computed as:

$$p_c = \frac{e^{x_c}}{\sum_{j=1}^{C} e^{x_j}} \tag{3}$$

where x_c is the output of the model corresponding to class c.

Given p_c as the probability of the predicted class and y_c as the true label, the cross entropy loss for that data point is defined as:

$$CE(p_c) = -y_c \log(p_c) \tag{4}$$

3.2.2 Focal Loss

While Cross Entropy is effective for many classification tasks, it may not perform as well in scenarios with significant class imbalance. In such cases, the model might become biased towards the majority class, often misclassifying the minority class.

To address this, Focal Loss was introduced as an enhancement over the standard Cross Entropy Loss [27]. It is specifically designed to give more importance to misclassified examples and is especially helpful for imbalanced datasets.

After obtaining the probability of each age class with softmax normalization as in (3), the Focal Loss for a true class c is defined as:

$$FL(p_c) = -\alpha(1 - p_c)^{\gamma} \log(p_c) \tag{5}$$

where p_c is the probability of the true class, α is a scaling factor for the loss and γ is a focusing parameter used to weigh down easy examples.

3.2.3 KL Loss

The proposed model leverages KL divergence (often referred to as the Kullback-Leibler divergence or relative entropy) as the loss function. KL divergence is a measure of how one probability distribution differs from a second, reference probability distribution. It's especially fitting for our problem since our model predicts a distribution over labels, rather than a singular label for each input.

For age-group detection, the KL divergence gauges the dissimilarity between the predicted label distribution and the true label distribution for each instance in the training set.

Given Q as the predicted probabilities for each instance, after softmax normalization, the true label for an instance with label c is represented as a one-hot encoded vector, P, defined as:

$$P = \begin{cases} 1 & if\ i = c \\ 0 & otherwise \end{cases}$$

The KL divergence is then computed as:

$$KL\ (P||Q) = \sum_{i=1}^{C} p_i \log(p_i/q_i)$$

where p_i and q_i are the true and predicted probabilities, respectively, for the i^{th} age group.

3.2.4 Focal-KL

The Focal-KL Loss is a hybrid loss that is a combination of the focal loss and the KL divergence loss, which attempts to leverage the benefits of both losses, where Focal Loss addresses the class-imbalance problem by giving more weight to the misclassified examples, while KL Divergence measures the divergence between two probability distributions, making it especially suitable when the model's predictions are distributions over labels.

To create a hybrid loss, we take a linear combination of the Focal Loss and KL Divergence:

$$\text{Focal} - \text{KL} = \boldsymbol{\lambda} \times \text{Focal\_Loss} + (1 - \boldsymbol{\lambda}) \times \text{KL}$$

where λ is a weighting coefficient in the range [0, 1] determining the contribution of each loss. A higher λ gives more weight to the Focal Loss, while a lower λ emphasizes the KL Divergence Loss.

4. EXPERIMENTS AND RESULTS

In this section, we evaluate our model's performance against various benchmarks, different loss functions and input durations to understand its strengths and potential areas of improvement.

To demonstrate the effectiveness of the proposed model, several experiments are performed. The first set of experiments compares the performance of a baseline model with 3 convolutional layers and no attention mechanism and the proposed model in age-group and gender detection. Next, we compare the performance of different loss functions on the proposed model. Finally, duration analysis is performed by performing tests on different durations of the model ranging from one to five seconds. The experiments are performed with a learning rate (1×10^{-6}) and a batch size of 32.

4.1 Self-attention Mechanism

The primary objective here is to discern the impact of incorporating a self-attention mechanism into our model as compared to a baseline model that lacks this feature. To investigate the efficacy of integrating a self-attention mechanism, we compare our proposed model against a baseline architecture. This baseline encompasses three convolutional layers, employs wav2vec for feature extraction and incorporates adaptive pooling. Notably, it lacks the self-attention mechanism characteristic of our proposed design. Both models were trained under identical settings using cross-entropy loss. As presented in Table 2, the inclusion of the self-attention mechanism manifests in marked improvements in age-prediction accuracies for both male and female categories. Conversely, the gender-recognition capability remains consistent across the two models, underscoring the specific advantages of self-attention in age-prediction tasks.

Table 2. Age and gender accuracy of the proposed and baseline models.

	Age Accuracy			Gender
	Male	Female	All	
Baseline	0.72	0.76	0.734	0.98
Proposed	0.76	0.83	0.78	0.98

4.2 Effect of Different Loss Functions

This sub-section aims to evaluate and compare how the model performs when trained with various loss functions, emphasizing the model's adaptability and optimization potential. To demonstrate the effects of the loss function on the model's performance, we evaluate the model with different loss functions; namely, CE, CE with focal loss, Kl divergence loss and a hybrid Kl with focal loss. As seen in Table 3, across the board, it's evident that the model is highly adept at gender classification, achieving an accuracy range of 0.98 to 0.99 regardless of the loss function used. This underscores the robustness of the architecture in distinguishing gender-based audio features.

For age-group detection, using the plain CE, we observed that the model had a higher accuracy for female speakers at 0.841 compared to male speakers at 0.796, yielding an overall average accuracy of 0.807. However, incorporating the focal loss, which is especially effective in addressing class imbalance, shows a marked improvement in performance for both genders. The gap between male and female accuracy narrows, with females achieving a commendable 89.5% accuracy. Switching to the KL divergence loss sees further improvements, especially for female speakers who achieve a 91.5% accuracy. The overall accuracy, taking into account both genders, reaches 86.7%, marking a substantial enhancement over the traditional CE loss.

Combining KL with focal loss produces results that are marginally better than using CE alone, but slightly lag when compared to using either the focal loss or KL divergence loss separately. This could indicate that while both focal and KL loss individually address certain nuances of the dataset, their combination may not necessarily be synergistic for this specific task.

Table 3. Age and gender accuracy of the proposed model using different loss functions.

Loss Function	Age Accuracy			Gender
	Male	Female	All	
CE	76	83	78	98
Focal	84.5	89.5	85.8	98.9
KL	85	91.5	86.7	98.9
Focal_KL	85.4	88.5	86.2	99

4.3 Duration Analysis

In this experiment, we explore how speech input duration influences age-prediction accuracy across different loss functions, aiming to identify the optimal speech duration for accurate predictions. Comparing the age-prediction accuracy of different loss functions at various durations of speech input, it can be seen in Figures 2, 3 and 4 that as the duration of speech input increases, the accuracy tends to increase for all loss functions. This suggests that having more speech data generally results in better age prediction. However, The KL loss seems to consistently provide the highest or one of the highest accuracies across different speech durations. It's especially dominant in the 1-second and 2-second durations. Similarly, Focal-KL shows an interesting trend, where it jumps to 89% at 2 seconds, leading all other methods, but it then aligns more closely with the rest at longer durations.

At 1 second, the KL loss seems to be the most effective with an accuracy of 40%, while other losses are somewhat close, between accuracies of 35% and 38%. However, starting at 2 seconds, there is a significant improvement with KL loss and Focal-KL loss providing the best performance with accuracies of 79.9% and 89%, respectively. At 3 seconds, all loss functions are in the mid-80s range with KL loss leading at 86.7%. The highest accuracy is achieved at the duration of 4 seconds, with KL loss slightly ahead at 88.6%. The performance plateaus at 5 seconds with KL still leading at 88.2% with other losses performing very closely.

In general, between 1 and 2 seconds, there is a large accuracy improvement, jumping from 38% to 89%.

However, between 3 and 5 seconds, there's very minimal improvement across all methods, suggesting a diminishing return of increased speech duration beyond 3 seconds for age prediction with the given model and dataset. Table 4 summarizes the accuracies achieved at 4 seconds.

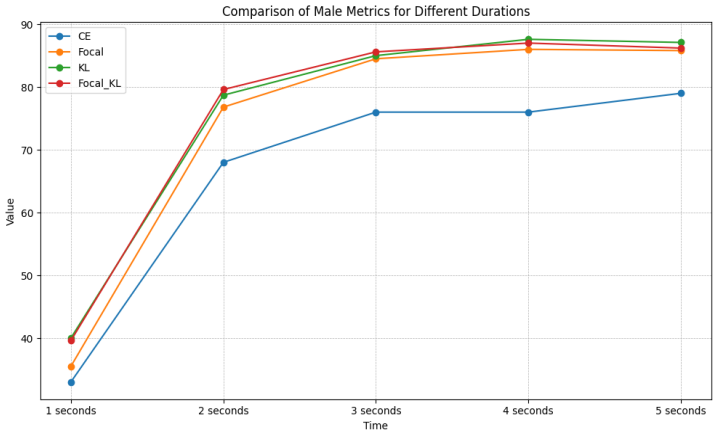

Figure 2. Comparison of male accuracies for different durations.

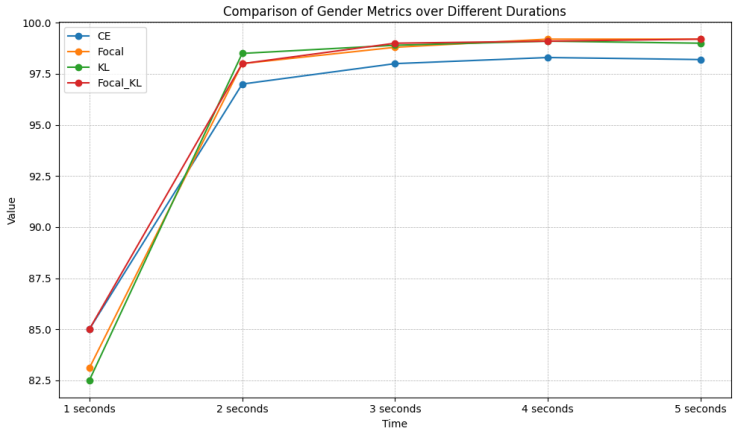

Figure 3. Comparison of female accuracies for different durations.

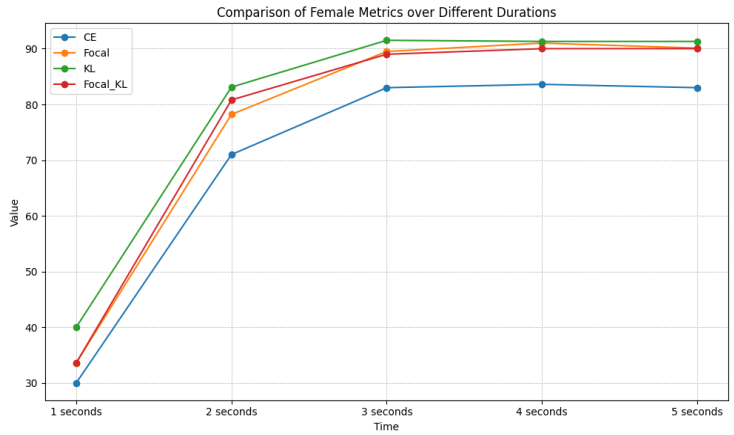

Figure 4. Comparison of gender detection for different durations.

Table 4. Age and gender accuracy of the proposed model at 4-second duration.

Loss Function	Age Accuracy			Gender
	Male	Female	All	
CE	76.7	83.6	78.5	98.3
Focal	86	91	87.3	**99.2**
KL	**87.6**	**91.3**	**89**	99.1
Focal_KL	87	90	87.7	99.1

4.4 Discussion

In this sub-section, we delve deeper into the results obtained from the experiments, aiming to extract insights and understand patterns in the model's performance. The series of experiments performed in this work not only establishes the effectiveness of our proposed model, but also uncovers intriguing insights regarding age and gender prediction from audio data.

Integrating the self-attention mechanism led to a discernible improvement in age-prediction accuracies for both genders. This enhancement particularly highlights the capacity of the self-attention mechanism to discern age-related attributes in audio data. Contrastingly, the gender-recognition performance remained consistent across the models, implying that the impact of the self-attention mechanism on gender prediction, given the current architectural choices, is relatively limited.

Notably, there's a distinct gender disparity in age accuracy when using the CE loss, pointing to potential inherent biases or differentiating features in the dataset. The introduction of alternative loss functions not only boosts the overall performance, but significantly narrows this gender discrepancy. This is indicative of the effectiveness of these losses in managing potential class imbalances. Combining KL loss and focal loss offered a slight improvement over the individual focal loss; however, it didn't outperform KL loss, suggesting that the performance improvement might be attributed to the KL part of the hybrid loss.

Duration analysis offered an understanding of the relationship between the amount of speech data and prediction accuracy. An evident rapid surge in accuracy with increased duration emphasizes the additional informative value extracted from longer speech samples. Yet, the performance plateau beyond 3 seconds hints at a saturation point, suggesting an optimal duration window that offers the maximum informational value without redundancy.

To provide a comparative perspective on the effectiveness of various methodologies in the field, Table 5 summarizes the classification accuracies achieved by different studies.

Table 5. Comparison of classification accuracies across different studies and numbers of classes.

Study	No. of Classes	Accuracy (All)	Accuracy (Gender)
H. Abdulmohsin et al. [28]	2	87.97%	-
Sánchez-Hevia et al. [29]	6	83.23%	98.24%
D. Kwasny et al. [30]	8	-	99.6%
A. Tursunov et al. [31]	6	73%	96%
Sánchez-Hevia et al. [32]	8	80%	98.14%
Proposed Method	8	89%	99.1%

Our experimental results showcase not only a high degree of accuracy in age and gender detection, but also a significant improvement over existing state-of-the-art methods. Compared to the latest reported accuracies in speaker age-group detection, as reported in Table 5, our model demonstrates a marked increase in precision, especially in distinguishing between closely adjacent age groups—a longstanding challenge in the field. The proposed model achieved an overall accuracy of 89% in age detection and 99.1% in gender detection. Differently from similar studies presented in Table 6 [28]-[29], [31], the age detection accuracy is achieved over eight age groups while similar studies divided the dataset into 2 or 6 classes. These results are notably superior to those of existing models, indicating the effectiveness of our approach in capturing and analyzing the nuanced features of speech that correlate with age and gender.

Figures 5 and 6 show the confusion matrix of the best-performing model with a 4-second duration of the speech and KL loss for age-group and gender prediction respectively. The confusion matrix analysis for age-group classification reveals a detailed performance of the model across various age brackets. For the "teens" group, the model correctly classified 81% of the samples, suggesting a reasonable accuracy, but leaving room for improvement. The model's performance peaks for individuals in their twenties, forties, fifties and sixties with accuracy rates of 89%, 89%, 93% and 94%, respectively. The "thirties" group witnesses a slightly lower accuracy at 87%. Remarkably, the model's efficacy ascends as it approaches the "seventies" age group, achieving a 97% accuracy. However, this trend takes a downturn for the oldest age bracket in the dataset. The "eighties and more" group observes a significant

decline in accuracy of 60%; however, 20% of the misclassified instances were misclassified as the adjacent age group seventies.

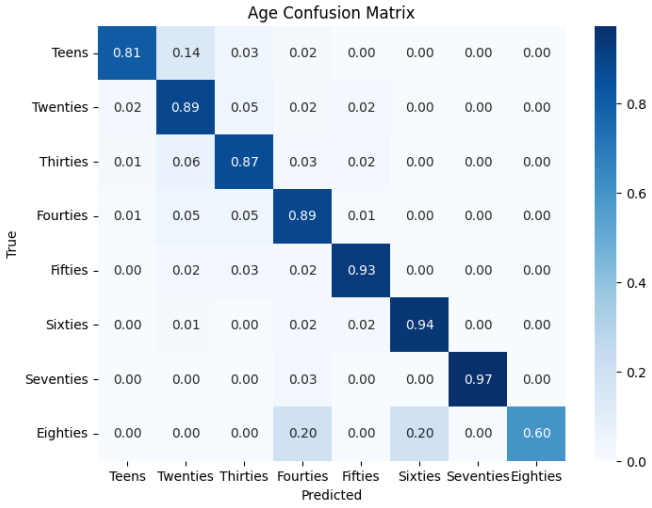

Figure 5. Confusion matrix of age-group prediction of the proposed model.

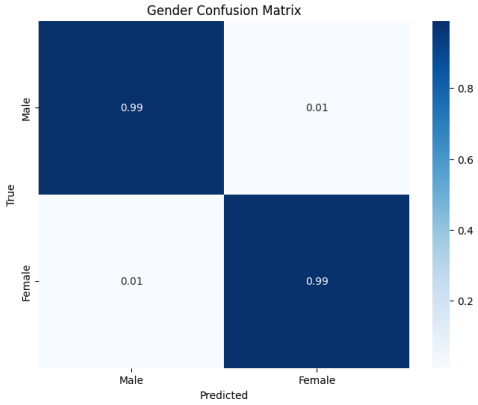

Figure 6. Confusion matrix of gender prediction of the proposed model.

Comparing the number of training and testing instances with the acquired accuracies (Figures 7 and 8) shows that there doesn't appear to be a direct linear relationship between dataset size and accuracy. Larger datasets (like "twenties") don't necessarily have the highest accuracy and smaller datasets (like "seventies") don't necessarily have the lowest accuracy. However, the sharp drop in accuracy for the "eighties and more" group suggests that a minimum threshold of data might be essential for achieving reasonable performance.

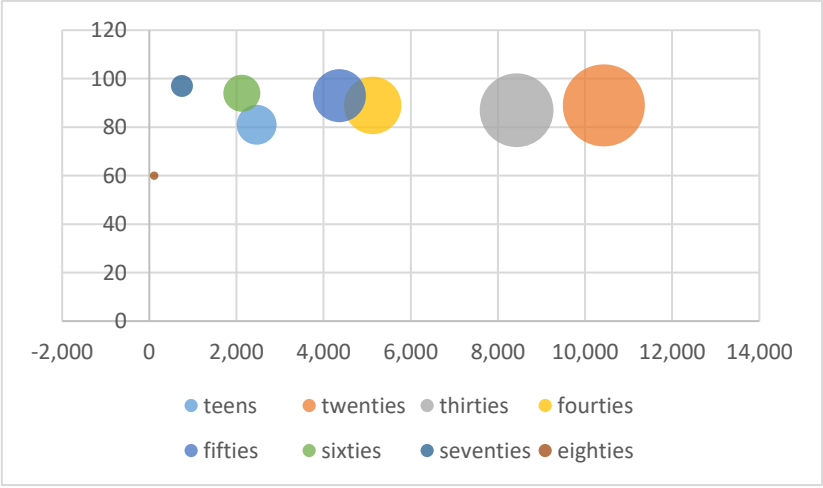

Figure 7. Correlation between obtained accuracies and number of instances in the training dataset.

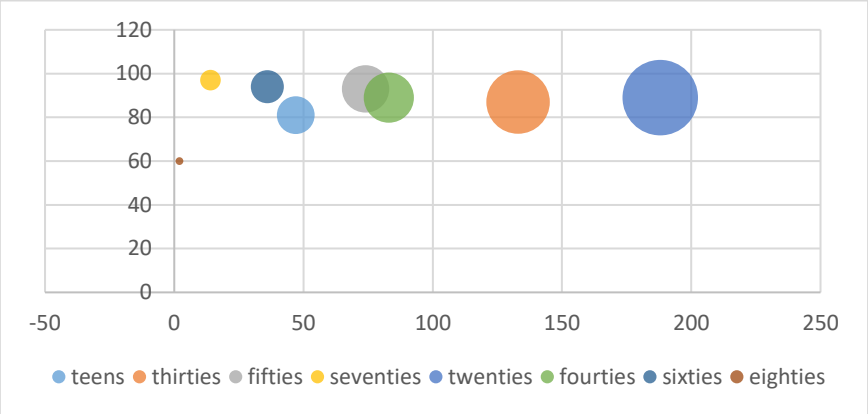

Figure 8. Correlation between obtained accuracies and number of instances in the testing dataset.

Unlike traditional approaches that rely on handcrafted features, our model facilitates an end-to-end learning process by utilizing the Wav2Vec2.0 for feature extraction, benefiting from rich, pre-trained representations of audio data. This unsupervized learning approach allows the model to leverage large amounts of unlabeled audio data, providing a robust foundation for understanding complex speech characteristics without the need for extensive manual feature engineering.

The integration of a self-attention mechanism within the CNN architecture enables the model to dynamically focus on the most informative parts of the audio signal. This aspect is particularly beneficial for age detection, where subtle variations in speech patterns can significantly impact accuracy. Our findings indicate that the self-attention mechanism contributes to a marked improvement in age-prediction accuracies for both male and female speakers.

The proposed model also demonstrates consistent performance across a range of speech durations, from short clips to longer utterances. This versatility suggests that the model can effectively extract and utilize relevant information from audio signals of varying lengths, enhancing its applicability in real-world scenarios where speech samples may not be uniformly sized.

While our proposed method demonstrates promising results in speaker age and gender detection, it is not without limitations. One of the main difficulties lies in the reliance on high-quality, diverse training data. The performance of our model, especially its ability to generalize across different accents, dialects and speech patterns, is heavily dependent on the breadth and depth of the dataset used for training. The Common Voice dataset, while being extensive, may not fully represent the global diversity of speech, potentially limiting our model's applicability in real-world scenarios across various languages and socio-linguistic backgrounds.

Additionally, the computational complexity of our model, driven by the sophisticated feature extraction with Wav2Vec2.0 and the self-attention mechanism, presents a challenge for deployment in low-resource environments or in real-time applications. The balance between model complexity and practical usability is a critical consideration, especially for applications requiring rapid processing or deployment on devices with limited computational capabilities.

Moreover, while our approach addresses age ambiguity to some extent, distinguishing between speakers of closely adjacent age groups remains a challenge. The subtle vocal variations that differentiate age groups may not always be captured or deemed significant by the model, particularly in cases where the training data lacks sufficient examples of such subtle differences.

5. CONCLUSIONS

Our study introduces a novel, end-to-end 1D CNN model for detecting speaker age and gender from speech signals, achieving an overall accuracy of 89% for age groups and a 99.1% accuracy in gender detection, thereby demonstrating significant improvements over traditional methods. This network architecture, built upon three convolutional layers, integrates a self-attention mechanism and leverages direct-speech representations from the advanced pre-trained wav2vec2.0 model, eliminating the need for manual feature extraction. Our evaluation, conducted on the Common Voice dataset comprised of 35,845 speech samples, not only yields promising results in age-group classification and gender

detection, but also showcases the model's versatility by accommodating variable audio lengths. This paves the way for its application in real-world scenarios, particularly enhancing user experiences in mobile devices and human-computer interaction domains where adaptability to varying speech inputs is crucial. The distinct influence of the loss function on model efficacy, with a marked preference for KL and the innovative focal-KL loss functions, underscores the nuanced approach required for optimal performance. Despite the robust performance of our model, the challenge of differentiating between adjacent age groups underscores the complexity of vocal age markers and highlights an avenue for future exploration. Delving deeper into neural-network architectures or innovative feature representations could unveil more granular age-related vocal characteristics. Moreover, expanding our dataset to encompass a broader spectrum of languages, dialects and recording conditions will be imperative for enhancing the model's generalizability and mitigating potential biases.

REFERENCES

[1] G. Assunção, P. Menezes and F. Perdigão, "Speaker Awareness for Speech Emotion Recognition," Int. J. of Online and Biomedical Engineering, vol. 16, no. 4, pp. 15-22, 2020.

[2] A. H. Poorjam and M. H. Bahari, "Multitask Speaker Profiling for Estimating Age, Height, Weight and Smoking Habits from Spontaneous Telephone Speech Signals," Proc. of the 2014 4th IEEE Int. Conf. on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, pp. 7-12, 2014.

[3] C. Müller, "Automatic Recognition of Speakers[2] Age and Gender on the Basis of Empirical Studies," Proc. of the 9th Int. Conf. on Spoken Language Processing (Interspeech 2006), pp. 2118–2121, paper 1031-Wed3CaP.11, DOI: 10.21437/Interspeech.2006-195, 2006.

[4] C. Müller and F. Burkhardt, "Combining Short-term Cepstral and Long-term Pitch Features for Automatic Recognition of Speaker Age," Proc. of the 8th Annual Conf. of the Int. Speech Communication Association, (Interspeech 2007), pp. 2277–2280, Antwerp, Belgium, 2007.

[5] S. B. Kalluri, A. Vijayakumar, D. Vijayasenan and R. Singh, "Estimating Multiple Physical Parameters from Speech Data," Proc. of the 2016 IEEE 26th Int. Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-5, Vietri sul Mare, Italy, 2016.

[6] S. Galgali, S. S. Priyanka, B. Shashank and A. P. Patil, "Speaker Profiling by Extracting Paralinguistic Parameters Using Mel Frequency Cepstral Coefficients," Proc. of 2015 IEEE Int. Conf. on Applied and Theoretical Computing and Communic. Technology (iCATccT), pp. 486-489, Davangere, India, 2015.

[7] A. A. Badr and A. K. Abdul-Hassan, "Estimating Age in Short Utterances Based on Multi-class Classification Approach," Computers, Materials & Continua, vol. 68, no. 2, pp. 1713-1729, 2021.

[8] I. Mporas and T. Ganchev, "Estimation of Unknown Speaker's Height from Speech," International Journal of Speech Technology, vol. 12, no. 4, pp. 149-160, DOI: 10.1007/s10772-010-9064-2, 2010.

[9] K. A. Williams and J. H. Hansen, "Speaker Height Estimation Combining GMM and Linear Regression Subsystems," Proc. of 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2013), pp. 7552-7556, 2013.

[10] H. Arsikere, G. K. F. Leung, S. M. Lulich and A. Alwan, "Automatic Estimation of the First Three Subglottal Resonances from Adults Speech Signals with Application to Speaker Height Estimation," Speech Communication, vol. 55, no. 1, pp. 51-70, DOI: 10.1016/j.specom.2012.06.004, 2013.

[11] A. A. Mallouh, Z. Qawaqneh and B. D. Barkana, "New Transformed Features Generated by Deep Bottleneck Extractor and a GMM-UBM Classifier for Speaker Age and Gender Classification," Neural Computing & Applications, vol. 30, no. 8, pp. 2581-2593, DOI: 10.1007/s00521-017-2848-4, 2018.

[12] O. Buyuk and M. L. Arslan, "Combination of Long-term and Short-term Features for Age Identification from Voice," Advances in Electrical and Computer Engineering, vol. 18, no. 2, pp. 101-108, 2018.

[13] R. Zazo et al., "Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks," IEEE Access, vol. 6, pp. 22524-22530, DOI: 10.1109/access.2018.2816163, 2018.

[14] S. B. Kalluri, D. Vijayasenan and S. Ganapathy, "A Deep Neural Network Based End to End Model for Joint Height and Age Estimation from Short Duration Speech," Proc. of ICASSP 2019 - 2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 6580-6584, Brighton, UK, 2019.

[15] M. Kaushik, V. T. Pham and E. S. Chng, "End-to-End Speaker Height and Age Estimation Using Attention Mechanism with LSTM-RNN," arXiv preprint arXiv: 2101.05056, 2021.

[16] S. Kwon, "1D-CNN: Speech Emotion Recognition System Using a Stacked Network with Dilated CNN Features," Computers, Materials & Continua, vol. 67, no. 3, pp. 4039-4059, 2021.

[17] U. H. Jaid and A. K. AbdulHassan, "End-to-End Speaker Profiling Using 1D CNN Architectures and Filter Bank Initialization," Int. J. of Online & Biomedical Engineering, vol. 19, no. 10, 2023.

[18] Mustaqeem and S. Kwon, "Optimal Feature Selection Based Speech Emotion Recognition Using Two-stream Deep Convolutional Neural Network," Int. J. of Intellig. Syst., vol. 36, no. 9, pp. 5116-5135, 2021.

[19] M. Z. Tarashandeh, A. Torkanloo and M. H. Moattar, "AgeNet-AT: An End-to-End Model for Robust

Joint Speaker Age Estimation and Gender Recognition Based on Attention Mechanism and Titanet," Proc. of the 2023 13th IEEE Int. Conf. on Computer and Knowledge Engineering (ICCKE), pp. 414-419, Mashhad, Iran, 2023.

[20] T. Gupta, D.-T. Truong, T. T. Anh and C. E. Siong, "Estimation of Speaker Age and Height from Speech Signal Using Bi-encoder Transformer Mixture Model," arXiv preprint, arXiv: 2203.11774, 2022.

[21] S. Si, J. Wang, J. Peng and J. Xiao, "Towards Speaker Age Estimation with Label Distribution Learning," arXiv preprint, arXiv: 2202.11424, 2022.

[22] S. Kwon, "Att-Net: Enhanced Emotion Recognition System Using Lightweight Self-attention Module," Applied Soft Computing, vol. 102, p. 107101, 2021.

[23] A. Galassi, M. Lippi and P. Torroni, "Attention in Natural Language Processing," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 10, pp. 4291-4308, 2020.

[24] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, S. Stüker and A. Waibel, "Very Deep Self-attention Networks for End-to-End Speech Recognition," arXiv preprint, arXiv:1904.13377, 2019.

[25] R. Ardila et al., "Common Voice: A Massively-multilingual Speech Corpus," arXiv: 1912.06670, 2019.

[26] A. Baevski et al., "Wav2vec 2.0: A Framework for Self-supervised Learning of Speech Representations," Advances in Neural Information Processing Systems, vol. 33, pp. 12449-12460, 2020.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," Proc. of the IEEE Int. Conf. on Computer Vision, pp. 2980-2988, 2017.

[28] H. A. Abdulmohsin, J. J. Stephan, B. Al-Khateeb and S. S. Hasan, "Speech Age Estimation Using a Ranking Convolutional Neural Network," Proc. of Int. Conf. on Computing and Communication Networks (ICCCN 2021), pp. 123-130, Springer, 2022.

[29] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso and M. Rosa-Zurera, "Age Group Classification and Gender Recognition from Speech with Temporal Convolutional Neural Networks," Multimedia Tools and Applications, vol. 81, no. 3, pp. 3535-3552, 2022.

[30] D. Kwasny and D. Hemmerling, "Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks," Sensors, vol. 21, no. 14, p. 4785, 2021.

[31] A. Tursunov, Mustaqeem, J. Y. Choeh and S. Kwon, "Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-attention Module through Speech Spectrograms," Sensors, vol. 21, no. 17, p. 5892, 2021.

[32] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso and M. Rosa-Zurera, "Convolutional-recurrent Neural Network for Age and Gender Prediction from Speech," Proc. of the 2019 IEEE Signal Processing Symposium (SPSympo), pp. 242-245, Krakow, Poland, 2019.

ملخص البحث:

يمكن أن يزودنا الصّوت بمعلوماتٍ غنية عن بعض الخصائص الشخصية للمتكلّم، بما في ذلك عمره وجنسه. وإنّ تقدير عمر المتكلّم وجنسه يوفر لنا مدىً واسعاً من التّطبيقات يمتدّ من التّحليل الشّرعي للصّوت إلى الإعلان المشخّص والرّصد المتعلّق بالرّعاية الصّحية وتفاعل الإنسان مع الحاسوب. ومع ذلك، يبقى التّحديد الدّقيق للعُمْر أمراً مَشوباً بالصُّعوبة والغموض. ويتعلّق الأمر بتشابه سِمات أصوات الأشخاص ذوي الأعمار المتقاربة إلى درجةٍ يصْعب معها تمييزها. ولمعالجة هذا الأمر، نقترح نموذجاً يستخدم ما يعرف بمجموعة بيانات الصّوت العام (Common Voice) لتحويل الصّوت الخام إلى تمثيلاتٍ للخصائص عالية الجودة. ومن ثمّ يجري إدخال هذه الخصائص إلى سلسلة من الشّبكات العصبية الالتفافية. وللتّغلُّب على غُموض العُمر، فإنّنا نعمل على تقييم آثار مجموعة متنوعة من دوالّ الفقْد على دقّة النّموذج في تقدير إشاراتٍ صوتية مختلفة الفترة الزمنية لبقائها.

وقد أثبتت التّجارب التي أجريناها على مجموعة بيانات (الصّوت العام) نجاعة النّموذج المقترح وتفوقّه على عددٍ من النّماذج المماثلة الواردة في أدبيات الموضوع؛ فقد حقّق النّموذج المقترح في تقدير العُمْر دقّةً وصلت إلى 87% للمتكلمين الذّكور، و 91% للمتكلمات الإناث، بدقةٍ إجمالية بلغت 89%، بينما بلغت الدّقة الإجمالية فيما يتعلّق بجنس المتكلّم 99.1%.