Voice Based Age and Gender Detection using Long Short - Term Memory

Nguyen Minh Nhut, Nguyen Thanh Trung, Nguyen Huu Tan,
Luong Dang Doanh, and Duc Ngoc Minh Dang
FPT University, Ho Chi Minh Campus, Vietnam
{nhutnmse184534, trungntse180355, tannhse180669, doanhldse184146}@fpt.edu.vn,
and ducdnm2@fe.edu.vn

Abstract

This research investigates the efficacy of Long Short-Term Memory (LSTM) networks for voice-based age and gender detection. Leveraging a large dataset of labeled voice samples, LSTM models were trained to predict age and gender attributes from raw voice signals. The study achieved notable accuracy rates, with age prediction accuracy reaching 61% and gender prediction accuracy achieving 93%. By employing LSTM networks, renowned for their ability to capture temporal dependencies in sequential data, the research demonstrates promising results in demographic attribute prediction from voice data. The findings suggest that LSTM-based approaches hold considerable potential for robust and accurate age and gender detection, with implications for various applications including virtual assistants, personalized healthcare, and social interaction platforms. Further exploration of model architectures and training strategies may enhance prediction accuracy and broaden the applicability of voice-based demographic prediction systems.

Index Terms

Voice-based, Age Classification Model, Gender Classification Model, Long Short-Term Memory, Deep Learning, MFCC.

I. Introduction

In recent times, there have been notable advancements in the field of voice-based recognition systems, specifically in the areas of distinguishing gender and age. These systems offer vast potential for a variety of applications, including personalized interactions in digital assistants and security measures [1]. The increasing prevalence of smart devices and the incorporation of voice interfaces across different sectors has resulted in a growing need for precise and effective voice recognition systems.

The capacity to accurately determine gender and age through voice signals demonstrates advancements in artificial intelligence and machine learning, while also presenting opportunities for cutting-edge applications in human-computer interaction [2]. Previously, identifying gender and age relied on visual cues or explicit user inputs, both of which were susceptible to biases and inaccuracies. However, with the emergence of advanced algorithms and the abundance of labeled data, voice-based recognition systems have attained significant levels of accuracy and dependability.

Utilizing spoken communication to express ideas is a fundamental aspect of human interaction. The tone of voice encompasses various attributes that differ from individual to individual, offering insights into emotions [3], mental states, age, and gender. This multi-faceted information can be applied in various contexts such as voice-activated security systems [1], automated speech systems, artificial intelligence assistants [4], Automatic Speaker Verification [5], emotion-based artificial intelligence systems [6], and voice-based input systems for database navigation. Additionally, voice-based gender and age identification can play a role in artificial intelligence-based security measures, crime resolution, and victim safeguarding. The diverse characteristics of speech hold significant value in numerous contemporary technologies. The waveform representation of a male audio signal is illustrated in Figure 2, while the waveform of a female audio signal is depicted in Figure 1.

Visual analysis of voice patterns (spectrograms) [7] can reveal differences between people, such as those caused by age and gender. As an individual grows older, the pitch and frequency of their voice [8] tend to decrease [9]. Through the use of annotated training data, supervised machine learning techniques can accurately predict characteristics

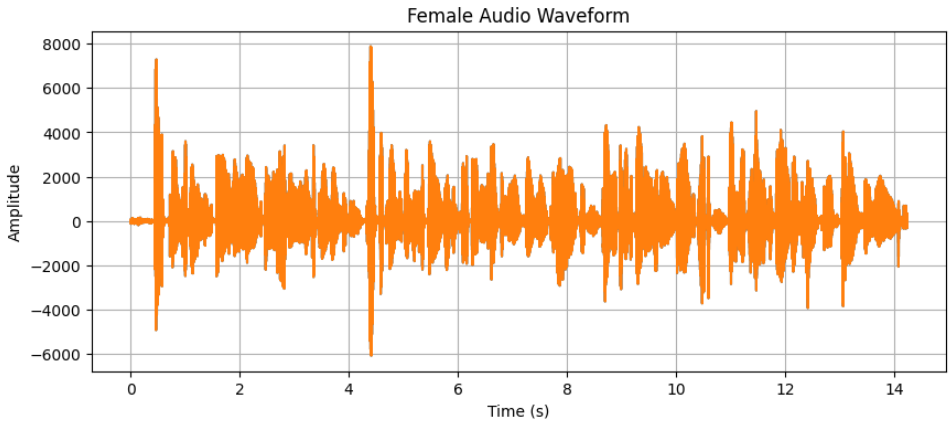

Fig. 1. Female Audio Signal as Waveform

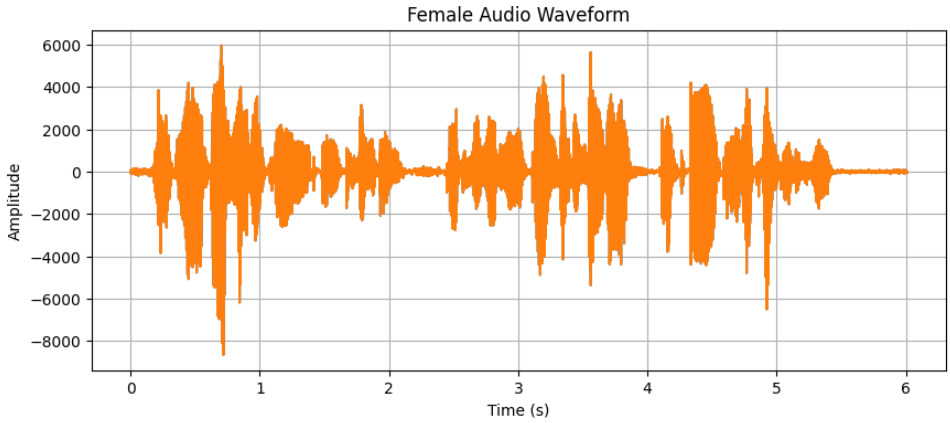

Fig. 2. Male Audio Signal as Waveform

that may otherwise remain unnoticed. The analysis of speech involves examining attributes like gender, age, and emotional tone to generate forecasts. The identification of these characteristics in spoken language holds the potential to enhance human-machine interactions significantly. For example, by categorizing audio recordings according to age and gender, businesses such as telecommunications companies can forecast customer demographics, age, gender, and emotions to tailor appropriate offerings.

Gender prediction through voice in Deep Learning is a potent technique for determining an individual's gender based on their voice characteristics. Depending on the model's complexity, it may incorporate various layers such as Convolutional Neural Network (CNN) [10], recurrent [11], temporal CNN [12], or fully connected layers. These models consist of layers designed to extract features from audio files, such as tone and pitch [6]. Once trained, the model can effectively predict gender in new audio files. Age prediction from voice recordings in Machine Learning is a field of research that utilizes machine learning algorithms to discern age from voice recordings. This process involves identifying traits like pitch, timbre, amplitude, and other acoustic features. Principal component analysis (PCA) [13] can be employed to uncover patterns in the dataset and identify significant features for accurate age prediction.

This project explores using Long Short-Term Memory (LSTM) networks to predict age and gender from voice recordings. The goal is to create reliable and accurate models that work well for different speakers and recording situations. We will train LSTMs on training datasets with labeled information to identify hidden patterns in voices that connect to age and gender. We will also investigate new approaches, like attention mechanisms and multi-tasking, to improve the accuracy of these voice-based demographic prediction systems.

The document is structured as follows: Section II offers a comprehensive review of established models and research about the subject matter. This section offers valuable perspectives on the issue, aiding in pinpointing areas for enhancement, identifying gaps in current understanding, and proposing alternative strategies. Moreover, it furnishes evidence supporting the efficacy of a proposed solution or highlighting potential challenges. Section III delineates the implemented model, while Section IV details the dataset utilized for both testing and training purposes. Section V encapsulates the report's discoveries and succeeds with a summary of findings and potential avenues for future exploration.

II. Related Work

This section discusses various approaches employed for voice-based Age and Gender Detection. S. Mavaddati [14] introduces a novel gender and age recognition system for telephone speech processing, aiming to identify individuals based on voice characteristics. The system utilizes generative incoherent models learned through sparse non-negative matrix factorization and atom correction post-processing. The algorithm involves training to associate atoms with signal classes and testing to evaluate classification performance. Mel-frequency cepstral coefficients are employed for feature representation, learned over data from male and female speakers using non-negative matrix factorization with sparsity constraint. Atom correction reduces coherence between different dictionary categories by replacing high-energy atoms related to other sets with low-energy bases, provided reconstruction error remains within a specified limit. Experimental results demonstrate the algorithm's superior performance over previous methods, particularly in noisy environments.

M. Abdollahi *et al.* [15] introduce a novel gender identification method utilizing adaptive multiresolution (MR) classification of Spectro-temporal maps derived from auditory-inspired representations of speech signals, including Mel-spectrogram, cochlea gram, and auditory spectrogram. Segments of utterances are represented as 2-D images and undergo MR decomposition before classification, employing feature extraction and classification within each MR subspace. A weighting algorithm combines the results into a global decision. The proposed method achieves high accuracy, up to 99%, surpassing common algorithms combining pitch and acoustical features for gender identification.

Human communication is important because it allows us to express our emotions, according to N. Sandeep Chaitanya *et al.* [16]. A powerful emotion that stems from one's environment or situation is called an emotion. Speech analysis is essential for fostering natural interactions between people and machines as well as for lowering feelings of isolation and alienation among people. Speech emotion detection involves taking a person's emotional state directly out of their speech. The goal of our project is to create a hybrid system that can analyze speech patterns to determine age, gender, and emotion. This is something the current system cannot accomplish because it uses different systems to determine age, gender, and emotion. We will be taking speech signals as input, which will be converted into a NumPy array and later classified using the SVM algorithm.

Vinayak Sudhakar Kone *et al.*'s research [17] underscores the importance of identifying gender and age from voice data across various applications. They propose a Machine Learning-driven approach, leveraging Deep Learning methods to enhance accuracy. Their study introduces a systematic grid search method for age estimation, integrating techniques like RobustScaler, Principal Component Analysis (PCA), and Logistic Regression to optimize prediction models. For gender detection, they utilize a sequential model with five hidden layers. Their experiments on the common voice dataset demonstrate promising results, achieving around 91% accuracy in gender prediction and 59% accuracy in age prediction.

Sandeep Kumar *et al.* [18] discusses the inherent human ability to easily identify gender from voice due to the acquisition of knowledge from the surrounding environment. However, this task becomes challenging for computers, necessitating the training of machines with relevant features and diverse datasets. The research paper proposes a model combining Naïve Bayes and deep learning methods for gender identification based on acoustic features of voice. Approximately 2000 voice samples from classmates and various websites are collected for model training and testing, with durations ranging from 3 to 8 seconds. The dataset consists of an equal distribution of male and female voices. The proposed model utilizes two classifiers, namely Neural Network and Naïve Bayes, achieving significant accuracies of 99% and 98%, respectively, showcasing the effectiveness of the approach.

III. Proposed Work

Our primary objective is to develop an efficient system capable of predicting the age and gender of a person without compromising their security or bypassing any security measures. Such a system can be beneficial in various applications and activities. Our main focus is to train a model that can accurately predict age and gender in the most efficient manner possible. Additionally, we aim to utilize this system for age-specific content access control, where the system can detect the age and gender of a user and grant or deny access to content based on predefined limitations. The Architecture of Voice based Age and Gender Detection is presented in Figure 3

The Long Short-Term Memory architecture (LSTM) incorporates memory blocks within the recurrent hidden layer. These memory blocks consist of memory cells with self-connections, allowing them to store the temporal state of the network. Additionally, the architecture includes special multiplicative units called gates, which are responsible for controlling the flow of information. In the original design, each memory block featured an input gate and an output gate [19].

LSTM networks have demonstrated their effectiveness and strength in handling sequential data by overcoming the difficulties of capturing long-term dependencies and mitigating the vanishing gradient problem [19].

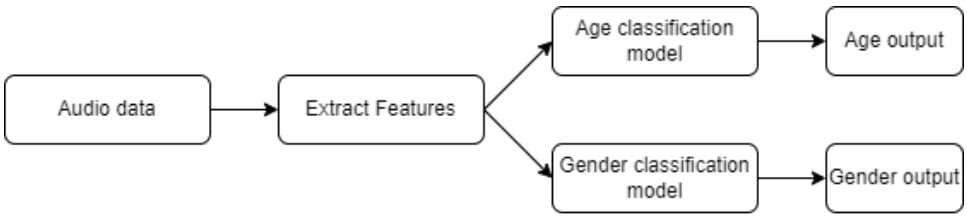

Fig. 3. Architecture of Voice based Age and Gender Recognition

A. Dataset

The Mozilla Voice Dataset [20] stands as a substantial compilation, comprising 356,921 meticulously curated items, designed to capture the intricate interplay between age and gender among Japanese speakers. This extensive dataset serves as a valuable resource for researchers, developers, and linguists, offering a comprehensive exploration of human speech. Within its vast array of recordings, one encounters a diverse spectrum of age groups, ranging from the gentle sounds of infants to the articulate discourse of adolescents and the seasoned wisdom of the elderly. Each age cohort contributes unique nuances to the dataset, enriching it with a variety of vocal qualities and linguistic features, thus highlighting the complexity of human communication.

The Mozilla Voice Dataset [20] comprises 356,921 entries sorted into three gender categories: male (44%), female (38%), and a classification labeled ”No information” (18%). This information is depicted in Figure 4.

Within this expansive dataset, age classifications span across seven distinct categories: 53% representing individuals aged between 20 to 29 years, 16% lacking specific age information, 10% falling within the range of 40 to 49 years, 8% spanning from 30 to 39 years, 7% encompassing individuals under the age of 20, 4% aged between 50 and 59 years, and 1% aged between 60 and 69 years. This distribution is visualized in Figure 5

Moreover, the Mozilla Voice Dataset places [20] equal emphasis on gender diversity, encompassing a wide range of gender expressions beyond traditional binaries. Masculine voices exude strength and determination, while feminine voices embody grace and eloquence. Additionally, non-binary and gender-nonconforming voices further enrich the dataset, challenging conventional gender norms and contributing to a more inclusive representation of human identity. Overall, the dataset serves as a valuable resource for exploring the intersection of age and gender in speech patterns, offering insights into the richness and diversity of human communication.

B. Feature Extraction

1) MFCC - Mel-Frequency Cepstral Coefficients: are adept at encapsulating key attributes within audio signals [21]. Their incorporation of both temporal and spectral data renders them invaluable for feature extraction purposes, particularly in domains like speech and music analysis. Widely utilized for extracting features, MFCCs effectively capture the distinctive qualities of vocal signals.

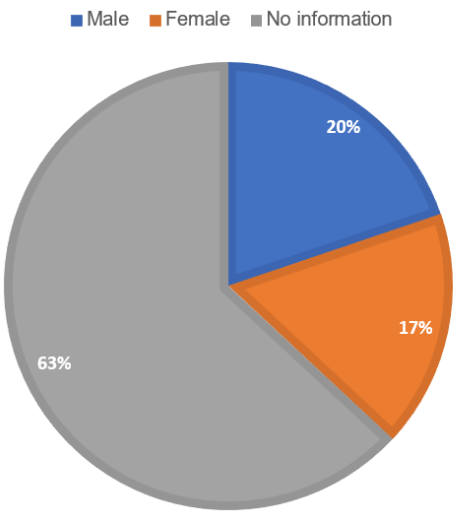

Fig. 4. Gender Dataset

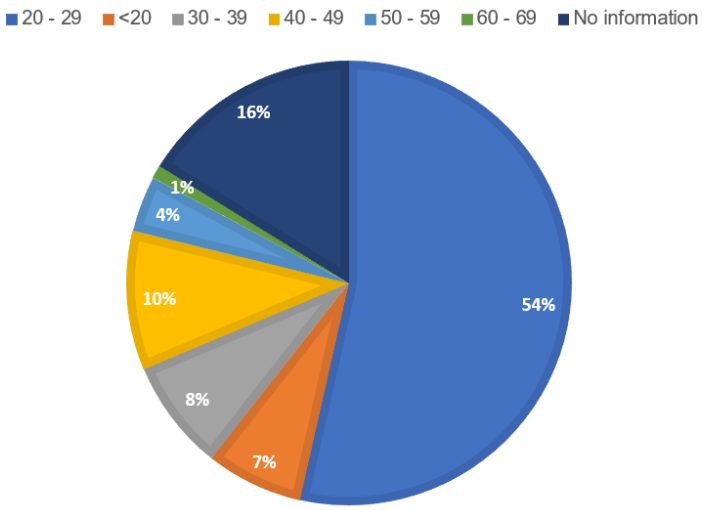

Fig. 5. Age Dataset

2) Delta MFCC: are used to represent the temporal information. One common technique allowing differentiation crossing trajectories is delta features [22]. They capture the temporal dynamics of the spectral features and can enhance the performance of machine learning models by incorporating temporal information into the feature representation.

3) Delta Delta MFCC: captures the rate of change of the delta coefficients over time, providing information about the acceleration or second-order temporal changes in the audio signal's spectral characteristics [22].

4) Shifted delta coefficients (SDC): are a variation of the standard delta coefficients used in audio signal processing. They can be used to improve language recognition [23]. Results using SDC features are comparable to using PPRLM and calculating these features is less computationally expensive. SDC features provide additional temporal features. They are created by stacking delta cepstra computed over several frames.

5) Multi-Pitch estimate: provides valuable information about the melodic content, intonation, and tonal characteristics of an audio signal [24]. It provides valuable information about the melodic content, intonation, and tonal characteristics of an audio signal.

6) Magnitude: Magnitude estimation, outlined by Stevens (1975) and further detailed by Stevens (2017) [25], stands as a foundational method in Psychophysics to assess how individuals perceive sensory stimuli. It involves participants assigning numerical values to physical stimuli based on their perceived magnitude. Renowned for its consistency, this technique is widely applicable across diverse sensory experiences such as brightness, loudness, and tactile sensation.

C. Proposed Gender Classification Model

To detect the gender of a voice sample from a speaker, the recurrent neural network model with 5 hidden layers which are made of multiple different layers of neurons that are connected in a manner that passes information from the input layer to the output layer as shown in Figure 6. The neural network's hidden layers consist of interconnected neurons that have specific roles. Each layer performs a distinct task and is made up of neurons that are activated using activation functions like ReLU or Softmax. These activation functions determine the output of the layer, allowing the network to process and transform the input data effectively.

The Rectified Linear Unit (ReLU) activation function is a type of function employed in Deep Learning. It assigns a value of 0 to negative inputs and retains positive inputs as they are. ReLU often outperforms other activation functions. In a specific model setup, after normalization, ReLU is applied in the first two LSTM hidden layers, each with 256 units and 0.3 dropout. Following these are two additional hidden layers, also utilizing ReLU, with 128 units and 0.3 dropout in one layer and 64 units and 0.3 dropout in the subsequent layer. The model concludes with a softmax activation function for gender classification.

This regularisation technique is aimed at preventing overfitting on the training dataset. It involves applying a dropout rate of 30% after each fully connected layer. Dropout randomly "drops out" a certain percentage of the neurons in the layer during training, forcing the network to learn more robust and generalized data representations.

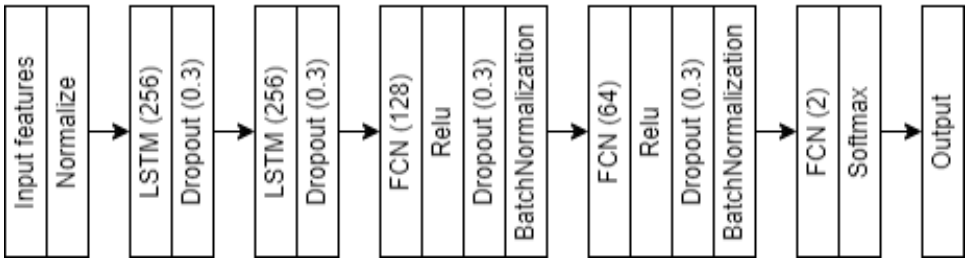

Fig. 6. Architecture of Gender Classification Model

D. Proposed Age Classification Model

Our sophisticated architecture for age classification from voice recordings starts with a thorough audio preprocessing step that makes use of the most recent audio analysis tools. The first step is crucial because it transforms voice signals into spectrograms, which are a structured format that shows patterns in the frequency and time domains.

Our custom model can handle the complex, wide frequency range that characterizes human speech with ease. The first step in the feature extraction process is onset detection, which is essential for identifying when speech sounds begin. After that, tempo analysis is used to measure the speech's rhythm and so capture each person's speech pattern.

The fundamental frequency, which is the lowest frequency of vocal sounds, is carefully examined. This research clarifies the speech's harmonic structure and provides a basis for deducing vocal characteristics associated with aging. Employing an exhaustive exploratory examination of a meticulously selected subset of a noteworthy speech dataset, we can discover the distribution of distinct age groups and their correlation with diverse extracted audio characteristics.

Every voice sample is carefully edited to a short clip to ensure consistency. For comparative analysis to be performed across the dataset, this consistency is essential.

Next, we add a multi-parameter feature set to the data. These settings turn the audio data into a format that is ideal for machine learning applications by distilling its essence. Examining the architecture in detail, the model

consists of a deep neural network with several layers, each using a selective dropout rate to prevent overfitting and a ReLU activation function to introduce non-linearity.

Our proposed architecture combines deep learning techniques with the natural harmonic and rhythmic subtleties of speech to create a powerful system for age estimation. This method not only expands on what can be done in the field of audio analytics, but it also sets the stage for further studies into voice markers associated with age Figure 7.

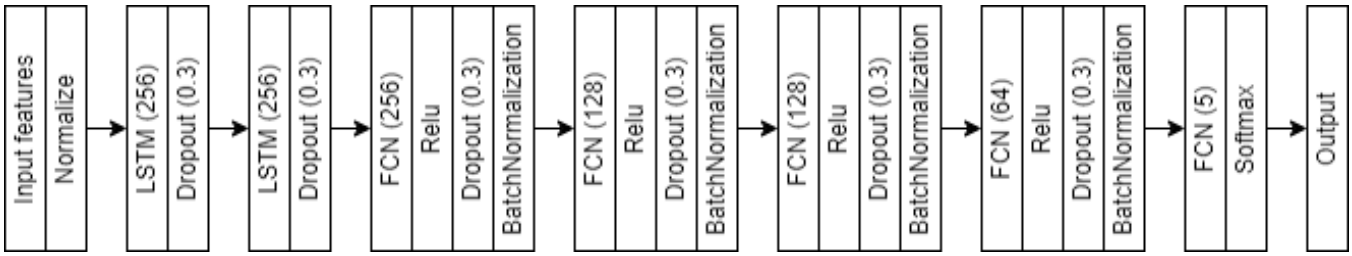

Fig. 7. Architecture of Age Classification Model

E. Performance Metrics

In this research, we utilize precision, recall, and F1-score as metrics for evaluating the performance of our classification models.

- **Precision:** Precision evaluates the percentage of accurately predicted positive instances among all instances identified as positive by the model. This metric is determined by dividing the number of true positives by the total of true positives and false positives. It is followed by Eq. 1.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

- **Recall:** Recall, known as sensitivity, evaluates the proportion of correctly predicted positive instances out of all true positive instances within the dataset. In Eq. 2, this metric is calculated by dividing the number of true positives by the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

- **F1-score:** The F1-score, depicted in Eq. 3, represents the harmonic mean of precision and recall. This metric offers a balanced assessment by considering both precision and recall, rendering it especially valuable in scenarios where the dataset exhibits imbalance.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

- **Macro average:** Macro average is used to describe a specific type of averaging technique utilized in multi-class classification tasks. It is commonly applied to compute performance metrics like precision in Eq. 4, recall in Eq. 6, or F1 score in Eq. 5, particularly in situations involving imbalanced datasets. In the case of the Macro average, the metric is computed individually for each class, and subsequently, the average is calculated across all classes.

$$\text{Macro Precision} = \frac{\text{Precision of Class 1} + \text{Precision of Class 2} + \ldots + \text{Precision of Class N}}{N} \tag{4}$$

$$\text{Macro F1-score} = \frac{2 \times (\text{Macro Precision} \times \text{Macro Recall})}{\text{Macro Precision} + \text{Macro Recall}} \tag{5}$$

$$\text{Macro Recall} = \frac{\text{Recall of Class 1} + \text{Recall of Class 2} + \ldots + \text{Recall of Class N}}{N} \tag{6}$$

- **Weighted average:** The weighted precision, recall, and F1-score functions are essential metrics in evaluating the performance of multi-class classification models. They incorporate weights assigned to each class to calculate a weighted average, reflecting the significance or contribution of individual classes to overall performance. Weighted precision assesses the accuracy of positive predictions, weighted recall measures the completeness of positive predictions, precision in Eq. 7 and balances recall in Eq. 8 and weighted F1-score is presented in Eq. 9, providing a comprehensive evaluation of model effectiveness while considering class-specific importance.

$$\text{Weighted Precision} = \frac{p_1 \times w_1 + p_2 \times w_2 + \ldots + p_N \times w_N}{\text{Total Weight}} \tag{7}$$

where p_i is the precision of class i and w_N is the weight of class i.

$$\text{Weighted Recall} = \frac{r_1 \times w_1 + r_2 \times w_2 + \ldots + r_N \times w_N}{\text{Total Weight}} \tag{8}$$

where r_N is the recall of class i and w_N is the weight of class i.

$$\text{Weighted F1 - Score} = \frac{f_1 \times w_1 + f_2 \times w_2 + \ldots + f_N \times w_N}{\text{Total Weight}} \tag{9}$$

where f_N is the F1-Score of class i and w_N is the weight of class i.

IV. Experience Result

A. Experiment settings

In this research, we use Long Short-Term Memory (LSTM) networks, recognized as a form of recurrent neural network architecture renowned for their capacity to apprehend temporal relationships in sequential data. Moreover, the application of LSTM models to predict age and gender characteristics directly from unprocessed voice signals is categorized within the realm of technical methods. The study investigates the effectiveness of these LSTM-based methodologies in discerning age and gender from voice data, presenting commendable accuracy levels in prediction.

The training process is conducted on a personal computer equipped with an Intel® Core™ i7-12650H processor, and an NVIDIA GeForce RTX 3060 GPU with 8GB of RAM.

B. Dataset

1) Dataset for Gender Classification Model: After The Mozilla Voice Dataset [20] cleansing procedures, we homogenized the dataset to solely include male and female classifications. Subsequently, we partitioned this refined dataset into subsets suitable for Gender Classification Modeling, totaling 8,000 items. This allocation comprised 5,400 entries designated for training purposes, 600 for validation, and 2,000 for testing. Notably, each subset maintained a balanced representation of genders, with male and female categories accounting for 50% and 50% of the data, respectively. It is represented through Figure 8 and in Table I.

TABLE I
DISTRIBUTION OF SAMPLES IN THE GENDER DATASET

Category	Train Dataset	Valid Dataset	Test Dataset
Male	2685	315	1000
Female	2715	285	1000
Total	5400	600	2000

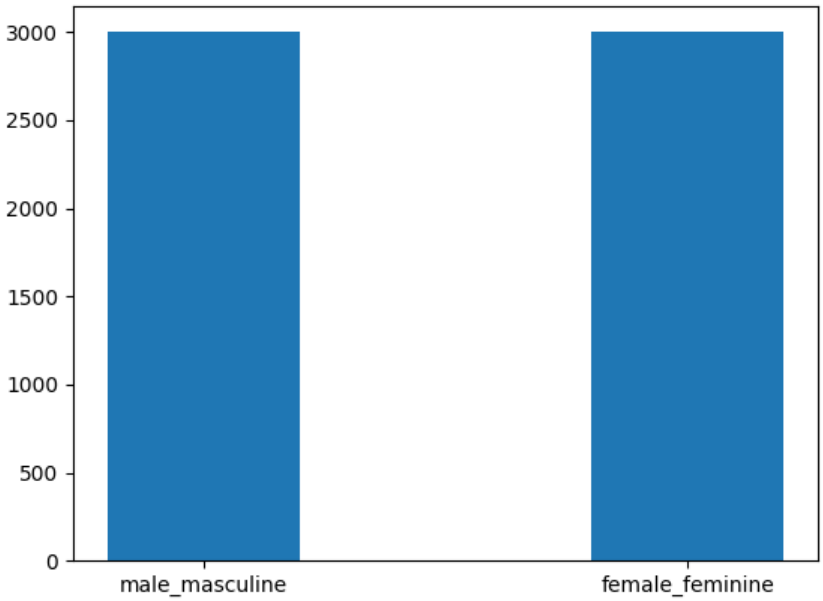

Fig. 8. Dataset for Training Gender Classification Model

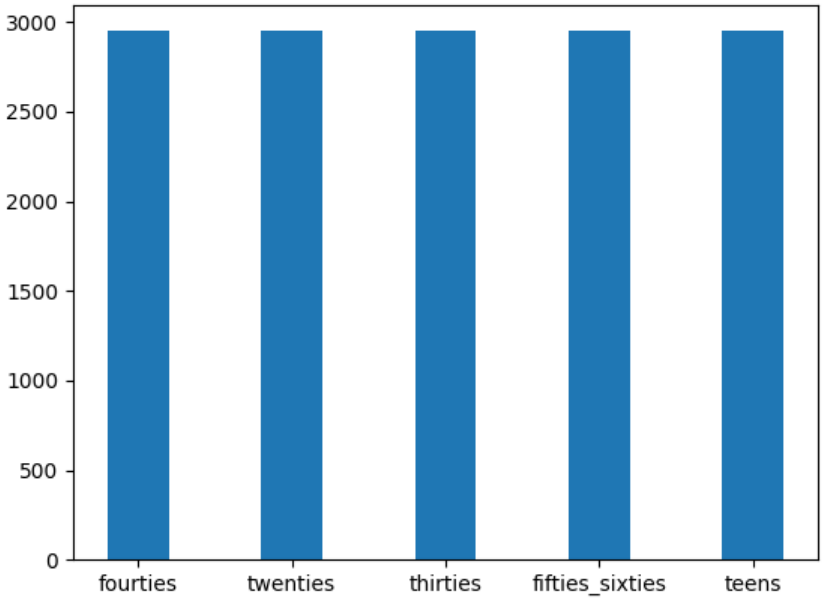

Fig. 9. Dataset for Training Age Classification Model

TABLE II
DISTRIBUTION OF SAMPLES IN THE AGE DATASET

Category	Train Dataset	Valid Dataset	Test Dataset
Teens	2658	293	158
Twenties	2655	296	158
Thirties	2647	304	158
Forties	2674	277	158
Fifties - Sixties	2645	306	158
Total	13279	1476	790

2) Dataset for Age Classification Model: Following the Mozilla Voice Dataset [20] refinement processes, we standardized the dataset to encompass only two age categories. Subsequently, we partitioned this refined dataset into subsets tailored for age-based modeling, totaling 8,000 items. This allocation included 13,279 entries designated for training purposes, 1,476 entries for validation, and 790 entries for testing. It is noteworthy that each subset retained a balanced representation across five age classifications: teens, twenties, thirties, forties, and fifties to sixties. It is represented through Figure 9, and Table I.

C. Extract Feature

The current feature extraction functionality supports a variety of features from an input audio file in Figures 10, 11, 12. These include:

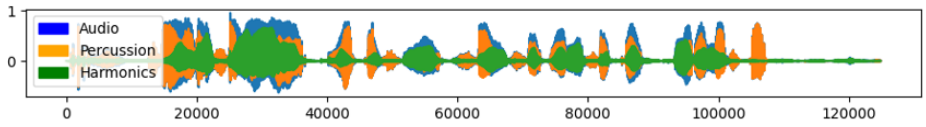

Fig. 10. Initial Loaded Audio

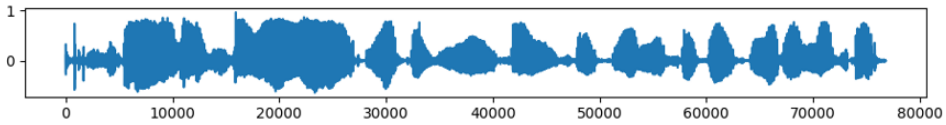

Fig. 11. Audio after split threshold

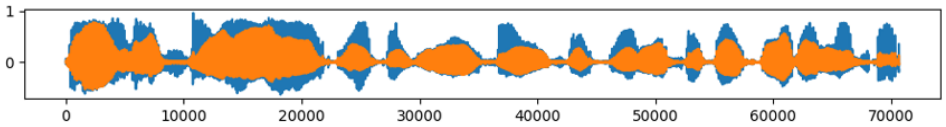

Fig. 12. Audio after trim threshold

- Mel-Frequency Cepstral Coefficients (MFCC) features, which capture the spectral characteristics of the audio signal, are illustrated in Figure 13.
- Delta MFCC features, which denote the rate of change of MFCC coefficients over time, are depicted in Figure 13.
- Delta-delta MFCC features, indicating the acceleration of MFCC coefficients, are shown in Figure 13.
- Shifted delta coefficients, providing information on the relative changes in the audio signal.

- Pitch estimate, offering insights into the fundamental frequency of the audio signal, is visualized in Figure 15.
- Magnitude estimate, representing the strength or intensity of the audio signal, is displayed in Figure 14.

This feature extraction process enables comprehensive analysis and characterization of audio signals, facilitating various applications such as speech recognition, speaker identification, and emotion detection.

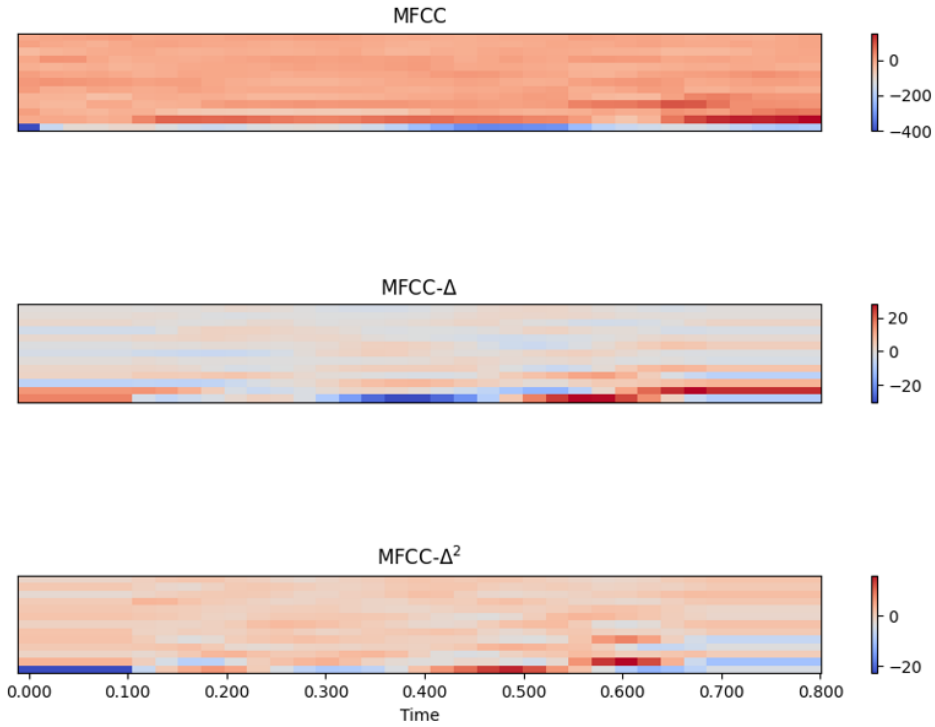

Fig. 13. MFCC and Deltas for audio

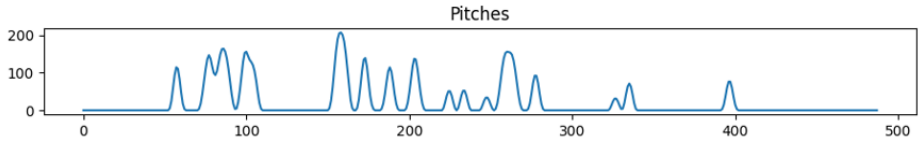

Fig. 14. Audio Magnitude

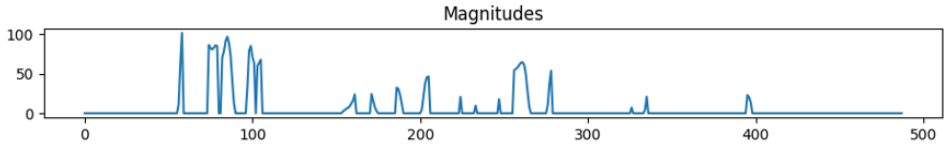

Fig. 15. Audio Pitches

D. Experience Result of Gender Classification Model

After training the Gender Classification Model for 60 epochs, spanning 5 minutes, the training process yielded promising results. The maximum accuracy achieved during training reached 92.9% in Figure 16, accompanied by a minimum loss of 0.169 in Figure 17. Meanwhile, the validation phase also exhibited notable performance, with a peak accuracy of 90% in Figure 16 and a minimum loss of 0.25 in Figure 17. These metrics indicate that the

model has successfully learned to classify gender with high accuracy and minimal loss, showcasing its efficacy in gender classification tasks.

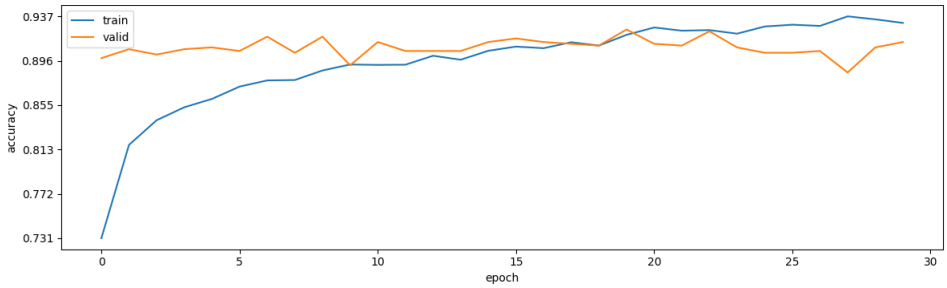

Fig. 16. Accuracy of Gender Classification Model performance

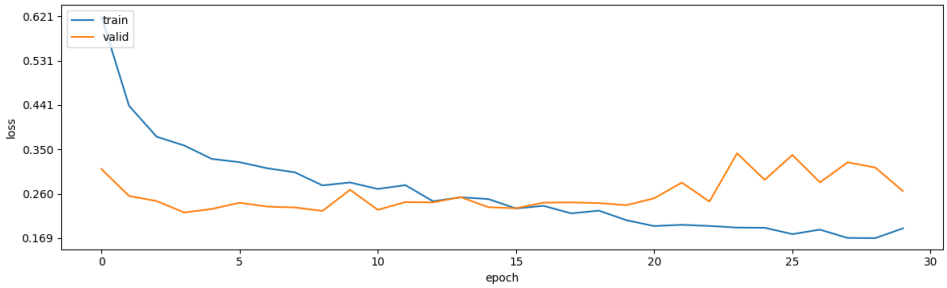

Fig. 17. Loss of Gender Classification Model performance

Afterward, we proceeded to predict using the Gender Classification Model on a test dataset, and the results obtained are as follows: Table III, and the confusion matrix normalize in Figure 18.

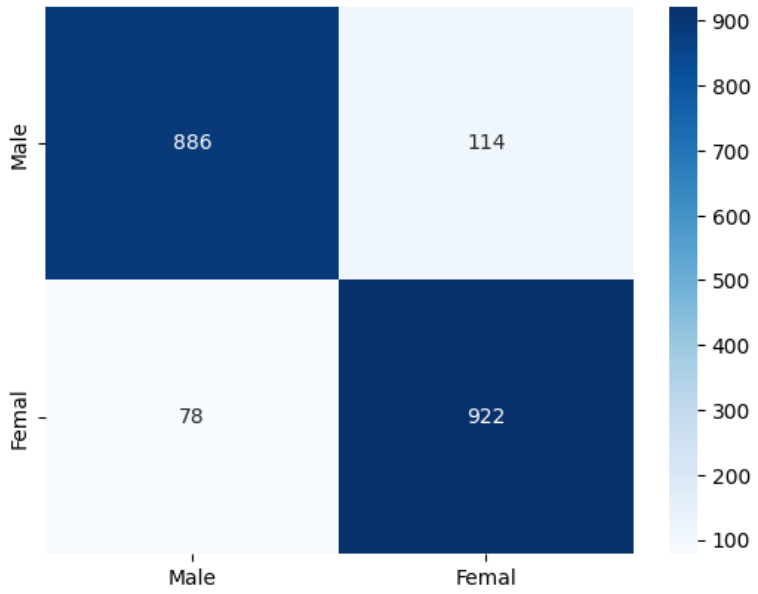

Fig. 18. Confusion Matrix for Gender Prediction

TABLE III
THE OVERALL ACCURACY OF GENDER CLASSIFICATION MODEL

Category	Precision	Recall	F1 - scored	Support
Male	0.91909	0.88600	0.90224	1000
Female	0.88996	0.92200	0.90570	1000
Accuracy			**0.90400**	**2000**
Macro average	**0.90452**	**0.90400**	**0.90397**	**2000**
Weighted average	**0.90452**	**0.90400**	**0.90397**	**2000**

E. Experience Result of Age Classification Model

Following the training of the Age Classification Model for 100 epochs over 60 minutes, notable results were obtained. The maximum accuracy achieved during training peaked at 80% in Figure 19, coupled with a minimum loss of 0.521 in Figure 20. Conversely, in the validation phase, the model demonstrated a maximum accuracy of 58.9% in Figure 19, with a minimum loss recorded at 1.1 in Figure 20. These metrics provide insights into the model's performance, indicating its ability to learn and generalize from the training data while highlighting areas for improvement, particularly in validation accuracy.

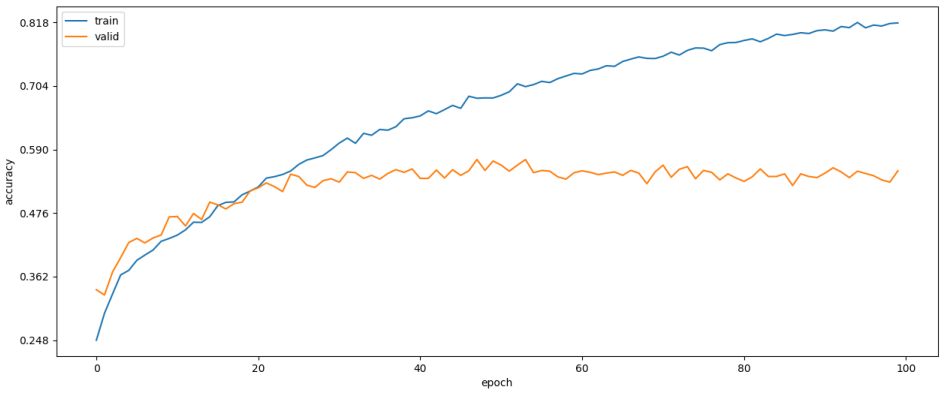

Fig. 19. Accuracy of Age Classification Model performance

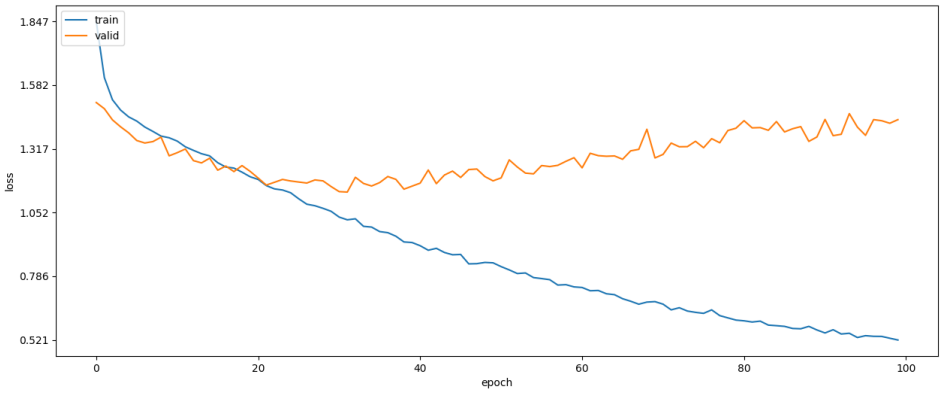

Fig. 20. Loss of Age Classification Model performance

Subsequently, we utilized the Gender Classification Model to predict on a test dataset, yielding the following results as depicted in Table IV. Additionally, the confusion matrix is illustrated in Figure 21, and the normalized confusion matrix is illustrated in Figure 22.

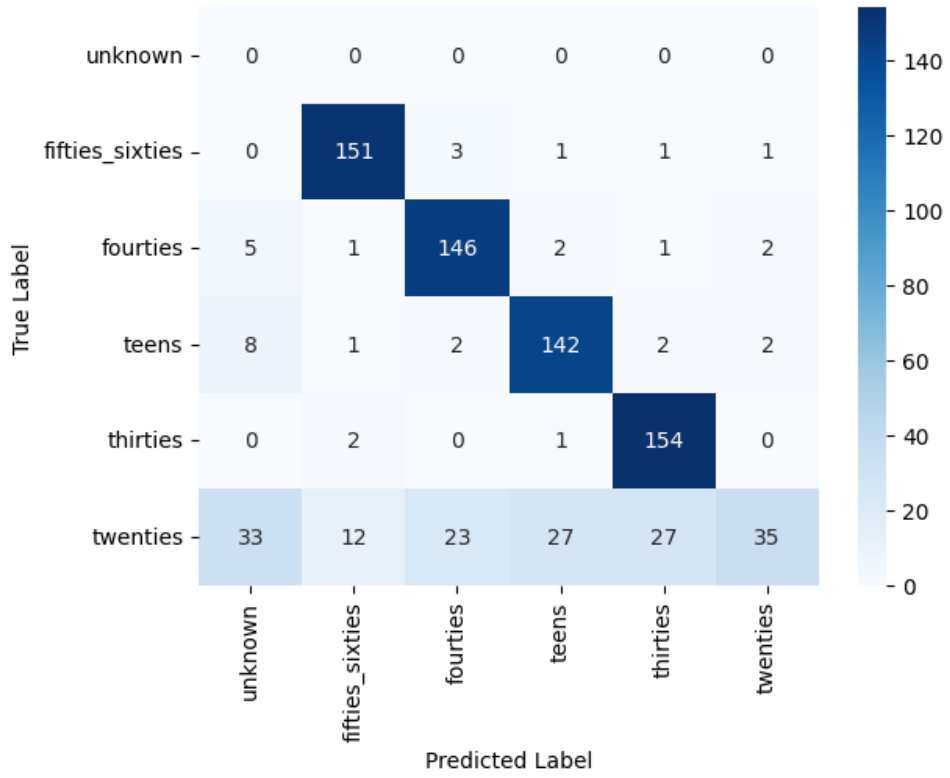

Fig. 21. Confusion Matrix for Age Prediction

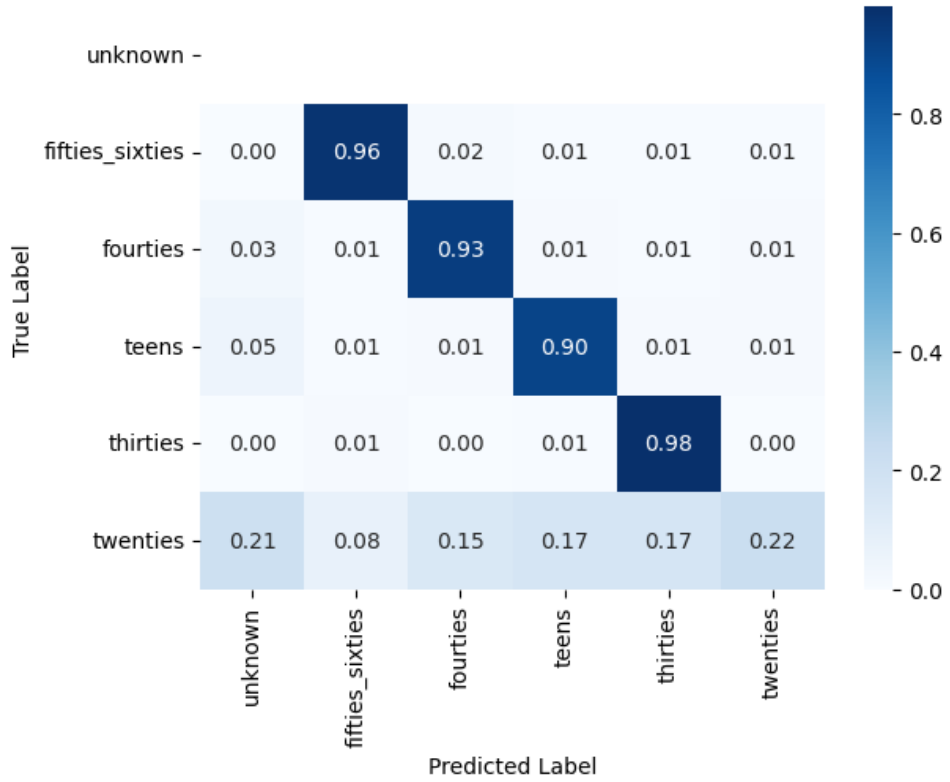

Fig. 22. Confusion Matrix Normalize for Age Prediction

TABLE IV
THE OVERALL ACCURACY OF AGE CLASSIFICATION MODEL

Category	Precision	Recall	F1 - scored	Support
Unknown	0.0	0.0	0.0	0
Fifties - Sixties	0.90419	0.96178	0.93210	157
Forties	0.83908	0.92994	0.88218	157
Teens	0.82081	0.90446	0.86061	157
Thirties	0.83243	0.98089	0.90058	157
Twenties	0.87500	0.22293	0.35533	157
Accuracy			**0.80000**	**785**
Macro average	**0.71192**	**0.66667**	**0.65513**	**785**
Weighted average	**0.85430**	**0.80000**	**0.78616**	**785**

F. Overall Experience Result

- **Gender Classification Model:** Both precision and recall are high for both classes, indicating that the model performs well in correctly identifying both males and females. The F1-score, which balances precision and recall, is also high for both classes, suggesting good overall performance. Overall, the model shows decent performance with an accuracy of 90%
- **Age Classification Model:** These metrics give an overall evaluation of the model's performance across all classes, with the weighted average giving more weight to classes with larger support. Overall, the model shows decent performance with an accuracy of 80%, but there may be room for improvement, especially in classes like "twenties" where precision, recall, and F1-score are low.

V. CONCLUSION AND FUTURE ENHANCEMENTS

The Voice Based Age and Gender Detection using Long Short-Term Memory has shown promising results in accurately identifying the gender and age of a speaker. The system utilizes a sequential model, which has been effective in predicting gender. Furthermore, by employing the grid search model, the system has achieved successful age prediction. In summary, the Voice Recognition System has demonstrated encouraging outcomes in both gender and age identification tasks. While the Voice Recognition System has exhibited promising results in age prediction, it does have some limitations. Specifically, the model struggled to accurately predict the age of voices with different accents. However, with further enhancements and improvements to the system, it has the potential to be utilized in a variety of applications that rely on voice-based age prediction. These improvements could address the limitations and enhance the system's accuracy and applicability in various scenarios.

REFERENCES

[1] X. Qiu, Z. Du, and X. Sun, "Artificial intelligence-based security authentication: Applications in wireless multimedia networks," *Ieee Access*, vol. 7, pp. 172 004–172 011, 2019.

[2] S. Ioffe and C. Szegedy, "International conference on machine learning," in *International conference on machine learning*, 2015, pp. 448–456.

[3] A. Umesh and R. Patole, "Automatic recognition identifying speaker emotion and speaker age classification using voice signal."

[4] A. Maedche, C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner, "Ai-based digital assistants: Opportunities, threats, and research perspectives," *Business & Information Systems Engineering*, vol. 61, pp. 535–544, 2019.

[5] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[6] G. Fairbanks and W. Pronovost, "An experimental study of the pitch characteristics of the voice during the expression of emotion," *Communications Monographs*, vol. 6, no. 1, pp. 87–104, 1939.

[7] W. Endres, W. Bambach, and G. Flösser, "Voice spectrograms as a function of age, voice disguise, and voice imitation," *The Journal of the Acoustical Society of America*, vol. 49, no. 6B, pp. 1842–1848, 1971.

[8] P. Jafarian and M. Sanaye-Pasand, "A traveling-wave-based protection technique using wavelet/pca analysis," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 588–599, 2010.

[9] S. E. Linville, "Source characteristics of aged voice assessed from long-term average spectra," *Journal of Voice*, vol. 16, no. 4, pp. 472–479, 2002.

[10] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*. Ieee, 2017, pp. 1–6.

[11] G. Son, S. Kwon, and N. Park, "Gender classification based on the non-lexical cues of emergency calls with recurrent neural networks (rnn)," *Symmetry*, vol. 11, no. 4, p. 525, 2019.

[12] M. A. Uddin, R. K. Pathan, M. S. Hossain, and M. Biswas, "Gender and region detection from human voice using the three-layer feature extraction method with 1d cnn," *Journal of Information and Telecommunication*, vol. 6, no. 1, pp. 27–42, 2022.
[13] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
[14] S. Mavaddati, "Voice-based age and gender recognition using training generative sparse model," *International Journal of Engineering*, vol. 31, no. 9, pp. 1529–1535, 2018.
[15] M. Abdollahi, E. Valavi, and H. A. Noubari, "Voice-based gender identification via multiresolution frame classification of spectro-temporal maps," in *2009 International Joint Conference on Neural Networks*. IEEE, 2009, pp. 1–4.
[16] N. S. Chaitanya, P. Shivani, N. Sahithi, M. Sravanthi, and J. Aditya, "Analyzing vocal patterns to determine gender, age and emotion," in *Proceedings of International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2020*. Springer, 2021, pp. 349–357.
[17] V. S. Kone, A. Anagal, S. Anegundi, P. Jadhav, U. Kulkarni, and S. Meena, "Voice-based gender and age recognition system," in *2023 International conference on advancement in computation & computer technologies (InCACCT)*. IEEE, 2023, pp. 74–80.
[18] S. Kumar, R. Mishra, and B. Tyagi, "Classifying human gender by learning the acoustic features of voice samples," *NeuroQuantology*, vol. 20, no. 7, p. 1600, 2022.
[19] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.
[20] "Mozilla voice dataset."
[21] S. G. Koolagudi, D. Rastogi, and K. S. Rao, "Identification of language using mel-frequency cepstral coefficients (mfcc)," *Procedia Engineering*, vol. 38, pp. 3391–3398, 2012.
[22] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for mfcc feature extraction," in *2010 4Th international conference on signal processing and communication systems*. IEEE, 2010, pp. 1–5.
[23] S. Fernando, V. Sethu, and E. Ambikairajah, "Eigenfeatures: An alternative to shifted delta coefficients for language identification," in *Proc. 16th Speech Science and Technology Conference (SST 2016)*, 2016, pp. 253–256.
[24] M. Christensen and A. Jakobsson, *Multi-pitch estimation*. Springer Nature, 2022.
[25] S. S. Stevens, *Psychophysics: Introduction to its perceptual, neural and social prospects*. Routledge, 2017.