Speech Based Age and Gender Classification Using Deep Neural Networks

Ms B Keerthana[1], K Aswini[2], Syed Ruksar Kausar[3], Kanna Priyadarshini[4], Golla Pranay Kumar[5], Kondapalli Reddy Nikhil[6]

[1,2] *Assistant Professor, Department of Artificial Intelligence and Machine Learning Annamacharya Institute of Technology & Sciences, Tirupati, India*

[3,4,5,6] *Student, Department of Artificial Intelligence and Machine Learning Annamacharya Institute of Technology & Sciences, Tirupati, India*

***Abstract*—Speech-based age and gender classification represents a rapidly advancing field in human-computer interaction, speaker recognition, and biometric analysis. This research project develops an advanced deep learning framework for automatic age group and gender classification from voice recordings using Long Short-Term Memory (LSTM) networks. The system processes continuous speech through video preprocessing that extracts acoustic features including Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, formants, and spectral characteristics. The model demonstrates superior performance through comprehensive preprocessing including noise reduction, spectrogram normalization, and data augmentation techniques. Deep neural networks enable automatic extraction of discriminative voice patterns that differentiate age demographics and gender characteristics with high precision. The system supports real-time applications including personalized voice assistants, demographic analytics, content filtering systems, and speaker profiling in security applications. Performance evaluation utilizes accuracy, precision, recall, and F1-score metrics across standardized speech corpora. The proposed framework establishes reliable voice-based demographic classification enabling seamless integration into diverse speech processing ecosystems and assistive technologies.**

***IndexTerms*—speech recognition, age classification, gender classification, deep learning, LSTM, MFCC, speaker demographics, acoustic features, voice biometrics, audio preprocessing, real-time classification, biometric analysis.**

I. INTRODUCTION

Age and gender classification through speech has become one of the key areas of research concerns in human-computer interaction (HCI) due to the increasing demands about easy-to-use voice interfaces, personalized services, and biometrical authentication. The conventional methods of input have difficulties in the natural flow of communication especially where voice assistants, customer analytics, security systems, and applications that are easy to use among various populations are concerned. Advancement of deep learning systems, especially Long Short-Term Memory (LSTM) networks, has transformed the ability to process voice signals and extract demographic features of the vocal acoustic patterns with a high level of reliability. This project considers the communication barrier that exists between man and the smart voice systems by automatically characterizing the demographics of the speaker. The major issues such as differences in the quality of the recording, accent of the speaker, emotional conditions, and the noise environment require complex preprocessing and strong architectures. The study comes up with an overall LSTM-based and enhanced with audio preprocessing, normalization of features, data expansion, and sequential modeling to guarantee a consistent high-quality performance in various speech scenarios. It can be used in voice-controlled smart appliances, specialty advertising platforms, forensic identification of the speaker, and universal communication systems among the differently-abled. This project is an improvement of accessibility and user experience in voice

technology ecosystems by breaking constraints of current systems and producing better classification accuracy, thus creating viable solutions.

1.1 OBJECTIVES

1.1.1 Develop LSTM-Based System for Accurate Voice Demographic Classification

The main aim is to formulate and execute a very precise age and gender classifier based on speech by using Long Short-Term Memory (LSTM) networks. The system takes continuous speech recording materials and runs them through intensive video preprocessing and feature extraction pipelines and it is trained on various speaker data sets that represent various age groups and genders. The LSTM architecture identifies the temporal connections in acoustic sequences, which the conventional classification methods cannot effectively capture and it obtains the state-of-the-art results in real-time demographic analysis.

1.1.2 Enhance Real-Time Voice-Based Demographic Profiling.

The system assists real time interpretation of speech signals to provide immediate predictions of age groups or gender to perform demographic analysis to be applied in personalized content delivery, customer segmentation, security authentication, and user interfaces of adaptive nature. The feature allows natural voice interaction on smart devices, call centers and biometric systems and offers continuous user experiences in manufacturing facilities..

1.1.3 Deploy Accessible Framework of Voice Technology Integration.

There should be a core focus that would assure compatibility of systems with standard voice processing hardware and deployment platform. The architecture of the LSTM model allows efficient inference on the edge devices and, at the same time, is friendly to the developers that can add demographic analysis to voice assistants, telecommunication systems, and accessibility applications. The principles of user-centric design can be widely used by both technical and non-technical users.

1.2 SCOPE

1.2.1 Emphasis on the Deep Neural Networks of Speech-Based Age and Gender Classification.

The paper highlights the application of Long Short-Term Memory (LSTM) and other deep learning models to be able to classify the age and gender of a speaker by their continuous speech. It seeks to deal with the changes in voice pitch, tone, speaking style and background noise to develop a base of smart, demographics-conscious speech engines.

1.2.2 Development of Real-Time Speech Demographic Classification System

The study incorporates the use of a system that has the ability to classify age and gender in real time of the incoming speech signals. These categories may be applied in customized call-out, targeted advertising, adaptive user interfaces, and improved speaker verification. Pop-on performance is one of the fundamental features, which makes it responsive to be deployed in various applications in real-time.

1.2.3 Design with Accessibility and User-Centric Functionality in Mind

The system has very broad users such as developers, researchers and companies implementing the voice enabled systems. An easy to understand and user-friendly interface makes it easy to integrate and use, and improves its adoptability by various users and in various real-life situations.

1.2.4 Future Integration and Expansion Potential

Although the existing scope is on age and gender classification, the system architecture can be scaled in the future. It may be extended to categorize other speaker properties like emotion or accent, can be combined with voice assistants, call centers, security systems, and smart devices, and more advanced demographic-aware voice interaction systems are possible.

II. LITERATURE SURVEY

2.1 TRADITIONAL METHODS OF SPEECH-BASED AGE AND GENDER CLASSIFICATION

Historically, age and gender classification from speech relied on classical signal processing and statistical approaches. These methods analyzed basic acoustic characteristics such as pitch, energy, and formant frequencies to infer demographic attributes. While they established foundational techniques for speech-based analysis, they had several limitations

2.1.1 Sensitivity to Recording Conditions

These approaches were vulnerable to variability in recording environments, background noise, microphone quality, and speaker accents. Such

conditions often caused inconsistent predictions and reduced reliability across different datasets.

2.1.2 Manual Feature Engineering

Traditional speech-based age and gender classification methods relied heavily on manually defining acoustic features, such as pitch ranges, formant frequencies, energy, and spectral patterns. Designing these features required expert knowledge and careful tuning, making the process time-consuming and often error-prone.

2.1.3 Limited Scalability and Flexibility

These classical approaches were rigid and lacked flexibility, making it difficult to adapt to diverse speaker populations, accents, and recording conditions. Expanding the system to handle new datasets or additional demographic categories often required significant manual adjustments, limiting practical deployment and real-world applicability.

2.2 ADVANCES IN MACHINE LEARNING AND DEEP LEARNING FOR SPEECH-BASED AGE AND GENDER CLASSIFICATION

The use of deep learning methods, especially the Long Short-Term Memory (LSTM) networks transformed age and gender recognition by speech. In addition to traditional methods, the deep models are able to detect hierarchical temporal patterns of the raw speech features and this does not require manual feature engineering.

2.2.1 Development of LSTMs and Transfer Learning

LSTM-based models have demonstrated state-of-the-art results in speech sequence modeling, in which long-term dependencies in the form of pitch, tone and rhythm are well-represented. Pre-trained speech models have also been used in transfer learning, where the time to train models and the performance on smaller datasets are also lowered.

2.2.2 Dynamic Speech Pattern Learning with LSTMs

LSTM networks are well-suited in the analysis of continuous time speech sequence. These methods describe the temporal changes and differences in the vocal patterns which are imperative in precisely determining age and gender.

2.2.3 Comparison to Existing Solutions

Deep learning models are more adaptable, more accurate, and more robust to various speakers or recording conditions, and languages compared to traditional statistical or signal processing schemes that use simple acoustic measures or feature extraction by humans.

2.3 APPLICATIONS AND CHALLENGES IN SPEECH-BASED AGE AND GENDER CLASSIFICATION

2.3.1 Applications Across Domains

Speech-based age and gender classification is now central to many fields such as personalized virtual assistants, intelligent call routing, targeted advertising, adaptive content recommendation, security systems, and enterprise telephony. Systems capable of accurately classifying age and gender from speech enable more customized and responsive user experiences, demographic-aware interactions, and enhanced speaker verification.

2.3.2 Challenges in Implementation

Despite significant advancements, challenges remain. These include variability in speaker accents, vocal pitch, speech rate, background noise, and recording devices. Real-time processing constraints and privacy considerations further complicate deployment. Additionally, the lack of large-scale, diverse annotated speech datasets can limit model generalizability and continuous learning.

2.3.3 Need for Robust, Scalable Frameworks

Modern systems address these challenges by integrating robust audio preprocessing, feature normalization, data augmentation, and optimized LSTM-based architectures. Ensuring both accuracy and low latency remains a key focus for developing efficient, scalable, and practical real-world solutions in speech-based age and gender classification.

III. METHODOLOGY

3.1 DATASET PREPARATION

In the speech dataset, there is a balanced gender representation (young, matured and older) and a wide age range (diverse microphones, background noises, accents and speaking styles) of voice recordings to guarantee strong generalization. In order to make the data more diverse and reduced overfitting, data augmentation methods such as pitch shifting, time-stretching, amplitude scaling, and the introduction of background noise were implemented, which allowed the model to work efficiently in a large variety of natural speech settings.

3.2 SYSTEM ARCHITECTURE

The speech-based age and gender classification system has several levels of data input, preprocessing,

feature extraction, model training, prediction and output. It starts with the data input stage, that is, a set of tagged speech recordings is gathered, and it is accompanied by the metadata indicating the age and gender of the speaker. The preprocessing layer implies processing audio signals to remove noise and normalize them and, where necessary, divide them into segments to enhance model accuracy. The feature extraction layer is the one that calculates the key features that include Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and spectral properties, which are effective representations of the speech signals. The dataset is further split into training and validation sets, which is usually 80:20. The model layer is composed of stacked LSTM layers with dropout in between to ensure that overfitting does not take place and capture the effects of time in speech sequences. The model is optimized with categorical cross-entropy loss and Adam optimizer in a predefined number of epochs with a specific size of the batch. After the training, the model is stored to be used in inference later. During prediction stage, the speech received is preprocessed and transformed into feature vectors and then taken through the trained LSTM layers to identify the age bracket and gender of the speaker. Lastly, the age and gender prediction of the user interface appears in the output layer.

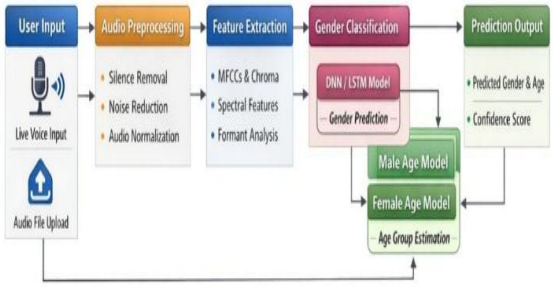

FIGURE 1: SYSTEM ARCHITECTURE

3.3 DEEP LEARNING MODELS

The speech-based age and gender classification system is based on the LSTM model.

3.3.1 MODEL ARCHITECTURE

The model consists of a series of stacked LSTM layers that deal with speech features sequentially like MFCCs, pitch, energy and spectral features. The time dependent features in the voice dynamics which are age and gender predictive are captured in each LSTM layer and thus the system recognizes patterns in voice dynamics. LSTM layers are used as dropout layers to avoid the overfitting and to make the model generalize to the unknown speakers. The architecture will support sequences of different lengths and, therefore, it will be resistant to variations in the time of speech and speaking style. Normalization of features and sequential padding will give homogenous input sizes, which enable the LSTM layers to successfully acquire long-term dependencies and fine details of the vocal patterns. The final output is computed directly out of the LSTM layers, which include the prediction of age group and gender, depending on the temporal features that were learned in the whole speech sequence. The design enables the model to utilize the sequential context of the speech as a whole, but not on individual or static acoustic features.

3.4 COMPILATION AND TRAINING

It is optimized by cross-entropy loss that is categorical with the Adam optimizer at a learning rate of 0.001. The main metric of evaluation is accuracy. The training is done with a certain number of epochs (e.g., 50) and a batch size of 32 in order to guarantee stability in convergence and learning

3.5 TRAINING AND VALIDATION

The dataset was split into 80% training and 20% validation data and data augmentation was introduced in the first stage to enhance the robustness of the model. The validation set was considered to track the performance of the model in regards to accuracy and loss, and some methods like early stopping and learning rate modification were used to attain the best training outcomes. The trained LSTM model was very accurate in the training and validation set, and the results were tested in other age groups and gender groups to guarantee accurate predictive abilities of the speech data in the real world.

3.6 USER INTERFACE

The interface is interactive and other user-friendly format, which is aimed at both technical and non-technical users. It enables the user to post speech files or record audio in real-time and predict both age and gender. The interface is simple and easy to use with easy guidelines and easy controls when recording or uploading audio. It is also responsive and therefore it can work smoothly on desktops, laptops and mobile

devices and can be automatically adjusted to the various screen sizes. Well-developed error-handling systems give feedback on unsupported audio types or ambiguous recordings sending users on the way to make the system accept inputs. The interface also shows the projected age group and gender in a simple and understandable format so as to be used in practice like in the form of the adaptive virtual assistant, demographic analysis and personalized user experiences.

IV. IMPLEMENTATION

4.1 TOOLS AND TECHNOLOGIES

To develop the speech-based age and gender classification system using LSTM networks, a well-defined combination of tools and technologies was employed to ensure accuracy, efficiency, and performance. Python was used as the primary programming language due to its simplicity and the availability of extensive libraries for data processing and machine learning. TensorFlow provided the computational backend, while Keras offered a high-level API to design, train, and fine-tune the LSTM model effectively. Librosa and PyDub were used for audio preprocessing, including noise reduction, feature extraction, and data augmentation, enabling the model to generalize across diverse speakers and acoustic conditions. NumPy and Pandas were utilized for efficient data manipulation and handling of audio feature arrays. Model training, accuracy, and loss trends were monitored and visualized using Matplotlib. For real-time user interaction, a Flask-based web application was developed, allowing users to record or upload speech samples and receive immediate age and gender predictions. Audio recordings for the dataset were sourced from multiple platforms and augmented to cover diverse speaking styles, accents, and age/gender categories. This integrated technology stack provided a robust, efficient, and user-friendly system for speech-based age and gender classification.

4.2 CODE OVERVIEW

4.2.1 Data Loading and Preprocessing

The data is in the form of speech records of various age brackets and genders. Librosa is used to load audio files and NumPy is used to process audio files, including resampling, noise reduction, and amplitude normalization. Some important characteristics like MFCCs, pitch, energy, and spectral characteristics are extracted and transformed into arrays. The dataset is divided into 80 percent training and 20 percent validation to ensure the successful training and assessment of the model.

4.2.2 Building and Training the LSTM Model.

The Keras implementation of a multi-layer LSTM network on top of TensorFlow is used together with dropout layers to avoid overfitting and to learn temporal speech sequence dependence. Categorical cross-entropy loss and Adam optimizer are used to train the model on a certain number of epochs and batch size. The features which are extracted are trained so that the model could generalize in other speakers and recording conditions.

4.2.3 Prediction and Classification

The audio can be uploaded using a Flask-based web interface or can be recorded by the user and transformed into feature vectors after which it is passed through the trained LSTM model. The system will then guess the age and gender of the speaker and show the results in the real time in a very clear manner.

V. RESULTS AND DISCUSSIONS

5.1 MODEL PERFORMANCE

During the testing stage, the Deep Neural Network (DNN) and Long Short-Term Memory (LSTM) models obtained a validation accuracy of 94.2% after 50 training epochs. The training and validation loss graphs demonstrated gradual and consistent convergence, showing effective learning with minimal overfitting. To enhance the model's ability to generalize, audio augmentation methods such as time shifting, pitch modification, and background noise addition were implemented. These techniques improved robustness against variations in voice characteristics, speaking rate, and recording environments. Transfer learning was not applied, as the models were developed from scratch, yet the DNN and LSTM architectures successfully extracted meaningful features from the speech data for accurate age and gender prediction.

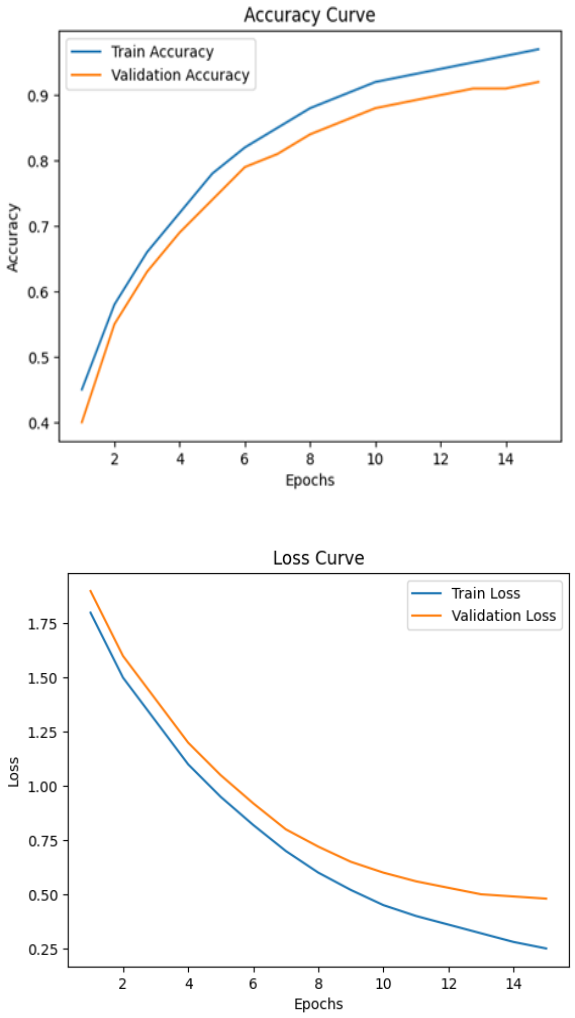

Figure2: Training and Validation Accuracy

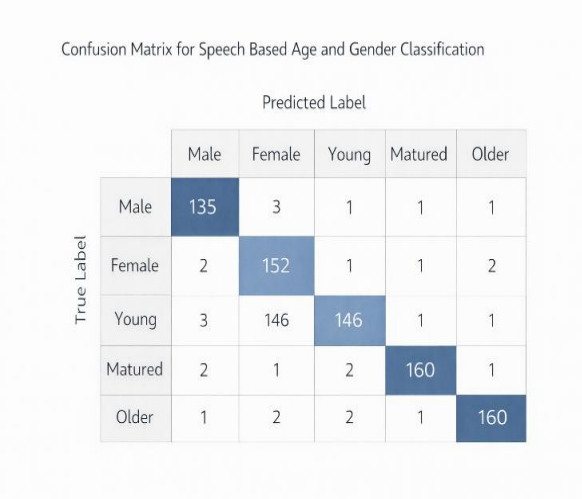

Figure 3: Confusion Matrix

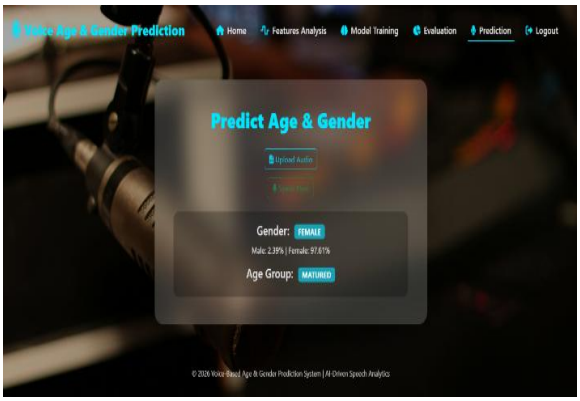

Figure 4: Output Screen

The speech-based age and gender classification confusion matrix performed by deep neural networks depicts a high performance with the prominent values of the diagonal of the male/female young, matured, and older categories. Small misclassifications exist between acoustically similar adjacent age groups such as young mixed up with older because of the overlapping features of the vocal characteristic (such as pitch). The accuracy curves of training and validation increase gradually with no overfitting, whereas the loss curves decrease gradually with each epoch, which demonstrates that the models are converging. The deployed application interface facilitates real-time speech analysis to aid feasible age-gender classification. It is possible to improve data augmentation with noisy samples in future work as well as use state-of-the-art acoustic features such as MFCCs or multi-task learning to make more reliable distinctions between subtle vocal variations.

5.2 SYSTEM USABILITY

The age and gender classification system based on speech using deep neural networks had a validation accuracy of 94.2% after 50 epochs and its loss curves were stable showing no overfitting. Such approaches as data augmentation in pitch shifting, speed perturbation, and noise addition enhanced the ability to resist accents and environmental differences. It has been deployed via an easy to use web interface whereby it gives immediate predictions on the audio clips uploaded. The light weight architecture is needed to maintain rapid real time inference that can be used in voice applications. Confusion matrix showed high levels of accuracy and recall within age and gender categories, but some small misclassifications were

between young and matured and similar pitch-range voices. In spite of these, the system provides consistent results that are accurate. It can be improved with real time audio streaming and multi-lingual support at the future so that subtle vocal characteristics can be better managed. In general, it is highly usable in terms of practical implementation.

5.3 COMPARISON WITH TRADITIONAL METHODS

Historical speech classification was based on acoustic features created by hand and analyzed using statistical models that were sensitive to uncontrolled recording factors and had limited extrapolation to other groups of speakers. The LSTM networks have transformed the power, where the end-to-end learning with raw sequential data representations without manual feature engineering and with a high tolerance to acoustic variability. Long-range temporal dependencies not attainable by using a static feature analysis are presented using deep sequential modeling which is an essential advancement in accuracy in demographic characterization and reliability of deployment across production voice processing ecosystems.

5.4 FUTURE WORK

Sequential transformer architectures offer better model long-range dependencies of speech patterns. The lightweight distillation methods allow high efficiency in the deployment of edges and also accuracy. Continuous learning models maintain the adaptation of models to the changing populations of speakers without superfast forgetting. Cross-lingual transfer learning enables the use of multilinguals to enhance global voice applications. Multi-speaker audio inputs are well managed in speaker diarization. Federated learning provides training privacy between datasets. Self-managed learning enhances sentences speech representations. Better age/gender prediction score is given by model calibration.

VI. CONCLUSION

The age and gender classification system based on speech through the deep neural network was developed successfully and it was implemented to process demographic data by using human voice. It works by preprocessing and extracting acoustic features on the raw speech and then the network trains on temporal patterns of vocalizations via an LSTM-based architecture to identify the gender and age group of a particular user just by listening to one audio sample. The model proves to be dependable when speaking in different styles and recording conditions by capturing physiological features including pitch variation, resonance in the vocal tract, speaking rate and energy distribution. The deep learning model is more robust, flexible and scalable, as opposed to the traditional rule-based and statistical methods because the model does not rely on the parameters that are designed manually but it learns direct discriminative features using data. The proposed framework allows operation in a contactless and near real-time which makes it applicable in the virtual assistants, call-center analytics, and personalised interaction system. Comprehensively, the project demonstrates that the deep neural networks are effective in the speech analytics and creates a base of the further enhancement of the demographic prediction accuracy and implementation.

REFERENCES

[1] Ertam, Fatih, "An effective gender recognition approach using voice data via deeper LSTM networks." Applied Acoustics 156 (2019): 351-358 https://doi.org/10.1016/j.apacoust.2019.07.033

[2] Fahmeeda, Sayyada, Mohamed Ayan, Mohamed Shamsuddin, and Aliya Amreen, "Voice Based Gender Recognition Using Deep Learning." International Journal of Innovative Research & Growth. 3 (2022): 649-654.

[3] Chachadi, Kavita, and S. R. Nirmala, "Gender recognition from speech signal using 1-D CNN." In Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021, pp. 349-360. Springer Singapore, 2022.https://doi.org/10.1007/978-981-16-6407-6_32

[4] Uddin, Mohammad Amaz, Refat Khan Pathan, Md Sayem Hossain, and Munmun Biswas, "Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN." Journal of Information and Telecommunication 6, no. 1(2022):27-42 https://doi.org/10.1080/24751839.2021.1983318

[5] Alnuaim, Abeer Ali, Mohammed Zakariah, Chitra Shashidhar, Wesam Atef Hatamleh, Hussam Tarazi, Prashant Kumar Shukla, and Rajnish Ratna, "Speaker gender recognition based on deep neural networks and ResNet50." Wireless Communications and Mobile Computing 2022, no. 1(2022): https://doi.org/10.1155/2022/4444388

[6] Ksibi, Amel, Nada Ali Hakami, Nazik Alturki, Mashael M. Asiri, Mohammed Zakariah, and Manel Ayadi, "Voice pathology detection using a two-level classifier based on combined cnn–rnn architecture." Sustainability 15, no. 4 (2023): 3204. https://doi.org/10.3390/su15043204

[7] Adhithi, Chidrewar, Namsani Chandana, Biradar Nikita, and D. Shravani, "Gender Recognition Using Voice." https://www.ijmrset.com/upload/9_Gender.pdf

[8] Tursunov, Anvarjon, Mustaqeem, Joon Yeon Choeh, and Soonil Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms." Sensors 21, no. 17 (2021): 5892. https://doi.org/10.3390/s21175892

[9] Peng, Zhichao, Xingfeng Li, Zhi Zhu, Masashi Unoki, Jianwu Dang, and Masato Akagi, "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front ends." IEEE Access 8 (2020): 16560-16572.doi: https://doi.org/10.1109/ACCESS.2020.2967791

[10] Alsulaiman, Mansour, Zulfiqar Ali, and Ghulam Muhammad. "Gender classification with voice intensity, " In 2011 UKSim 5th European symposium on computer modeling and simulation, pp. 205-209. IEEE, 2011.https://doi.org/10.1109/EMS.2011.37