IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204054 |

Exploring the Challenges and Advancements in Male and Female Speech Recognition: A Comprehensive Review

Akanksha[1], Sumit Dalal[2], Rohini Sharma[3]

Student, Dept. of ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India[1]

Assistant Professor, Dept. of ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India[2]

Assistant Professor and Corresponding Author, GPGCW, Rohtak, Haryana, India[3]

ABSTRACT: Speech recognition technology has witnessed significant advancements in recent years, yet it still encounters challenges, particularly in accurately distinguishing between male and female voices. This review examines the current state of male and female speech recognition systems, highlighting the underlying complexities and exploring the methodologies employed to address them. We delve into the physiological and sociolinguistic factors influencing speech production and how they impact recognition accuracy. Additionally, we discuss the role of machine learning algorithms, including deep neural networks, in enhancing gender classification in speech recognition systems. Furthermore, we analyze the implications of gender bias and its mitigation strategies within speech recognition technologies. By synthesizing recent research findings, this review offers insights into the progress made and the avenues for future advancements in male and female speech recognition.

KEYWORDS: Speech recognition, male speech, female speech, gender classification, machine learning, deep neural networks, gender bias

I. INTRODUCTION

Speech recognition technology has become an integral part of our daily lives, revolutionizing how we interact with devices and systems. From virtual assistants to automated customer service systems, speech recognition enables seamless communication between humans and machines. However, despite significant progress, accurately distinguishing between male and female voices remains a persistent challenge in speech recognition systems. The ability to correctly identify the gender of the speaker is crucial for personalized user experiences and tailored responses. Moreover, gender recognition plays a vital role in various applications, including forensic analysis, gender-specific marketing, and voice-controlled devices [1].

Male and female speech exhibit distinct characteristics stemming from physiological and sociolinguistic differences. These differences pose unique challenges for speech recognition algorithms, which must effectively capture and interpret gender-specific features to achieve accurate classification. Physiologically, variations in vocal tract length, fundamental frequency (pitch), and resonance contribute to differences between male and female voices. Sociolinguistic factors such as intonation, pitch range, and speaking style further differentiate male and female speech patterns.

This review aims to provide a comprehensive overview of the current state of male and female speech recognition, highlighting the underlying complexities and recent advancements in the field. We explore the physiological and sociolinguistic factors influencing male and female speech production, examine the methodologies employed for gender classification in speech recognition systems, and discuss the role of machine learning algorithms, particularly deep neural networks, in enhancing gender recognition accuracy. Furthermore, we address the implications of gender bias in speech recognition technologies and discuss potential mitigation strategies.

By synthesizing recent research findings and insights from both academia and industry, this review aims to contribute to a deeper understanding of the challenges and opportunities in male and female speech recognition. Understanding the nuances of gender-specific speech patterns is essential for the development of more inclusive and effective speech

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204054 |

recognition systems that cater to diverse user populations. Moreover, advancements in this area have the potential to unlock new applications and further improve human-machine interaction in various domains.

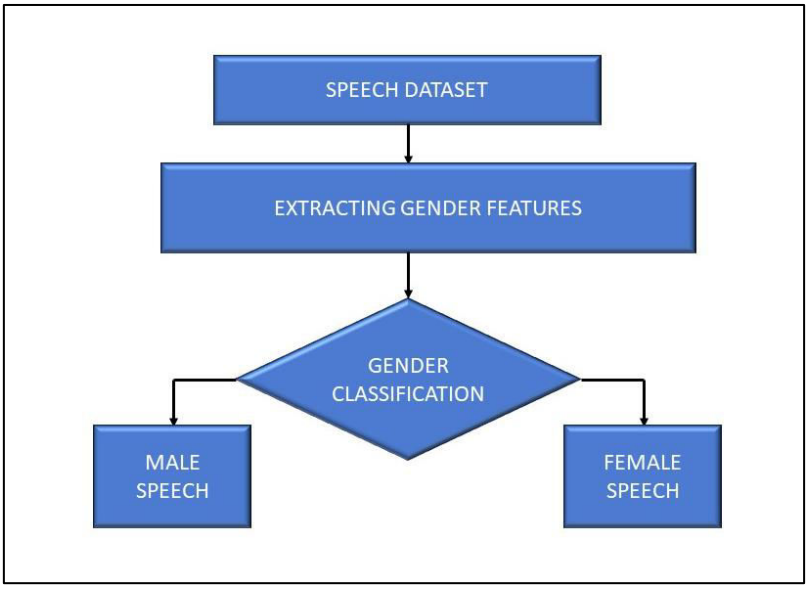

Figure 1: Framework of Speech based Gender Recognition

II. LITERATURE SURVEY

Computer simulations of human emotion perception and comprehension are used in speech emotion recognition. Using a microphone sensor, it takes the emotion's auditory characteristics out of the speech signals that have been gathered and determines the correspondence between these characteristics and human emotion. Human–computer interaction makes extensive use of speech emotion recognition.

A comparison is made between multiple methods for identifying gender through the evaluation of voice signals captured over telephone channels, including the Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Support Vector Machine (SVM) [2] [3]. The noise may cause the speech identification to be less accurate. A strong model for deep neural networks and bidirectional LSTM-based domain identification for acoustic dialogue [4]. The text-dependent system outperforms the text-independent system in terms of accuracy. The age of the speaker also affects how accurately their gender is identified. When it comes to the gender classification of younger speakers, it is more practical than for older speakers. In 2012, the MFCC was created and used to determine gender. Many changes have been made to the MFCC since it was developed in order to enhance system performance. Gender identity in several domains has also been investigated using MFCC [5]. SVM is typically utilised for gender identification in binary classification. In terms of class separation technique accuracy, it performs the best. According to an article, among SVM kernels, the Gaussian radial basis function SVM has the highest efficiency [6]. The author demonstrated how, in participating teams assessed for comparable languages, variations, and dialects at the Third Workshop on NLP, deep learning and conventional machine learning models differed significantly [7].

III. METHODOLOGIES

We begin by reviewing the fundamental methodologies employed by speech-based gender recognition systems. This includes an overview of acoustic features used to capture gender-specific characteristics in speech signals, such as pitch, formants, and prosodic cues. Furthermore, we explore the role of machine learning algorithms, including support vector machines, Gaussian mixture models, and deep neural networks, in extracting discriminative features and classifying gender. Additionally, we discuss statistical modeling techniques, such as hidden Markov models, and their application in gender recognition.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204054 |

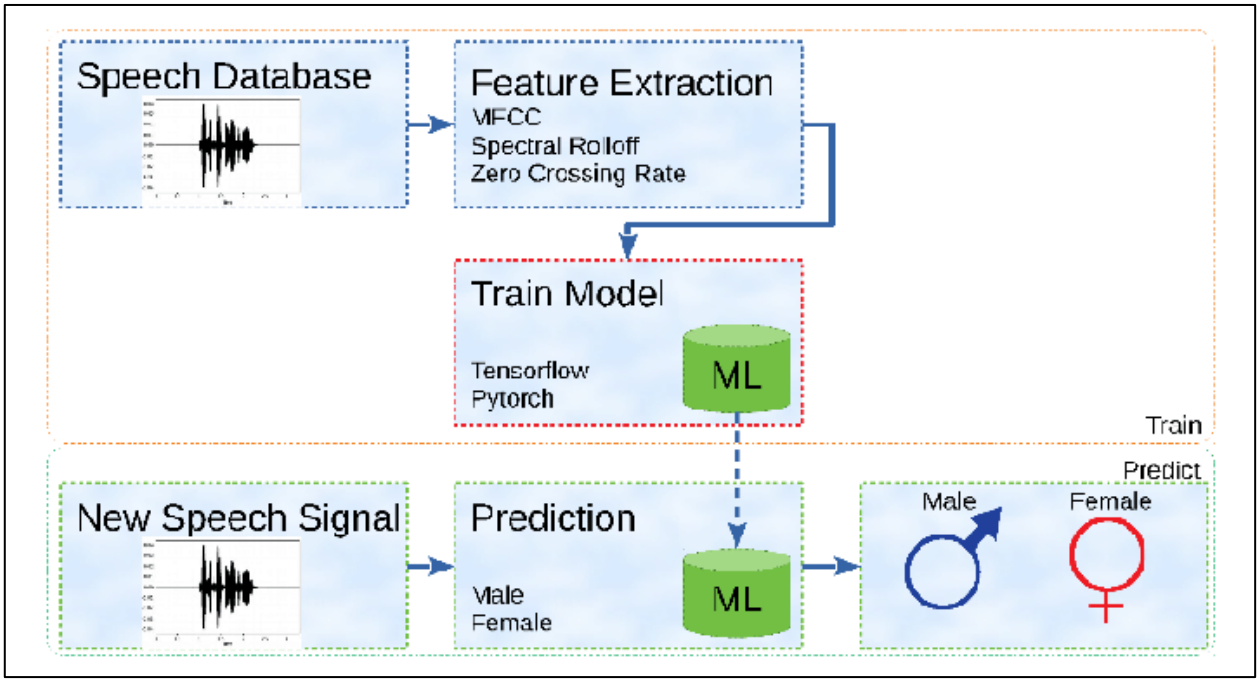

Figure 2: Machine learning based speech recognition [8].

The primary idea behind the gender identification system is to take the features out of the voice signals and make a determination by contrasting the characteristics with feature vectors that have been saved. There are two stages to the gender identification framework: the phases of training and testing, respectively.

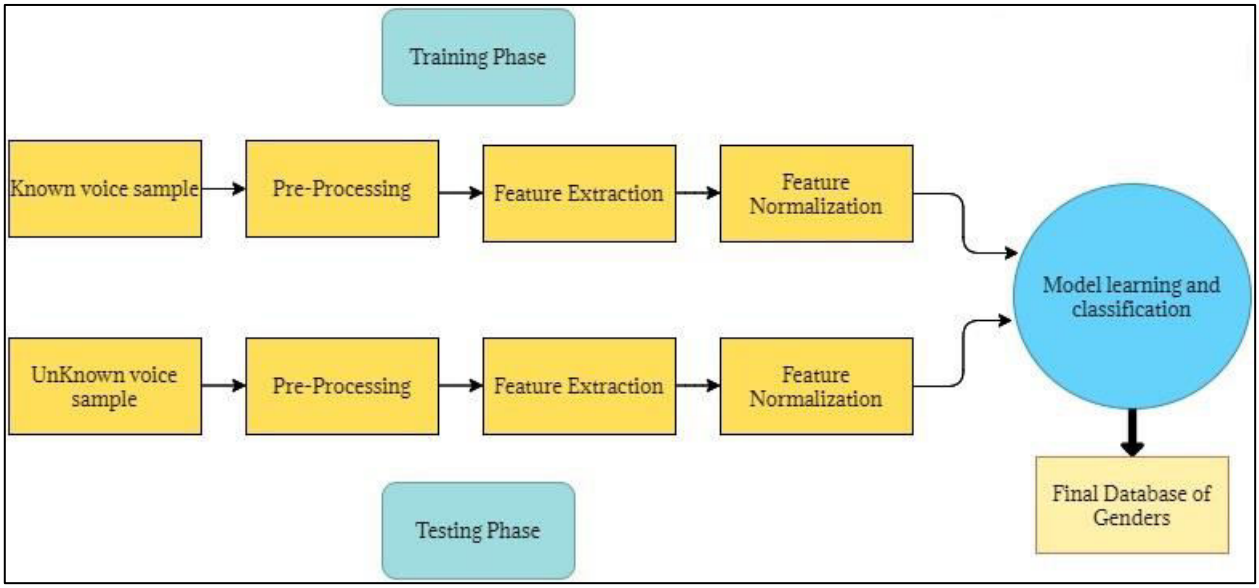

Figure 3: Speech Signals Processing

Feature Extraction of Voice Samples:

The great accuracy of the gender recognition algorithm can be attributed in large part to the extracted feature selection. Because the collected features include useful information about the speakers, they are a crucial and significant input for the classifier. Reducing the search space for the classifier is the primary goal of gathering and removing information from the audio signals. Because quasi-frequency analysis and human ear function are comparable, short-term spectral voice signals are needed for voice signal analysis. The mel frequency scale is also necessary for the study of the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204054 |

auditory nerve. There is less noise in the speech signals' frequency-domain features than in their time-domain features [9].

Mel Frequency Cepstral Coefficients (MFCC)
Davis and Mermelstein created MFCC[10]. Information regarding the speakers' many attributes is available from MFCC. Therefore, voice signal analysis uses MFCC, a characteristic of voice signals, to determine gender. The gender of the speakers is determined by the unique feature known as MFCC in the suggested model. Mel-frequency Cepstrum exhibits a response akin to that of the human auditory system and is located on a frequency range on the mel scale that is equally spaced.

PERFORMANCE METRICS: Next, we evaluate the performance of various gender recognition systems across different datasets and conditions. Performance metrics such as accuracy, precision, recall, and F1-score are essential for assessing the effectiveness of gender recognition algorithms. We compare the performance of different systems in terms of their ability to accurately classify gender across diverse speaker populations, speech styles, and languages. Additionally, we consider factors such as computational complexity and real-time processing requirements.

IV. CHALLENGES

Despite significant advancements, speech-based gender recognition systems face several challenges and limitations. Speaker variability, including differences in pitch, accent, and speaking style, can impact recognition accuracy and reliability. Moreover, linguistic diversity and cultural biases pose additional challenges, particularly in multilingual and multicultural contexts. We discuss how different systems address these challenges through feature normalization techniques, speaker adaptation strategies, and data augmentation methods. Furthermore, we examine the implications of gender bias in training data and the potential consequences for fairness and inclusivity in gender recognition systems.

V. CLASSIFICATION ALGORITHM

There are similarities between the categorization process and the supervised learning process. The speakers' gender classes are divided using the categorization techniques. The most difficult task in getting high gender identification efficiency is choosing the classifier. To determine the gender of the speakers, the classifier compared the stored features of the training voice signals with the features of the tested speech signals. Gender can be determined using a variety of classification methods, including SVM, GMM, LDA, RNN, and HMM. The RNN-BiLSTM, LDA, and SVM algorithms are used as classifiers in the proposed work.

Support Vector Machine (SVM)
SVM is an extremely effective algorithm for using voice cues to determine gender. Fixing the hyperplane in accordance with the characteristics that distinguish the genders is the primary goal of the SVM. The SVM is capable of doing binary classification with the use of the hyperplane [11]. The term "support vector" refers to the data points that are located close to the hyperplane. The support vector's separation from the other accessible data points in the vicinity of the hyperplane is difficult. The unknown samples are classified based on the margin value. The line that is perpendicular to the hyperplane is known as the margin [12]. SVM can be used with appropriate kernels, such as polynomial, radial basis function, and multilayer perceptron, to categories the nonlinear data.

Linear Discriminant Analysis (LDA)
Typically, LDA is used to separate data between two or more labels. One hyperplane is used to classify labels linearly if there are two classes. In multiple discrimination, however, multiple hyperplanes are needed to separate the classes. The following guidelines are followed in the development of the hyperplane: (i) There is a maximum distance between the two labels; and (ii) there should be a minimum variation in the feature values across both labels [13].

Recurrent Neural Networks
An algorithm for nonlinear classification that functions like a human brain is called an artificial neural network, or ANN. The procedure involves periodic adjustments to the biases and weights based on the input signals during the training phase. This procedure is repeated until there is little difference in the bias and consequence values [14]. Three layers make up a traditional ANN: an input layer, an output layer, and one hidden layer. The RNN classification algorithm belongs to the ANN family of algorithms. Speech signals and time series are examples of sequential data that RNN is particularly adept at processing. The RNN unit's output loops back to itself before moving on to the following

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204054 |

unit. The RNN algorithm takes two sorts of inputs: (a) current input and (ii) input that has been applied before. The RNN algorithm relies heavily on the previous input sequence to predict the subsequent input [15]. The RNN's limited memory is one of its shortcomings. RNN can improve long-term memory capacity by utilizing the long-short term memory (LSTM). Multiple LSTM cells are combined to form LSTM-RNN. These cells provide for more efficient control over the flow of information inside the network. Three different types of gates are found in LSTMs: input, forget, and output gates. The LSTM cell has the ability to control the information's movement through gate operation. In contrast to a BiLSTM layer, which can operate in both forward and backward directions, an LSTM layer can only operate in the forward direction. Two LSTM layers are combined to form BiLSTM. While the second layer of LSTMs can work in the opposite direction of the first, the first layer can only act in the forward direction. Capturing the future and previous input features for the particular time arrangement is the BiLSTM's primary goal. The two factors that determine the network's behavior are (i) the input at that moment and (ii) the output from a recent past input.

VI.SPEECH EMOTION RECOGNITION MODELS

The gender-based emotion detection approach has shown to be reliable and effective since it integrates gender data into the emotion recognition process. Research has indicated that emotion detection systems that are gender-neutral or gender-mixed perform worse than systems that are gender-specific.

CNN Model

CNN is a prime example of deep learning algorithms; they have demonstrated exceptional performance in voice emotion perception and have made significant progress in natural speech processing tasks including language translation and speech recognition.

A completely connected layer and two convolution layers make up CNN. The activation function is "Relu," the convolution step is 1, and the window length of the convolution kernel is 5. The prediction results are produced after the softmax activation layer, with each convolution layer's output being shifted to one dimension for batch normalization. By normalizing the output of the preceding layer, batch normalization lowers the internal covariance drift in the feature graph. Overfitting may be lessened by the irregular impact. Figure 4 depicts the CNN organizational structure.

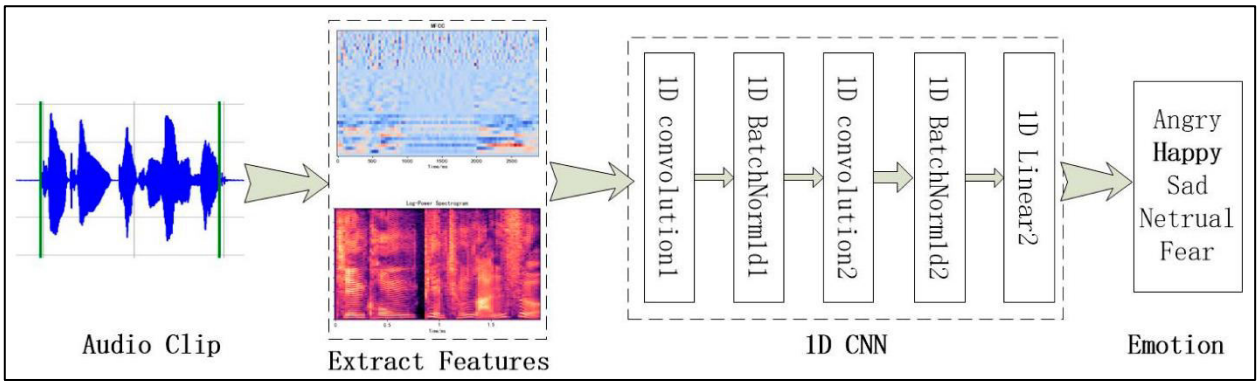

Figure 4: CNN Model for Speech recognition [16].

CNN's optimizer chooses "Adam," and cross-entropy is the loss function. Every time the parameters are altered during training, input neurons are arbitrarily disconnected with a chance of 0.3 to avoid overfitting. Five-fold cross-validation is utilised in the training and testing process, with 80% of the dataset being used for training and 20% for testing.

BiLSTM Model

BiLSTM directly chooses the 256-dimensional feature outputs of the hidden layer for batch normalization after using a layer of bidirectional LSTM to obtain the characteristics of the hidden layer [17]. The characteristic graph's internal covariance drift can be lessened, and the regularization impact that results can lessen overfitting, by normalizing the output of the preceding layer. The characteristics of the input are then down sampled and the feature space's dimension is decreased using a full connection layer. The structure of BiLSTM is shown in Figure 5.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204054 |

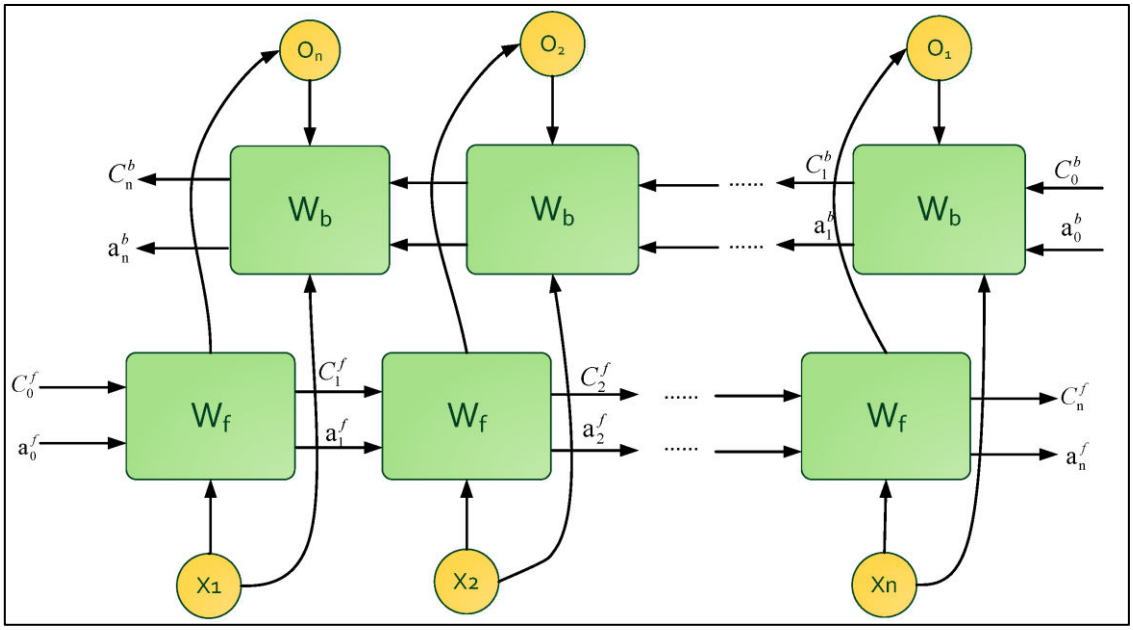

Figure 5: Edifice of BiLSTM [16].

VII. CONCLUSION

In conclusion, this comparative analysis provides valuable insights into the methodologies, performance, and challenges of speech-based gender recognition systems. By understanding the strengths and weaknesses of different approaches, stakeholders can make informed decisions regarding the selection and deployment of gender recognition technology. Moreover, this analysis guides future research directions aimed at advancing the state-of-the-art in speech-based gender recognition while addressing the ethical and societal implications of gender bias. The types of classifiers used determine the system's performance and the accuracy of gender identification. The accuracy result changes in accordance with modifications made to the classification methods. The recall values exhibit variance in correlation with the variation in the quantity of voice samples used for training and testing. The paper's primary contributions are its investigation of the influence weights of several speech emotion elements in speech emotion recognition across genders and its gender-based classification. The male and female speech data are separated using the MLP model, which also determines the gender of the original speech. The acoustic disparities in speech produced by men and women are examined.

REFERENCES

[1] A. Raahul, R. Sapthagiri, K. Pankaj, and V. Vijayarajan, "Voice based gender classification using machine learning," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 263, no. 4, p. 42083.
[2] K.-H. Lee, S.-I. Kang, D.-H. Kim, and J.-H. Chang, "A support vector machine-based gender identification using speech signal," *IEICE Trans. Commun.*, vol. 91, no. 10, pp. 3326–3329, 2008.
[3] R. R. Rao and A. Prasad, "Glottal excitation feature based gender identification system using ergodic HMM," *Int. J. Comput. Appl.*, vol. 17, no. 3, pp. 31–36, 2011.
[4] S. Rathor and S. Agrawal, "A robust model for domain recognition of acoustic communication using Bidirectional LSTM and deep neural network.," *Neural Comput. Appl.*, vol. 33, no. 17, pp. 11223–11232, 2021.
[5] M. A. Nasr, M. Abd-Elnaby, A. S. El-Fishawy, S. El-Rabaie, and F. E. Abd El-Samie, "Speaker identification based on normalized pitch frequency and Mel Frequency Cepstral Coefficients," *Int. J. Speech Technol.*, vol. 21, pp. 941–951, 2018.
[6] E. Ramdinmawii and V. K. Mittal, "Gender identification from speech signal by examining the speech production characteristics," in *2016 International conference on signal processing and communication (ICSC)*, 2016, pp. 244–249.
[7] S. Malmasi, M. Zampieri, N. Ljubešić, P. Nakov, A. Ali, and J. Tiedemann, "Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task," in *Proceedings of the third workshop on NLP for similar languages, varieties and dialects (VarDial3)*, 2016, pp. 1–14.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204054 |

[8] T. J. Sefara and A. Modupe, "Yorùbá gender recognition from speech using neural networks," in *2019 6th International conference on soft computing & machine intelligence (ISCMI)*, 2019, pp. 50–55.
[9] S. G. Koolagudi, Y. V. S. Murthy, and S. P. Bhaskar, "Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 167–183, 2018.
[10] M. Gupta, S. S. Bharti, and S. Agarwal, "Gender-based speaker recognition from speech signals using GMM model," *Mod. Phys. Lett. b*, vol. 33, no. 35, p. 1950438, 2019.
[11] B. Jena, A. Mohanty, and S. K. Mohanty, "Gender recognition of speech signal using knn and svm," 2020.
[12] M. Gupta, S. S. Bharti, and S. Agarwal, "Support vector machine based gender identification using voiced speech frames," in *2016 fourth international conference on parallel, distributed and grid computing (PDGC)*, 2016, pp. 737–741.
[13] C. Castaldello *et al.*, "A model-based support for diagnosing von Willebrand disease," in *Computer Aided Chemical Engineering*, vol. 40, Elsevier, 2017, pp. 2779–2784.
[14] M. K. Reddy and K. S. Rao, "Excitation modelling using epoch features for statistical parametric speech synthesis," *Comput. Speech Lang.*, vol. 60, p. 101029, 2020.
[15] L. Jasuja, A. Rasool, and G. Hajela, "Voice Gender Recognizer Recognition of Gender from Voice using Deep Neural Networks," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 319–324.
[16] L.-M. Zhang, Y. Li, Y.-T. Zhang, G. W. Ng, Y.-B. Leau, and H. Yan, "A Deep Learning Method Using Gender-Specific Features for Emotion Recognition," *Sensors*, vol. 23, no. 3, p. 1355, 2023.
[17] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details