International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 3, May-Jun 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Comparative Study of Speech Recognition Accuracy for Male and Female Voices

Karambir[1], Kirti Hooda[2]

[1]PG Student, ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India
[2]Assistant Professor, ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India

ABSTRACT

Even if speech recognition technology has advanced significantly in recent years, there are still issues with it, especially when it comes to correctly differentiating between male and female voices. The present state of male and female voice recognition systems is examined in this paper, along with the roots of the issues and the approaches used to overcome them. We explore both the physiological and social language aspects of speech production and their effects on the precision of recognition. We also go over how deep neural networks and other machine learning methods can improve gender classification in speech recognition systems. Additionally, we examine the effects of gender bias and methods for reducing it in speech recognition software. This review provides insights into the achievements made in male and female voice recognition as well as future directions by combining the results of previous studies.

KEYWORDS: *speech recognition, male speech, female speech*

How to cite this paper: Karambir | Kirti Hooda "A Comparative Study of Speech Recognition Accuracy for Male and Female Voices" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-3, June 2025, pp.694-699, URL: www.ijtsrd.com/papers/ijtsrd81092.pdf

IJTSRD81092

INTRODUCTION

Our everyday lives now revolve around technology that recognizes speech, which has completely changed the way we use gadgets and systems. Speech recognition allows for smooth human-machine communication in a variety of applications, including automated customer support systems and virtual assistants. But even with great advancements, speech recognition systems still struggle to reliably differentiate between male and female voices. Accurately determining the speaker's gender is essential for customized user encounters and answers. Furthermore, gender recognition is essential for a number of uses, such as voice-activated gadgets, gender-specific promotion, and forensic investigation [1].

Due to physiological and sociolinguistic variations, male and female speech differs from one another. These variations present particular difficulties for speech recognition systems, which need to accurately classify speech by efficiently capturing and interpreting gender-specific characteristics. Male and female voices vary physiologically due to variances in vocal tract length, fundamental frequency (pitch), and resonance. Male and female speech patterns are further distinguished by sociolinguistic elements like intonation, pitch range, and speaking style.

The goal of this paper is to present a thorough analysis of the state of male and female voice recognition today, emphasizing both recent developments and underlying complexity. We investigate how male and female vocalization is influenced by physiological and sociolinguistic factors, look at the methods used to classify gender in speech recognition systems, and talk about how machine learning algorithms—in particular, deep neural networks—can improve the precision of gender recognition. We also talk about possible mitigating techniques and the effects of gender bias in speech recognition techniques.

This review seeks to advance knowledge of the possibilities and challenges in male and female voice recognition by combining the most recent research findings with insights from academia and industry. In order to create more inclusive and efficient voice recognition systems that serve a variety of user

demographics, it is imperative to comprehend the subtleties of gender-specific speech patterns. Furthermore, developments in this field could open up new uses and enhance human-machine interaction across a range of fields.

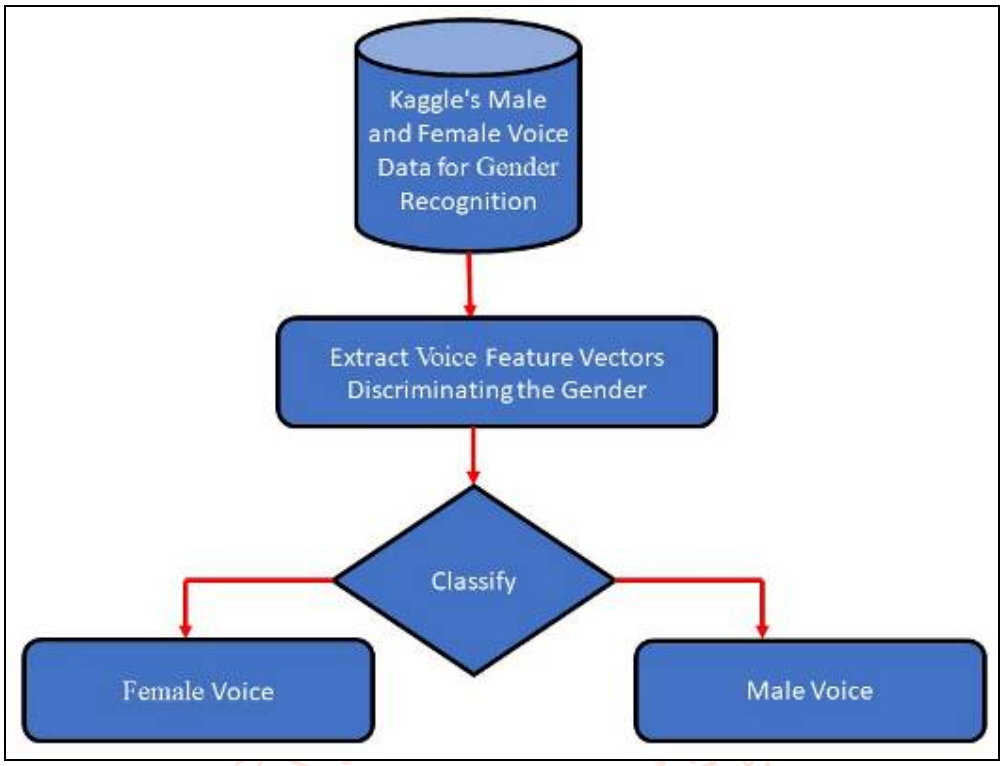

Figure 1: Framework of Speech based Gender Recognition

LITERATURE SURVEY

Speech emotion recognition uses computational models of human emotion perception and understanding. It extracts the auditory features of the emotion from the collected speech data using a microphone sensor and establishes the relationship between these features and human emotion. Speech emotion recognition is widely used in human–computer interaction.

The Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Support Vector Machine (SVM) are among the several techniques for determining gender based on the analysis of voice data recorded over telephone channels [2] [3]. The speech recognition may be less precise as a result of the noise. Bidirectional LSTM-based domain classification for auditory discourse and a robust model for DNN [4]. In terms of accuracy, the text-dependent system performs better than the text-independent method. The speaker's age has an impact on how correctly their gender is determined as well. Compared to elder speakers, it is more practicable to classify younger speakers according to their gender.

The MFCC was developed in 2012 and is used to identify gender. Since its creation, the MFCC has undergone numerous modifications to improve system performance. MFCC has also been used to examine gender identification in a number of domains [5]. In binary categorization, SVM is commonly used for gender identification. It is the most accurate in terms of class separation technique. An article claims that the Gaussian radial basis function SVM is the most efficient SVM kernel [6]. The author illustrated the considerable differences between deep learning and traditional machine learning (ML) models that the groups being evaluated for comparable tongues, dialects, and changes at the Third Workshop on NLP. [7].

METHODOLOGIES: We start by going over the basic techniques used by systems that recognize gender based on voice. This provides a summary of the acoustic features—such as pitch, formants, and prosodic cues—that are utilized to record gender-specific traits in speech signals. We also investigate how ML technologies, such as DNNs, SVM, and Gaussian mixture models, can be used to retrieve discriminatory characteristics and categorize gender. We also go over the use of statistical modeling methods in gender recognition, including hidden Markov systems.The main concept of the gender identification system is to extract the features from the voice signals and compare them with previously stored feature vectors to determine the gender. The gender identification framework is divided into two steps, which are training and testing.

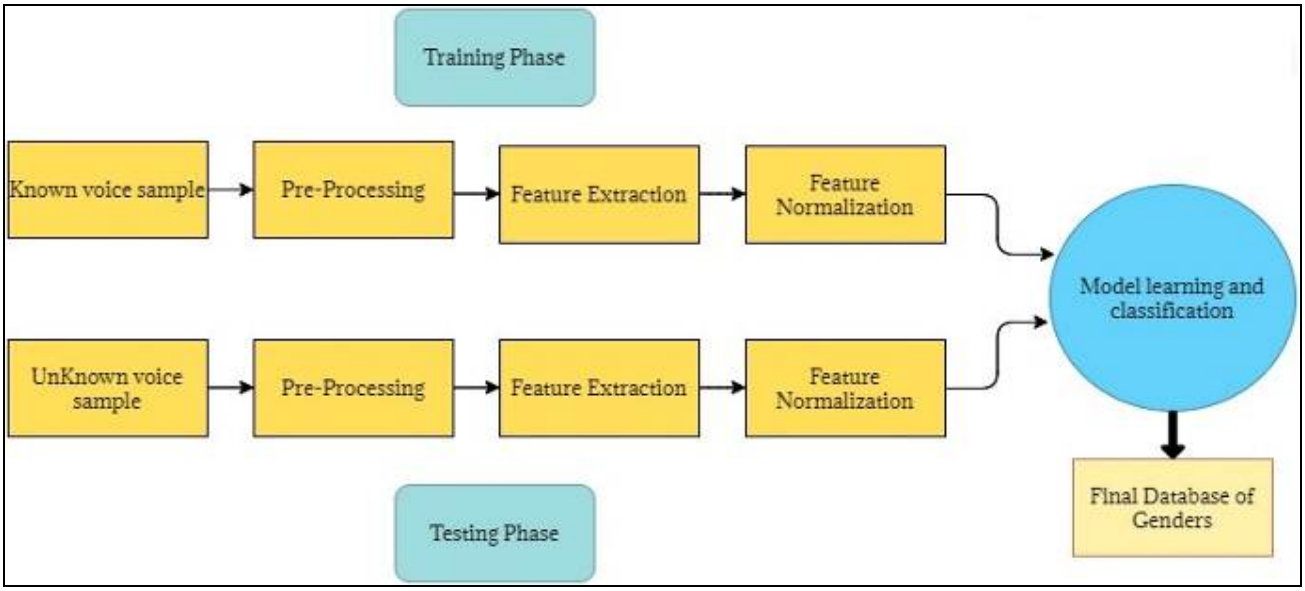

Figure 2: Speech Signals Processing [8]

Extracting Features from Voice Samples:

The derived feature selection (see figure 3) is largely responsible for the male or female recognition algorithm's excellent precision. The gathered features are an important and vital input for the classifier since they include valuable information about the speakers. The main objective of collecting and eliminating details about the audio signals is to minimize the classifier's search space. Voice signal analysis requires short-term spectral voice signals due to the similarity between quasi-frequency analysis and human ear function. The study of the auditory nerve also requires the use of the Mel frequency scale. The frequency-domain features of the voice signals have less noise than the time-domain aspects [9].

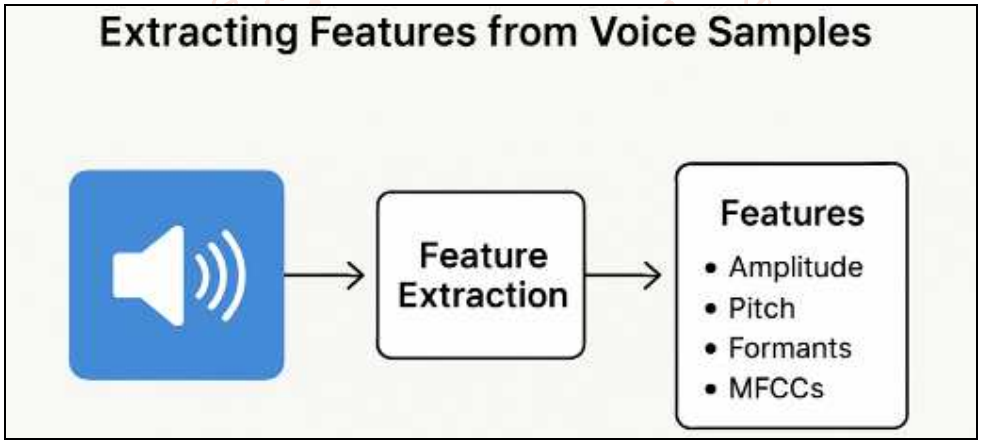

Figure 3: Process of extracting features from a voice sample

Mel Frequency Cepstral Coefficients (MFCC)

MFCC was founded by Davis and Mermelstein [10]. The MFCC provides information on the speakers' numerous qualities. Thus, MFCC, a feature of voice signals, is used in voice signal analysis to identify gender. In the proposed approach, the distinctive feature called MFCC determines the speakers' gender. Mel-frequency Cepstrum is situated on a frequency range on the mel scale that is equally spaced and responds similarly to the human auditory system.

PERFORMANCE METRICS: The efficacy of the various gender recognition systems under varied settings and datasets is then assessed. For evaluating the efficacy of gender recognition computations, performance metrics including accuracy, precision, recall, and F1-score are crucial. We evaluate how well various systems perform in regards to correctly classifying gender across a range of speaker groups, languages, and speech patterns. We also take into account elements like the need for real-time processing and complexity of computation.

CHALLENGES: Speech-based gender identification systems still face a number of obstacles and restrictions despite tremendous progress. Reliability and dependability of recognition can be impacted by speaker variation

which includes variations in pitch, accent, and speaking style. Further challenges are presented by linguistic variation and cultural prejudices, especially in multicultural and multilingual settings. We go over how various systems use data augmentation techniques, speaker adaption tactics, and feature normalization techniques to overcome these obstacles. Additionally, we look at the effects of gender bias in training data and how it might affect gender recognition systems' inclusion and fairness.

CLASSIFICATION ALGORITHM

The classification process and the supervised learning approach share several commonalities. The classification techniques are used to classify the speakers into different gender groupings. Selecting the classifier is the most challenging step in achieving high recognition of gender efficiency. The classifier compared the stored attributes of the training voice signals with the features of the tested speech signals in order to identify the speakers' gender. A number of classification techniques, including as SVM, GMM, LDA, RNN, and HMM, can be used to identify gender. The proposed study uses the RNN-BiLSTM, LDA, and SVM algorithms as classifiers.

Support Vector Machine (SVM)

SVM is a very powerful algorithm for identifying gender from speech cues (figure 4). The main objective of the SVM is to fix the hyperplane according to the traits that differentiate the genders. By using the hyperplane, the SVM can do binary classification [11]. The data points near the hyperplane are referred to as the "support vector". It is challenging to distinguish the support vector from the other available data points near the hyperplane. The margin value is used to categorize the unidentified samples. The margin is defined as the line perpendicular to the hyperplane [12]. To classify the nonlinear data, SVM can be utilized with suitable kernels, including multilayer perceptrons, polynomial, and radial basis functions.

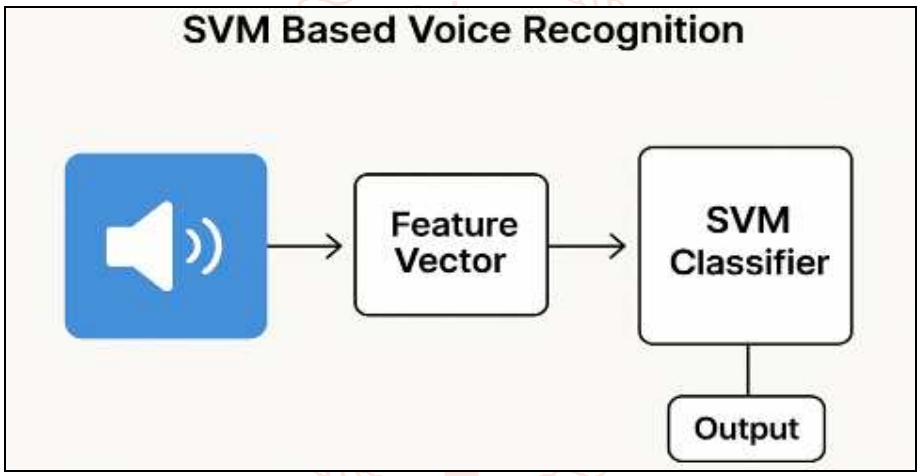

Figure 4: Speech recognition using SVM

Linear Discriminant Analysis (LDA)

To segregate data between two or more labels, LDA is usually utilized. If there are two classes, labels are linearly classified using a single hyperplane. Yet, many hyperplanes are required to divide the classes in numerous ways. When creating the hyperplane, the following rules are adhered to: (i) The two labels have a maximum separation, and (ii) the feature values for each label should vary as little as possible [13].

Recurrent Neural Networks

An artificial neural network, or ANN, is a nonlinear classification system that works similarly to the human brain. Throughout the training phase, the process entails making periodic modifications to the weights and biases in response to the input signals. Until there is minimal variation in the bias and consequence values, this process is repeated [14]. An input layer, an output layer, and one hidden layer comprise a conventional ANN. The ANN family of algorithms includes the RNN classification method. RNN is especially good at processing sequential data, such as time series and speech signals. After looping back to itself, the output of the RNN unit advances to the next unit.

Two different types of inputs are accepted by the RNN algorithm: (a) current input and (ii) previously applied input. To forecast the next input, the RNN algorithm mostly depends on the prior input sequence [15]. One of the drawbacks of the RNN is its little memory. By making use of the long-short term memory (LSTM), RNN can increase long-term memory capacity. LSTM-RNN is created by combining many LSTM cells. More effective control over the information flow inside the network is made possible by these cells. LSTMs contain three distinct kinds of gates: input, forget, and output gates.

Via gate operation, the LSTM cell may regulate the flow of information. An LSTM layer is limited to forward operation, whereas a BiLSTM layer can function in both forward and backward directions. BiLSTM is created by combining two LSTM layers. Only forward motion is possible for the first layer of LSTMs, whereas the second layer can operate in the opposite direction. The main objective of the BiLSTM is to capture the past and future input properties for the specific temporal arrangement. Both (i) the current input and (ii) the output from a recent previous input influence how the network behaves.

SPEECH EMOTION RECOGNITION MODELS

Because it incorporates gender information into the emotion recognition process, the gender-based emotion detection method has proven to be dependable and successful. Studies have shown that gender-neutral or gender-mixed emotion detection systems are less effective than gender-specific ones.

CNN Model

CNN is an outstanding instance of DL algorithms; they have made great strides in natural speech processing jobs including translating languages and speech identification, and they have shown remarkable results for voice emotion detection. CNN consists of two convolution layers and a fully connected layer. The convolution step is 1, the activation function is "Relu," and the convolution kernel's window length is 5. Following the softmax activation layer, the output of each convolution layer is switched to one dimension for batch normalization, yielding the prediction outputs. Batch normalization reduces the internal covariance drift in the feature graph by normalizing the result of the previous layer. The uneven impact may reduce overfitting.

BiLSTM Model

After employing a bidirectional LSTM layer to acquire the hidden layer's properties, BiLSTM selects the hidden layer's 256-multidimensional feature outputs for batch normalization [16]. By normalizing the output of the previous layer, the internal covariance drift of the specific graph can be reduced, and the resulting regularization effect can reduce overfitting. A full connection layer is then used to reduce the dimension of the feature space and down sample the features of the input.

CONCLUSION: To sum up, this comparative study offers insightful information about the approaches, capabilities, and difficulties of speech-based gender recognition systems. Stakeholders are better equipped to choose and use gender recognition tools if they are aware of the advantages and disadvantages of various strategies. Furthermore, this analysis directs future research efforts to address the ethical and societal ramifications of gender bias while increasing the state-of-the-art in speech-based gender recognition. The system's functionality and the precision of gender identification are determined by the classifier types that are employed. As adjustments are made to the classification techniques, the precision result varies accordingly. There is a correlation between the recall values and the fluctuation in the number of voice samples utilized for testing and training. The main contributions of the paper are its gender-based classification and its examination of the influence weights of several speech emotion aspects in speech emotion recognition across genders. In addition to identifying the original speech's gender, the MLP model separates the speech data into male and female categories. Men's and women's speech acoustic differences are investigated.

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*, IOP Publishing, 2017, 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)*, IEEE, 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.

[8] R. S. Akanksha, Sumit Dalal, "Exploring the Challenges and Advancements in Male and Female Speech Recognition: A Comprehensive Review," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 12, no. 4, pp. 3136–3142, 2024.

[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)*, IEEE, 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)*, IEEE, 2020, pp. 319–324.

[16] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.