Computer Systems Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.046730

ARTICLE

Comprehensive Analysis of Gender Classification Accuracy across Varied Geographic Regions through the Application of Deep Learning Algorithms to Speech Signals

Abhishek Singhal* **and Devendra Kumar Sharma**

Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Delhi–NCR Campus, Ghaziabad, 201204, India

*Corresponding Author: Abhishek Singhal. Email: abhisheksinghal.srm@gmail.com

Received: 13 October 2023 Accepted: 12 December 2023 Published: 20 May 2024

ABSTRACT

This article presents an exhaustive comparative investigation into the accuracy of gender identification across diverse geographical regions, employing a deep learning classification algorithm for speech signal analysis. In this study, speech samples are categorized for both training and testing purposes based on their geographical origin. Category 1 comprises speech samples from speakers outside of India, whereas Category 2 comprises live-recorded speech samples from Indian speakers. Testing speech samples are likewise classified into four distinct sets, taking into consideration both geographical origin and the language spoken by the speakers. Significantly, the results indicate a noticeable difference in gender identification accuracy among speakers from different geographical areas. Indian speakers, utilizing 52 Hindi and 26 English phonemes in their speech, demonstrate a notably higher gender identification accuracy of 85.75% compared to those speakers who predominantly use 26 English phonemes in their conversations when the system is trained using speech samples from Indian speakers. The gender identification accuracy of the proposed model reaches 83.20% when the system is trained using speech samples from speakers outside of India. In the analysis of speech signals, Mel Frequency Cepstral Coefficients (MFCCs) serve as relevant features for the speech data. The deep learning classification algorithm utilized in this research is based on a Bidirectional Long Short-Term Memory (BiLSTM) architecture within a Recurrent Neural Network (RNN) model.

KEYWORDS

Deep learning; recurrent neural network; voice signal; mel frequency cepstral coefficients; geographical area; gender

1 Introduction

In accordance with Webster's phrasebook, "Speech" is defined as the act of conveying sentiments and ideas through spoken words. Speech signals are carriers of distinct characteristics that encode valuable information about speakers, including aspects such as their health status, gender, and emotional disposition [1,2]. These characteristics are dynamic and contingent upon the real-time movements of the speaker's vocal tract. Notably, paralinguistic features of the speaker can be discerned through a comprehensive analysis of speech signals, even in instances where the speaker remains concealed. Among these characteristics, gender holds a pivotal role as a fundamental attribute of the

speaker. Furthermore, it is noteworthy that attributes inherent to speech samples exhibit variations that correspond to gender distinctions. The endeavor to identify gender through the analysis of speech signals is an intricate task, given its wide-ranging applicability across diverse and demanding domains. Significantly, the accurate determination of a speaker's gender finds relevance in applications spanning voice recognition, automated speaker identification, legal proceedings, targeted phone-based advertising, computerized language learning systems, healthcare, gender-centric telephonic surveys, and human-robot interactions, among others [3–6].

In the ever-evolving landscape of modern computing, the existence of a robust gender identification model is imperative [7]. This is particularly true due to the manifold advantages presented by speech signal analysis in contemporary scenarios, which encompass healthcare systems, military telephony infrastructure, and support for individuals with disabilities. The multifaceted applications of speech signal analysis have consequently thrust this field into the spotlight, emerging as a fertile ground for researchers. It is important to acknowledge that gender identification within speech signal analysis is a formidable challenge, primarily owing to the myriad factors influencing the task. These factors include pitch variations, frequency modulations, magnitude changes in speech signals as influenced by emotional states, environmental noise, and more. Gender identification systems are typically categorized as either gender-dependent or gender-independent, with the former often achieving higher accuracy than the latter [8]. To attain optimal accuracy in gender identification models, it is imperative to minimize the search space for features within the gender identification system [9]. The construction of a gender identification model involves the utilization of speech samples, the extraction of pertinent features from speech signals, and the deployment of classification algorithms. Notably, the selection of the appropriate classifier and the extraction of relevant features pose considerable challenges due to the inherent variability in speech signals [4,10,11]. The process of gender identification is inherently divided into two phases: (a) the training phase and (b) the testing phase. In the initial training phase, features are extracted from known speech samples, subsequently forming the basis of the search space. The testing phase entails the extraction of features from unknown speech samples, with the gender classification algorithm rendering decisions based on the extracted features and the feature space established during the training phase.

This article endeavors to systematically compute and contrast gender identification accuracy across varied geographical regions. Employing meticulous analysis, it seeks to unveil potential disparities in accuracy rates, shedding light on the impact of regional linguistic nuances on gender identification algorithms. By scrutinizing diverse datasets from distinct locales, this research contributes valuable insights into refining and optimizing gender identification models for enhanced cross-cultural effectiveness. The findings aim to foster a nuanced understanding of the multifaceted influences that geographical variations may exert on the performance of gender identification systems. To achieve this, Mel Frequency Cepstral Coefficients (MFCCs) with 12 coefficients are employed as extracted features, in conjunction with a deep learning algorithm. Deep learning algorithms have demonstrated superior performance when compared to alternative methods, as evidenced in the existing literature [12,13]. These deep learning architectures typically comprise input layers, one or more hidden layers, and an output layer [14]. In the study, the RNN–BiLSTM algorithm is used as the gender classification mechanism. The speech samples used in the research encompass recordings from speakers outside the Indian subcontinent, sourced from open repositories, and live-recorded speech samples from typical Indian speakers. Importantly, the speech samples are text-independent. The decision-making process of the simulation is represented through a confusion matrix. The structure of this article is organized as follows: Section 2 provides a concise overview of pertinent prior work in the literature by esteemed researchers. Section 3 elucidates the methodology employed for the analysis of speech signals and

the intricacies of the gender classification algorithm. Section 4 presents the findings arising from the analysis of speech signals. Finally, Section 5 offers concluding remarks.

2 Related Work

The cornerstone of any speech signal analysis system lies in its ability to accurately identify the gender of the speaker [15]. Over the years, various classification algorithms have been employed to assess gender identification accuracy, including the Gaussian Mixture Model (GMM), Support Vector Machine (SVM), and Hidden Markov Model (HMM), all of which have been applied to the analysis of speech signals [16–18]. The SVM classification algorithm, for instance, has proven to be effective in speaker gender classification, achieving an accuracy rate exceeding 90% [19]. It is noteworthy that the accuracy of gender identification varies between text-independent and text-dependent systems. In the context of text-independent systems, accuracy tends to be lower. Moreover, gender identification accuracy is also influenced by the speaker's age, with the classification of older speakers proving to be more complex compared to their younger counterparts when analyzing speech signals [20,21]. The presence of noise in the speech signal can further degrade the accuracy of gender identification [1,11].

The introduction of Mel Frequency Cepstral Coefficients (MFCC) in 2012 marked a significant development in gender classification of speakers. Subsequent refinements have sought to enhance the performance of classification systems. MFCC has emerged as a valuable extracted feature for identifying the gender of speakers across various domains [22]. The I-vectors classification algorithm, rooted in the concept of embedding, served as the foundation for several prominent classification methods prior to the advent of deep learning [23]. In recent investigations, gender identification has seen the utilization of the bidirectional Recurrent Neural Network (RNN) classification algorithm, incorporating recurrent gated units, resulting in a reported classification accuracy of 79% [24,25]. Furthermore, both naïve Bayes and RNN classifiers have achieved notable classification accuracies of 62.3% and 78.8%, respectively [26,27]. This marks a substantial shift towards the adoption of deep learning techniques for gender identification, given their demonstrably superior accuracy in comparison to traditional machine learning approaches. Deep Neural Networks (DNN) have emerged as particularly effective in learning intricate data sequences, attaining classification accuracies as high as 95.4% [28]. This paradigm shift underscores the heightened efficacy of deep learning models in significantly advancing the accuracy and reliability of gender identification systems, presenting promising avenues for further research and application in diverse domains. Additionally, it has been observed that DNN-based classification models consistently outperform other approaches [29,30]. In the contemporary landscape of gender classification based on speech signal analysis, Convolutional Neural Networks (CNN), DNNs, and RNNs have become prevalent choices [31]. These sophisticated models have revolutionized the field, providing robust solutions for accurate gender identification in a variety of applications.

The fine-tuned ResNet 50 on gender data demonstrated exceptional accuracy at 98.57%, surpassing traditional machine learning methods and prior works with the identical dataset. This signifies the superior performance of deep learning models, particularly ResNet 50, in advancing gender classification tasks [32]. The amalgamation of MFCC and mel feature sets exhibits superior accuracy, achieving an impressive 94.32%. This combination underscores the effectiveness of integrating diverse feature sets in enhancing accuracy and robustness, making it a promising approach for advanced signal processing tasks [33]. A Multi-Output based 1D Convolutional Neural Network was employed for gender and region recognition using a consolidated dataset comprising TIMIT, RAVDESS, and BGC datasets. The model achieved a gender prediction accuracy of 93.01% and a region prediction

accuracy of 97.07%. Outperforming conventional state-of-the-art methods, this approach excels in both individual and combined datasets for gender and region classification [34].

3 Materials and Methods

The functionality of the lungs plays a crucial role in activating and supporting speech signals. Speech signals are generated through the rhythmic movements of the vocal tract, and the articulation of these movements can significantly impact the characteristics of these signals [35]. In order to effectively analyze speech signals, it is imperative that they exhibit stability. To achieve this stability, segmenting the speech signals into smaller, minute fragments becomes necessary. Segmentation allows for the isolation of specific portions of the speech signal, facilitating the extraction of relevant features. Typically, the initial part of a speech signal often contains noise or silence signals, while the distinctive characteristics of the speaker are predominantly found in the remaining part of the signal. The proposed model adopts Mel Frequency Cepstral Coefficients (MFCCs) as the extracted features from the speech signals. These MFCCs serve as informative representations of the speech signals. In conjunction with the MFCCs, the classification algorithm employed in the model is based on a Recurrent Neural Network–Bidirectional Long Short-Term Memory (RNN–BiLSTM). This choice of algorithm is known for its effectiveness in handling sequential data and is well-suited for the analysis of speech signals. The results of the classification process are presented in the form of a confusion matrix. This matrix provides a comprehensive overview of the model's performance in categorizing and identifying the speakers based on the extracted features from the segmented speech signals.

3.1 About the Database

In the current study, the analysis is conducted using speech signals in the .mp3 format. Specifically, the speech samples are distributed into two distinct categories: those originating from speakers outside the Indian continent and live speech recordings from Indian citizens, both of which serve as valuable data sources for the research. Speech signals from speakers outside the Indian continent have been obtained from the widely accessible website https://commonvoice.mozilla.org/en/datasets. These signals are collected in the .mp3 format and serve as an essential component of the dataset. In parallel, the live speech samples are recorded from Indian citizens at a sampling rate of 44.1 kHz. These recordings were conducted in various environments, encompassing both indoor and outdoor settings, to ensure the diversity and representativeness of the dataset. The speech samples from Indian citizens include individuals from typical and ordinary backgrounds, covering a spectrum of linguistic diversity. These samples comprise both the English language and the national language of India, which is "Hindi".

To facilitate the research, the speech samples are categorized into two distinct sets, each serving a specific purpose. Category 1, designated for training purposes, exclusively contains speech samples related to speakers from outside the Indian continent. In contrast, Category 2 is populated with speech samples from Indian citizens. Each category serves as a dedicated training dataset, allowing the system to learn and adapt to the distinct speech characteristics of these two groups, as shown in Table 1a. For the testing phase, the speech samples are further divided into four distinct sets, each with its unique attributes. Set 1 comprises speech samples from speakers outside the Indian continent, with English as the language of choice. Set 2 consists of speech samples from Indian citizens, with English as the language of expression. Set 3 retains speech samples from Indian citizens, but with Hindi as the chosen language. Lastly, Set 4 incorporates a mix of speech samples, featuring contributions from both Indian citizens and speakers from outside the Indian continent. The distribution of speech samples

within these sets is detailed in Table 1b, providing a clear overview of the number of samples available for testing across each set. This comprehensive dataset ensures that the analysis is well-rounded and capable of addressing various linguistic and geographical variables.

Table 1a: Datasets for training purposes of the classification algorithm

S. no.	Category number	Geographical area	Language	Number of speech samples	
				Male	Female
1.	Category–1	Outside indian continent	English	400	400
2.	Category–2	Indian continent	English and Hindi	280	280

Table 1b: Datasets for testing purposes of the classification algorithm

S. no.	Set number	Geographical area	Language	Number of speech samples	
				Male	Female
1.	Set 1	Outside Indian continent	English	40	40
2.	Set 2	Indian continent	English	40	40
3.	Set 3	Indian continent	Hindi	40	40
4.	Set 4	Outside Indian continent and Indian continent	English and Hindi	40	40

3.2 The Layout of the Classification Model

The schematic representation of the speech signal processing workflow is visually depicted in Fig. 1. The gender identification model is divided into two fundamental phases: training and testing. At the core of both these phases lies the pivotal step of feature extraction, a process of paramount importance in the execution of both training and testing procedures. Within the feature extraction process, the inherent characteristics of the speech signal undergo a transformation into numerical values. These numerical representations are effectively organized into combinations known as feature vectors. This essential step serves to expound upon the valuable information and distinctive traits contained within the speech samples. In essence, it is the feature extraction process that enables the model to capture and interpret the salient attributes embedded in the speech data.

During the training phase, a total of 80 speech samples from each gender category are employed for the analysis of speech signals. These samples serve as the training data for the classification model, with the primary objective being the differentiation of genders into the categories of female and male. Feature vectors, derived from the speech signal characteristics, are instrumental in training the classification model. They encapsulate the distinctive attributes necessary for the model to effectively discern between female and male speakers. Subsequently, in the testing phase, a similar process of feature extraction is carried out, mirroring the steps taken during the training phase. Feature extraction from the testing samples follows the same methodology as in the training phase, ensuring consistency in the data representation. With the feature-extracted data in hand, the classification algorithm is applied. Leveraging the insights gained from the training phase, the proposed model employs this classification algorithm to render decisions regarding the gender of the speakers. In essence, the model

leverages the knowledge gained from the training data to classify speakers as either female or male during the testing phase, based on the extracted features from the speech signals.

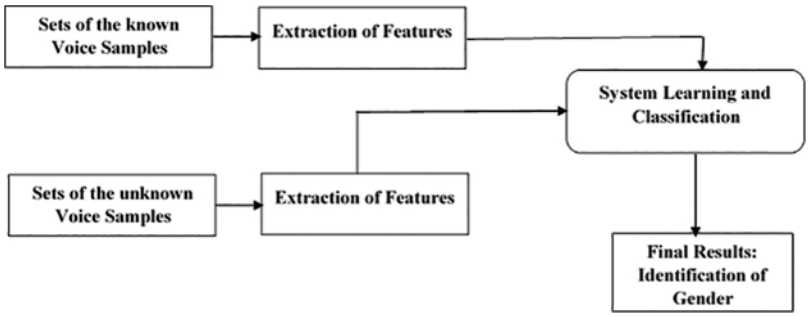

Figure 1: Layout of the classification model

3.3 Feature Expulsion from the Speech Samples

Achieving high accuracy in the gender identification system hinges significantly on the careful selection of extracted features, often referred to as "expulsed traits". These expulsed traits play a pivotal role in enhancing the performance of classification algorithms. They encapsulate valuable information about the speakers, effectively summarizing the distinguishing characteristics present in the speech signals. The primary objective behind the extraction and collection of these important traits from the speech signals is to streamline and optimize the search space for the classification algorithms [36]. By distilling the relevant information into feature vectors, the system reduces the complexity and dimensionality of the data, making it more amenable to analysis. This focused and efficient representation of the data ensures that the classification algorithms can work with a more manageable and discriminative set of features, ultimately contributing to higher accuracy in the gender identification process. In essence, the choice of expulsed traits serves as a critical determinant in the success of the gender identification system.

3.3.1 Mel Frequency Cepstral Coefficients (MFCC)

The process of reducing the dimensionality and complexity of speech signals is commonly referred to as "trait expulsion" within the context of this study. The efficacy and behavior of the classification algorithm largely hinge on the specific features extracted from the speech signals. In the proposed model, Mel Frequency Cepstral Coefficients (MFCCs) is utilized as the expulsed trait of the speech signals. This selection is motivated by the widespread use of MFCCs in the field and their ability to encapsulate a wealth of information pertaining to the speakers [21].

The pre-emphasis stage marks the commencement of Mel Frequency Cepstral Coefficient (MFCC) extraction, strategically implemented to bolster the efficacy of the identification model. Voice signals encompass both high-frequency and low-frequency components. Through pre-emphasis, the potency of high-frequency signals is heightened, facilitating a refined and improved outcome in the subsequent stages of the identification model. This meticulous process addresses the nuanced composition of voice signals, optimizing their representation for enhanced accuracy in the extraction of MFCCs and, consequently, in the overall performance of the identification model. The mathematical representation for the pre-emphasis process is expressed by Eq. (1).

$$y(n) = x(n) - 0.95x(n-1) \tag{1}$$

where, x(n) is input to the pre-processing process and y(n) is the output.

The subsequent phase in Mel Frequency Cepstral Coefficient (MFCC) extraction is framing, wherein voice signals undergo segmentation into smaller, manageable segments. This segmentation proves instrumental in capturing the stationary features inherent in voice signals. The samples are effectively segregated into N segments, facilitating a structured representation. Following framing, a windowing operation is executed to eliminate discontinuities within the samples. Upon the application of the window function, voice signals can be accurately calculated, as delineated in Eq. (2). This systematic process, encompassing framing and windowing, ensures the precise delineation and preparation of voice signals for subsequent stages in the MFCC extraction process.

$$B(m) = A(m) * W(m) \tag{2}$$

where W(m) is the hamming window [37]

The signal edges are now smoothed, thanks to the preceding steps. Utilizing Fast Fourier Transform (FFT) allows for the identification of the spectrum within the voice signals. This transformative stage shifts the voice signals from the time domain to the frequency domain, capturing essential spectral information. The output from the FFT process is subsequently channeled through the mel filter bank, aligning with the logarithmic nature of the human auditory response. Eq. (3) delineates the process through which the output of the Mel filter bank is obtained, encapsulating critical information essential for further analysis and feature extraction in the Mel Frequency Cepstral Coefficient (MFCC) extraction process.

$$F(Mel) = 2595 * \ln\left(1 + \frac{f}{100}\right) \tag{3}$$

where f is the input frequency in Hz and F(Mel) is the frequency of the output signal [38].

In the final step, the Discrete Cosine Transform (DCT) is applied to extract Mel Frequency Cepstral Coefficients (MFCC) from the voice signals. This transformative step converts the frequency-domain voice signal back into the time domain, completing the cycle of feature extraction. DCT plays a pivotal role in condensing and capturing the essential characteristics of the voice signals, resulting in a compact representation of features crucial for subsequent analysis. The integration of DCT finalizes the MFCC extraction process, providing a comprehensive and informative representation of the distinctive traits inherent in the original voice signals. These MFCCs represent distinctive traits of the speech samples. The collective assembly of these distinctive traits is recognized as 'acoustic vectors' [5,9,39–41].

3.4 Classification Algorithm

The process of classifying speakers based on their gender relies on the utilization of classification algorithms. The choice of which classification algorithm to employ is a pivotal aspect of this process, as the accuracy in gender identification can significantly differ based on the algorithm selected. Deep learning stands as a fundamental and potent approach for data representation [42,43]. Its multilayered architecture offers remarkable efficiency in achieving classification results. In the context of this article, the power of RNN–BiLSTM, a recurrent neural network with bidirectional learning capabilities, is used to perform the task of gender classification for human subjects [44,45]. This advanced algorithm leverages its bidirectional learning capacity to enhance the precision and robustness of gender identification within the scope of the research.

3.4.1 Recurrent Neural Networks

An Artificial Neural Network (ANN), often referred to as a non-linear classification algorithm, mimics the functioning of the human brain. During the training phase of the classification process, the network continuously adjusts its weights and biases based on input signals. This iterative process persists until the variations in weight and bias become imperceptible [46–48]. An ANN typically consists of three layers: the input layer, the hidden layer situated in the middle, and the output layer. Within the realm of ANNs, the Recurrent Neural Network (RNN) classification algorithm holds a prominent place. RNN is employed when dealing with sequential input data, relying on both the current and previously applied inputs as its primary inputs [49,50]. However, RNN possesses a limitation in terms of memory capacity. To address this issue, Long Short-Term Memory is employed to enhance the memory of the classification algorithm. The RNN–LSTM classification algorithm is constructed by combining several LSTM cells. A distinctive feature of LSTM is that it can operate in either a forward or backward direction. For bidirectional learning, the Bidirectional LSTM (BiLSTM) is utilized, which comprises two LSTM layers, each serving a different and opposing function. The first LSTM layer functions in the forward direction, while the second operates in the backward direction. The underlying principle of BiLSTM lies in its ability to capture features from both past and future inputs, thereby enhancing the algorithm's performance. In the BiLSTM architecture, one layer generates the forward hidden state $\underset{h}{\rightarrow}$ and cell state $\underset{c}{\rightarrow}$ for the input, while the other layer produces the reverse order of the hidden state $\underset{h}{\leftarrow}$ and cell state $\underset{c}{\leftarrow}$. Both $\underset{h}{\rightarrow}$ and $\underset{h}{\leftarrow}$ are concatenated to create an output sequence for the BiLSTM layer, as expressed in Eq. (4). Similarly, the combination of both $\underset{c}{\rightarrow}$ and $\underset{c}{\leftarrow}$ forms an output sequence for the BiLSTM layer, as illustrated in Eq. (5) [51]. This dual-layer approach allows the BiLSTM to capture contextual information in both forward and reverse directions, enhancing its ability to understand dependencies and patterns within the input sequence.

$$y_t = W_{\underset{h}{\rightarrow}t} \underset{h}{\rightarrow}_t + W_{\underset{h}{\leftarrow}t} \underset{h}{\leftarrow}_t + b_y \tag{4}$$

$$y_t = W_{\underset{c}{\rightarrow}t} \underset{c}{\rightarrow}_t + W_{\underset{c}{\leftarrow}t} \underset{c}{\leftarrow}_t + b_y \tag{5}$$

4 Results and Discussion

Speech signals encapsulate various attributes of the speaker, offering insights into their physical and mental state, including factors such as gender, sentiments, health status, and age. Among these, gender identification, derived from speech signal analysis, finds numerous practical applications. In the current study, Mel Frequency Cepstral Coefficients (MFCC) are harnessed as discriminative features within the speech signals. These MFCCs are integrated into an RNN–BiLSTM classifier to gauge the precision of gender identification within the system. It is imperative to acknowledge that the accuracy of gender identification for speakers hinges upon several factors, including the volume of testing samples, the linguistic characteristics, and the accents prevalent among the speakers. The training dataset is stratified into two primary categories, reflecting distinct geographical areas. Category 1 encompasses speech samples from speakers located outside the Indian subcontinent, whereas Category 2 comprises speech samples from Indian speakers. Similarly, the testing dataset is stratified into four distinct sets, each characterized by diverse accents and languages in the speech samples, enabling a comprehensive assessment of gender identification accuracy. The evaluation of classifier performance can be conducted using a confusion matrix, as outlined in Eq. (6), to ascertain the degree of identification accuracy achieved.

$$\text{Overall Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100 \quad (6)$$

where TP is True Positive, TN is True Negative, FN is False Negative and FP is False Positive.

In order to elucidate and deliberate upon the ultimate objective of the proposed endeavor, the simulation's results have been expounded upon in a comprehensive manner, delineated across four distinct sections, namely: Case–1, Case–2, Case–3, and Case–4. Case–1 expounds upon the accuracy of gender identification, calculated through the utilization of training samples sourced externally from India. Case–2 delineates the accuracy of gender identification, ascertained through the employment of training samples sourced specifically from Indian citizens. The juxtaposition of the outcomes of Case–1 and Case–2, as evaluated against the respective testing datasets, is elucidated within Case–3. The conclusive findings of this study are intricately examined and discussed within the confines of Case–4.

Case–1: Analysis of gender identification accuracy for Category–1

Table 2 presents the computed values for gender identification accuracy following the analysis of speech signals within Set 1. To derive the values in Table 2, the classification model undergoes training using speech samples from Category 1. Specifically, Set 1, comprising downloaded speech samples from regions beyond the Indian subcontinent, is designated for testing purposes. Notably, the number of testing samples within Set 1 is systematically altered to provide clarity regarding the outcomes as detailed in Table 2.

Table 2: Gender identification accuracy for Set 1–Category–1

Set 1	No. of testing samples
	40
Gender identification accuracy	88.78

Table 3 displays the computed values for gender identification accuracy concerning the testing samples within Set 2. For the generation of these values, speech samples are utilized from Category 1 as the training dataset to ascertain the gender of speakers within Set 2. Notably, Set 2 comprises recorded speech samples from typical Indian speakers conversing in the English language. The diversity in the quantity of speech samples within Set 2 was intentionally introduced to facilitate a comprehensive simulation. The results demonstrate that gender identification accuracy varies in response to alterations in the number of testing samples, as detailed in Table 3. Moreover, Tables 4 and 5 provide the computed values of gender identification accuracy for Set 3 and Set 4, respectively. Set 3 comprises recorded Hindi speech samples from speakers within the Indian continent. In contrast, Set 4 encompasses speech samples from speakers of mixed backgrounds, including those from both the Indian continent and beyond. These observations reveal that gender identification accuracy exhibits fluctuations corresponding to variations in the accent of testing samples across all testing sets, as indicated in Tables 4 and 5.

Table 3: Gender identification accuracy for Set 2–Category–1

Set 2	No. of testing samples
	40
Gender identification accuracy	77.59

Table 4: Gender identification accuracy for Set 3–Category–1

Set 3	No. of testing samples
	40
Gender identification accuracy	82.47

Table 5: Gender identification accuracy for Set 4–Category–1

Set 4	No. of testing samples
	40
Gender identification accuracy	83.97

Case–2: Analysis of gender identification accuracy for Category–2

The calculation of gender identification accuracy is based on the training samples from Category 2, which exclusively pertain to the Indian continent. The speech samples within CategorAnalysis of gender identification accuracyy 2 encompass both Hindi and English languages. To ensure a clear and fair comparison of the training samples, the testing datasets used align with those in Case 1. Tables 6 and 7 present the computed gender identification accuracy for the testing samples within Set 1 and Set 2, respectively. It is important to note that the gender identification accuracy fluctuates in response to changes in the accent of testing samples. This observation underscores the significance of the quantity of testing samples in determining the accuracy of gender identification, as revealed in Tables 6 and 7.

Table 6: Gender identification accuracy for Set 1–Category–2

Set 1	No. of testing samples
	40
Gender identification accuracy	59.97

Table 7: Gender identification accuracy for Set 2–Category–2

Set 2	No. of testing samples
	40
Gender identification accuracy	96.73

Similarly, gender identification accuracy has been computed for Sets 3 and 4. The results for gender identification accuracy in Set 3 are detailed in Table 8, where the speech samples are from

the Indian continent, specifically in the Hindi language. On the other hand, Set 4 comprises mixed speech samples in terms of both language and continent. Consistent with the prior observations, the identification accuracy exhibits variations that correlate with changes in the accent of testing samples. These fluctuations are elaborated upon in Table 6 through Table 9, further emphasizing the significance of the quantity of testing samples in determining the accuracy of gender identification across different scenarios.

Table 8: Gender identification accuracy for Set 3–Category–2

Set 3	No. of testing samples
	40
Gender identification accuracy	98.89

Table 9: Gender identification accuracy for Set 3–Category–2

Set 4	No. of testing samples
	40
Gender identification accuracy	87.33

Case–3: Category wise comparison of gender identification accuracy for different testing datasets

Fig. 2 provides a comparative analysis of the average computed accuracy for gender identification between the two categories of training samples in the context of testing samples within Set 1. This visual representation clearly demonstrates that, as depicted in Fig. 2, the average accuracy for gender identification in Set 1 is notably higher for Category 1 compared to Category 2. In a similar vein, Fig. 3 presents a comparison of the average gender identification accuracy for Set 2. Fig. 3 emphasizes the distinction between the two training sample categories. In this instance, the visual representation shows that the average gender identification accuracy for Category 2 surpasses that of Category 1. These figures effectively illustrate the differential performance of the two training sample categories in the context of gender identification accuracy for the respective testing sample sets.

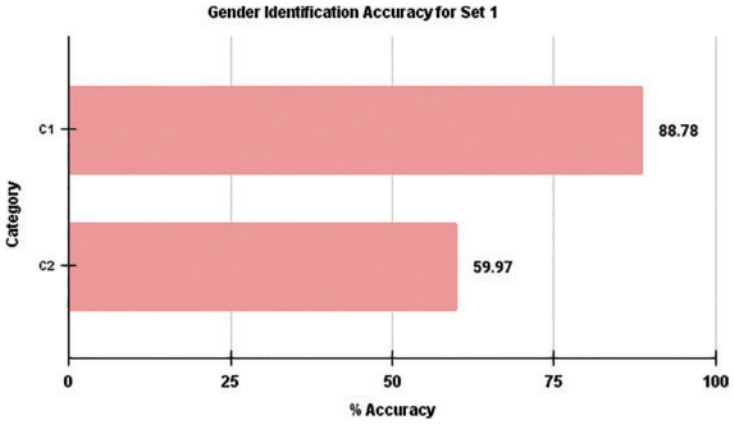

Figure 2: Comparison of average gender identification accuracy of Set 1 for Category–1 and Category–2

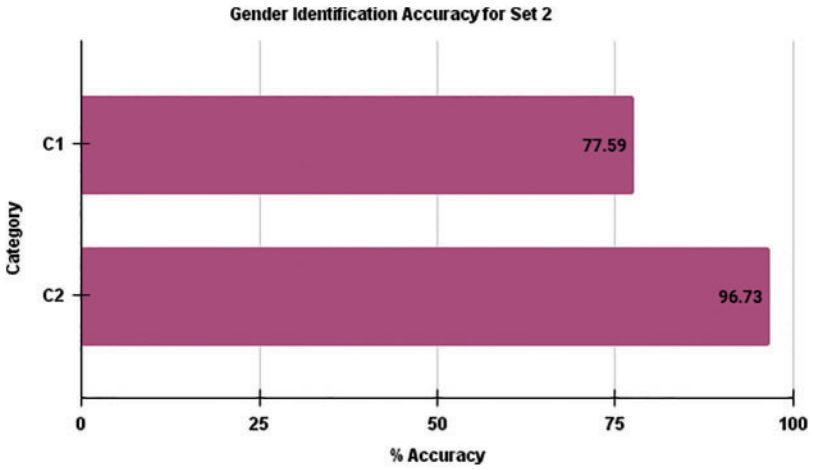

Figure 3: Comparison of average gender identification accuracy of Set 2 for Category–1 and Category–2

Fig. 4 illustrates the comparison of average gender identification accuracy between the two categories of training samples for recorded Hindi language in Set 3. In this analysis, it is evident that Category 2 outperforms Category 1, achieving a higher accuracy level for gender identification. Similarly, Fig. 5 provides a comparison of the calculated average accuracy of gender identification between the two categories of training speech samples for Set 4, which contains mixed language samples, including Hindi and English. Consistently with the previous observations, Fig. 5 reaffirms that Category 2 consistently exhibits higher gender identification accuracy compared to Category 1. Taken together, the findings presented in Figs. 4 and 5 emphasize the superior performance of Category 2 in terms of gender identification accuracy across various testing sample sets.

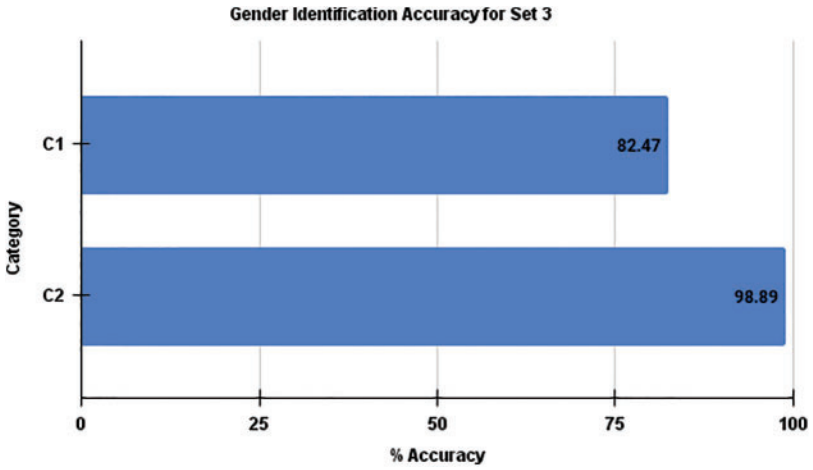

Figure 4: Comparison of average gender identification accuracy of Set 3 for Category–1 and Category–2

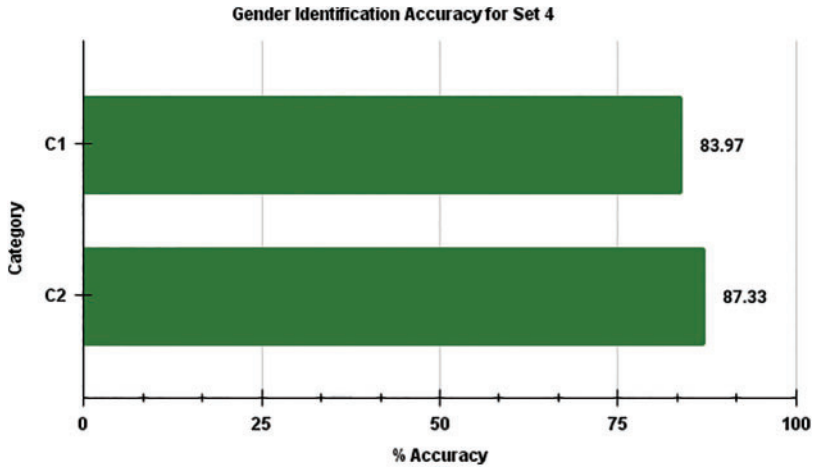

Figure 5: Comparison of average gender identification accuracy of Set 4 for Category–1 and Category–2

Case–4: Comparison of gender identification accuracy for Category–1 and Category–2

The overall average gender identification accuracy for each category has been computed by considering all four testing datasets. Fig. 6 provides a comprehensive comparison of gender identification accuracy between Category 1 and Category 2. The findings of the study unveil a consistent gender identification accuracy of 85.73% for Category 2, surpassing the accuracy of 83.20% observed in Category 1. This disparity can be ascribed to the inherent linguistic capabilities of speakers within each category. The linguistic diversity within Category 2 potentially contributes to a more distinct and identifiable pattern for the gender identification algorithm. These results underscore the significance of linguistic characteristics in refining and optimizing gender identification systems, emphasizing the need for a nuanced understanding of language dynamics for more accurate and culturally sensitive models. Notably, speakers in Category 2 possess the ability to pronounce a wider range of phonemes, including 52 Hindi phonemes and 26 English phonemes, which contributes to their superior performance. In contrast, Category 1 speakers are proficient in pronouncing only 26 English phonemes. Moreover, the analysis underscores that gender identification accuracy is influenced by both the geographical area of the speaker and the language they speak. The speaker's accent can also vary depending on their pronunciation abilities with different phonemes. As a conclusion, enhancing the accuracy of the system can be achieved by mitigating noise during the recording of speech signals. In summary, Fig. 6 demonstrates the favorable performance of Category 2 in gender identification, highlighting the interplay of geographical location, language, and phonemic proficiency as factors influencing accuracy in this context.

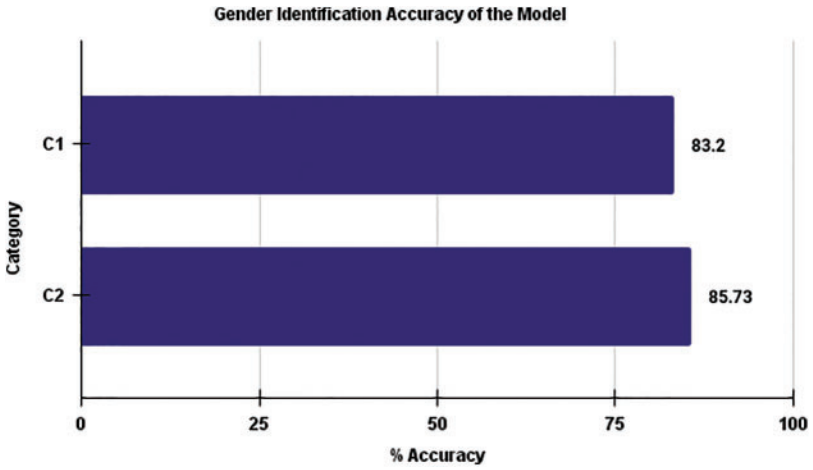

Figure 6: Comparison of overall average gender identification accuracy for Category–1 and Category–2

5 Conclusion

The presented article focuses on computing gender identification accuracy through speech signal analysis. The comparison of gender identification accuracy is conducted based on both the language spoken by the speakers and their geographical location. To achieve this objective, Mel Frequency Cepstral Coefficients (MFCCs) are utilized as a feature extracted from the speech signals, and an RNN–BiLSTM deep learning algorithm is employed for gender classification. The speech samples are systematically organized based on language and geographical origin to facilitate algorithmic testing. The training dataset is categorized into two groups: Category 1 comprises speech samples from speakers outside of India, while Category 2 consists of speech samples from Indian citizens. The results demonstrate that Category 2 speakers achieve an average gender identification accuracy of 85.73%, surpassing the accuracy achieved for Category 1 speakers. The highest gender identification accuracy is observed for Indian citizens speaking in English, with an accuracy rate of 98.89%, whereas the lowest accuracy is noted for speakers from other continents, at 59.97%. These findings underscore the influence of phonemic proficiency on gender identification accuracy, as Indian speakers are capable of articulating 52 phonemes in Hindi and 26 in English, contributing to their superior performance. Furthermore, the article highlights the potential for further improving gender identification accuracy by incorporating additional factors such as speaker sentiments, health status, and age during speech sample recording. These attributes can offer valuable insights into distinguishing gender characteristics and further enhance the precision of the identification process. In conclusion, the study underscores the multifaceted nature of gender identification through speech signal analysis and the potential for leveraging additional features to refine the accuracy of the classification system. To enhance gender identification accuracy, speaker emotion is incorporated, recognizing the distinct emotional characteristics of males and females. Employing a majority rule strategy in the confusion matrix further improves system accuracy. Quality voice samples are pivotal; thus, minimizing noise during recording is essential. Additionally, the speaker's health condition, intertwined with age, is a crucial parameter influencing voice signals. Proposing a novel hybrid model for gender classification incorporating age, emotion, and health condition could provide a comprehensive and sophisticated framework, potentially advancing the accuracy and depth of gender identification systems.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding study for this study.

Author Contributions: Study conception and design: Abhishek Singhal and Devendra Kumar Sharma, data collection: Abhishek Singhal, analysis and interpretation of results: Abhishek Singhal and Devendra Kumar Sharma; draft manuscript preparation: Abhishek Singhal and Devendra Kumar Sharma. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data is openly available in a public repository. The data that support the findings of this study are openly available at https://commonvoice.mozilla.org/en/datasets. Recorded data is not available due to ethical restrictions. Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Agrawal and S. Rathor, "A robust model for domain recognition of acoustic communication using Bidirectional LSTM and deep neural network," *Neural Computer & Application*, vol. 33, pp. 11223–11232, 2021.

[2] S. J. Chaudhari and R. Kagalkar, "Methodology for gender identification, classification and recognition of human age," in *Int. Journal of Computer Applications*, pp. 5–10, 2015.

[3] E. Ramdinmawii and V. K. Mittal, "Gender identification from speech signal by examining the speech production characteristics," in *2016 Int. Conf. on Signal Processing and Communication (ICSC)*, Noida, India, pp. 244–249, 2016. https://doi.org/10.1109/ICSPCom.2016.7980584

[4] V. Y. Mali and B. G. Patil, "Human gender classification using machine learning," *International Journal of Engineering Research & Technology*, vol. 8, pp. 474– 477, 2019.

[5] V. Passricha and R. K. Aggarwal, "Convolutional support vector machines for speech recognition," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 601–609, 2019.

[6] V. Patil, R. Vineetha, S. Vatsa, D. K. Shetty, A. Raju *et al.,* "Artificial neural network for gender determination using mandibular morphometric parameters: A comparative retrospective study," *Cogent Engineering*, vol. 7, pp. 1–12, 2020.

[7] O. S. Faragallah, "Robust noise MKMFCC-SVM automatic speaker identification," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 185–192, 2018.

[8] A. Acero and X. Huang, "Speaker and gender normalization for continuous-density hidden Markov models," in *1996 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing Conf. Proc.*, Atlanta, GA, USA, pp. 342–345, 1996. https://doi.org/10.1109/ICASSP.1996.541102

[9] R. Djemili, H. Bourouba and A. Korba, "A combination approach of Gaussian mixture models and SVMs for speaker identification," *The International Arab Journal of Information Technology*, vol. 6, no. 5, pp. 490–497, 2009.

[10] M. K. Mishar and A. K. Shukla, "A survey paper on gender identification system using speech signal," *International Journal of Engineering and Advanced Technology*, vol. 6, pp. 165–167, 2017.

[11] M. Gupta, S. S. Bharti and S. Agarwal, "Gender-based speaker recognition from speech signals using GMM model," *Modern Physics Letters B*, vol. 33, no. 35, pp. 1950438, 2019.

[12] A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1421–1432, 2014.

[13] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[14] O. Mamyrbayev, A. Toleu, G. Tolegen and M. Nurbapa, "Neural architectures for gender detection and speaker identification," *Cogent Engineering*, vol. 7, no. 1, pp. 1727168, 2020.
[15] S. J. Chaudhari and R. M. Kagalkar, "Automatic speaker age estimation and gender dependent emotion recognition," *International Journal of Computer Applications*, vol. 117, no. 17, pp. 0975– 8887, 2015.
[16] K. H. Lee, S. Kang, D. H. Kim and H. Chang, "A support vector machine based gender identification using speech signal," *IEICE Transactions on Communications*, vol. E91-B, no. 10, pp. 3326–3329, 2008.
[17] R. R. Rao, "Source feature based gender identification system using GMM," *International Journal on Computer Science and Engineering*, vol. 3, no. 2, pp. 586–593, 2011.
[18] R. R. Rao and A. Prasad, "Glottal excitation feature based gender identification system using ergodic HMM," *International Journal of Computer Applications*, vol. 17, pp. 31–36, 2011.
[19] J. Villalba., N. Chen, D. Snyder, D. Garcia-Romero, A. McCree *et al.,* "State-of-the-art speaker recognition for telephone and video speech," in *Proc. of the Interspeech 2019*, Graz, Austria, 1488-1492, 2019.
[20] C. Muller, F. Wittig and J. Baus, "Exploiting speech for recognizing elderly users to respond to their special needs," in *Proc. of Europspeech*, Geneva, Switzerland, pp. 1305–1308, 2003.
[21] H. J. Kim, K. Bae and H. S. Yoon, "Age and gender classification for a home-robot service," in *RO-MAN 2007–The 16th IEEE Int. Symp. on Robot and Human Interactive Communication*, Jeju, Korea (South), pp. 122–126, 2007. https://doi.org/10.1109/ROMAN.2007.4415065
[22] 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," *International Journal of Speech Technology*, vol. 21, pp. 921–941, 2018.
[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.
[24] B. Bsir and M. Zrigui, "Bidirectional LSTM for author gender identification," *in ICCCI 2018: Computational Collective Intelligence*, pp. 393–402, 2011.
[25] L. Stout, R. Musters and C. Pool, "Author profiling based on text and images," in *Proc. of the Ninth Int. Conf. of the CLEF Association (CLEF 2018)*, Avignon, France, 2018.
[26] B. Bsir and M. Zrigui, "Gender identification: A comparative study of deep learning architectures," *Intelligent Systems Design and Applications*, vol. 941, pp. 792–800, 2018.
[27] D. Bhardwaj and R. K. Galav, "Identification of speech signal in moving objects using artificial neural network system," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 4, pp. 418–424, 2020.
[28] I. E. Livieris, E. Pintelas and P. Pintelas, "Gender recognition by voice using an improved self-labeled algorithm," *Machine Learning & Knowledge Extraction*, vol. 1, no. 1, pp. 492–503, 2019.
[29] M. Buyukyilmaz and A. Osman, "Voice gender recognizer using deep learning," in *Int. Conf. on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016)*, Xiamen, China, pp. 409–441, 2016.
[30] R. V. Sharan and T. J. Moir, "Robust acoustic event classification using deep neural networks," *Information Sciences*, vol. 396, pp. 24–32, 2017.
[31] D. Ö Batur and N. Aydin, "An optimal feature parameter set based on gated recurrent unit recurrent neural networks for speech segment detection," *Applied Sciences*, vol. 10, no. 4, pp. 1273, 2020.
[32] A. A. Alnuaim, M. Zakariah, C. Shashidhar, W. A. Hatamleh, H. Tarazi *et al.,* "Speaker gender recognition based on deep neural networks and ResNet50," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 4444388, 2022.
[33] K. Chachadi and S. R. Nirmala, "Voice-based gender recognition using neural network," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces*, pp. 741–749, Singapore: Springer, 2022.
[34] 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.

[35] R. Saloni, K. Sharma and A. K. Gupta, "Classification of high blood pressure persons vs normal blood pressure persons using voice analysis," *International Journal Image, Graphics and Signal Processing*, vol. 6, no. 1, pp. 47–52, 2013.
[36] 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," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 167–183, 2018.
[37] O. S. Faragallah, "Robust noise MKMFCC-SVM automatic speaker identification," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 185–192, 2018.
[38] A. Bala, A. Kumar and N. Birla, "Voice command recognition system based on MFCC and DTW," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7335–7342, 2010.
[39] A. A. Abdulsatar, Y. V. Davydov, V. V. Yushkova, A. P. Glinushkin and V. Y. Rud, "Age and gender recognition from speech signal," *Journal of Physics: Conf. Series*, vol. 1410, pp. 012073, 2019.
[40] G. Dobry, R. M. Hecht, M. Avigal and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1975–1985, 2011.
[41] R. Solera-Urena, A. L. Garcia-Moral, C. Pelaez-Moreno, M. Martinez-Ramon and F. Diaz-de-Maria, "Real-time robust automatic speech recognition using compact support vector machines," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1347–1361, 2012.
[42] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Proc. of the Interspeech*, Makuhari, Japan, pp. 2362–2365, 2010.
[43] Y. Niu, D. Zou, Y. Niu, Z. He and H. Tan, "A breakthrough in speech emotion recognition using deep retinal convolution neural networks," arXiv:1707.09917. 2017.
[44] M. W. Bhatti, Y. Wang and L. Guan, "A neural network approach for human emotion recognition in speech," in *Proc. of IEEE Int. Symp. Circuits Syst. (ISCAS)*, Vancouver, BC, Canada, vol. 2, pp. 81, 2004.
[45] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar *et al.,* "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.
[46] M. K. Reddy and K. S. Rao, "Excitation modelling using epoch features for statistical parametric speech synthesis," *Computer Speech & Language*, vol. 60, pp. 101029, 2020.
[47] Y. Liu, L. He, J. Liu and M. T. Johnson, "Introducing phonetic information to speaker embedding for speaker verification," *Journal on Audio, Speech, and Music*, vol. 19, pp. 1–17, 2019.
[48] A. Greco, A. Saggese, M. Vento and V. Vigilante, "A Convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff," *IEEE Access*, vol. 8, pp. 13077–130781, 2020.
[49] L. Jasuja, A. Rasool and G. Hajela, "Voice gender recognizer recognition of gender from voice using deep neural networks," in *2020 Int. Conf. on Smart Electronics and Communication (ICOSEC)*, Trichy, India, pp. 319–324, 2020. https://doi.org/10.1109/ICOSEC49089.2020.9215254
[50] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," arXiv:2012.03411. 2020.
[51] G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," in *2016 Int. Joint Conf. on Neural Networks (IJCNN)*, Vancouver, BC, Canada, pp. 3412–3419, 2016. https://doi.org/10.1109/IJCNN.2016.7727636