ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20220302182704
Original scientific paper

Voice-Based Gender Recognition Model Using FRT and Light GBM

Priya KANNAPIRAN*, Mohamed Mansoor Roomi SINDHA

Abstract: Voice-based gender recognition is vital in many computer-aided voice analysis applications like Human-Computer Interaction, fraudulent call identification, etc. A powerful feature is needed for training the machine learning model to discriminate a gender as male or female from a voice signal. This work proposes the use of a gradient boosting model in conjunction with a novel Cumulative Point Index (*CPI*) feature computed by Forward Rajan Transform (*FRT*) for gender recognition from voice signals. Firstly, voice signals are preprocessed to remove the nonsignificant silence period and are further framed and windowed to make them stationary. Then *CPI* is computed using the first coefficients of *FRT* and concatenated to form a feature set, and it is used to train the Light Gradient Boosting Machine (LightGBM) to recognize the gender. This approach provides better accuracy and faster training compared with the state of the art techniques. Experimental results show the primacy of the FRT_{CPI} over other standard features used in the literature. It is also shown that the proposed features, in combination with LightGBM, provide better accuracy of 95.26% with a less computational time of 2.25 s over the challenging large datasets like Speech Accent Archive, Voice Gender Dataset, Common Voice, and Texas Instruments/Massachusetts Institute of Technology corpus.

Keywords: gender recognition; forward rajan transform; LightGBM; voice signal

1 INTRODUCTION

Gender recognition is an increasingly important research direction with the continuous advancement of personalization and intelligence. It aims to develop Human-Computer Interaction (HCI), targeted advertising, fraudulent call prevention, voice analysis of forensic applications, cryptography, multimedia retrieval, etc., [1-3]. Image-dependent or speech/voice-dependent approaches are widely used in the gender recognition system. In an image-dependent system, facial features with additional information, viz., eyebrows, body structure, dressing sense, and hairstyle, provide important visual cues to recognize the gender. They rely on distinguishable features between the face images of males and females. These features are extracted by computational methods and classified using various classifiers. Gender recognition from face images comprises preprocessing feature extraction and recognition. The first step follows preprocessing procedures like resizing, geometry alignment, filtering, noise removal, histogram equalization, etc.; in the feature extraction phase, the optimal facial features are extracted using shape, color, textural, and structural-based techniques.

The features used in image dependant gender recognition are Raw pixels, Haar-based features, Independent Component Analysis (IDA), Principal Component Analysis (PCA), Local Binary Pattern (LBP), Gabor, Discrete Cosine Transform (DCT), Scale Invariant Feature Transform (SIFT), Histogram Oriented Gradients (HOG), Weber's Law Descriptor (WLD), etc., The classifiers such as Decision tree, Support Vector Machine (SVM), ensembles of Radial Basis Functions (RBFs), Adaboost, Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Gaussian Mixture Model (GMM) and Random Forest are frequently used in this gender recognition system [4].

The success rate of gender recognition is limited by factors like illumination variation, pose, expression, occlusion, ethnicity, and age. Moreover, the occluding objects like hats, spectacles, jewelry, and glasses also affect the success rate of gender recognition. Besides, the image quality, like noise and resolution during image acquisition, also affects the recognition rate. Gender recognition based on voice is proposed to overcome these issues for better recognition accuracy [5].

Most gender recognition systems rely on acoustic or spectral features, such as fundamental frequency, signal length, amplitude, mean, standard deviation, skew, Mel Frequency Cepstral Coefficient (MFCC), spectral centroid, spectral flatness, entropy, energy, short time energy, formant, Zero Crossing Rate (ZCR), etc. These features are classified using deep learning and machine learning paradigms [6-8]. Various machine learning paradigms employed to identify the gender from voice signals are SVM, K Nearest Neighbor (KNN), Gradient Boosting (GB), Gaussian Mixture Modeling (GMM), and Decision Tree [6-8]. Humans can identify their gender by hearing their voices within seconds, but machines require 2 to 3 seconds of voice samples [9]. The iCST voting algorithm proposed by Ioannis et al. [10] is based on an ensemble semi-supervised self-labeled algorithm that is self-taught and shows better performance than base learners and other supervised learning algorithms. Manish Gupta et al. [11] presented work on gender-based speaker recognition using the GMM model, concluding from experimental results that a 1.12-second voice signal without silence is enough to recognize gender.

LindasawaMuda et al. [12] proposed a voice recognition method using MFCC features and a Dynamic Time Warping (DTW)- matching algorithm. Jamid et al. [13] used MFCC features and five schemes recognition algorithms to find the gender of the telephonic speech signal, and Chaudhary, et al. [14] combined pitch, energy, and MFCC features and used an SVM classifier to recognize gender. The acoustic features of MFCC and first-order and second-order derivative functions of MFCC are called delta, and delta-delta features are classified by SVM for the Vowel database [15]. Archana et al. [16] presented an Artificial Neural Network (ANN) model for classifying MFFC features with frame energy for the real-time audio database. Apeksha shewalker et al. [17] evaluated Recurrent Neural Network (RNN), LSTM, and Gated Recurrent Unit (GRU). The LSTM network shows better results in word error rate, and GPU yields faster optimization. Aravind et al. [18] proposed a voice recognition system and described issues related to the large dataset. Manjunath et al. [19] presented an acoustic event classification model using new features extracted from spectrogram blocks. A hybrid model using 1-D Stationary Wavelet Transform (SWT) for signal denoising and reconstruction was given to ANN for gender recognition developed by Yasin et al. [20]. Anna V. Kuchebo et al. [21] presented a deep learning network to classify gender from speech using the Mel spectrogram of speech of the Mozilla voice dataset. Mohammad Amaz et al. [22] extracted three-layer features from the acoustic features such as pitch, spectral flat, spectral entropy, MFCC, and Linear Predictive Coding (LPC). One Dimensional convolutional neural network classified these features into male or female gender.

It is easy for individuals to recognize gender, but for machines, it is difficult due to environmental factors, accessories, and low image quality. As a result, gender recognition from voice is being pursued in the research field to provide explicit gender recognition. Many methods for identifying gender from voice signals have been proposed; however, generating optimal feature sets and developing high-performance classification techniques has proven to be complicated. Most research has been conducted on small databases. This means there is a need for strong feature representation and classification in the recognition of gender in large datasets. In this paper, section 2 describes dataset collections. The proposed methodology of gender recognition is explained in section 3. In section 4, the results of the experimental work are enlarged and deal with state-of-the-art techniques. The conclusion and future direction are explained in section 5.

The significant contributions of the proposed work are as follows:

- Collected the popular, challenging, large voice dataset for gender recognition.
- Performed frame analysis on the input voice signal to determine the length of the shortest voice signal.
- Introduction of new 1-D *FRT* feature and selection of Cumulative Point Index (*CPI*)-based *FRT* features for gender recognition
- Recognition of gender by LightGBM gradient boosting framework with FRT_{CPI} feature.
- Performance assessment on collected dataset against various machine learning algorithms and state-of-the-art techniques.

2 DATABASES

In contrast to contemporary research that uses small datasets, the proposed study makes use of publicly available large datasets such as Speech Accent Archive (SAA) [23], Common Voice (CV) [24], TIMIT Acoustic Phonetic Continuous Speech Corpus (TIMIT-APCSC) [25], and Voice Gender Dataset (VGD) [26, 27]. The SAA database contains a large set of speech accents in 250 native languages, and it is downloaded from Kaggle.com. The voice files in this database have native and non-native English-speaking people reading the same paragraph. The CV database is freely accessed through kaggle.com and created for usable voice technology for machines. The TIMIT speech data is designed for acoustic-phonetic studies in speech recognition systems and is accessible through Deepai website. The VGD contains the sample speech from the gender recognition datasets such as Telecommunications & Signal Processing Laboratory Speech Database at McGill University and VoxForge Speech Corpus. These databases aim to facilitate research into Automatic Speech Recognition (ASR). Except for the SAA database, the collected voice datasets have enormous files and unequal gender distribution. The details of the collected voice databases are given in Tab. 1.

Table 1 Collected voice database details

Database Name	Sampling Rate / kHz	Number of voice samples		
		Male	Female	Total
SAA	44.1	1102	1057	2159
CV	48	961	290	1021
TIMIT	44.1	438	192	630
VGD	44.1	355	235	570

2.1 Ethical Clearance

This work on collected speech samples uses public domain databases like SAA, CV, VGD, TIMIT Acoustic Phonetic Continuous Speech Corpus (TIMIT-APCSC) and does not force threat or risk to the human beings exempting from the ethical concerns; however, we have added the details regarding the availability of the agreement and licensing terms of use of each of the database for the benefit of reader given below.
SAA: https://accent.gmu.edu/about.php
TIMIT: https://catalog.ldc.upenn.edu/license/ldc-non-members-agreement.pdf
CV: https://www.mozilla.org/enUS/MPL/license-policy
VGD: http://www.voxforge.org/home
http://www-mmsp.ece.mcgill.ca/Documents/Data/TSP-Speech-Database/Licence.html

3 PROPOSED METHODOLOGY

The proposed work presents a computational framework for classifying a person's voice signal as male or female using innovative *FRT* features. Fig. 1 depicts the method's overall flow. Pre-processing, frame and windowing, feature extraction, and recognition are among the four processes included in this proposal. Preprocessing techniques include silence removal [28] and pre-emphasis [29], which removes the undesired signal. Consequently, framing [30] and windowing techniques [31] are also implemented to make the input signal stationary. Then the *CPI*-based *FRT* features [32] are extracted from the framing and windowing signal, which provides the signal's sparse representation [33]. Light GBM is used to classify the extracted features in the training phase, coupled with a ground truth label. The gradient boosting framework improves performance, memory utilization, and accuracy on unbalanced datasets. The query voice signal is male or female in the testing phase.

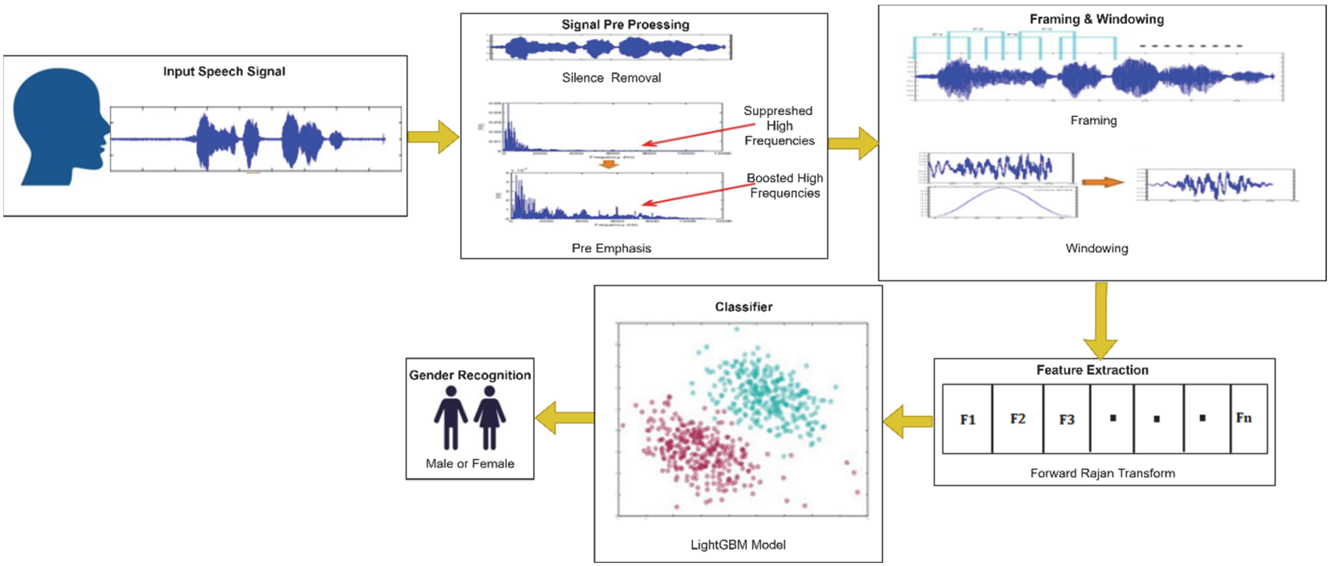

Figure1 Proposed framework for gender recognition

3.1 Hypothesis

The data present in the audio stream is presumed to comprise both silence and voice signals in this study. Eq. (1) is the mathematical formulation of the proposed system.

$$I(n)=S(n)+\rho_0(n) \tag{1}$$

where $I(n)$ is the audio signal, $S(n)$ is the voice signal and $\rho_0(n)$ is the silence signal.

$$\mathrm{H}_0\colon I(n)=\rho_0(n) \tag{2}$$

$$\mathrm{H}_1\colon I(n)=S_1(n)+\rho_0(n) \tag{3}$$

$$\mathrm{H}_2\colon I(n)=S_2(n)+\rho_0(n) \tag{4}$$

The null hypothesis H_0 represented in Eq. (2) denotes the audio stream that contains a silent signal that does not require any further processing. The voice signal $S_1(n)$ and $S_2(n)$ in the alternate hypothesis H_1 and H_2 represents male and female respectively, shown in Eqs. (3) and (4).

$$F_m(y) \underset{\mathrm{H}_2}{\overset{\mathrm{H}_1}{\lessgtr}} \theta_m \tag{5}$$

The proposed method decides the hypothesis H_1 or H_2 based on the θ_m denoted in Eq. (5) obtained from the Light GBM algorithm discussed in section 3.5 below

3.2 Signal Pre-Processing

Signal pre-processing steps include silence removal to remove the unwanted signal and pre-emphasis to boost the magnitude of higher frequency energies.

3.2.1 Silence Removal

Non-voice content, such as silence or noise, may be present in the input voice signal. This is not important because it does not contain any information. The energy level of non-voice and noise are very low when compared with the voice signal. Hence removing silence and noise is an essential process to improve the performance of the proposed method. In this work, the thresholding algorithm based on short-time energy (E) (Eq. (6)) and spectral centroid (C_S) (Eq. (7)) is used to remove the non-voice content and noise from the input voice signal $I(n)$.

$$E=\frac{1}{N}\sum_{n=1}^{N}\left|I(n)\right|^2 \tag{6}$$

$$C_S=\frac{\sum_{k=1}^{K}(k+1)I(k)}{\sum_{k=1}^{K}I(k)} \tag{7}$$

where N is the number of samples in the signal. Where $I(k), k\in[1,K]$ are the Discrete Fourier Transform (DFT) coefficients of the input signal.

3.2.2 Pre-Emphasis

The higher frequency component of voice has low energy values, making it difficult for humans to discern speech. The pre-emphasis process boosts the higher frequency energies from the silence removed signal $\hat{I}(n)$ using a first-order FIR filter with coefficient alpha (α) shown in Eq. (8). It reduced the high spectral dynamic range and flattens the spectrum.

$$y(n)=\hat{I}(n)-\alpha\hat{I}(n-1) \qquad 0<\alpha<1 \tag{8}$$

3.3 Framing and Windowing

Voice is generally a nonstationary signal with prosody, phoneme, and vocal tract variations. Over time the signal characteristics change to reflect the changes in speech sound. It is, therefore, critical to change it to stationary by converting it into a short period as a frame. The pre-emphasized signal is divided into frames of N samples. Initially, the number of samples in each frame (N) is calculated by multiplying the frame time T_d by the sampling frequency f_s as given in Eq. (9), followed by calculating the number of samples in each frame overlapping with M samples given in Eq. (10). Fig. 2 shows the framing and windowing function of the voice signal.

$$N = T_d \times f_s \tag{9}$$

$$M = \frac{N \times T_s}{T_d} \tag{10}$$

$$N_f = \text{floor}\left(\frac{\text{length}\left(y(n)\right) - N}{M+1}\right) \tag{11}$$

$$s(n) = \left\{s_1 \middle| s_2 \middle| s_3 \ldots . s_n\right\} \tag{12}$$

Then the adjacent frames are split by using M samples ($N > M$) which are the different samples between the consequence frame, and the number of frames (N_f) in each voice signal is calculated by Eq. (11). Finally, the framed signal $s(n)$ is derived as in Eq. (12).

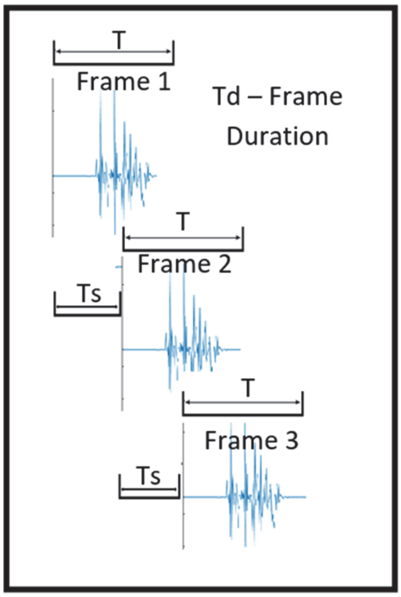

Figure 2 Framing and windowing of pre-emphasized signal

$$S_w(n) = s(n) \times w(n) \text{ where } n = 1, 2 \ldots . N_f \tag{13}$$

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

The windowing technique or tapered function is used on short-time signals after framing to eliminate spectral distortions in the signal represented in Eq. (13). According to this technique, if a certain interval is selected, it will return a value that is both zero and non-zero value at the outside and inside of that interval respectively. A major effect of windowing is that the discontinuities of the frequency response are converted into transition bands between values on either side of the discontinuity. Various windowing techniques are available, including rectangular, hanning, hamming, Blackman, and others. The hamming window $w(n)$ was chosen for its superior performance in changing the signal regularly and minimizing the discontinuities in the edges, as shown in Eq. (14). The framed signal $s(n)$ is multiplied by the hamming windowing $w(n)$ function to get the hamming window signal $S_w(n)$.

3.4 Feature Extraction by Forward Rajan Transform

Forward Rajan Transform (*FRT*) is used to extract the feature, which provides both loss and lossless sparse representation to achieve compactness. The sparse representation of the signal is $X \epsilon R\text{^}N$, where X is the signal vector, and this vector consists of a smaller number of non-zero coefficients and a large number of zero coefficients. The non-zero coefficients and locations are used to reconstruct the sparse signal effectively. This is known as a sparse approximation and serves as the foundation for transform coding. This transform is developed based on Decimation in Time Fast Fourier Transform (DIF-FFT). It generates an output sequence with key values highly related to the input sequence. Using these key values and output sequences, the input sequence is obtained by reverse RT. Both the input and output sequences have the same length (D). The condition for RT is the length of the input should be a power of two. This method is repeated until no more divisions are applied, and the number of stages G is given as Eq. (15).

$$G = \log_2 D \tag{15}$$

First, the sequence $S_w(n)$ is divided into two halves with a $D/2$ length, as shown in Eqs. (16) and (17).

$$S_{w1}(m) = \\ S_w(n) + S_w\left(n + \frac{D}{2}\right); 0 \le m \le \frac{D}{2}; 0 \le n \le \frac{D}{2} \tag{16}$$

$$S_{w2}(m) = \\ \left|S_w(n) - S_w\left(n + \frac{D}{2}\right)\right|; 0 \le m \le \frac{D}{2}; \frac{D}{2} \le n \le D \tag{17}$$

Then, arithmetical operations like summation and difference are applied to these sequences to produce the output sequence with the $D/2$ length. Then each $D/2$ length sequence is divided into two $D/4$ length sequences. This process is continued until it reaches the G stage. The *FRT* is applied to all the sequences. It is the permutation invariant of RT. The input of *FRT* is an integer, real or complex, and rational. It is applicable for both positive and negative values. The product of the R matrix and hamming window signal (S_w) gives the *FRT* coefficients as represented in Eq. (17). Generally, 23% of the RT coefficients have zero values, and many of the remaining values are nearly zero, representing signal sparsity. For example, the DC signal of amplitude 2 and length 8 gives the RT sequence as [16, 0, 0, 0, 0, 0, 0, 0]. This analysis shows that the first coefficient, the Cumulative Point Index (*CPI*), has a value of 16, and the remaining values are zero. This *CPI* value represents the feature characteristics of that signal to prove its sparsity. The first value in this spectrum is the *CPI*.

$$FRT = R_D \times (S_w)_{D\times 1} \tag{18}$$

where $R_D = \begin{bmatrix} I_{\frac{D}{2}} & I_{\frac{D}{2}} \\ -e_k I_{\frac{D}{2}} & e_k I_{\frac{D}{2}} \end{bmatrix}_{D\times D}$ and

$$e_k = \begin{cases} -1;\text{ where } k = 1 \text{ for } S_w\left(n + \dfrac{D}{2}\right) < S_w(n) \\ \qquad 0;\text{ otherwise} \end{cases}$$

$I_{D/2}$ the identity matrix. e_k is the key matrix of the Rajan transform.

$$FRT = \{C_1,\ C_2,\ \ldots..C_D\} \tag{19}$$

The *FRT* coefficients are obtained from Eq. (18), expressed as in Eq. (19), where the first coefficient (C_1) has been taken as *CPI*.

$$\hat{x}_f = \left[CPI_{1f}, CPI_{2f}, \ldots\ldots CPI_{qf}\right] \tag{20}$$

$$\bar{x}_i = \frac{1}{q}\sum_{n=1}^{q}\hat{x}_{qf}\, i = 1,2\ldots..N_f \tag{21}$$

$$FRT_{CPI} = \left[\left(\bar{x}_1, \bar{x}_2, \ldots..\bar{x}_{N_f}\right)\right] \tag{22}$$

Then Eq. (20) represents the *CPI* of the *FRT* feature that is extracted from each hamming windowed signal, where q is the ratio of the length of frame and length of *FRT*. The mean *CPI* value gives the CPI-based *FRT* ($\bar{x}_i$) features of one frame (Eq. (21)). So that the length of FRT_{CPI} features is the number of frames per signal obtained by Eq. (22).

3.4 Classification Using LightGBM

The extracted FRT_{CPI} features are classified by a gradient-boosting machine learning framework called LightGBM. It uses a tree-based learning algorithm (Gradient Boosting Decision Tree Algorithm - GBDT) and performs better than XGBoost [34]. It can handle a larger data size and takes lower memory to run. The use of LightGBM in this proposed methodology is due to its ability to modify the algorithm and increase the accuracy of the training model with unbalanced class databases. Since the collected databases have unbalanced class databases except for the SAA database, it is a series of linear combinations of sub-model. It uses the regression tree as a sub-model and adds the sub-model one by one to reduce the learner's loss function.

Initialize the LightGBM model with an initial loss function θ_0.

$$F_0(y) = \arg\min\left(\sum_{i=1}^{n} L\left(FRT_{CPI}, \theta_0\right)\right) \tag{23}$$

Then compute the negative gradient using Eq. (24).

$$\theta_{im} = -\left[\frac{\partial L\left(FRT_{CPI})_{\theta}, F(y_i)\right)}{\partial F(y)}\right] \tag{24}$$

Initialize $m = 1$ to M. M is the tree number, and θ_{im} is the residual or error. Fit a base learner $h_m(y)$ represents in Eq. (25) by updating residual error θ_{im}.

$$h_m(y) = \{y_i, \theta_{im}\} \tag{25}$$

$$\theta_m = \arg\min\sum_{i=1}^{n} L\left(x_i, F_{m-1}\left((FRT_{CPI})_i\right) + \theta\right) \tag{26}$$

where X is, the input vector L is the loss function, and θ is the value used to initialize the gradient boosting. The θ is dependent on the loss function. Update the model to split the features for the best solution using Eq. (26) and repeat this step until the convergence.

$$F_m(y) = F_{m-1}(y) + \theta_m h_m(y) \tag{27}$$

The extracted FRT_{CPI} features are classified by LightGBM and expressed as $F_m(y)$ in Eq. (27).

4 RESULTS AND DISCUSSION

In this work, 4380 voice samples were used to investigate a compact 1D feature based on *FRT* for gender recognition. The collected data are 2159 SAA files, 1021 CV files, 630 TIMIT files, and 570 VGD files. This algorithm was tested using Matlab 2020a installed on an Asus Rog Strix G731GT laptop with a nine-generation Intel i7 processor and NVIDIA GeForce GTX 1650 graphics card.

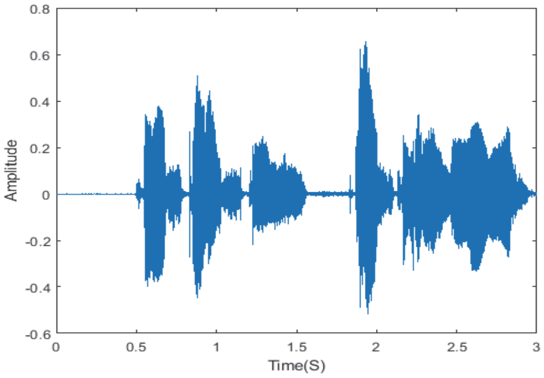

Figure 3 Input voice waveform

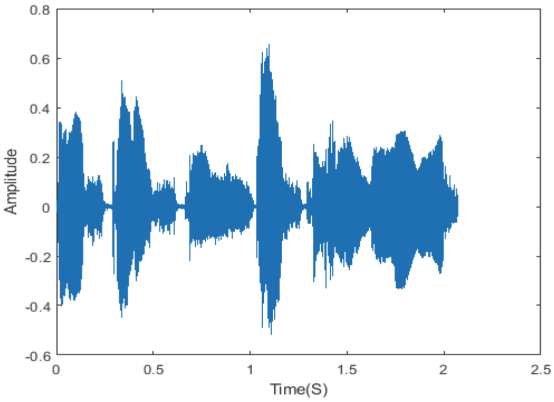

Figure 4 Silence removal waveform

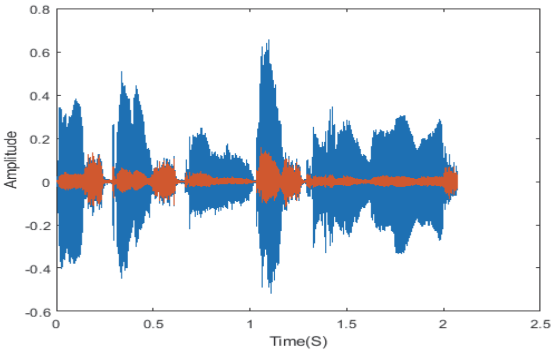

Figure 5 Pre-emphasised waveform

The voice signal is nonstationary, and it consists of noise interferences and silence or pause signals, as shown in Fig. 3. The preprocessing step removes the noise interference and silence by the thresholding algorithm. Fig. 4 shows the silence removed signal filtered by a pre-emphasis with a coefficient alpha of 0.95. The typical value of alpha is 0.95 to 0.98. This technique normalizes the signal by changing the signal's amplitude within the frame while not affecting the duration of the signal as shown in Fig. 5. The subsequent framing process segments the pre-emphasized signal into frames of varying framing lengths (10 ms, 20 ms, and 30 ms) and overlapping percentages (40, 50, and 60 percent).

The *FRT* features are extracted with these different combinations of framing lengths with frame overlapping and classified by the LightGBM classifier with default hyperparameter values. The hyperparameters such as maximum bin, number of leaves, learning rate, and number of iterations are used to avoid overfitting problems and improve the classification accuracy. Tab. 2 shows the experimental results of the performance of the proposed algorithm against collected databases. This experiment aims to find the shortest length of voice samples for which the accuracy obtained is stable. Initially, the different combinations have been executed in 1-second voice, achieving accuracy in the range of 88-95%. When the number of samples is increased to 1.15-second voice samples, it performs an accuracy from 89 to 96.45 at 20 ms framing length and 50% overlap.

Following frame analysis, the 1.15 s voice signal with 20 ms overlapping and 20% framing is selected for further processing. The hamming window function is applied to each framed signal to get a tapered signal. Then the features are extracted from the tapered signal using FRT.

Table 2 Experimental results of the performance of algorithm against collected databases

LightGBM Parameters: Max-bin (30), Number of leaves (31), Learning Rate (0.1), Numer of Iteration (100)			SAA	CV	TIMIT	VGD
Length of Voice / s	Framing / ms	Overlapping / %	Accuracy / %	Accuracy / %	Accuracy / %	Accuracy / %
1	10	40	88.23	91.30	94.30	90.14
		50	89.75	92.2	95.54	89.12
		60	88.92	93.02	96.80	88.12
	20	40	89.42	92.77	94.71	89.25
		50	89.78	93.22	93.85	89.40
		60	90.10	92.34	92.47	89.54
	30	40	88.23	92.22	96.25	89.75
		50	89.24	92.29	94.47	90.47
		60	90.04	92.79	95.10	90.32
1.15	10	40	90.14	93.90	95.80	90.12
		50	90.15	92.50	96.15	89.47
		60	90.42	93.42	96.42	89.45
	20	40	89.24	94.10	95.47	90.78
		50	**90.46**	**94.50**	**96.45**	**91.00**
		60	89.54	93.40	94.5	90.14
	30	40	89.05	94.05	95.25	90.47
		50	90.24	93.20	95.9	89.14
		60	88.57	93.50	96.05	90.14

For example, in the SAA database, the tapered signal has 120 shorts with an input sequence of $a(n) = \{17, 70, 5, 19, 54, 7, 32, 25\}$ with length 8 and it comprises 960 samples. For each length of sequences, the 3-stage *FRT* feature analysis was performed and it produces *FRT* features $A(K) = \{299, 13, 67, 1, 133, 5, 67, 47\}$. Out of all features, the mean, minimum, maximum, and FRT_{CPI} features are considered for the classification input. Among these, the FRT_{CPI} feature is classified as it contains more information for sparse representation. Finally, the 120 shorts have 120 FRT_{CPI} features, and the mean of this CPI coefficient is taken as the final FRT_{CPI} feature information. As a result, the gender recognition algorithm uses the CPI of the *FRT* feature. Since each frame consists of one coefficient, the length of FRT_{CPI} is equal to the number of frames shown in Tab. 3.

Table 3 Proposed input feature parameter

System	Frame size	Frame shift	Window	Type	No of Features
Proposed Methodology	20 ms	10 ms (50% Overlap)	Hamming	FRT_{CPI}	Number of Frames (N_f)

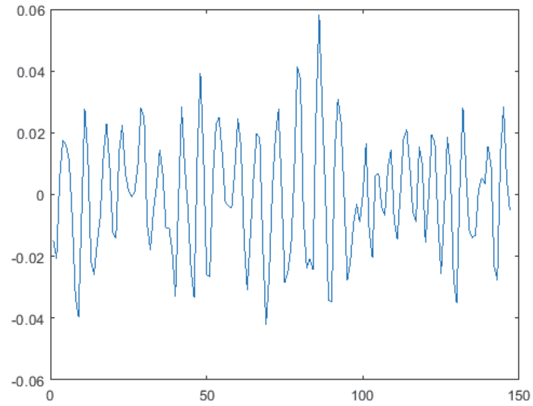

Figure 6 FRT_{CPI} - female

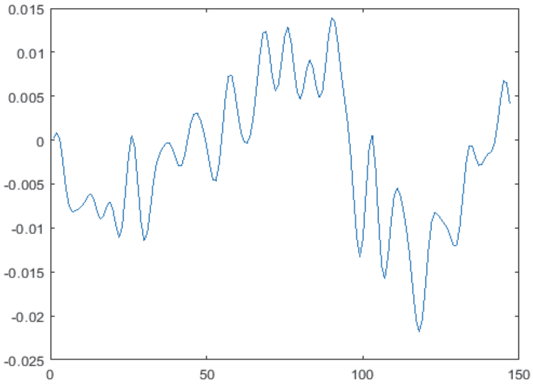

Figure 7 FRT_{CPI} - male

FRT_{CPI}'s probability distribution function illustrates the distinct variations between male and female voice signals. For female voice signals, the samples in FRT_{CPI} are equally distributed over the 0 to 1.5 amplitude range, as seen in Fig. 6. Male FRT_{CPI} coefficients, on the other hand, showed amplitudes ranging from 0 to 0.2, as seen in Fig. 7. There are only a few samples with values greater than 0.2.

After the complete analysis of FRT_{CPI} features, 80% of the extracted features are taken for training, and the remaining 20% of the features have been taken for testing. Then these features are classified by the LightGBM classifier to recognize gender. The hyperparameter of the LightGBM is shown in Tab. 4. Among these hyperparameters, maximum bin, number of leaves, learning rate, and number of iterations are the significant parameters for preventing overfitting and improving classification accuracy.

Table 4 LightGBM classifier parameters

Parameters	Tuning Factors
Objective	Binary
Metric	Auc
is_unbalance	False (for CV) / True (for SAA, VGD, TIMIT)
num_trees	20
min_data_in_leaf	20
num_rounds	5000
boosting_type	Gbdt
max_depth	25
min_child_samples	100
feature_fraction	0.5
bagging_freq	10
Seed	42
Verbosity	100
device Type	GPU

After tuning the significant parameters and keeping the remaining parameters as default, the decision tree, as shown in Fig. 8, was constructed with 504 training samples for the TIMIT database. Using this decision tree, LightGBM tried to classify the given feature as male or female.The empirical values of tuning parameters, as shown in Tab. 5, achieve better classification accuracy when the maximum bin, number of leaves, learning rate, and number of iterations is 63, 150, 0.0001, and 10000. The improved classification accuracy for SAA, CV, TIMIT, and VGD is 91.76, 96, 99.21, and 93.5%, respectively.

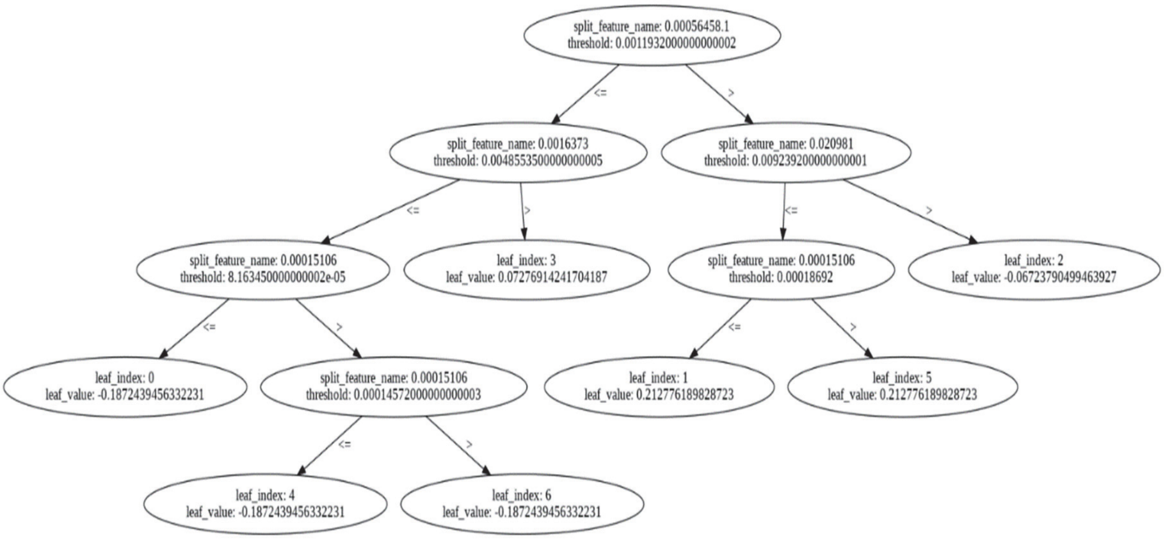
split_feature_name: 0.00056458.1
threshold: 0.0011932000000000002
<=
>
split_feature_name: 0.0016373
threshold: 0.0048553500000000005
split_feature_name: 0.020981
threshold: 0.009239200000000001
<=
>
<=
>
split_feature_name: 0.00015106
threshold: 8.163450000000002e-05
leaf_index: 3
leaf_value: 0.07276914241704187
split_feature_name: 0.00015106
threshold: 0.00018692
leaf_index: 2
leaf_value: -0.06723790499463927
<=
>
<=
>
leaf_index: 0
leaf_value: -0.1872439456332231
split_feature_name: 0.00015106
threshold: 0.00014572000000000003
leaf_index: 1
leaf_value: 0.212776189828723
leaf_index: 5
leaf_value: 0.212776189828723
<=
>
leaf_index: 4
leaf_value: -0.1872439456332231
leaf_index: 6
leaf_value: -0.1872439456332231

Figure 8 Decision tree from training samples

Table 5 Classifier performance of tuning parameter

Maximum bin	Num of leaves	Learning Rate	Number of Iterations	Accuracy / %			
				SAA	CV	TIMIT	VGD
40	50	0.01	100	90.24	95.24	98.25	92.7
45	100	0.001	1000	91.45	95.78	98.48	92.7
631.	150	0.0001	10000	91.76	96.00	99.21	93.5

The performance evaluation metrics of gender recognition based on voice have been evaluated and analyzed on the training model. The confusion matrix of the LightGBM classifier defines the classifier's performance for the four databases, as shown in Figs. 9, 10, 11, and 12. Class 1 is labeled as female, and class2 is labeled as male.

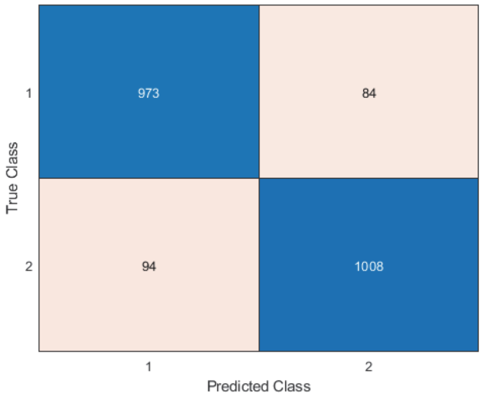

Figure 9 Confusion matrix for SAA

In Fig. 9, 973 female voice (class1) is correctly classified as female, and 1008 male voice (class2) is correctly classified as male for the SAA database. If positive samples are classified correctly, it is called True Positive (TP); otherwise, it is termed False Positive (FP).

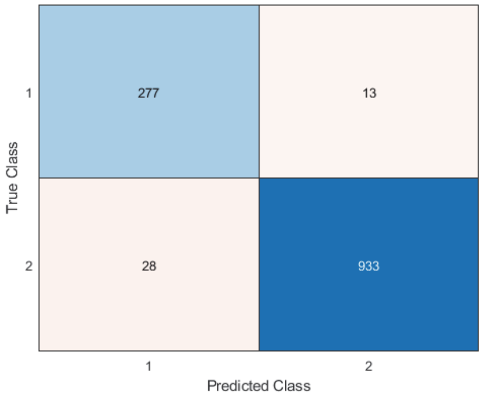

Figure 10 Confusion matrix for CV

If negative samples are classified as a correct class, it is called True Negative (*TN*); otherwise, it is called False Negative (*FN*). Using *TP*, *FP*, *TN*, and *FN*, four performance measures are used to predict the classifier performance: *precision*, *recall*, *F*1 *score*, and *specificity*. The SAA database's *precision*, *recall*, *specificity*, and *F*1 *score* are 0.92, 0.92, 0.91, and 0.92, respectively. The performance metrics of other databases such as CV, TIMIT, and VGD are also listed in Tab. 6. The overall classification performance of the classifier is above 91%.

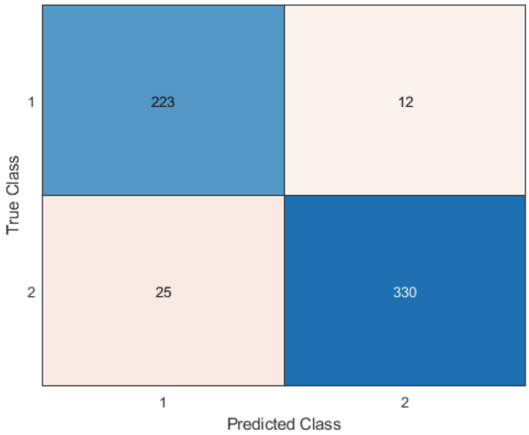

Figure 11 Confusion matrix for TIMIT

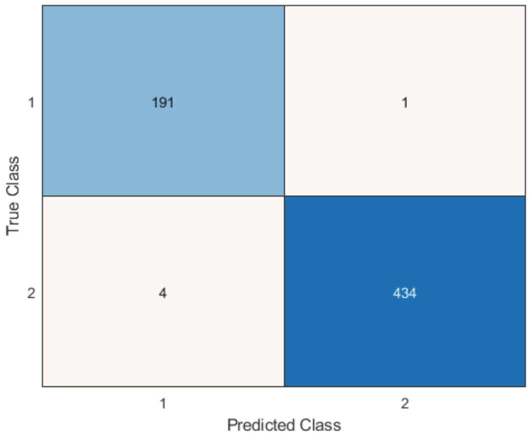

Figure 12 Confusion matrix for VGD

Table 6 LightGBM classification performance

Database Name	$Precision=\frac{TP}{(TP+FP)}$	$Recall=\frac{TP}{(TP+FN)}$	$Specificity=\frac{TN}{(TN+FP)}$	$F1Score=\frac{2\times(Precision\times Recall)}{Precision+Recall}$
SAA	0.92	0.92	0.91	0.92
CV	0.91	0.96	0.97	0.93
TIMIT	0.98	0.99	0.94	0.99
VGD	0.89	0.95	0.93	0.92

The performance evaluation of the gender recognition by LightGBM for the collected datasets is compared with other machine learning classifiers such as SVM, DA, NB, DT, KNN, and Ensemble. Among these classifiers, LightGBM attains high performance as shown in Tab. 7. The performance of the proposed method is evaluated with results outstripping state-of-the-art, as shown in Tab. 8. According to this study's findings, various feature extraction methods are employed for gender classification. From this comparison, it is noticed that the classification accuracy is strongly dependent on the features. The FRT_{CPI} features classified with the LightGBM classifier achieve a maximum classification rate, different from the state-of-the-art techniques.

Table 7 Performance evaluation of the LightGBM classifier with existing machine learning classifier

	Accuracy / %			
Classifier Name	SAA	CV	TIMIT	VGD
SVM	85	85	89	87
DA	77.5	79	72	74
NB	76	74	80	82
DT	86	84	79	82
KNN	89	81	78	79
Ensemble	89	81	85	87
Light GBM	91.76	96	99.21	93.5

The performance of the proposed method is compared against five existing gender recognition methods on collected prominent databases, including different ethnic voices, with all the variability as shown in Tab. 9. The accuracy rate of the proposed method for the SAA database is 12.36% higher than that of the mean accuracy rate of five existing gender recognition methods. Likewise, the proposed system attains a gain of 3.98%, 7.52%, and 2.98% compared with state-of-the-art techniques and achieves a reasonable gender recognition rate.

Tab. 9 presents the computational expenses (train and test time in second) for classifying speech samples from the collected database. The train and test time of the existing algorithm ranges from 3.6 s to 19.2 s, whereas the proposed method consumes a computational time of 2.25 s. The proposed method can perform less computational time than other existing methods.

Table 8 Accuracy of the proposed method against the state-of-the-art techniques

Reference	Database	Features	Classifier	Accuracy / %
Apeksha et al. (2015)	Real-Time Audio Dataset [18]	MFCC + Entropy + Frame Energy	ANN	80.40
Lakhan Jasuja et al. (2020)	Korean Telephonic speech [8]	MFCC	SVM	90
Yasin pir et al. (2019)	Michigan University database [21]	One Dimensional Wavelet	Neural Network	94
Pashwa et al. (2018)	TIMIT [16]	Pitch, MFCC, Energy	SVM	96.45
Archana et al. (2016)	Vowel Data [17]	MFCC, Delta, Delta-Delta	SVM	93.4
Proposed Method	TIMIT	FRT_{CPI}	LightGBM	99.9
	SAA			91.76
	VGD			93.5
	CV			96

Table 9 Comparison of computational expenses of proposed method versus existing methods (train and test time in seconds)

Database Gender Recognition Methods	SAA		CV		TIMIT		VGD	
	Accuracy / %	Time / s	Accuracy / %	Time / s	Accuracy / %	Time / s	Accuracy / %	Time / s
Apeksha et al (2015)	82.3	33.5	89.2	27.4	97.5	15.3	91	12.7
Archana et al (2016)	77.3	6.8	87.4	4.4	98.3	1.85	90.1	1.44
Pashwa et al (2018)	79	15	88.6	12.2	97.4	5.59	89	4.2
Yasin pir et al (2019)	81.2	42.7	90.2	34.3	98	23.5	87	19.8
Lakhan Jasuja et al (2020)	77.4	7.6	87	3.4	93.4	2.92	90.5	1.76
Proposed Method	**91.8**	**4.5**	**96**	**2.3**	**99.9**	**1.4**	**93.5**	**0.9**

5 CONCLUSIONS

A novel and powerful feature based on a Forward Rajan Transform (*FRT*) for gender recognition is proposed. The proposed method performs gender recognition on large datasets, namely SAA, VGD, CV, and TIMIT, by preprocessing, novel feature extraction, and recognition. To generate an efficient gender recognition system, preprocessing steps such as pre-emphasis, framing, and windowing are applied to the audio signal. The novel feature descriptor Cumulative Point Index of *FRT* (FRT_{CPI}) is used to extract the time domain features from the preprocessed and framed voice signal. Then the features are analyzed and classified with different machine learning classifiers; these LightGBM classifiers achieve higher accuracy, implying the feasibility of FRT_{CPI} features for recognizing voice signals as male or female. The proposed method achieves average recognition accuracy of 95.26%, with less computational time of 2.25 s, compared with the recent approaches. As a result, the proposed method is highly recommended for gender recognition based on voice signals.

Acknowledgments

The authors express their gratitude to the Thiagarajar College of Engineering (TCE) for supporting us in carrying out this research work. Also, the financial support from TCE under the Thiagarajar Research Fellowship scheme (File. no: TRF/Jul-2021/04) is gratefully acknowledged.

6 REFERENCES

[1] Bocekci, V. G. & Yildiz, K. (2016). Classification of Textures Using Filter Based Local Feature Extraction. *MATEC Web of Conferences*, 75,03001. https://doi.org/10.1051/matecconf/20167503001

[2] Poh, N. & Korczak, J. (2001, June). Hybrid biometric person authentication using face and voice features. *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, Berlin, Heidelberg, 348-353. https://doi.org/10.1007/3-540-45344-X_51

[3] Poornima, S., Sripriya, N., Preethi, S., & Harish, S. (2021). Classification of Gender from Face Images and Voice. *Intelligence in Big Data Technologies - Beyond the Hype*. Springer, Singapore, 115-124. https://doi.org/10.1007/978-981-15-5285-4_11

[4] Gupta, S. K. & Nain, N. (2022). Review: Single attribute and multi attribute facial gender and age estimation. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-12678-6

[5] Lin, F., Wu, Y., Zhuang, Y., Long, X., & Xu, W. (2012). Human Gender Classification: A Review. *International Journal of Biometrics, Inderscience Enterprises Ltd.* https://doi.org/10.48550/arXiv.1507.05122

[6] Kushwah, S., Singh, S., Vats, K., & Nemade, V. (2019). Gender Identification Via Voice Analysis. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. https://doi.org/10.32628/CSEIT1952188

[7] Zhong, B., Liang, Y., Wu, J., Quan, B., Li, C., Wang, W., Zhang, J., & Li, Z. (2019). Gender Recognition of Speech based on Decision Tree Model. *Advances in Computer Science Research (ACSR), 90*. https://doi.org/10.2991/iccia-19.2019.91

[8] Jasuja, L., Rasool, A., & Hajela, G. (2020). Voice Gender Recognizer Recognition of Gender from Voice using Deep

Neural Networks. *International Conference on Smart Electronics and Communication (ICOSEC)*, 319-324. https://doi.org/10.1109/ICOSEC49089.2020.9215254

[9] Kulkarni & Vaishali (2013). Speaker Identification Using Orthogonal Transforms and Vector Quantization.

[10] Ioannis, E. Livieris, E. P., & Panagiotis, P. (2019). Gender Recognition by Voice Using an Improved Self-Labeled Algorithm. *Machine Learning and Knowledge* Extraction, *1*, 492-503. https://doi.org/10.3390/make1010030

[11] Gupta, M., Bharti, S. S., & Agarwal, S. (2019). Gender-based speaker recognition from speech signals using GMM model. *Modern Physics Letters B*, 33(35), 1950438. https://doi.org/10.1142/S0217984919504384

[12] Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal Of Computing*, *2*(3). https://doi.org/10.48550/arXiv.1003.4083

[13] Ahmad, J., Fiaz, M., Kwon, S., Sodanil, M., Vo, B., & Baik, S. W. (2015). Gener Identification using MFCC for Telephone Applications-A Comparative Study, *International Journal of Computer Science and Electronics Engineering (IJCSEE)*, *3*(5). https://doi.org/10.48550/arXiv.1601.01577

[14] Chaudhary, S. & Sharma, D. K. (2018). Gender identification based on voice signal characteristics. 2018 *International conference on advances in computing, communication control and networking (ICACCCN)*, 869-874. https://doi.org/10.1109/ICACCCN.2018.8748676

[15] Pashwa, A. & Aggarwal G. (2016). Speech feature extraction for gender recognition. *International Journal of Image, Graphics and Signal processing*, *8*(9), 17. https://doi.org/10.5815/IJIGSP.2015. 09.03

[16] Archana, G. S. & Malleshwari, M. (2015). Gender identification and performance analysis of speech signals. *2015 Global conference on communication technologies (GCCT)*, 483-489. https://doi.org/10.1109/GCCT.2015.7342709

[17] Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, *9*(4), 235-245. https://doi.org/10.2478/jaiscr-2019-0006

[18] Ganapathiraju, A., Hamaker, J. E., & Picone, J. (2004). Applications of support vector machines to speech recognition. *IEEE Transactions on Signal Processing*, *52*(8), 2348-2355. https://doi.org/10.1109/TSP.2004.831018

[19] Mulimani, M. & Koolagudi, S. G. (2018). Acoustic Event Classification Using Spectrogram Features. *TENCON, IEEE Region 10 Conference*, 1460-1464. https://doi.org/10.1109/TENCON.2018.8650444

[20] Yasin Pir, M. & Idris Wani, M. (2019). A Hybrid Approach to Gender Classification using Speech Signal. *IJSRSET*, *6*(1). https://doi.org/10.32628/IJSRSET196110

[21] Anna, V., Kuchebo, V., Bazanov, V., Kondratev, I., & Kataeva, A. M. (2021). Convolution Neural Network Efficiency Research in Gender and Age Classification from Speech. *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. https://doi.org/10.1109/ElConRus51938.2021.9396365

[22] Uddin, M. A., Pathan, R. K., Hossain, M. S., & Biswas, M. (2021). Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN. *Journal of Information and Telecommunication*, 1-16. https://doi.org/10.1080/24751839.2021.1983318

[23] Weinberger, S. (2015). *Speech Accent Archive*. George Mason University.

[24] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common Voice: A Massively-Multilingual Speech Corpus. https://doi.org/10.48550/arXiv.1912.06670

[25] Garofolo, L., Lamel, F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. *Web Download. Philadelphia: Linguistic Data Consortium*. https://doi.org/10.35111/17gk-bn40

[26] See http://www.tsp.ece.mcgill.ca/Documents/Data
TSP Speech Database, Electrical & Computer Engineering McGill University.

[27] See http://www.voxforge.org.

[28] Giannakopoulos, T. (2022). Silence removal in speech signals. *MATLAB Central File Exchange.*

[29] Vergin, R. & O'Shaughnessy, D. (1995). Pre-emphasis and speech recognition. *Proceedings 1995 Canadian Conference on Electrical and Computer Engineering*, *2*, 1062-1065. https://doi.org/10.1109/CCECE.1995.526613

[30] Musaed, A., Zulfiqar, A., Muhammad, I., & Wadood, A. (2016). Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System. *Hindawi Publishing Corporation Mobile Information Systems*, 12.

[31] Bhatnagar, A. C., Sharma, R. L., & Rajeshkumar, L. (2012, July). Analysis of Hamming Window Using Advance Peak Windowing Method. *International Journal of Scientific Research Engineering &Technology (IJSRET)*, *1*(4), 015-020.

[32] Starck, J.-L., Murtagh, F., & Fadili, J. M. (2010). *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. New York: Cambridge University Press.

[33] Ekambaram Naidu Mandalapu, & Rajan, E.G. (2009). Rajan Transform and its Uses in Pattern Recognition. *Informatica*, *33*(2009) 213-220.

[34] Alzamzami, F., Hoda, M., & El Saddik, A. (2020). Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A comparative Evaluation. *IEEE Access*, *8*.

Contact information:

Priya KANNAPIRAN, PhD Student
(Corresponding author)
Thiagarajar College of Engineering,
Department of Electronics and Communication Engineering,
Madurai, Tamil Nadu, India - 625015
E-mail: priya5586@gmail.com

Mohamed Mansoor Roomi SINDHA, Prof. Dr.
Thiagarajar College of Engineering,
Department of Electronics and Communication Engineering,
Madurai, Tamil Nadu, India - 625015
E-mail: smmroomi@tce.edu