Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN

Mohammad Amaz Uddin, Refat Khan Pathan, Md Sayem Hossain and Munmun Biswas

Department of Computer Science and Engineering, BGC Trust University Bangladesh, Chittagong, Bangladesh

ABSTRACT
Analysing the human voice has always been a challenge to the engineering society for various purposes such as product review, emotional state detection, developing AI, and much more. Two basic grounds of voice or speech analysis are to detect human gender and the geographical region based on accent. This study presents a three-layer feature extraction method from the raw human voice to detect the gender as male or female, as well as the region from where that voice belongs. Fundamental frequency, spectral entropy, spectral flatness, and mode frequency have been calculated in the first layer of feature extraction. On the other hand, Mel Frequency Cepstral Coefficient has been used to extract the features in the second layer and linear predictive coding in the third layer. Regular voice contains some noises which have been removed with multiple audio data filtering processes to get noise-free smooth data. Multi-Output-based 1D Convolutional Neural Network has been used to recognize gender and region from a combined dataset which consists of TIMIT, RAVDESS, and BGC datasets. The model has successfully predicted the gender with 93.01% and region with 97.07% accuracy. This method works better than usual state-of-the-art methods in separate datasets along with the combined dataset on both gender and region classification.

ARTICLE HISTORY
Received 23 August 2021
Accepted 4 September 2021

KEYWORDS
Gender recognition; region of accent detection; MFCC; 1D CNN

1. Introduction
The human voice is the most compatible medium for interaction among human beings. When sound comes out from a vocal throat it carries much regional, biological, and surrounding atmospheric data. Using those kinds of information, we can find out human language, gender, age, accent, emotional and present state. In general, the human ear is a natural sound analysing system. It has an exceptional ability to classify gender, age, region, emotional state and much more based on some attributes of the human voice like loudness, frequency, etc. (Livieris et al., 2019). A machine can't do the same things...
but it can be possible by converting the voice into a digital signal and choosing the right features from them by using a different machine and deep learning algorithms.

Gender recognition is a technique to determine gender categories by the speaker’s voice signal analysis (Keyvanrad & Homayounpour, 2010). As gender recognition is a significant task, it increases the performance of applications such as human–computer interaction, emotion recognition, person identification, sorting of telephone calls by gender categorization, online advertisements involving voice (Alsulaiman et al., 2011; Chaudhary & Sharma, 2018; Pahwa & Aggarwal, 2016). Moreover, Human voices have different accents due to differences in culture, history, and geographical locations. Generally, accents are commonly different in the quality of the voice, the distinction of vowels and consonants, pronunciation, and stress. If two people from two different countries speak the same language, their pronunciation is not the same. A native English speaker and a non-native English speaker have many differences in their speaking features. Most probably their speech acoustic features, spectral features would not be the same for the native and non-native English speakers. Nowadays, we can see the more effectiveness of the gender and region recognition system in many advanced fields like commercial areas, criminal investigation, robotics, security system, and more.

In this paper, we extended our previous work (INISTA-2020) (Uddin et al., 2020) on gender recognition where we have extracted features from three datasets in two layers and classified the gender with SVM and KNN. In this extended work, we not only recognize the gender from audio datasets but also find out the regions of datasets voice actors with three layers of feature extraction method. To the best of our knowledge, we couldn’t find any related research which has classified regions from voice accent. The proposed method is based on the three-layer architecture of the feature extraction process. Before feature extraction, we used multiple filtering methods to remove the noise from the audio. The first layer consists of the fundamental frequency, spectral entropy, spectral flatness, mode frequency where the fundamental frequency is calculated by autocorrelation function. The Mel Frequency Cepstral Coefficient (MFCC) has calculated in the second layer using the linear interpolation function. At last, in the third layer Linear predictive coding (LPC) has been calculated. For training these features data 1D CNN has been used. The main contribution of this paper is working with mixed datasets from three regions to detect gender and region using a neural network model and also introduced an optimal way of feature extraction using three-layer architecture.

This paper is arranged in the following sections: Section 2 has the review of related research in this area, Section 3 describes the dataset, Section 4 contains our proposed feature extraction method with many sub-sections, Section 5 has the details of our proposed CNN model, training and testing process, and Section 6 contains the result analysis and comparison with literature. Finally, Section 7 concludes our work.

2. Literature review

Our modern life is fully dependent on technology. As many technologies are growing rapidly like machine learning, deep learning, big data, and various algorithms are improving day by day (Alsulaiman et al., 2011; Chaudhary & Sharma, 2018; Keyvanrad & Homayounpour, 2010; Livieris et al., 2019; Pahwa & Aggarwal, 2016), we can see the development of machine learning for research work as well as its popularity is become high to
apply in many fields. As machine learning is a subgroup of artificial intelligence, it learns from experience or data by developing an algorithm to teach a computer system to make decisions on various problems like speech recognition, image processing, etc. (Holzinger, 2019). Machine learning has various types of algorithms that help the researcher to distribute or examine the data. Some deep learning classifiers like CNN, multi-layer perception (MLP), recurrent neural network (RNN), long short-term memory (LSTM), artificial neural network (ANN), Hidden Markov model (HMM), etc. and different types of machine learning classifiers like support vector machine (SVM), k-nearest neighbours (KNN), Random Forest (RF), etc. have been used in many classification and regression problems to solve it. To recognize or classify the gender from human voice several researchers used several kinds of machine learning and deep learning techniques. However, gender recognition from the human voice is not an easy task without developing an ideal model.

To develop such a model for classification problems, one must choose the right feature from the human voice by which both gender and region of a person can be determined easily. There are most common features like MFCCs, Mel-scaled power spectrogram (Mel), power spectrogram chroma (Chroma), spectral contrast (Contrast), and tonal centroid features (Tonnetz) employed for gender recognition using 1D CNN model (Alkhawaldeh, 2019). Fundamental frequency, Energy, Spectral flatness, entropy, intensity, zero-crossing rate, etc. features are also used to categorize gender and region. After collecting the features from human voice, we then combined them with the proper labels to prepare a training set. After that, Machine learning techniques are utilized to construct a good model for recognizing gender. Although all ML techniques not performed well, so we had applied multiple techniques and measure their performance to select an optimal classification technique from them.

However, the researcher conducted a multitude of experiments using multiple methods and datasets to measure the efficiency of their proposed methods. A supervised machine learning was implemented (Majkowski et al., 2019) using a two-layer neural network for gender classification. In this study, they worked with 20 types of features with minimum, maximum, and average values of them. Every audio data has some noise problem which can make a problem to extract the right features. As a result, we can’t predict a problem perfectly without good features. To solve this noise problem needs a pre-processing process for feature extraction. Pre-emphasis, frame blocking, the hamming window are used in (Alkhawaldeh, 2019; Yusnita et al., 2017) as a pre-processing phase for removal of the noise.

For a robust gender detection system, researchers are trying to find the most distinctive features (Alhussein et al., 2016). For example, Simpson’s rule is used in (Alsulaiman et al., 2011) to measure the intensity of the human voice measuring area under the normalized curve. An Arabic digit database and a manual threshold were used to recognize the gender. MFCC, energy entropy and frame energy are feed to ANN and SVM and the performance of the SVM is better than the ANN (Archana & Malleswari, 2015). A gender classification model is developed in (Chaudhary & Sharma, 2018; Gaikwad et al., 2012) by using pitch, energy, MFCC as features, and SVM as a classification model. Both studies used a small number of features for their experiment. Multi-layer perceptrons (MLP), Gaussian mixture model (GMM), learning vector quantization (LVQ), and vector quantization (VQ) are used as a classifier in (Djemili et al., 2012). The IViE corpus was
used for their experiment and MFCC was used to extract the features. Although the researcher used multiple classifiers to classify the gender all classifiers were not performed well. Fatih Ertam used a three-stage model (Ertam, 2019) for gender classification. For this study, 10 most effective parameters: IRQ, meanfun, sfm, sd, median, Q25, Q75, mode, centroid, and meandom were selected. A deeper Long Short Term Memory (LSTM) network structure was used to predict the gender which performed a good accuracy of 98.4%. A non-linear parameter-based method was developed in (Yingle et al., 2008). Fractal dimension and fractal complexity are used as a feature where entropy is used to calculate the fractal complexity. The experimental results showed that the non-linear parameter-based method performed effectively well compared with the traditional method. A set of low-level time-domain features (Ghosal & Dutta, 2014) such as Mean of Short-Time Energy, Standard deviation of Short-Time Energy, Mean of Zero-Crossing Rate, Standard deviation of Zero-Crossing Rate, and Spectral flux calculated from the voice for discriminating the gender. Two simple classification techniques, RANSAC and neural-net were used for their study. Noisy speech has been used in (Zeng et al., 2006) for testing the proposed system for gender categorization. Using the pitch and relative spectral perceptual linear predictive coefficients the proposed system performed 95% against the noisy speech.

The researchers not only worked with adult people’s voices they also used children’s voices to predict gender. In (Chen et al., 2010), researchers proposed a gender recognition system for children of multi-age groups. To implement the system cepstral peak prominence, spectral harmonic magnitude, harmonic-to-noise ratio are used as features. One of the most dominant and most researched features Mel coefficients and its first and second derivatives (Pahwa & Aggarwal, 2016) used to determine the gender from the 46 speech sample dataset containing the Hindi vowels. Moreover, the proposed method is capable of working well with vowels, although the researchers didn’t mention anything about the model that it can perform well or bad for larger sentences. Another vowel-based dataset was used in (Yusnita et al., 2017) to investigate the gender using a different order of LPC coefficients.

3. Dataset

In this experiment, we have used three audio datasets named The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and BGC (details shown in Table 1). Every dataset was recorded in English. In TIMIT, every speaker spoke 10 sentences. Besides, RAVDESS is an emotion-based audio dataset where every actor recorded 60 audio files. On the other hand, BGC is our self-made audio dataset for this experiment where every actor has 50

| Name  | Number of Audio | Actor | Gender | Region       |
|-------|-----------------|-------|--------|--------------|
| TIMIT | 320             | 32    | 16     | 16           | US            |
| RAVDESS | 1440           | 24    | 12     | 12           | North America |
| BGC   | 300             | 6     | 3      | 3            | Bangladesh    |
4. Proposed method

In the audio datasets, there are many noise and distortion available along with the original usable voice which needs to be filtered out. In the pre-processing part of the voice analysis, we have used many filter techniques to get good quality data from those voices. After filtering, we used our proposed three-layer feature extraction method to get the feature set from raw audio data. In the first layer, we have used fundamental frequency, spectral entropy, spectral flatness, mode frequency. In the second layer, we have used the linear interpolation method for mapping the audio data into a range. After mapping the data, MFCC has been used to extract the features. Those MFCC features have been normalized using the Z-score normalization. At last, in the third layer, LPC has been calculated for our research. Once features are extracted, the new feature dataset is ready to be processed for the Convolutional Neural Network (CNN) training. The whole process structure has been shown in Figure 1.

4.1. Pre-processing

High pass filter: It attenuates those signals whose frequencies are below a certain cut-off frequency and passes the above frequencies. In this study, a high pass filter has been used to eliminate the unwanted sounds or noise those near to the audible range or below. For
the high pass filter, we used 0.05 as a value of Normalized passband frequency which was specified as a scaler in the interval (0, 1).

**Z-score normalization**: The main purpose of normalization is to change the value of data using a common scale without distorting or changing its original nature. Generally, it is used when data have different ranges. In this study, we have been used the Z-score normalization technique to reduce this outlier issue with Equation (1). It determines the data point from the mean ($\mu$) value in terms of the standard deviation ($\sigma$) value.

$$Z_{\text{score}} = \frac{x - \mu}{\sigma}$$  \hspace{1cm} (1)

Here $x$ is the value, $\mu = 0$ is the mean value of the feature, and $\sigma = 1$ is the standard deviation of the feature.

**Savitzky–Golay filter**: It is used to smooth the signal whose frequency span is large. It is also known as digital smoothing polynomial filters. It is applied to increase the precision of the data without distorting or changing the original signal and also reduce the least-square error in fitting a polynomial to the frames of noisy data. In the Savitzky–Golay filter, the polynomial order must be smaller than the frame size where the frame size must be odd. In this study, we used 3 polynomial order and the frame size 21 to smooth the data. The audio data pre-processing is shown in Figure 2.

4.2. Feature extraction

**Fundamental frequency**: The fundamental frequency is one of the most important features in speech recognition and describes the actual physical phenomenon of the signal. It is also known as the approximate frequency of the quasi-periodic voiced speech signal. In this research, we used the autocorrelation method which is a time-domain method to estimate the fundamental frequency. It measures the correlation between two values of the speech signal changes as their separation changes (Ramdinmawii & Mittal, 2016). The mathematical representation of the autocorrelation function is in the

![Figure 2. Audio data pre-processing.](image)
following equation:

$$A(k) = \frac{1}{T} \sum_{t=0}^{T-1} s(t)s(t+k), \quad k = [0, T - 1]$$  \hspace{1cm} (2)

Here, $k$ represents the lag and $A(k)$ is an autocorrelation function at lag $k$. $s(t)$ is the speech signal which is defined for all $t$. $T$ represents the window size of a speech signal. Following Equation (2), when the lag value is 0, it represents the highest value. After measuring the autocorrelation values, those values are used to measure the maximum, minimum, and average value of the fundamental frequency using the following equation:

$$f_0 = \frac{1}{T}$$  \hspace{1cm} (3)

Spectral entropy: It is a measure of the spectral power distribution of a signal. It is one of the most important features in the speech recognition field. By spectral entropy, the normalized power distribution of a signal is treated as a probability distribution in the frequency domain (Shen et al., 1998). The probability distribution is calculated with the following equation:

$$p(k) = \frac{S(k)}{\sum_{k=1}^{N} S(k)}, \quad \text{for } k = 1 \text{ to } N$$  \hspace{1cm} (4)

Here, $p(k)$ is the probability distribution. $S(k)$ is the $k^{th}$ power spectrum where $S(k) = |X(k)|^2$. $X(k)$ is the discrete Fourier transform of the signal. The power spectrum and probability distribution are needed to compute the spectral entropy with Equation (5). For this experiment, we used the maximum, minimum, and average values of spectral entropy.

$$SE = \sum_{k=1}^{N} P(k) \log_2 P(k), \quad \text{for } k = 1 \text{ to } N$$  \hspace{1cm} (5)

Spectral flatness: It is a feature of acoustic signals which is useful in digital signal processing (Madhu, 2009). Generally, the ratio of the geometric mean to the arithmetic mean of a power spectrum is known as spectral flatness. It is mainly used to quantify the noise-like or tone-like nature of a sound signal. Using spectral flatness, we can find out the flat or non-flat position of a signal. If the ratio of arithmetic mean and geometric mean is 1 and a power spectrum is perfectly flat, then it means the arithmetic mean and geometric mean are equal. Although, the ratio can’t be more than 1, because the geometric mean is always greater or equal than the arithmetic mean.

Mode frequency: For this study, we computed the mode frequency from the human voice. It is a value that has more occurrences in the signal. It is also known as the most frequent value in the speech signal. When multiple values are occurring equally in the speech signal, the mode frequency will be the smallest value among them.

Mapping: We used the linear interpolation method (Linear Interpolation, 2021) to construct new data points within the given range of known data points. Mainly, it provides interpolated one-dimensional (1D) data. In this research, each audio speech data was mapped into 6000 values. The audio data mapping using the linear interpolation method is shown in Figure 3. For two known data points $(x1, x2)$ and $(y1, y2)$ the linear
The interpolation function is shown in the following equation:

\[ y = y_1 + (x - x_1) \times (y_2 - y_1)/(x_2 - x_1) \] (6)

**Mel Frequency Cepstral Coefficient (MFCC):** MFCC is a representation of a short-term power spectrum of sound which is extracted from the speech signal to perform the speech recognition tasks. It is one of the most popular features in the speech recognition system (Pahwa & Aggarwal, 2016; Practical Cryptography, n.d.). The MFCC feature extraction method is based on many steps.

Frame blocking and windowing is the first step to extract the MFCC feature from the speech signal. In this step, the audio speech signal is converted into a short frame to calculate the coefficient or power spectrum. In this research, the sample rate of every audio is 16,000 Hz. The audio signal is divided into frames of 30 ms. This means we got the frame length for a 16,000 Hz audio signal 16,000*0.03 = 480 (round value) frames. The number of divided frames overlapped the adjacent windows with an overlap length of 20 ms (16,000*0.02 = 320 samples). The next step is applying the Discrete Fourier Transform (DFT) to convert each windowed frame from the time domain into a frequency domain to obtain a frequency spectrum with (7).

\[ X[f] = \sum_{t=0}^{L-1} x[t] e^{-j2\pi ft/L}, \quad t = 0, \ldots, L - 1 \] (7)

Here, \( f \) is the frequency domain, \( L \) represents the length of the sequence to be transformed and \( t \) represents the time domain.

After applying the DFT, the Mel filter bank has been calculated. It can be executed in the time domain and frequency domain. Generally, a Mel filter bank is executed in the frequency domain. However, MFCC designs half overlapped triangular filters based on band edges to transmute the frequency information to mimic the non-linear sound

---

**Figure 3.** Audio data mapping (linear interpolation).
that a human perceived. The band edges of the filter bank are described as a nonnegative increasing row vector in the range [0, sample rate/2] and also determinate in Hz.

In MFCC, the non-linear rectification used quondam to the discrete cosine transform (DCT) which is specified as Log. Most probably, it is used to transform Mel frequency coefficients to provide a set of Cepstral coefficients. Log reduces the acoustic variants that are not noteworthy for speech recognition. It produces an L-by-M matrix of features where L = number of analysis frames of the speech signal is partitioned into and M = the number of coefficients returned per frame. In this experiment, MFCC produces 14 coefficients per frame (total of 35 frames available). After computing the MFCC coefficients we have used the Z-score normalization method to normalize those MFCC features which improved the recognition accuracy.

Linear predictive coding (LPC): It is one of the most significant techniques in speech recognition (Kim, n.d.). It is also known as source filter modelling for signal processing because it has a sound source that goes through a filter and produces a signal (8).

\[ X(n) = H(n) \ast E(n) \] (8)

Here, \( E(n) \) is the sound source which models the vocal cords, \( H(n) \) is the filter that vocal tract and \( X(n) \) is the resulting signal. In LPC, it assumes the filter is a \( p \)th order all-pole filter with a transfer function to modelled the vocal tract transfer function shown in the following equation:

\[ H(n) = \frac{1}{1 - \sum_{j=1}^{p} a_j n^{-j}} \] (9)

Here, \( a_j \) is the filter coefficients and \( p \) is the order. Over the analysis frame of signal, the speech signal is assumed as constant which is approximated in (10).

\[ \hat{S}_k = \sum_{j=1}^{p} a_j S_{k-j} \] (10)

Here \( S_k \) is the speech signal. Minimizing the mean square filter prediction error between \( S_k \) and \( \hat{S}_k \) the filter coefficients \( a_j \) can be found. In this experiment, we used 16 orders of LPC that provides 17 coefficients for every audio.

5. Training and validation of CNN model

5.1. Creating CNN model

To train our dataset we have used one-dimensional CNN model with Keras library in Python. The concatenated output data shape of the three layers are 8 (1st), 490 (2nd) and 17 (3rd) make the model’s input shape of (515, 1) which is taken by the first Convolution layer with filter size 64, kernel size 2 and ReLu activation. We have used multiple convolution layers with filter sizes of 64, 128, 512, 128 and kernel sizes as 2 and 3. Along with the convolution layer, there are many batch normalization and max-pooling layer of pool sizes 2 and 1. To avoid overfitting, we have used dropout layers at 10–30% rate. After the flatten layer, the model is divided into output gender and region (two different output types) that has dense layer of 2 and 3 units with Softmax activation. Finally, the model has been compiled with Adam optimizer and
Mean Squared Error (MSE) as a loss calculator. The model has been illustrated in Figure 4 and details have been shown in Table 2.

5.2. Model validation

We have given 200 epochs with batch size 20 to train the dataset with 70–30 train-test set. After 200 epochs the training accuracy is 99.45% for gender and 99.65% for the region. On the other hand, the validation accuracy is 93.01% accuracy for gender with 0.06 MSE loss and 97.07% for the region with 0.02 MSE loss. The detailed graph of 200 epochs of accuracy and loss is shown in Figure 5.

6. Result analysis

We have tested our proposed method on a combined dataset. To combine the datasets we have used three audio datasets: TIMIT, RAVDESS, and BGC, where the TIMIT dataset is the most widely used for gender recognition and the RAVDESS is the most widely used emotional dataset. We have used a multi-output-based 1D CNN model which gives good accuracy to predict the gender and region shown in Figure 1. For this work, we used 1433 audio files as training set and 615 audio files as a testing set from the combined dataset. The confusion matrices are shown in Table 3 for gender and Table 4 for the region on the combined dataset. We also tried applying many optimizers and loss functions. Among them MSE works better comparing with others and using this MSE with many optimizers’ functions comparison has been done. In Table 5, some results are using MSE with Adam, SGD, Adamax, Adagrade, RMSprop and Adadelta optimizers. Among them, Adam works better and gives higher accuracy.

The precision, recall, and F1 score of the combined dataset for multi-output-based 1D CNN show the excellent performance of our proposed method. For computing these values, we need to analyse the confusion matrix. The confusion matrix describes the performance of a classification model based on test data. When a class is actually positive and also classified as positive, it is called True Positive (TP). As well as, when a class is not negative and also classified as not negative, it is called True Negative (TN). If a class is negative
Table 2. Details of CNN model architecture.

| Layer (type)            | Output Shape       | Parameter  | Connected to              |
|-------------------------|--------------------|------------|---------------------------|
| input_1 (Input Layer)   | [None, 515, 1]     | 0          | input_1                   |
| conv1d_1 (Conv1D)       | (None, 515, 64)    | 192        | conv1d_1                  |
| dropout_1 (Dropout)     | (None, 515, 64)    | 0          | dropout_1                 |
| batch_normalization_1 (Batch Normalization) | (None, 515, 64)    | 256        | batch_normalization_1     |
| max_pooling1d_1 (MaxPooling1D) | (None, 257, 64)    | 0          | max_pooling1d_1           |
| conv1d_2 (Conv1D)       | (None, 257, 128)   | 24,704     | conv1d_2                  |
| dropout_2 (Dropout)     | (None, 257, 128)   | 0          | dropout_2                 |
| batch_normalization_2 (Batch Normalization) | (None, 257, 128)   | 512        | batch_normalization_2     |
| max_pooling1d_2 (MaxPooling1D) | (None, 128, 128)   | 0          | max_pooling1d_2           |
| conv1d_3 (Conv1D)       | (None, 128, 512)   | 131,584    | conv1d_3                  |
| batch_normalization_3 (Batch Normalization) | (None, 128, 512)   | 0          | batch_normalization_3     |
| max_pooling1d_3 (MaxPooling1D) | (None, 128, 128)   | 196,736    | max_pooling1d_3           |
| dropout_3 (Dropout)     | (None, 128, 128)   | 0          | dropout_3                 |
| batch_normalization_4 (Batch Normalization) | (None, 128, 128)   | 512        | batch_normalization_4     |
| max_pooling1d_4 (MaxPooling1D) | (None, 64, 128)    | 0          | max_pooling1d_4           |
| flatten_1 (Flatten)     | (None, 8192)       | 0          | flatten_1                 |
| dense_1 (Dense)         | (None, 64)         | 524,352    | dense_1                   |
| dropout_2 (Dropout)     | (None, 64)         | 0          | dropout_2                 |
| dense_2 (Dense)         | (None, 2)          | 130        | dense_2                   |
| dense_3 (Dense)         | (None, 3)          | 195        | dropout_2                 |

Figure 5. Training and validation performance of proposed model, (a) for gender and (b) for region.
and classified as positive, it is called False Positive (FP). Along with this, when a class is positive and classified as negative, it is called False Negative (FN).

Precision: Precision is the ratio of TP and total predicted positive observation.

\[ \text{Precision} = \frac{TP}{TP + FP} \]

Recall: It is the ratio of TP and total positive observations in the actual class.

\[ \text{Recall} = \frac{TP}{TP + FN} \]

F1 score: F1 score is the weighted average of precision and recall.

\[ \text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

Tables 6 and 7 show the precision, recall, and F1 score for gender and region.

Throughout this study, it is observed that several kinds of feature extraction models are used for gender classification in many research. For this, the researchers used different types of datasets such as short vowel datasets (Pahwa & Aggarwal, 2016; Yusnita et al., 2017), digit datasets (Ertam, 2019) and long (Chaudhary & Sharma, 2018) sentence-based datasets. Though the proposed method for those audio datasets performed well, but in longer sentences, there is no clarification to work well which we solved in this study. Besides, the characteristics of the human voices are separate in different regions because of geographical differences. For this reason, the same number of features
might not work well for different region data. In Chaudhary and Sharma (2018) and Archana and Malleswari (2015), the proposed models ensured good performance using the lower number of features for a particular region dataset to determine the gender. But those models would not perform well in other region datasets or combined datasets because of the lower number of features. To solve this problem, we have worked with a combined dataset and two-layer architecture to extract enough number of features. We have detected the gender and region of user’s from this dataset. The outcome of the proposed method is shown in Table 8, compared with the other existing methods of gender recognition. But a comparison of region detection couldn’t be possible as we were unable to find any related work in this field.

7. Conclusion

In this study, we utilized a three-layer feature extraction method for gender and region classification from the human voices. We have extracted Fundamental Frequency, Spectral Entropy, Spectral Flatness, and Mode Frequency in the first layer of features. Besides, linear interpolation function has been used to map the audio data and extract the features using MFCC in the second layer. On the other hand, LPC has been calculated in the third layer. To get smooth and noise-free data for the right feature extraction several filtering and normalization processes such as High pass filter, Z-score normalization, and Savitzky–Golay Filter have been used in the pre-processing step. Moreover, multi-output-based 1D CNN has been utilized in this study to clarify the recognition rate of gender and region. For this work, we have used a combined dataset that consists of three English language-based audio datasets recorded in three different regions. The proposed method successfully performed better on both gender and region classification.
Our future work is focused on working with more regional datasets and improving the prediction accuracy of our work. Moreover, we want to implement our methodology in a real-time system.

Notes on contributors

Mohammad Amaz Uddin is a researcher as well as a software developer. Since 2017, he is involved in research work. The main objectives of his research work are focused on the analysis of large datasets and the detection of right patterns. He always tries to work in multiple research fields such as IoT, Sound Processing, etc. have been published in many scientific journals. His fields of interest are deep learning, machine learning, IoT, image processing, and sound processing. Along with this, now he is working as a software ASP.NET developer.

Refat Khan Pathan is a researcher in computer vision and image processing related topics. Along with image processing, he also worked with numeric complex data processing such as genetic data of COVID-19, IoT and Big Data. His main research interest is in complex data, image processing and time series data processing. He also has software development experience.

Md Sayem Hossain is a full-time software engineer and a researcher. Since 2019 he has involved himself in many research works and also got one successful publication in 2020. Overall, his research, he worked with audio much. He tries to find unique patterns among the data and develop algorithms to overcome the problem. Also, he worked with many professors to learn their working methods and for developing the methods of his own.

Munmun Biswas is currently a lecturer at BGC Trust University Bangladesh. She is a researcher having many publications in the field of IoT, Computer Vision, Numeric data processing, Time series data processing, Sound processing and Cloud Computing. She has extreme research interest in IoT, Machine Learning and Big Data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Mohammad Amaz Uddin http://orcid.org/0000-0002-0416-5676
Refat Khan Pathan http://orcid.org/0000-0002-3773-0950
Md Sayem Hossain http://orcid.org/0000-0002-3911-6567
Munmun Biswas http://orcid.org/0000-0001-8370-8773

References

Alhussein, M., Ali, Z., Imran, M., & Abdul, W. (2016). Automatic gender detection based on characteristics of vocal folds for mobile healthcare system. *Mobile Information Systems*, 2016, 1–12. https://doi.org/10.1155/2016/7805217

Alkhalaf, R. S. (2019). DGR: Gender recognition of human speech using one-dimensional conventional neural network. *Scientific Programming*, 2019, 1–12. https://doi.org/10.1155/2019/7213717
Alsulaiman, M., Ali, Z., & Muhammad, G. (2011, November). Gender classification with voice intensity. In 2011 UKSim 5th European symposium on computer modeling and simulation (pp. 205–209). IEEE. https://doi.org/10.1109/EMS.2011.37

Archana, G. S., & Malleswari, M. (2015, April). Gender identification and performance analysis of speech signals. In 2015 Global conference on communication technologies (GCCT) (pp. 483–489). IEEE. https://doi.org/10.1109/GCCT.2015.7342709

Chaudhary, S., & Sharma, D. K. (2018, October). Gender identification based on voice signal characteristics. In 2018 International conference on advances in computing, communication control and networking (ICACCCN) (pp. 869–874). IEEE. https://doi.org/10.1109/ICACCCN.2018.8748676

Chen, G., Feng, X., Shue, Y. L., & Alwan, A. (2010, September). On using voice source measures in automatic gender classification of children’s speech [Paper presentation]. Eleventh Annual Conference of the International Speech Communication Association, pp. 26–30.

Djemili, R., Bourouba, H., & Korba, M. C. A. (2012, May). A speech signal based gender identification system using four classifiers. In 2012 International conference on multimedia computing and systems (pp. 184–187). IEEE. https://doi.org/10.1109/ICMCS.2012.6320122

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

Gaikwad, S., Gawali, B., & Mehrotra, S. C. (2012). Gender identification using SVM with combination of MFCC. Advances in Computational Research, 4(1), 69–73.

Ghosal, A., & Dutta, S. (2014, February). Automatic male-female voice discrimination. In 2014 International conference on issues and challenges in intelligent computing techniques (ICICT) (pp. 731–735). IEEE. https://doi.org/10.1109/ICICICT.2014.6781371

Holzinger, A. (2019). Introduction to MAchine learning & Knowledge Extraction (MAKE). Machine Learning and Knowledge Extraction, 1(1), 1–20. https://doi.org/10.3390/make1010001

Keyvanrad, M. A., & Homayounpour, M. M. (2010, May). Improvement on automatic speaker gender identification using classifier fusion. In 2010 18th Iranian conference on electrical engineering (pp. 538–541). IEEE. https://doi.org/10.1109/IRANIANCEE.2010.5507010

Kim, H. S. (n.d.). Linear predictive coding is all-pole resonance modeling. Center for Computer Research in Music and Acoustics, Stanford University.

Livieris, I. E., Pintelas, E., & Pintelas, P. (2019). Gender recognition by voice using an improved self-labeled algorithm. Machine Learning and Knowledge Extraction, 1(1), 492–503. https://doi.org/10.3390/make10100030

Madhu, N. (2009). Note on measures for spectral flatness. Electronics Letters, 45(23), 1195–1196. https://doi.org/10.1049/el.2009.1977

Majkowski, A., Kołodziej, M., Pyszczak, J., Tarnowski, P., & Rak, R. J. (2019, September). Identification of gender based on speech signal. In 2019 IEEE 20th International conference on computational problems of electrical engineering (CPEE) (pp. 1–4). IEEE. https://doi.org/10.1109/CPEE47179.2019.8949078

Pahwa, 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

Practical Cryptography. (n.d.). Mel Frequency Cepstral Coefficients (MFCC) tutorials. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

Ramdinmawii, E., & Mittal, V. K. (2016, December). Gender identification from speech signal by examining the speech production characteristics. In 2016 International conference on signal processing and communication (ICSC) (pp. 244–249). IEEE.https://doi.org/10.1109/ICSPCom.2016.7980584

Shen, J. L., Hung, J. W., & Lee, L. S. (1998, November). Robust entropy-based endpoint detection for speech recognition in noisy environments [Paper presentation]. Fifth international conference on spoken language processing, Vol. 98, pp. 232–235.

Uddin, M. A., Hossain, M. S., Pathan, R. K., & Biswas, M. (2020, August). Gender recognition from human voice using multi-layer architecture. In 2020 International conference on innovations in
Yingle, F., Li, Y., & Qinye, T. (2008, June). Speaker gender identification based on combining linear and nonlinear features. In 2008 7th World congress on intelligent control and automation (pp. 6745–6749). IEEE. https://doi.org/10.1109/WCICA.2008.4593952

Yusnita, M. A., Hafiz, A. M., Fadzilah, M. N., Zulhanip, A. Z., & Idris, M. (2017, November). Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal. In 2017 7th IEEE international conference on control system, computing and engineering (ICCSCE) (pp. 372–377). IEEE. https://doi.org/10.1109/ICCSCE.2017.8284437

Zeng, Y. M., Wu, Z. Y., Falk, T., & Chan, W. Y. (2006, August). Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech. In 2006 International conference on machine learning and cybernetics (pp. 3376–3379). IEEE. https://doi.org/10.1109/ICMLC.2006.258497