# Gender Recognition for Urdu Language Speakers using Composite and Multi-Layer Feature Approaches with Fuzzy Logic

A. Ghafar[1], N. Shah[2], M.M. Iqbal[3]

*[1,2] Department of Information System, SBE, University of Management & Technology, Lahore, Pakistan,*
*[3] University of Engineering and Technology, Taxila, Pakistan*

[1]abdul.ghafar@umt.edu.pk

***Abstract-*** Gender recognition by voice is one of the most demanding phenomena in speech analysis. With the increasing use of digital communication channels, many speech analysis techniques are being used to identify gender by acoustic features. In this paper, an algorithm is presented to develop a tool using Praat Script to classify Urdu language speaker's gender by analyzing sound features with various speech processing techniques broadly categorized into composite and multi-layer feature approaches. Euclidean distance and Naïve Bayes are implemented to compare cumulative feature vector containing fundamental frequency ($F_0$), formants, and MFCC coefficients, with a base vector of aforementioned sound features those are obtained through supervised training using Texas Instruments and Massachusetts Institute of Technology (TIMIT) speech corpus. Techniques are further refined to get a more accurate outcome by applying fuzzy logic rule-base on the aggregated result. The algorithm is also designed to make it efficient in terms of processing time, accuracy, and reliability by eliminating frames having undefined $F_0$ and removing outliers. Use of fuzzy logic rule-base returns 100% accuracy in gender recognition, whereas individually multi-layer feature approach achieves 98% accuracy as compared to the composite approach which returns just 77% with sample dataset of 133 Urdu language speakers' voices obtained through Pakistani Urdu dramas.

***Keywords-*** Gender Recognition by Voice, Gender Recognition of Urdu Speaker, Acoustic & MFCCs Analysis, Naïve Bayes Classifier, Fuzzy Logic

## I. INTRODUCTION

Recognizing the speaker's gender automatically with his or her voice by using different sound features such as pitch, formants, and MFCC coefficients are important in many real-time applications. Automatic gender recognition can be used to understand user needs by identifying their gender and to respond accordingly in different interactive voice response systems. It can be effective in mobile healthcare systems, where specific treatment is needed based on the patient's gender. It is helpful in automatic call response system, to perform basic gender-specific operations such as marketing gender-specific products and transferring the call to a male or female operator. It also plays a vital role in speech recognition in different multimedia applications, smart human-computer interactions, biometrics social robots, and security systems.

The automatic gender recognition process is implemented a number of times in the past by using various sound features. The most important in gender identification is the fundamental frequency or pitch which is considered by many researchers as the main sound feature. Other more common sound features are MFCC coefficients, which are used to get more reliable results in gender identification, but system training must be required with a specific classifier. Using artificial intelligence techniques in gender recognition by sound becomes popular to reduce the error level in identifying the speaker's gender.

### *1.1. Background*

Gender identification by using fundamental and formant frequencies is performed in [39], they calculate values of these sound features with the help of two lists of 10 words spoken by 76 speakers and those values are still followed by researchers as base values in gender classification.

According to them, fundamental frequency and first three formants are higher in female as compare to male, and they established the mean value of pitch for female 223 Hz and men 131 Hz. Reference [22] has used source filter synthesizer and frequency shifting to test male and female voices and confirmed both fundamental, and formant frequencies are required to identify speaker's gender with 98.8% accuracy, but $F_0$ is more important than formant frequency. They also identified that the fundamental frequency is 131 Hz for male and 220 Hz for female. The value of fundamental frequency is greatly affected by the speaker's gender, and its value is higher for female, whereas the impact of speaking style in gender recognition is more important than spoken language [1].

Formant frequencies are other important sound features, used to identify speaker's gender by using vowels in speech and they show accuracy for sustained vowels by using Euclidean distance in [5]. Reference [22] analyses that only formant frequencies to identify gender is not good so it should be combined with fundamental frequency, so reference [6] and [23] prefer fundamental frequency as well as formats to identify speaker's gender, whereas formants are not suitable to identify gender when entire length of speech is considered instead of vowel part [7].

As stated by many types of research, both pitch and formant frequencies are not enough to identify gender accurately, another important sound feature in gender identification is MFCC introduced by Davis & Mermelstein in 1980. It is a set of cepstral coefficients from the short term power spectrum of sound to identify gender more accurately as compared to other sound features. MFCC is a parametric representation of acoustic data of speech waveform and becomes more effective when combined with pitch and formant frequencies [8]. Due to its larger size, it must be used with different classifiers such as neural network, GMM, Naïve Bayes, and SVM. $F_0$ estimation may be difficult in a noisy environment, but when it is merged with MFCC, it shows more accuracy than former [9]. MFCC is a more robust technique even under noisy conditions, and it is used in the presence of white noise by using a sparse model training procedure in [10]. Use of Gammatone Frequency Cepstral Coefficients (GFCC) in [11], a modified form of MFCC, gives more accurate results even in an entirely noisy environment. Reference [16] extracts 12-MFCC coefficients with GMM classifiers with expectation maximization algorithm to identify gender, whereas [46] develops gender identification by using 39 MFCC coefficients with GMM model and i-vector. Transformed MFCC feature matrix is classified with DNN in [16]. Traditional MFCC is extended with EMD and complete ensemble EMD with SVM in [14], and another extension in MFCC by multi-taper windowing function performed in [15]. MFCC coefficients are criticized in [18] and [19] due to computational complexity by extracting large size data. MFCC also fails to produce accurate results if training and testing are performed in different environment and conditions. Therefore it must be combined with a pitch to eliminate its limitations [18]. Two-stage classification in [18] is used to set maximum $F_0$ of male as lower boundary and minimum $F_0$ of female as upper boundary and classified doubtful speaker by MFCC coefficients with GMM model. The same approach is also implemented in [31] by MFCC coefficients with the HMM model. Two phases approach is also followed in [20] to identify gender by using hot-vector of 12 MFCC, 12 Delta MFCCs and 12 Delta-Delta MFCC coefficients with DNN-HMM classifiers in the first phase.

### *1.2. Specific Research Motivation*

It is found through a literature review that most of the work related to automatic gender identification using voice is done with English and other well-known natural languages for testing as well as training. However, significant work on languages which are spoken is limited geographical regions or by a specific community, is not well defined. Urdu is one of those languages, where more research is needed in automatic gender identification. It is the national language of Pakistan and spoken in Pakistan as well as some others parts of sub-continent [21]. It is used to test the proposed & developed system to identify the Urdu language speaker's gender.

This research is an extension and implementation of two-stage classification recommended in [18] and [20] by adding the third layer. At the same time, research is also used to compare it with other approaches, as followed in [76] and many others by combining different sound features in gender recognition. The purpose of the research is also to compare several gender identification techniques with different sound features such as fundamental frequency, formant frequencies, and MFCC coefficients, to check their accuracy level and to aggregate their results to minimize error level by using Euclidean distance and Naïve Bayes classifier with fuzzy logic rule-base.

Gender recognition techniques are broadly categorized in Composite and Multi-Layer Feature Approaches. In Composite approach, different sound features, individually or collectively, are used to identify gender, whereas Multi-Layer Feature approach performs gender identification by layering the sound features in the parent-child relationship. Both of the above techniques are aggregated with Fuzzy Logic rule-base, to reduce the error level in gender identification in approaches above.

The remainder of this research is organized as follows. A detail description of related work to identify sound features in Section 2. The details about the proposed & developed a system in this research are available in Section 3, whereas Section 4 describes the procedure to produce test data for proposed & developed system testing. System results are presented in Section 5. The conclusions are given in Section 6.

## II. RELATED WORK

Sound is an analogue signal as shown in Fig. 1, which needs to be converted into a digital signal by pulse code modulation, then Fast Fourier transformation is used to convert time domain signal into a frequency domain [31]. There are some basic features in sound, those can be used in gender identification systems to recognize the speaker's gender.

The difference between male and female is due to physiological, acoustical, and perceptual parameters which lead toward acoustical differences, so extracted

Fig. 1: Analogue signal of sound

acoustic features can be used to classify gender as male or female [5]. Dimensions of the larynx, vocal folds, and muscles that control their fluctuations are different for men and women which becomes a reason for the difference in sound features in male and female [23].

Reference [5] describes the sharp difference of frequencies at resonant peaks in male and female voices due to longer vocal tracts in men. The male larynx is average 1.7 times larger than female, which results in 1.7 times lower pitch in male as compared to female. The male vocal tract is average 1.2 times longer than a female vocal track, so it produces a formant spectrum 1.2 times lower than female, and the ratio of male-female $F_0$ is 1.76. In many other types of research, it is found that vocal cords of females are shorter and thinner than males, so $F_0$ becomes higher for female.

*2.1. Speech Processing*

Various speech processing techniques are extensively used for analysis. A number of researchers have developed various functions or formulas to calculate sound feature values or to use some software such as Praat to perform analysis. Followings are various functions in sound processing techniques.

*2.1.1. Framing & Windowing*

According to various theories, sound data always has a finite length of non-stationary signals, which are not desirable for automatic processing, so the sound must be divided into frames to remove the quasi-stationary feature from the signal and to make them flatten. The vocal tract is not normally changing more than fifty times per second, so a frame of 20 ms, will be stationary [17] and will result in more accuracy [6]. Reference [53] identifies unvoiced signals and noise from an original speech by using framing and shows reasonable improvements in phoneme recognition.

As argued by [17], [2], windowing is one of the important functions used to remove the abrupt change on both ends of the frame, avoid loss of information, and gives more accurate feature values. Frames can be overlapped to remove biases in them and to achieve accurate and reliable results, by using windowing function [6]. Several frames are processed in a single window, so its size must be larger than frame length. There are different types of windowing function used by researchers. Hamming window of 32-40 ms is used to minimize signal discontinuities in frames in [26], [3], and by many other researchers, whereas Hanning window with 40 ms frame, is used to identify peak cepstral value in [6]. Due to biases in other types of windows, the Gaussian window is a better choice for formant analysis [25].

*2.1.2. Unvoiced Frames and Noise Removal*

Reference [47] and [48] mention that zero crossing rate (ZCR) is normally lower for the voiced frame and higher for the unvoiced frame and [47] also describe that amplitude of unvoiced frame is lower than the amplitude of voiced frame. Reference [29] describes that only voiced frames are required for further processing to extract feature values, and [23] describes that voice frames must be required to identify accurate pitch value. Unvoiced frames, as shown selected sections in Fig. 2, contain no valuable information and return UNDEFINED feature values in Praat tool [25].

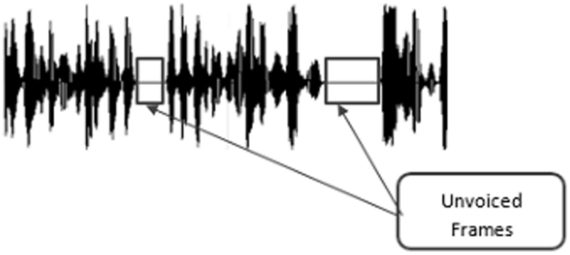

Fig. 2: Unvoiced frames in the sound signal

The sound that can be accepted for input to identify male is normally between 65-300 Hz and to identify female is normally between 100 to 600 Hz [57]. Reference [8] mentioned that sound below 75 Hz or above 600 Hz is not a good candidate for analysis, so any sound that is above or below the threshold is considered as noise and must be removed for further analysis.

*2.1.3. Pre-emphasis*

Pre-emphasis is set of various processing functions performed on the original sound before extracting sound features. Use of proper pre-filtering removes glottal wave shape effects and radiation characteristics and creates a more stable spectrum for smooth processing in gender identification [18]. Reference [51] and [52] describe that low-pass and band-pass filters are used to create a flatter spectrum and used to remove noise by eliminating higher and lower frequency components.

*2.2. Feature Extraction*

Feature extraction is the most important phase in speech processing to identify a speaker's gender and used to extract different features such as fundamental frequency, formant frequencies, and MFCC. Feature extraction phase will identify the various sound features with the help of following techniques used by various researchers in automatic gender recognition.

*2.2.1. Pitch Detection*

Pitch is acoustics or vocal cord vibration as a function of time, that is defined in [25] as "$F_0$ of the specific periodic signal in period $T_0$ as $F_0$= $1/T_0$". As stated in [19] and [25], the time autocorrelation function is accurate, noise-resistance, and robust method to detect the pitch of a voiced frame, then maximizing them to detect the highest value of frequency. Reference [13] criticizes that due to errors in double frequencies and half frequency and consuming hardware capacity greatly ACF is not better, and describes that Average Magnitude Difference Function (AMDF) has low computational cost as compared to it. AMDF is another variation of ACF analysis, instead of performing calculations for correlation function, it is calculated by a variation of original and delay signal to take the absolute magnitude of the signal [27]. Cepstrum Pitch Detection separates the excitation, and vocal tract contributions using homomorphic transformation in specified voiced frame and pitch is computed by identifying the first dominant peak in the cepstrum [54]. Simple Inverse Filtering, also called linear predictive analysis, was used in [28] and also discussed by [24], produces flatten spectrum except for unvoiced frames and transformed in a major peak used to determine $F_0$. Reference [12] suggests a data-driven method for monophonic pitch identification using a deep convolutional neural network on the time-domain signal.

*2.2.2. Formant Frequencies Detection*

Resonant peaks in the sound signals spectrum due to vowel sounds are called formant frequencies ($F_1$, $F_2$, $F_3$), and they are also important to identify the speaker's gender. To identify formant frequencies accurately, vowel sounds must be identified in the soundtrack. Linear Prediction Coding method as described in [5] and [50] is to determine resonance peaks of a frame by applying filter coefficients. Spectral-band method used in [23] for each frame is the average effective frequencies of the corresponding output of the bandpass filter. Cepstral Analysis method as described in [30], which is an improvement in the LPC algorithm and the most common method of identifying the peaks in the spectrum to find formant frequencies, especially for vowel segments. Prediction Parameter for each frame is converted to the cepstral coefficient, then Peak Picking algorithm is used to extract formant frequencies. Autoregressive parameters are obtained by the Berg algorithm, which is used to estimate the value of formants from roots of the LPC polynomial for each frame.[8]

*2.2.3. MFCC Coefficients Detections*

MFCC is a set of various coefficients collected from sound data and used to identify a speaker's gender more accurately. Reference [2] and [14] describe the following method to find MFCC coefficients. First of all, by using the pre-emphasis, framing and humming windowing function, perform the fast Fourier transfer or Discrete Fourier transformation in (1) by converting each frame from the time domain into frequency domain spectrum.

$$Y(w) = FFT[h(t) * x(t)] \tag{1}$$

The power spectrum is mapped onto the Mel scale by applying the bank of triangular bandpass filters with the linear distribution of frequencies with (2).

$$f_{mel} = 2595\ log_{10}\left(1 + {}^{f_{mel}}/_{700}\right) \tag{2}$$

The new energy distribution is obtained after applying the Mel scale. Discrete cosine transform is applied on the logarithm of the average of the new power spectrum, converting back to the time domain, to find the MFCC coefficients (3), where n is the number of MFCC coefficient.

$$c_n = \sqrt{\frac{2}{K}} \sum_{j=0}^{N} (\log_m j) \cos\left(\frac{\pi n}{K}(j - 0.5)\right) \tag{3}$$

A modified MFCC in [14], empirical mode decomposition, EMD-base MFCC, is extracted by applying discrete cosine transform on log power values calculated over some specific bands of intrinsic mode functions (IMF). Another variation in MFCC suggested in [15], is extracted by applying multi-taper window functions with Thomson multi-taper, Multipeak multi-taper, or Sinusoidal weighted cepstrum estimator methods. Transformed MFCCs are extracted in [16] by applying first and second derivatives on input features to obtained bottleneck features, and phoneme labels are used to create transformed features, which are used to generated transformed MFCC, it represents the prosodic features in addition to spectral features.

*2.3. Classifiers*

Use of classifiers is the core of any decision making or selection of class based on a given parameter. A classification problem occurs when an object needs to be assigned to a predefined class based on a number of observed attributes related to that object [34]. Several classifiers are used to classify data in supervised and unsupervised methods, some of the classifiers used in this proposed & developed system, are described below.

The distance formula is the simplest way to classify data by a certain threshold, so an object that is near to that threshold value will be classified in the same group. A number of different distance formulas such as Euclidean, Mahalanobis, Manhattan, and Bhattacharyya are described in [32]. Euclidean distance, one of the most common distance formulae in various researches such as [26], [3] and many others to classify speaker's gender, is the length of line between two points or more than two points in n-dimensional space [33]. Naïve Bayes is a very simple and effective probabilistic classifier based on Bayesian decision theory and used as statistical discriminant analysis [49]. It is used to group data in different clusters by using a supervised learning method. It assumes that there is no dependency of any feature value on other feature values, so it is suited for classifications using independent features [55] and for training the system with lesser data [56]. The Gaussian function is used to get the maximum likelihood of single or k-variables with means of given clusters. Class probability is updated at the end which makes the next iteration more accurate [35]. Fuzzy logic, an artificial intelligence technique, is a collection of rules to classify the input data. It is a very effective technique to test multiple parameters to accurately classify given input data with the given base value ranges by using rule-base. Rule-base IF-THEN conditions are generated for a number of classification variables to group input data [36]. Reference [49] mentions that the classification process can be improved by implementing the fuzzy logic rule-base with the various machine learning or feature selection algorithms.

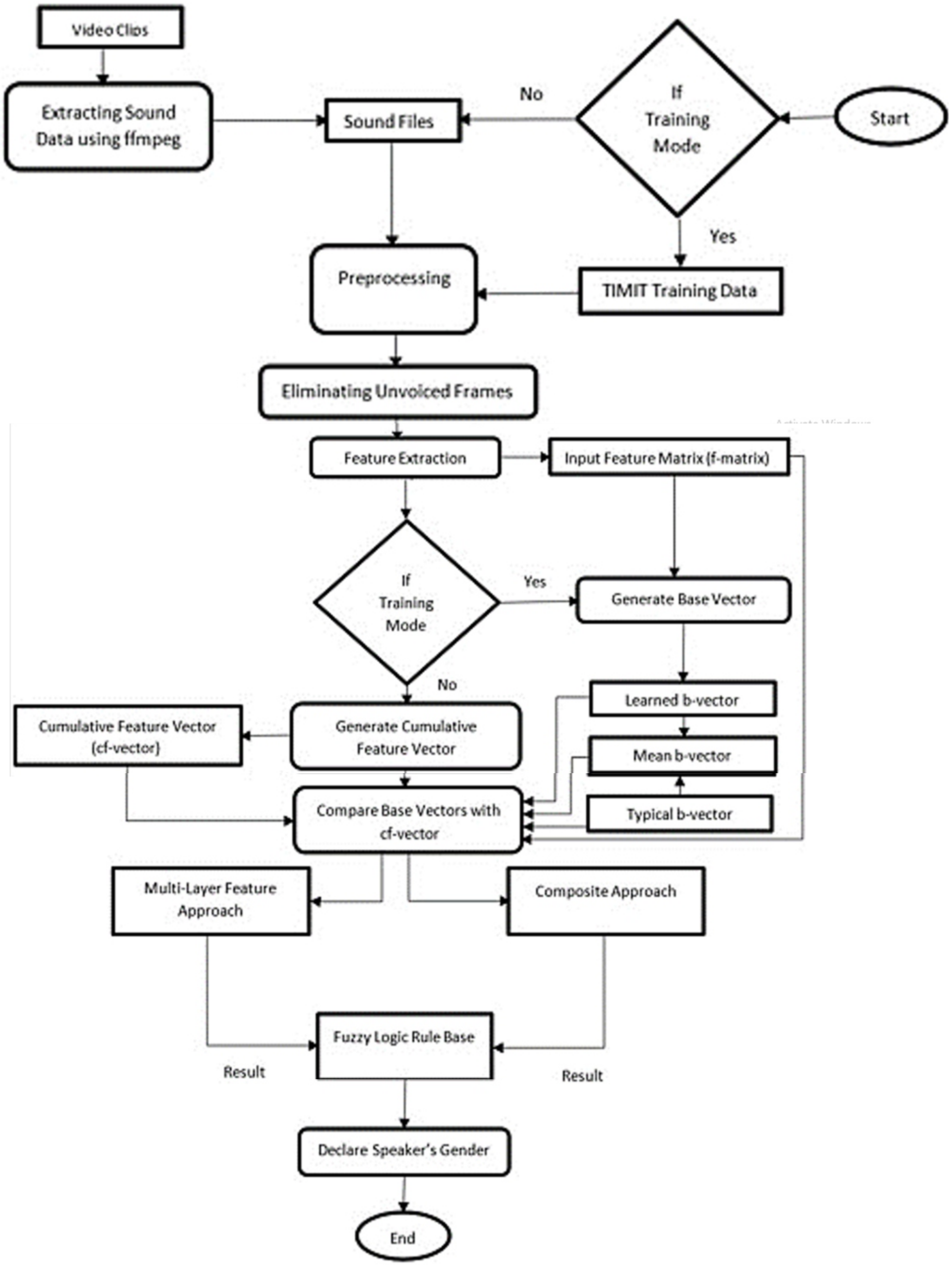

Fig. 3: Gender recognition system framework

## III. PROPOSED & DEVELOPED SYSTEM

The proposed & developed system, as shown in Fig 3, will try to identify a speaker as male or female automatically based on the fundamental frequency as well as other important features such as formant frequencies, and 12-MFCC coefficients by using various gender identification techniques by applying fuzzy logic rule-base with Naïve Bayes and Euclidean distance classifiers.

### *3.1. System Framework*

As shown in Fig. 3, sound files will be loaded into memory, which will be extracted from video clips by using ffmpeg tool. After loading sound files, the system will perform some basic pre-processing on the sound data to extract aforementioned sound features and store them in feature matrix (f-matrix). In training mode, the same features will be used to create the Base Vector (b-vector), which will be compared with the cumulative feature vector (cf-vector) in testing mode. In the testing mode, the system will follow different techniques, grouped into Composite and Multi-Layer Feature approaches, to classify speaker by using Euclidean distance and Naïve Bayes classifier. Results obtained through the above techniques will be aggregated by Fuzzy logic rule-base to decide about the speaker as male or female finally.

### *3.2. System Training*

Before executing system testing, the system will be trained by using TIMIT speech corpus [37], set of 1,357 female and 3,250 male sound files of 630 English speakers of eight different dialect regions of United States, recorded at a sampling rate of 16 KHz for research purpose. Training dataset will be provided in group or gender-wise for load balancing. At the end of the training, b-vector will be generated and stored in a text file which will be further used to recreate b-vector in testing mode. In testing, b-vector will be used to classify cf-vector generated by Urdu language speakers. One of the significant hurdles to creating a b-vector from Urdu speakers is the lack of resources to record Urdu speakers' voice in a clean environment. So, b-vector will be created by using TIMIT speech corpus [37] that is in the English language. It is verified in [1] that spoken language does not significantly affect the mean value of fundamental frequency.

### *3.3. Base Vector*

### *3.3.1. Typical B-Vector*

For fundamental frequency and formants, two types of base values, Typical b-vector, and Learned b-vector are used to compare with cf-vector. By analyzing previous researches, Typical b-vector is configured as given in Table 1 section A, for fundamental frequency identified in [38] and formants frequencies identified in [39].

TABLE I: BASE VALUES FOR FUNDAMENTAL AND FORMANT FREQUENCIES OF MALE AND FEMALE

| Gender | $F_0$ | $F_1$ | $F_2$ | $F_3$ |
|---|---|---|---|---|
| Section A: Typical [a] Base Values in Hz | | | | |
| Male | 120 | 500 | 1425 | 2382 |
| Female | 210 | 578 | 1695 | 2782 |
| Section B: Learned [b] Base Values in Hz | | | | |
| Male | 118.23 | 520.58 | 1433.49 | 2362.71 |
| Female | 201.16 | 564.21 | 1545.11 | 2505.53 |
| Section C: Mean [c] Base Values in Hz | | | | |
| Male | 119.02 | 510.27 | 1429.16 | 2372.21 |
| Female | 205.64 | 571.07 | 1620.16 | 2643.74 |

[a] Values based on [39] & [38]
[b] Generated through system training data
[c] Average of Typical and Learned Base values

### *3.3.2. Learned B-Vector*

Learned b-vector is derived after analyzing the training data by the supervised method from TIMIT speech corpus [37]. Learned b-vector is created as shown in Table 1 section B, after finishing comprehensive training.

### *3.3.3. Mean B-Vector*

In case avoid biasness in base values selection, the system will also calculate Mean b-vector from Typical b-vector and Learned b-vector as given in Table 1 section C. There is no significant difference among these base values, but the system will test all of these to confirm their importance and accuracy in gender identification.

### *3.4. Praat Scripting & Third-Party Tools*

Praat scripting is used to analyze sound files for sound features such as pitch, formant frequencies, MFCC coefficients, and other related sound features. The proposed & developed algorithm to identify speaker's gender is implemented in Praat script except image slicing and sound extraction which is done in VSDC free video editor and ffmpeg.

### *3.4.1. Basics of Praat Tool*

As the description found in [8], Praat tool is based on the solid theories of researchers so that Praat can be used in the sound analysis for advance research.

Autocorrelation method as described in [25], is used in identifying the fundamental frequency by using 10 ms frame with 40 ms long Gaussian window. Amplitude below 0.03 threshold is considered silent, and the threshold value must be below 0.45. Pitch floor is set to 75 Hz, and its upper limit is 500 Hz [8]. To find $F_0$, the system will set frame size equal to 50 ms, range of pitch between 75 Hz and 600 Hz and remaining values as default.

For formant analysis, Praat extracts minimum 5 formant frequencies per frame by computing LPC coefficients with Berg algorithm. The maximum frequency for an adult female is 5500 Hz and 5000 Hz for male. To flatten the spectrum, Praat applies an inverted low-pass filter with a slope of +6db /octave, frequencies below 50 Hz are not enhanced, frequencies around 100 Hz are amplified by 6 dB, frequencies around 200 Hz are amplified by 12 dB, and so on. As the female frequency is above male frequency so that the system will be configured maximum level of frequency by 5,500 Hz [8].

For MFCC, Praat extracts a number of MFCC coefficients frame by frame by using Gaussian windowing function with constant sampling from MFCC object, before that MelSpectrogrm object is created from the sound object. Praat performs all of the calculation in extracting MFCC coefficients by using theories of Davis & Mermelstein [8].

### *3.4.2. System Interface*

The proposed & developed system will be a functional Praat script to classify the speaker's gender by using his or her voice. A simple interface will be provided to perform various operations related to gender recognition as shown in Fig. 4. Each selected operation will be performed all of its related activities automatically without user interaction and returned results at the end.

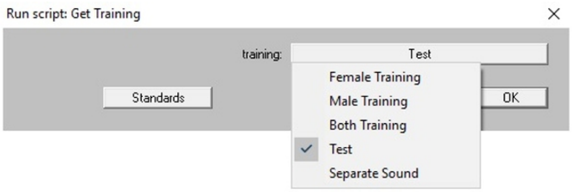

Fig. 4: Main functions available in the proposed & developed system

### *3.4.3. Input Data*

Input data for this script may be in video or sound files of variable sizes of less than a minute to two minutes. Input files are just needed to be copied in the relevant folders and system will read from those folders to perform its processing. Sound files may be in mp3 or wave format, whereas video files may be avi, dat, mp4, or wmv format. If video files are available, Extract Sound option will automatically extract sound files (mp3 format) in Sound folder by using ffmpeg tool.

### *3.4.4. Pre-processing*

Noise is one the major ingredient in sound, which invalidates results, especially fundamental frequency. So first pre-processing function is to remove background music or noise from input sound by using Spectral subtraction method with bandpass filter between the range of 75 Hz and 10 KHz [8]. For high precision sound data and more accurate sound features for reliable gender recognition, the sound is resampled with 10 KHz with the standard precision value of 50 [8]. After performing the above operations, the system will generate new sound without abruptness in it.

### *3.4.5. Framing and windowing function*

Each sound signal will be logically divided into equal and finite sized frames before the actual feature extraction process. The system will be configured 20 ms frame, during which signal can be assumed stationary as in [6], also shown in Fig. 5. The total length of the sound file is divided by frame size to get N frames to be processed. To reduce the distortion effect by framing, Praat objects are also be configured with window size. In the proposed & developed system, a window size of 150 ms, will be used to extract MFCC coefficients.

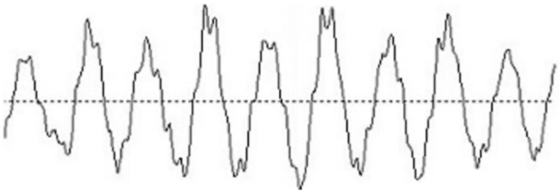

Fig. 5: Frame contains a stationary signal

### *3.4.6. Elimination of Unvoiced Frames*

To get reliable feature values, the frame must be identified as voiced or unvoiced. The system will perform the two-folded method of elimination of unvoiced frames, intensity detection, and Pitch detection. There are a number of features, those can be used to identify unvoiced frames, but the pitch is one the ideal method to identify unvoiced frames. Praat pitch object returns UNDEFINE value if there is no sound or above a threshold. So, frame with UNDEFINE pitch will be declared as an unvoiced frame and will be ignored for further processing, which is shown by the missing graph in Fig. 6. Sound intensity method will ignore frames if it is sound intensity is below 50dB, as there is no valuable voice data below this level [8].

If there are T unvoiced frames, the remaining frames to be analyzed are R = N – T, which are shown in Fig. 7 as modified sound waveform without unvoiced frames. On the basis of these unvoiced frames, all features are ignored even if someone has a reasonable value.

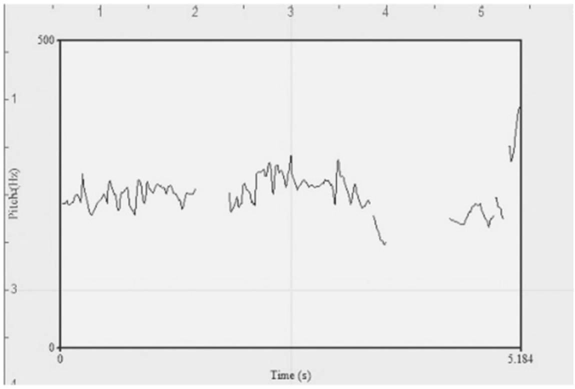

Fig. 6: Undefined fundamental frequency

Fig. 7: Regenerated sound signal

*3.4.7. Praat Objects*

TABLE II DEFAULT VALUES TO CREATE PRAAT OBJECTS

| Praat Object | Parameter | Value |
|---|---|---|
| Fundamental Frequency | Frame Size<br>Frequency Range | 20 ms<br>75-600 Hz |
| Formant Frequencies | Frame Size<br>Window Size<br>Frequency Range | 20 ms<br>40 ms<br>75-600 Hz |
| MFCC Coefficients | Frame Size<br>Window Size<br>No of coefficients | 50 ms<br>150 ms<br>12 |

In Praat scripting, different types of objects are created to extract various sound features. The sound object is the main object, which is used to create other objects such as Pitch object, Formant object, MFCC object, and other objects to analyze other sound features. For proposed & developed system, these objects are created by using constant values given in Table 2. These objects have a powerful set of parameters and functions to extract sound features, which will be used to populate f-matrix in next step. First of all, the sound object will be generated by resampled sound. Then pitch object with autocorrelation function as shown in (4), will be created from sound object to find the fundamental frequency. To find formant frequencies, formant object with Berg method will be created from the sound object as shown in (5). The same sound object will be used to create MelSpectrogram object with Davis & Mermelstein algorithm as shown in (6). MelSpectrogram object will be used to create an MFCC object to extract 12 MFCC coefficients.

To Pitch (ac): 0.02, 75, 10, "on", 0.03, 0.45, 0.01, 0.35, 0.14,600 (4)
To Formant (burg): 0.02, 5, 5500, 0.04, 50 (5)
To MelSpectrogram: 0.15, 0.05, 50, 50, 0 (6)

*3.4.8. Feature Extraction*

Feature extraction is the most important and critical phase in gender identification and any error while extracting feature, will lead to false results [6]. These sound features, as discussed before, are fundamental frequency, four formant frequencies, and 12 MFCC coefficients for all of the voiced frame in set R generated in the last step. To extract value for each of the above features, *get* function is executed on a specific Praat object created in the previous step. At the end of the feature extraction process, f-matrix will be populated with an array of mean values of all sound features for each voiced frame. During this process, unvoiced frames are assigned an array of zero values and ignored in further processing.

*3.4.9. Cumulative Means of F-matrix and Comparing with B-vector*

Next step will be to get cumulative means of all sound features from f-matrix which contains means of all sound features for voiced frames. Before finding cumulative means, f-matrix will be arranged in ascending order to identify outliers base on fundamental frequency due to its importance in gender identification. Outliers will be identified by finding the difference between first and third quartile to find the interquartile range IQR as (7) that will be multiplied with 1.5 to remove outliers and added the result in the third quartile to get the top range of valid data. To find the bottom range, the previous result will be subtracted from the first quartile, which is then used to remove the outliers from the bottom side [40].

$$IQR = q_3 - q_1 \tag{7}$$

All voiced frames have mean values within IQR boundaries defined previously. These mean values will be used to calculate the mean values of all features to populate cf-vector. Geometric mean or arithmetic mean can be used to find all of the cumulative means.

After that, cf-vector will be compared with b-vector by using Euclidean distance and Gaussian Naïve Bayes function to classify the speaker's gender. The proposed & developed algorithm will use a mixture of different approaches to identify the speaker's gender. It is also to compare the performance and efficiency of all techniques and to find the best method to identify a speaker's gender with a minimum level of error.

*3.5. Gender Identification Techniques*

To identify a speaker's gender as shown in Fig. 8, the proposed & developed system will perform various gender identification techniques categorized into Composite and Multi-Layer Feature approaches.

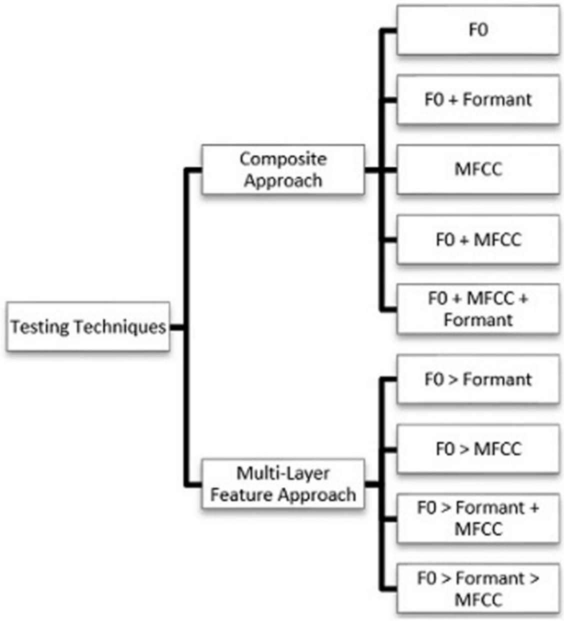

Fig. 8: Gender identification techniques used in a proposed & developed system

*3.5.1. Composite Approach*

In a composite approach, one or more sound features are used collectively to identify a speaker's gender as in [6], distance formula is performed with fundamental and formant frequencies. The different number of sound features are grouped as shown in Fig. 8, to identify gender in this approach. The cumulative result is obtained by Gaussian function to classify speaker.

*3.5.2. Multi-Layer Feature Approach*

Multi-Layer Feature approach, which is an extended form of two-stage classification in [18] and [20] with little modification by defining the marginal overlapping region. Sound features are arranged in layers as shown in Fig. 8 with their importance in identifying the speaker's gender. Due to the importance of fundamental frequency as cited by various researchers, it is the first layer in all techniques. Parent layer performs its processing by using Euclidean distance as well as Gaussian function and transfers control to child layer if the parent layer is unable to identify the speaker's gender. Otherwise, it returns the speaker's gender. This approach is expected to perform better than former as less calculation is required to identify the speaker's gender. Its performance is better than the composite approach as per test results, but it will be preferred in some cases where Multi-Layer Feature approach returns invalid results.

*Setting marginal overlapping region*: By using base values, a marginal range is defined for the overlapping region by setting 20% increase in male base value for lower boundary and 20% decrease in female base value for an upper boundary as shown in Table 3. If the sound feature of the speaker is within the overlapping region, the speaker will be declared as neutral, and the child layer is used to decide about the same speaker's gender.

TABLE III
OVERLAPPING FREQUENCY RANGE FOR NEUTRAL GENDER

| Feature | Overlapping Region [a] | Male | Female | Neutral Gender |
|---|---|---|---|---|
| $F_0$ | 90 | <= 138 Hz | >= 192 Hz | > 138 Hz & < 192 Hz |
| $F_1$ | 92 | <= 541.4 Hz | >= 596.6 Hz | > 541.4 Hz & < 596.6 Hz |
| $F_2$ | 250 | <= 1561 Hz | >= 1861 Hz | > 1561 Hz & < 1861 Hz |
| $F_3$ | 1649 | <= 2840.8 Hz | >= 3831.2 Hz | > 2840.8 Hz & < 3831.2 Hz |

[a] It is a doubtful range of frequency to detect the gender of the speaker.

*3.6. Use of the Euclidean Distance Formula & Naïve Bayes Classifier*

In either approach mentioned above, the system will compare cf-vector with b-vector generated in the training phase by using Euclidean distance and Naïve Bayes classifier. Euclidean distance will classify fundamental and formant frequencies in cf-vector, whereas Naïve Bayes classifier will classify cf-vector itself by using b-vector. Multi-variable Euclidean distance among fundamental frequency and first three formants is calculated by the formula for male (8) and female (9) on a two-dimensional plot [41]. The value of D in (10), which is the difference between male and female distance values, will decide speaker's gender, so positive value of D will confirm cf-vector as male otherwise female [2].

$$d_m = \sqrt{(f1 - bf1_m)^2 + (f2 - bf2_m)^2 + (f3 - bf3_m)^2} \tag{8}$$

$$d_f = \sqrt{\left(f1 - bf1_f\right)^2 + \left(f2 - bf2_f\right)^2 + \left(f3 - bf3_f\right)^2} \tag{9}$$

$$D = d_m - d_f \tag{10}$$

Reference [42] describes multi-variable Gaussian Naïve Bayes function to calculate the probability of male and female by using b-vector with features in cf-vector by using (11), where k is variable in b-vector, and v is the value of cf-vector. The function which returns higher value will decide the speaker's gender.

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \qquad (11)$$

*3.7. Applying Fuzzy Logic Rule-base*

After getting the results from all of the techniques in the previous section, results will be aggregated by applying Fuzzy logic rule-base as shown in Table 4 to declare speaker as male or female. Fuzzy logic rules are developed in the manner discussed in [43] for gender identification with different energy-related parameters. Another objective of this result aggregation with Fuzzy logic rule-base is to minimize the error level in gender identification in the above techniques.

TABLE IV
FUZZY LOGIC RULE-BASE TO AGGREGATE GENDER RECOGNITION APPROACHES

| Approach | Fuzzy logic rule-base |
|---|---|
| Composite approach rules | If MFCC result equals to $F_0$+MFCC and $F_0$+Formants+MFCC<br>If $F_0$ result equals to $F_0$+MFCC and $F_0$+Formants+MFCC |
| Multi-Layer Feature approach rules | If $F_0$>MFCC result equals to $F_0$>Overall and $F_0$>Formants>MFCC<br>IF $F_0$>MFCC result equals to $F_0$>Overall or $F_0$>Formants>MFCC |
| Aggregate rules | IF Composite returns Neutral, then the result will be Multi-Layer Feature<br>IF Multi-Layer Feature returns Neutral, then the result will be Composite |

## IV. System Test Data

The objective of this study is to identify Urdu language speakers by using voice, collected through different Pakistani Urdu TV Drama Serial episodes downloaded from Youtube.com. To maintain the quality of test data, a number of videos are downloaded to confirm the quality of the audio track, clarity of language, and diversity of various speakers.

*4.1. Slicing Video Data*

As download data is not suitable for testing, so the first requirement will be to slice the video into less than 2 minutes clip to make them suitable for system input. VSDC Free Video Editor [44] is a free tool to slice videos downloaded from Youtube.com by inserting markers on the cut points. Placing a marker is important and difficult task to cut the clip for required input. Two major problems are identified during this procedure. Firstly, extracting the single speaker dialog is not easy as the second speaker starts his or her dialog before the end of first one, and secondly selecting slices with a minimum amount of background music as there is loud background music in different scenes. To perform testing, video slices must have one speaker with minimum background music. To avoid the above issues, all slices are tested manually and copied them in Raw-Video folder for further processing.

*4.2. Separation of Sound Track from Video using ffmpeg*

As proposed & developed system will require sound files to recognize the speaker's gender, so the next step will be to extract sound from videos in Raw-Video folder. Extract Sound option or extractsound.bat file is executed to separate sound from video slices and to store them automatically in Sound Folder. To test system accuracy in gender identification, sound files must be arranged in male and female speakers by having the same set of folders created manually in Raw-Video folder.

This step will not be required if test data is already available in sound files. User will just copy files in Sound Folder, from the there system will automatically load those files for further processing.

*4.2.1. Extract Sound Option*

It will just initiate a batch file which contains ffmpeg commands as shown in (12) in [45], to extract sound from video files of different formats discussed before. Ffmpeg tool is used to manipulate video files in several ways such as extracting sound, slice, or merge the video. This option is configured to create mp3 format sound files which are smaller in size as compare to wav files, as well as keeping the quality of sound.

```
ffmpeg" -i "ts01.dat” -ab 160k -ac 2 -ar 44100 -vn
"ts01.mp3"
```
(12)

## V. System Results

*5.1. Test Sample Data*

The system is tested on a set of 133 Urdu language speakers' voice files in the required format, which have been extracted from Pakistani Urdu TV Dramas Serials discussed in Section 4. There are 82 different female actresses and 51 of male actors' voices in the dataset.

*5.2. Results Obtained*

Test results as shown in Table 5, are obtained after processing all of 133 voices of Urdu language speakers to identify their gender by various techniques in the proposed & developed algorithm. As per individual technique, fundamental frequency is most important in identifying gender with 97% accuracy with learned b-vector with Euclidean distance but when used along with MFCC with Naïve Bayes classifier, it gives 100% results for each of the base vectors as shown in Fig 9, whereas [9] shows 95.1% accuracy with $F_0$ and 98.4%

TABLE V: TEST RESULTS OF GENDER RECOGNITION OF URDU LANGUAGE SPEAKERS

| Gender Recognition Techniques | Typical Base | | | Learned Base | | | Mean Base | | |
|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Overall | Male | Female | Overall | Male | Female | overall |
| **Composite approach** | | | | | | | | | |
| $F_0$ | 94 | 96 | 95 | 94 | 100 | 97 | 94 | 98 | 96 |
| Formants | 73 | 29 | 51 | 59 | 38 | 48.5 | 67 | 33 | 50 |
| $F_0$+Formants | 73 | 34 | 53.5 | 63 | 60 | 61.5 | 67 | 43 | 55 |
| MFCC | 92 | 63 | 77.5 | 92 | 63 | 77.5 | 92 | 63 | 77.5 |
| $F_0$+MFCC | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| $F_0$+Formant+MFCC | 98 | 100 | 99 | 98 | 100 | 99 | 98 | 100 | 99 |
| Overall | 88 | 70 | 79 | 84 | 77 | 80.5 | 86 | 73 | 79.5 |
| **Multi-Filter Feature approach** | | | | | | | | | |
| $F_0$>Formants | 100 | 89 | 94.5 | 100 | 94 | 97 | 100 | 90 | 95 |
| $F_0$>MFCC | 100 | 98 | 99 | 100 | 100 | 100 | 100 | 99 | 99.5 |
| $F_0$>Overall | 100 | 95 | 97.5 | 100 | 98 | 99 | 100 | 96 | 98 |
| $F_0$>Formants>MFCC | 100 | 94 | 97 | 100 | 98 | 99 | 100 | 95 | 97.5 |
| Overall | 100 | 94 | 97 | 100 | 98 | 99 | 100 | 95 | 97.5 |
| **Fuzzy logic rule-base** | | | | | | | | | |
| Composite Approach | 98 | 100 | 99 | 98 | 100 | 99 | 98 | 100 | 99 |
| Multi-Layer Feature Approach | 100 | 91 | 95.5 | 100 | 95 | 97.5 | 100 | 93 | 96.5 |
| Overall Fuzzy logic | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

accuracy when merged it with MFCCs. In the same way, [11] shows 72.9% accuracy with GFCC in a noisy environment as compared to 60.4% accuracy with MFCC, [14] shows 99.6% accuracy by using extended MFCC coefficients, and [15] shows 99.55% accuracy by using Thomson multi-taper MFCC method with SVM classifiers. Formants frequencies are one of the weakest gender identification techniques in our proposed & developed system as shown in Fig. 9, as they are extracted its values from the full length of speech instead of vowel part, which is also mentioned by [7], and it is shown by [6] that formats frequencies and pitch identify gender nearly 97% accurately. In multi-layer feature approach, our proposed & developed system achieves 100% accurate results with $F_0$ layered with MFCCs by using Naïve Bayes classifier with learned b-vector, whereas [18] obtains 98.65% accuracy and [31] shows 98.92% accuracy in their two-stage classification model.

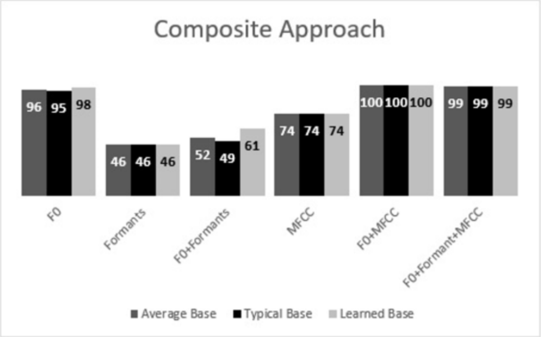

Fig. 9: Results of different techniques to identify a speaker's gender in Composite approach

The performance of multi-layer feature approach is also affected by formants frequencies as shown in Table- 5.

On average all of the techniques in multi-layer feature approach, give more accurate results than the techniques in composite approach and collectively multi-layer feature approach out-performs composite approach and gives 98% accurate results as compared with later which gives 77% only. As shown in Fig. 9 and 10, the accuracy of individual technique with learned b-vector is more accurate as compared with other base vectors. There are only 5 wrongly identified genders with learned b-vector as compare with typical b-vector which wrongly identifies 8 genders as shown in Fig. 12. Due to mean value biasness, none of the approaches is able to give 100% overall accuracy except MFCC with fundamental frequency in the composite approach as shown in Table 4.

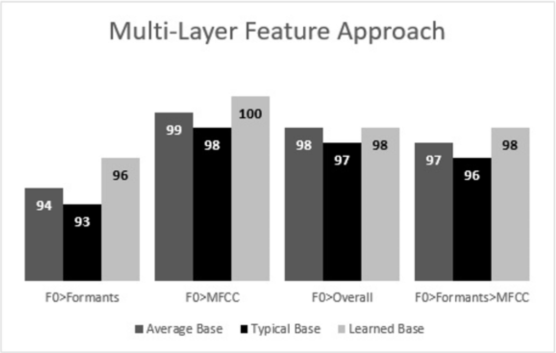

Fig. 10: Results of different techniques to identify a speaker's gender in Multi-Layer Feature approach

As our proposed & developed algorithm implements fuzzy logic rule-base to decide about the speaker's gender finally and its positive impact on the performance of the individual approach is shown in Table 5. However, in female voice identification error level is increased in a multi-layer feature approach due to the tendency of female voices toward the male voice. The composite approach shows more accuracy than multi-layer feature approach by using fuzzy logic, as shown in Fig. 11.

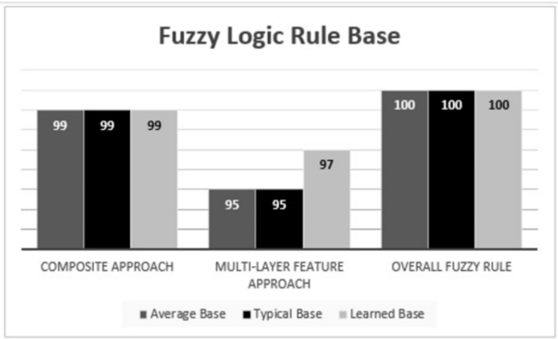

Fig. 11: Improvement in gender identification by using fuzzy logic rule base

The aggregate fuzzy logic rule-base shows 100% accuracy in recognizing the speaker's gender with sample data as shown in Fig. 12. Whereas transformed MFCC feature matrix with DNN classifier shows 89% accuracy in [16], as well as the use of the artificial neural network (ANN), shows 98.4% accuracy in gender identification in [9].

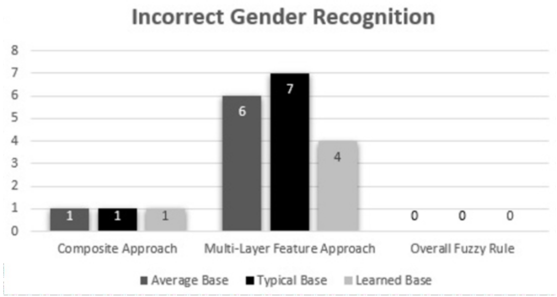

Fig. 12: Speaker is identified as invalid gender in different approaches

Our proposed & developed method of aggregate fuzzy logic rule-base, minimizes the error level in gender identification and shows 0% error level with the sample data, whereas [46] develops gender identification by using 39 MFCC coefficients with GMM model and i-vector and shows 4.29% error rate in i-vector as compare 5.93% in GMM.

The results show that fundamental frequency is the major sound feature in determining speaker's gender and its role can never be ignored, but there must be some other sound features such as formants and MFCC coefficients, to decide speaker's gender more accurately. It is shown in results that MFCC coefficients are more important than formants in identifying the speaker's gender and the accuracy in gender identification with the fundamental frequency and MFCC is greater than other techniques. Artificial intelligence techniques such as fuzzy logic are important in reducing the error level in the correct identification of gender, and proper fuzzy logic rule-base can be used to identify gender more accurately in all of the gender identification and classification techniques. Along with all of gender identification techniques, use of proper classifier is important in training the system so that that gender will be identified without any ambiguity in test mode. Naïve Bayes classifiers are better in classifying gender than Euclidean distance but later is good in setting boundaries to identify neutral gender in multi-layer feature approach.

### *5.3. Importance of Results*

Test results represent that multi-layer feature approach with fuzzy logic is better in identifying speaker's gender as compare to other techniques and there is no difference in comparing base values generated with any language speakers, but system training is important in identifying speaker's gender more accurately against using base values generated by other researchers.

## VI. CONCLUSIONS

In this paper, an algorithm is devised to identify Urdu language speaker's gender by comparing different sound processing techniques grouped into composite and multi-layer feature approaches by classifying various sound features such as fundamental frequency, formats, and MFCC coefficients, with the help of Naïve Bayes, Euclidean distance, and fuzzy logic rule-base. The proposed & developed algorithm gives 100% accurate results with the application of fuzzy logic rule-base on aggregated results of composite and multi-layer feature approaches, whereas individual approach gives 98% accuracy with multi-layer feature approach and 77% accuracy with the composite approach. It is also identified from the results that fundamental frequency is the core acoustic feature to identify the speaker's gender, whereas MFCC coefficients give more accuracy in gender recognition. The obtained results from sample data are overall satisfactory in identifying speaker's gender, but there is still a number of areas to be analyzed in future such as extracting formant frequencies from vowel parts of speech, base vector generated through native language speakers, and effect of other classifiers to identify speaker's gender.

## ACKNOWLEDGMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## REFERENCES

[1] P. Arantes, A. Eriksson and S. Gutzeit, "Effect of language, speaking style and speaker on long-term f0 estimation," in *Interspeech*, Stockholm, 2017.

[2] Y. Lokapavani and A. Akila, "Cloud based Organizational Information Access System using Voice Authentication," *International Journal of Pure and Applied Mathematics,* vol. 118, no. 5, pp. 555-563, 2018.

[3] S. S. Nidhyananthan, K. Muthugeetha and V. Vallimayil, "Human Recognition using Voice Print in LabVIEW," *International Journal of Applied Engineering Research*, *Applied Engineering Research,* vol. 13, no. 10, pp. 8126-8130, 10 Number 2018.

[4] V. Singhal, S. Jain and M. Parida, "Train Sound Level Detection System at Unmanned Railway Level Crossings," *European Transport \ Trasporti Europei,* no. 68, Paper no 3, pp. 1-18, 2018.

[5] D. G. Childers and K. Wuand, "Gender recognition from speech. Part I: Coarse analysis," *The Journal of the Acoustical Society of America,* vol. 90, no. 4, pp. 1828-1840, 1991.

[6] K. Rakesh, S. Dutta and K. Shama, "Gender Recognition using Speech Processing Techniques in LabVIEW," *International Journal of Advances in Engineering & Technology,* vol. 1, no. 2, pp. 51-63, 2011.

[7] A. M. Abbasi, M. A. Channa, M. A. Memon, S. John, I. Ahmed and Kamlesh, "Acoustic Characteristics of Pakistani English Vowel Sounds," *International Journal of English Linguistics,* vol. 8, no. 5, pp. 27-34, 2018.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer Program]," 11 May 2018. [Online]. Available: http://www.fon.hum.uva.nl/praat/. [Accessed 15 10 2018].

[9] A. S. Jerzy SAS, "Gender Recognition using Neural Networks and ASR Techniques," *Journal of Medical Informatics & Technologies,* vol. 22, pp. 179-187, 2013.

[10] S. Mavaddati, "Voice-based Age and Gender Recognition based on Learning Generative Sparse Models," *International Journal of Engineering,* vol. 31, no. 9, pp. 1529-1535, September 2018.

[11] B. Ayhan and C. Kwan, "Robust Speaker Identification Algorithms and Results in Noisy Environments," in *International Symposium on Neural Networks*, 2018.

[12] J. W. Kim, J. Salamon, P. Li and J. P. Bello, "CREPE: A Convolutional Representation for Pitch Estimation," in *International Conference of Acoustics Speech and Signal Processing, 2018 IEEE International Conference*, 2018.

[13] Z. Han and X. Wang, "A Signal Period Detection Algorithm based on Morphological Self-Complementary Top-Hat Transform and AMDF," *Information,* vol. 10, no. 1, p. 24, 2019.

[14] G. Alipoor and E. samadi, "Robust Speaker Gender Identification using Empirical Mode Decomposition-based Cepstral Features," *Asia-Pacific Journal of information Technology and Multimedia,* vol. 7, no. 1, pp. 71-81, June 2018.

[15] S. Besbes and Z. Lachiri, "Multitaper MFCC Features for Acoustic Stress Recognition from speech," *International Journal of Advanced Computer Science and Applications,* vol. 8, no. 3, pp. 446-451, 2017.

[16] Z. Qawaqneh, A. A. Mallouh and B. D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," *Knowledge-Based Systems,* vol. 115, pp. 5-14, 2017.

[17] J. S. Suresh and S. A. Thorat, "Language Identification System using MFCC and SDC Feature," *JournalNX,* pp. 113-119, 13 April 2018.

[18] Y. Hu, D. Wu and A. Nucci, "Pitch-based Gender Identification with Two-stage Classification," *Security and Communication Networks,* pp. 1-28, 2011.

[19] H. Hadi and C. Liming, "Voice-based Gender Identification in Multimedia Applications," *Journal of Intelligent Information Systems,* pp. 179-198, 2005.

[20] D. Kawai, K. Yamamoto and S. Nakagawa, "Lyric recognition in monophonic singing using pitch-dependent DNN," in *IEEE ICASSP*, 2017.

[21] "Pakistan Population Census 2017," 2017. [Online]. Available: http://www.pbs.gov.pk/sites/default/files//tables/POPULATION%20BY%20MOTHER%20TONGUE.pdf. [Accessed 5 9 2018].

[22] J. M. Hillenbrand and M. J. Clark, "The role of f0 and formant frequencies in distinguishing the voices of men and women," *Attention, Perception, & Psychophysics,* vol. 71, no. 5, pp. 1150-1160, 2009.

[23] S. Omelchenko, "development of the method of automatic determination of the speaker gender on the basis of joint evaluation of frequency moments of basic tons and formant frequencies," *Information and control system,* pp. 29-33, 2018.

[24] S. P. Dubagunta, B. Vlasenko and M. M. Doss, "Learning Voice Source Related Information for Depression Detection," *IEEE International Conference on Acoustics, Speech, and Signal*

*Processing,* 2019.

[25] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise of a sampled sound," in *Proceedings Institute of Phonetic Sciences 17*, Amsterdam, 1993.

[26] A. Maurya, D. Kumar and R. Agarwal, "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach," *Procedia Computer Science,* vol. 125, pp. 880-887, 2018.

[27] M. A. Nasr, M. Abd-Elnaby, El-Fishawy, A. S. El-Rabaie and S. El-Samie, "Speaker identification based on normalized pitch frequency and Mel Frequency Cepstral Coefficients," *International Journal of Speech Technology,* pp. 1-11, 17 September 2018.

[28] M. A. Nwachuku, "Inverse Filtering Techniques in Speech Analysis," *Nigerian Journal of Technology,* vol. 1, no. 1, pp. 38-42, June 1975.

[29] E. D. SENEVIRATHNA, W. Nishan and JAYARATNE, "Audio Music Monitoring: Analyzing Current Techniques for Song Recognition and Identification," *GSTF Journal on Computing (JoC),* vol. 4, no. 3, pp. 23-34, January 2018.

[30] M. S. S. Safya Bhore, "A Comparative study of formant estimation," *International Journal of Advance Research in Electronics and Communication Engineering,* vol. 4, no. 12, pp. 2879-2882, 2015.

[31] R. Phoophuangpairoj and S. Phongsuphap, "Two-stage Gender Identification Using Pitch Frequencies, MFCCs and HMMs," in *IEEE International Conference on Systems, Man, and Cybernetics*, Hong Kong, 2015.

[32] M. Gomathy, K. Meena and K. Subramaniam, "Gender Clustering and Classification Algorithms in Speech Processing: A Comprehensive Performance Analysis," *International Journal of Computer Applications,* vol. 51, no. 20, pp. 9-17, August 2012.

[33] A. Howard, Elementary Linear Algebra Chapter 3, 10 ed., John Wiley & Sons, 2010.

[34] G. P. Zhang, "Neural Networks for Classification: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics—PART C: Applications and Reviews,* vol. 30, no. 4, pp. 451-462, November 2000.

[35] M. Kirk, Thoughtful Machine Learning, Chapter 4, 1st ed., United State of America: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2015.

[36] G. Chen and T. T. Pham, Fuzzy Sets, Fuzzy Logic and Fuzzy Control Systems, Chapter 3, New York: CRC Press, LCC, 2001, pp. 90-98.

[37] J. S. Garofolo, W. M. F. Lori F. Lamel, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren and V. Zue, *TIMIT-Acoustic-Phonetic Continuous Speech Corpus,* Philadelphia: Lignuistic Data Consortium, 1993.

[38] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults, Unpublished manuscript," Stockholm, Sweden, 1995.

[39] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *The Journal of the Acoustical Society of America,* vol. 24, no. 2, pp. 175-184, March 1952.

[40] W. Navidi, Statistics for Engineers and Scientists, Chapter 1, 4, Ed., New York: McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121, 2015, pp. 33-37.

[41] M. Kirk, Thoughtful Machine Learning, Chapter 8, First, Ed., O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2015, pp. 156-157.

[42] G. Bonaccorso, Machine Learning Algorithms, First ed., Birmingham: Packt Publishing Ltd, 2017, pp. 118-128.

[43] A. Khan, V. Kumar and S. Kumar, "Speech Based Gender Identification Using Fuzzy Logic," *International Journal of Innovative Research in Science, Engineering and Technology,* vol. 6, no. 7, pp. 14344-14351, July 2017.

[44] VSDC, "VSDC free video editor," 2018. [Online]. Available: http://www.videosoftdev.com/support/video-editor-help. [Accessed 18 9 2018].

[45] F. Korbel, "ffmpeg documentation," 2009. [Online]. Available: https://www.ffmpeg.org/documentation.html. [Accessed 23 8 2018].

[46] A. Kanervisto, V. Vestman, M. Sahidullah, V. Hautamäki and T. Kinnunen, "Effects of gender information in text-independent and text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speechand Signal Processing*, New Orleans, USA, 2017.

[47] Z. Ali, M. S. Hossain, G. Muhammad and M. Aslam, "New Zero-Watermarking Algorithm Using HurstExponent for Protection of Privacy inTelemedicine," *IEEE Access,* vol. 6, pp. 7930-7940, December 2018.

[48] A. H. A. Absa, M. Deriche, M. Elshafei-Ahmed, Y. M. Elhadj and B.-H. Juang, "A Hybrid Unsupervised Segmentation Algorithm for Arabic Speech Using Feature Fusion and a Genetic Algorithm (July, 2018)," *IEEE Access,* vol. 6, pp. 43157-43169, 2018.

[49] R. Manikandan and R. Sivakumar, "Machine learning algorithms for text-documents classification: A review," *International Journal of Academic Research and Development,* vol. 3, No. 2, pp. 384-389, March 2018.

[50] W. S. M. Sanjaya, D. Anggraeni and I. P. Santika, "Speech Recognition using Linear Predictive Coding (LPC) and Adaptive Neuro-Fuzzy (ANFIS) to Control 5 DoF Arm Robot," in *ICCSE, Bandung: IOP Conference*, Bandung, 2017.

[51] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. B. Butler and J. Wilson, "Practical Hidden Voice Attacks against Speech andSpeaker Recognition Systems," *The Network and Distributed System Security Symposium (NDSS),* pp. 1-15, 2019.

[52] G. Allwood, X. Du, M. M. Webberley, A. Osseiran and B. J. Marshall, "Advances in Acoustic Signal Processing Techniques for Enhanced Bowel Sound Analysis," *IEEE Reviews in Biomedical Engineering,* vol. 12, pp. 240-253, 2019.

[53] A. Bhowmick, A. Biswas and M. Chandra, "Performance evaluation of psycho-acoustically motivated front-end compensator for TIMIT phone recognition," *Pattern Analysis and Applications,* pp. 1-13, April 2019.

[54] S. R. Dadiri and Y. B., "Estimation of fundamental frequency from singing voice using harmonics," in *Interspeech*, 2018.

[55] Z. Wu, Q. Xu, J. Li, C. Fu, Q. Xuan and Y. Xiang, "Passive indoor localization based on CSI and naive bayes classification," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 48, no. 9, p. 566–1577, 2018.

[56] M. Irfan, W. B. Zulfikar, C. N. Alam and M. A. Ramdhan, "Design of expert system for owning motorcycle with Naive Bayes classifier," in *IOP Conf. Series: Materials Science and Engineering*, 2018.

[57] M. M. Armstrong, A. J. Lee and D. R. Feinberg, "A house of cards: bias in perception of body size mediates the relationship between voice pitch and perceptions of dominance," *Animal Behaviour,* vol. 147, pp. 43-51, January 2019.