Chapter 12
Nearest Neighbor Classification Approach for Bilingual Speaker and Gender Recognition

Samrudhi Mohdiwale and Tirath Prasad Sahu

12.1 Introduction

Speech is the primary form of human communication. From the evolution of telephone, speech is transformed into electric signals by the transducer in order to increase the reachability and enhance modes of communication [1]. Processing of audio signal is important to understand the path between the speaker and listener. The initial step of communication is a thought which is transformed into words, sentences, and phrases according to the grammar of the particular language. A thought comes according to the situation of the surrounding; this excites nerves of the brain to generate electrical signals which further excites the vocal chords and muscles of the vocal tract. This results in vibration as a change of pressure in the vocal tract and lip movement depends on that process of pressure change. Finally, lip movement transfers the generated thought in the form of speech over the space [2]. Space is a medium of communication and it has certain characteristics which are important specially in the area of forensics to recognize the background. Space contains different undesired sound that affects the intelligibility of speech. In order to improve speech intelligibility, noise should be removed from the recorded sequence. The enhancement of speech signal by noise reduction is very effective if enhancement approach is modified according to the type of noise [3].

In the era of handset, speech signal is transmitted in the form of electric signal. These are transformed and decoded by the transducer. Now in the digital era of communication analog to digital converter has introduced to digitally transmit and process the speech signals. Digital technology with its high speed and low cost with reduced power consumption replaces the huge part of analog-based technology. The

S. Mohdiwale (✉) · T. P. Sahu
National Institute of Technology, Raipur, Chhattisgarh, India
e-mail: tpsahu.it@nitrr.ac.in

G.R. Sinha (ed.), *Advances in Biometrics*,
https://doi.org/10.1007/978-3-030-30436-2_12

Fig. 12.1 Analog to digital conversion of speech signal

term digital speech signal processing defines the process of change of analog speech to digital speech that is having desired properties required for further applications. Analog to digital conversion of speech is shown in Fig. 12.1.

Speaker Recognition and Reason for Its Popularity

Speaker recognition thought is initiated from a clue on how a human can communicate with the device without any physical contact such as a keyboard or mouse. The machine can understand digital information and in order to personalize the machine, the speaker should recognize correctly. Everyone in the world wants a sophisticated life with a personalized robot who can work on their single command. Research is nowadays going toward that direction.

As we have seen speech is the most common and effective way of communication between human beings. A speaker plays a very important role in communication to transfer information. Identity of a speaker has a significant impact on research area, since it is a billion-dollar industry and has an excellent opportunity in various application fields. In the digital arena wide-range applications of speaker recognition include personal assistant in mobile and other devices, robotic control, security and forensics, health, and education [1–5]. Smart industries now work on classification of gender and age group of speakers to advertise accordingly for their product [4, 5]. Automatic speaker recognition system verifies the speaker identity with the samples of known identity. Speech is a nonstationary signal, hence hard to process; to work on speech signals small samples of few milliseconds are taken which make them stationary for a short duration of time. Precise speaker recognition focuses on feature extraction and classification methods to enhance the existing models. A variety of features are investigated to classify speakers, Few of them are Pitch intonation, lexical information, prosody which is termed as high-level feature. These are easy to extract from humans but complex to extract via machines. Acoustic spectral features, vocal track length and resonance, and glottal flow are low-level features easy to extract by short-time Fourier transform (STFT) or mel frequency cepstrum coefficient (MFCC), etc. [6]. In industrial use of speech signal processing audio indexing, baking authentication, information retrieval, remote monitoring, and forensics are some popular areas which explore audio processing [4, 5]. Even though it's very popular, it has many research challenges that excite researchers to explore the area.

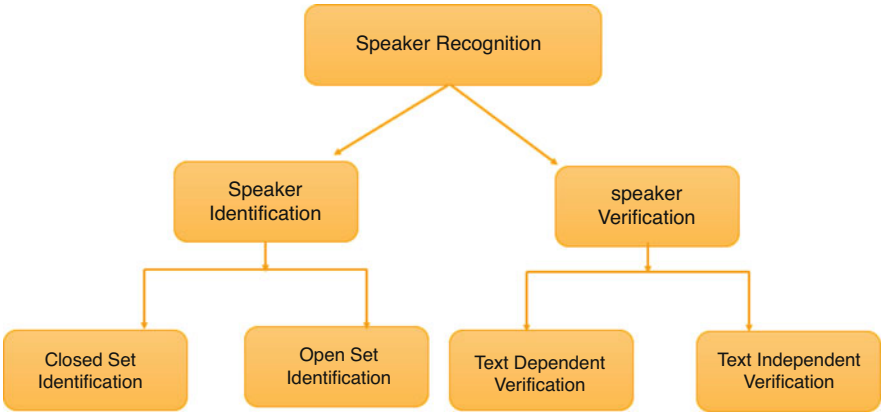


Fig. 12.2 Variants of speaker recognition

12.1.1 Variants of Speaker Recognition

Till now we studied about speaker recognition. Speaker recognition also known as speaker biometrics is a comprehensive term which is used for a procedure to identify a person, by recognizing him with his/her voice. It consists of a variety of branches which are either directly or indirectly related to each other. Speaker recognition variants are shown in Fig. 12.2.

Speaker identification and verification are two variants in which speaker recognition can be classified. The details for each are described below:

Speaker Identification
Speaker identification refers to the process of identifying the unknown speaker. In the process of identification speaker's voice is compared against the set of available speakers. Since for P available speaker it compares one voice against P voices, hence it is also represented as 1:P method [7].

For better understanding let's take an example of dataset that contains five voices: A, B, C, D, and E. If any one of them, suppose A, speaks the identification system, compare the voice of speaker A with all five voices. The identification model produces the output by comparing A's voice with available set of voices based on best similarity of voice.

The identification system of a speaker can be classified into two types:

1. Closed Set Identification

   In this type of identification, the speaker is one of the claimants from the group of members. No new member is introduced for identification. If the unknown speaker is not available in the database, the identification system still assumes the unknown voice is one of the speaker's voices from the available database of speech. This process consumes more time than speaker verification system because it is having multiple comparisons of voices in available database to

identify the person. This type of identification is used to find the person that belongs to which group in the set of available groups. One disadvantage of this process is that unknown person cannot be identified [2].

2. Open Set Identification

   In this type of identification system unknown voices can be introduced. If any unknown speaker speaks, the identification model shows the result "no match found" or "speaker is unknown." The total number of speakers is also unknown in this process [2].

Speaker Verification

Speaker verification refers to the method of verifying the identity of the speaker who is being spoken (test speaker) by nonspeech techniques known as content-dependent techniques. The voice of the speaker is matched with the similar voice present in the dataset; the similar voices are very less generally one or two, so the number of comparisons involved in verification task is less than identification task. This reduces the time complexity of the system [7]. Speaker verification can be done based on text-based features or voice-based features. So speaker verification can be classified into two types:

1. Text-Dependent Speaker Recognition

   This type of speaker recognition requires the speaker to say the exactly given or commanded password. This type of speaker recognition has a robust approach. It only applies to the speaker verification branch. Most of the other branches are in a passive manner; therefore text-dependent speaker cannot be used in various branches. Another use of this type of speaker recognition is that the liveness problem of text-dependent system can be fixed by text-prompted modality [2, 7].

2. Text-Independent Speaker Recognition

   It is one of the most versatile modalities and most feasible modalities which can be used in all branches of speaker recognition. This recognition system is independent of text being spoken by speaker and verify the identity of speaker by glottal features irrespective of text. It includes different categories into it. It also faces a few problems; one of the common problems faced by it is the possibility of a poor coverage of a part of speech. Another problem faced by it is the liveness assessment problem. It is also one of the common problems faced in text-independent system [2, 7].

12.2 Applications of Speech Processing and Speaker Recognition

Nowadays speech recognition (SR) technology industry has developed a broad range of commercial products in which speech processing technology is used efficiently and proves to be very useful. The historical uses of SR technology include dialing numbers by speech, call routing, response of a call with voice

commands, speech to text conversion for fast data entry, voice-controlled devices, games, voice search options, transcription, and robotics. These applications are limited for speech with limited training samples. As the demand of easy and handled resources increases big data came into picture. Computational power with GPU support is also available to handle big data. So by using these facilities personalized assistance on your mobile and PC is introduced; Google now, Siri, and Cortana are the recent developments in the area of SR. Home automation, robotics, and machine translation are growing area nowadays. Few applications are described in detail below.

1. Security

   Security of belongings and property is very much essential. Many companies and residents use different kinds of security systems which are UI-based ID/password for protection. But hacking of this ID/password is happening very frequently and this system is tagged as unsecure. This increases demand of such a system which will not depend on any ID/password. Speaker's voice comes out as the solution to this problem. In speaker recognition system identification of the speaker as well as its verification is based on biometric features of vocal system of humans which will be safer than previous security systems. On the other hand, speaker recognition-based systems are also helpful in criminal cases for investigation of fraud against the voice of the speaker [4]. For practical purpose the voice-based security systems should be built with care of acoustical mismatch at verification phase.

   If the case arises of twins, one study shows that features of twin voices are also unique, so this can be used in administration system [8]. One of the parametric representations of speech is mel frequency cepstral coefficient which gives better results in speaker verification systems for speaker recognition task [9].
2. Crime investigation

   As security issue is discussed above one of the major areas of application of speaker recognition system is for crime investigation. Although there are many tests such as Norco test available, speaker recognition will be very easy and less expensive. This could be possible against voice samples available at crime place or its mobile phone recording for investigation.
3. Personalized robots

   Various patients suffer from very serious disorders and are not able to work by their own. At this time personalized robots are very helpful which understand the direct command just by voice and work will be done with command only. Recently this feature is also available for everyone. Alexa is one of the best example of it.
4. Medical field

   In medical fields this technology is used for medical transcription, electronic medical record, and many more [10]. Medical science frequently deals with patients and nowadays distance medicine can also be supplied based on question answering (QA) with patients. This QA session in text is a very tedious task for doctors in their busy schedule so speech recognition systems will be helpful for

them. At the same time speaker's voice is also helpful to detect many problems. So the speech processing is applied in this area.

5. Education

   Speech recognition software is in demand in the market as it can transcript the speech into text. This is very helpful for the students who are not able to listen. The transcripted lectures can be displayed to them. These lectures also work as notes for the students to revise the concepts [11].

6. Smart homes

   Today the homes are assumed to be smart enough to save energy and time. In smart homes all the devices are work with owner's command. This is all in one application of speaker recognition in which privacy as well as personalized robot is covered.

12.3 Limitations of Speaker Recognition

1. *Not applicable for inarticulate persons*

   As voiceless persons are unable to speak, one cannot obtain speech from them; hence speaker recognition project is not applicable to those persons. This is one of the major drawbacks of speaker recognition devices.

2. *Health issues may create verification problem*

   Speaker verification directly depends on speaker's voice so any health issue such as throat problem, cold, and cough may cause problem at the time of real-time speaker verification.

3. *Effect of aging on voice*

   As the age of person increases from childhood to adulthood, the voice along with its pitch and modulation also change. This has become a major challenge for voice biometric systems.

12.4 Issues and Challenges

- *Nonstationary Signal Processing*

  Acoustic signal changes its frequency with respect to time. The speech signal didn't have gaps between the spoken words; this make it difficult to determine the word boundaries. To solve the problem of nonstationary nature of acoustic signal processing, speech signal is divided into small time stamp (generally in ms) in which speech is assumed to be stationary [12].

- *Lack of Relevant Data*

  Voice biometrics is one of the latest fields of research but it's hard to find speech dataset of the same speaker from his childhood to old age. Similarly, in the field of linguistics relevant data for bilingual or multilingual speaker with

proficiency in both the languages is not available, especially for native languages. This has become a challenge for speaker recognition.

- *Noise Removal*

  The speech itself is a very challenging domain due to its nonstationary nature but noise makes it more challenging. As noise constantly corrupts the speech signal and makes it difficult to process, an efficient method is important for noise removal to get better and optimal performance in speech processing task. Various methods have been proposed for noise removal [8, 9].
- *Multilingual Speaker Recognition*

  Text-independent speaker recognition model is considered as independent of language being spoken but its performance will be affected by multilingual trail conditions [15]. Since the large English corpora are available with multiple speaker recordings with different degradation conditions, therefore it shows better performance for English than any other languages. Lack of multilingual dataset restricts the model and is one of the causes of performance degradation [6].
- *Psychological State of Mind*

  Speech is directly affected by thoughts that came out of the mental state such as emotion, stress, anxiety, health, etc. These psychological states change the modulation of the speech signal and affect the feature of voice that characterizes different speakers [4, 16, 17].

Motivation

Speaker recognition system is a part of voice biometrics. In real-time scenario, this voice biometric system is not yet commercially implemented due to its reduced performance compared to other biometric systems. These are not yet fully reliable but the study reveals that speaker recognition-based systems are highly accessible and acceptable with ease of remote monitoring and low cost [13, 14]. To make this system reliable a perfect model is required which will provide effective preprocessing with faster response time. The system should conform to any situation. These requirements are the source of inspiration to provide better solutions for preprocessing techniques. India is a country of diversity. Many speakers are multilingual and most of the population is using a mix of two or more languages (e.g., Hinglish) instead of any single language, which increases the curiosity about the effect of language on automatic speaker recognition model. Different research on language-independent model encourages us to work in the area of speaker recognition.

Various classification methods such as support vector machine (SVM), Gaussian mixture model-universal background model (GMM-UBM), deep neural network (DNN), and deep belief networks (DBF) have been used for speaker classification [14, 15]. k-nearest neighbor (k-NN) classifier is also one of the effective classifiers with ease of implementation and good generalization ability. Ensemble methods are used to improve the instability of classifier as well as improve k-NN by subspace method by using sensitivity of input space. This improves overall performance of k-NN classification [4].

The rest of the paper is organized as follows. Section 12.2 represents related works, Sect. 12.3 describes proposed methodology for gender and speaker recognition, Sect. 12.4 represents results of classification, and Sect. 12.5 gives the conclusion of the presented method.

12.5 Related Work

Von Kempelen was the first one who demonstrated that human speech production system could be modeled. The first device based on human speech synthesis made in 18th century which is also called as speaking machine. The machine responds to all sounds and words other than Rex but it was found that this also responds to nonspeech signals which have 500 Hz energy. Rex was unable to reject the words which are not available in the vocabulary. Further in 1958 Duley created a classifier which works on spectra rather than formants. Grammar probabilities are added by Dene and from the recognition of few words or sounds the speech recognition concept evolved and enhanced in each decade [21]. Every individual has some specific vocal characteristics which may depend on linguistic means or utterances or may not. Vocal tract, articulator movement, gender, and pitch are some characteristics which allow the listener to identify the speaker. Researchers are more attracted toward speaker recognition due to its potential applications in the area of intelligence, fraud detection, authentication, and many more. Speaker and gender recognition is related area in our current work. Abbas Khosravani et al. [6] proposed multilingual speaker recognition with probabilistic linear discriminant analysis (PLDA) for text-independent speaker recognition category. The paper uses DNN-HMM approach for voice activity detection; i-vectors have been used as features with language-independent PLDA approach to minimize inconsistency in calculated features. Rosa Gonzalez Hautamaki et al. [22] discussed the effect of voice altered by age and its effect on speaker recognition; this shows fundamental frequency relatively less affected than other formant frequencies. While considering both male and female speakers fundamental frequency increases in both the cases but more increment shows in female voice during alteration. Vocal tract outline also has a significant impact of voice disguised in first four formant frequencies. Gang Liu et al. [23] proposed a novel approach to noisy and multi-session enrolled data with five back ends for speaker recognition system with extrinsic and intrinsic back end for extremely discriminative speaker recognition model. Authors also explore more comprehensive set of features for small dimensional i-vector models with diversify contents of information before modelling. Saeid Safavi et al. [24] work on children voice for classification of gender, age, and speaker identification. I-vector-based GMM-UBM and GMM-SVM models are compared for performance evaluation. Spectrum region with most significant information is investigated. This shows 0.9–2.6 kHz frequency band is for gender identification which is the second formant location while 1.6–3 kHz range is useful for age group identification. The human brain collectively interprets the speaker and speech while listening to any

voice. To work simultaneously with these two Zhiyuan Tang et al. [25] proposed multitask recurrent model with collaborative joint training framework. To enhance performance of individual task simultaneously collaborative tasks interchange information accordingly. This method presents another neural network approach for classification. Eduardo Lleida et al. [26] discussed special issues in the area of speaker recognition. The paper focused on issues like spoofing, channel mismatch, spoofing countermeasures, and short speech utterances. Also discussed are state-of-art technologies in speech recognition field. Ankur Maurya et al. [27] present MFCC-GMM approach of speaker recognition with Hindi speech signal. It also discussed the challenges faced during recognition process including psychological and physiological challenges.

12.6 Proposed Method

Bilingual dataset is created to perform experiments on change in language of the same speaker. MFCC feature is extracted and classification is done using the methods described below. Detailed description of methodology is presented in subsequent sections.

12.6.1 Dataset

The dataset for bilingual speaker and gender recognition is created for the research. The dataset consists of 20 speakers; ten male and ten female speaker voices have taken for the research. Voice samples are taken via different mobile phones and in different environments to check the performance of the model in different real-time environments. Speakers are said to record samples in Hindi and English language with given text. The 2-D distribution of dataset is shown in Fig. 12.3. Each color represents MFCC distribution of different genders for gender recognition shown in Fig. 12.3a and MFCC distribution of 20 speakers for speaker recognition shown in Fig. 12.3b. The figure shows distribution of first two consecutive features for both gender and speaker recognition. For all 16 features this combination can be judged.

12.6.2 Flow Chart

The flowchart below shows the flow of proposed method for speaker and gender recognition task (Fig. 12.4).

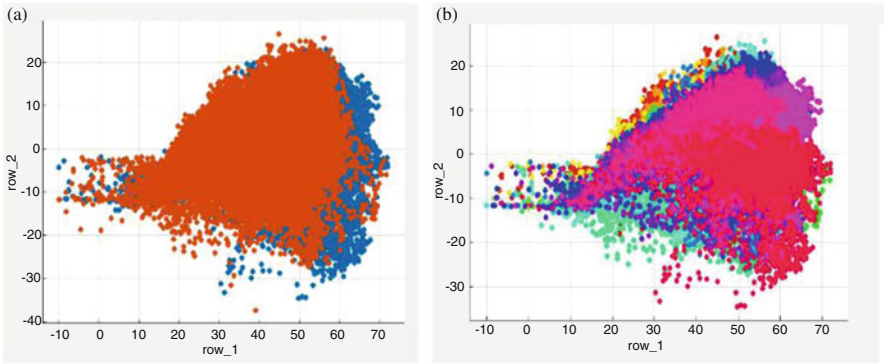


Fig. 12.3 Feature distribution of gender and speaker recognition, respectively. (**a**) Original dataset: gender classification, (**b**) Original dataset: speaker classification

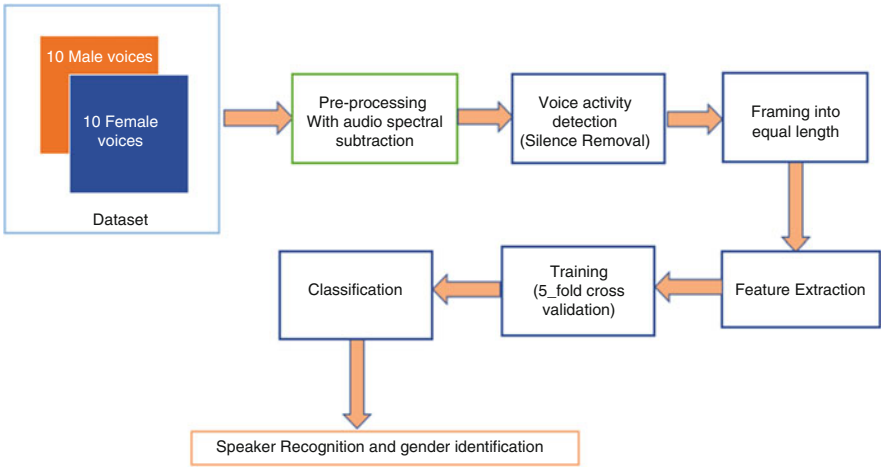


Fig. 12.4 Flowchart of proposed method

12.6.3 Methodology

In this section each phase of proposed method is discussed in detail. Starting from preprocessing, how it is performed for the given dataset, method for silence removal used, framing, feature extraction, and classification methods employed are presented below.

- *Preprocessing*

The initial step of speech processing is the removal of noise from speech signal so that noise cannot degrade the recognition performance. For preprocessing of speech spectral subtraction method is employed. To perform spectral subtraction speech

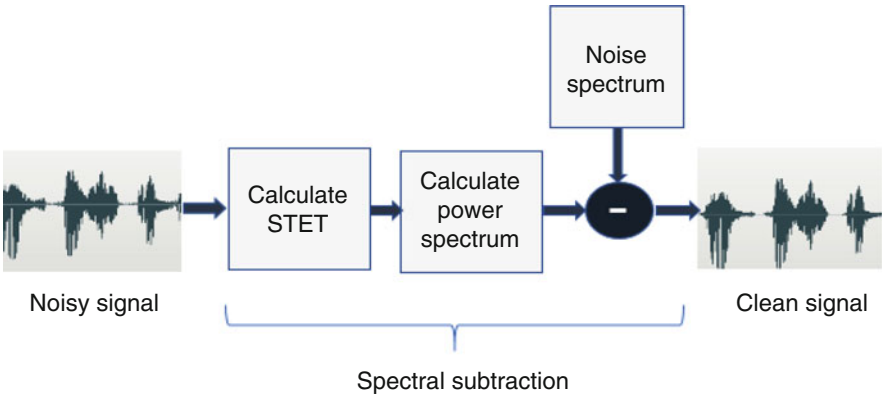


Fig. 12.5 Preprocessing using audio spectral subtraction

is assumed to be uncorrelated with additive noise and have zero mean [28]. The process of spectral subtraction is shown in Fig. 12.5. In this method, filtering of signal gives the less attenuation to high SNR than low SNR of speech sample.

As noise is assumed to be additive for spectral subtraction method, noise spectrum can be subtracted from noisy speech to get the clean speech. Noisy signal is first divided into frames of small window length typically 400 samples with different windowing techniques such as hamming window, hanning window, etc., to make the signal stationary within short time stamp. Short-time Fourier transform or discrete Fourier transform is calculated on each frame further in the process of spectral subtraction. Magnitude and phase spectrum are obtained from these transforms; magnitude transform is used to estimate the power spectrum of the signal. It is assumed that few initial frames are silence frames where no speech is observed; these frames are used for noise spectrum calculation. Now we have both speech and noise spectrums which can be subtracted and clean speech has obtained [33].

- *Silence Removal*

Silence in the speech gives many unnecessary observations which lead to misclassification of data points and affect the performance of the system. Voice activity detection using zero crossing rate is the popular method for silence removal.

Zero crossing rate is the rate of change of sign of the speech signal or rate of crossing the time axis (0–midpoint). Noise has property of maximum zero crossing rate but as it is removed in the previous step only silence points exist in the signal which follows the time axis with value zero. In the silence time stamp zero crossing rate becomes zero and this leads to silence removal from the given dataset (Fig. 12.6).

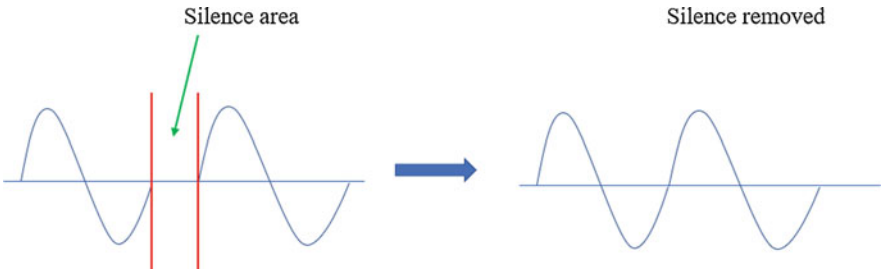


Fig. 12.6 Silence removal

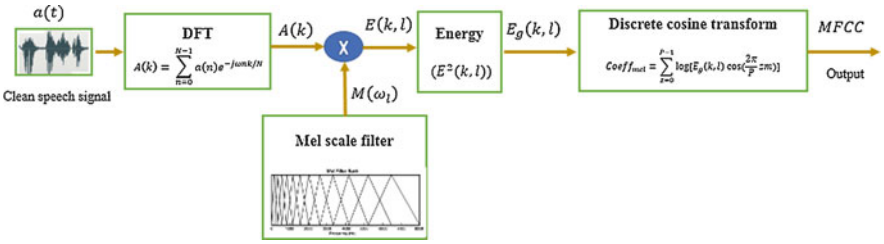


Fig. 12.7 Procedure for MFCC feature extraction

- *Framing*

After silence removal samples obtained having unequal length, to maintain sample size framing is applied. In previous case as speech is nonstationary signal, it's difficult to deal with this signal in continuous domain. To make the signal stationary, split the signal in the interval of 5 s each and apply windowing in each sample to make small samples of milliseconds to extract the features without loss of information due to aperture effect and discontinuities [29].

- *Feature Extraction*

Features are important since the information they produce are the backbone of model for classification purpose. mel frequency cepstral coefficient (MFCC) is the most relevant and effective feature in the field of speech processing [30]. MFCC extracts noncorrelating properties of cepstrum and also compensates channel distortion. Speech sample a(t) is transformed to DFT sample A(k) and windowed by mel scale $M(\omega_l)$. These are further squared to calculate energy. Logarithm of this energy is taken with its discrete cosine transform to get MFCC feature vector. Block diagram for MFCC calculation is shown in Fig. 12.7.

After framing of signal, the discrete Fourier transform (DFT) is calculated. This is important for calculation of energy of the speech signal. Periodogram is an important measure which represents the list of frequencies present in the speech. This function is similar to cochlea of human ear which vibrates and the brain understands the frequency with the help of vibration of different areas of cochlea.

For automatic speaker recognition individual frequency is not important, so a set of frequency named bins is used to calculate the average energy present in the speech. The formation of bins is based on mel scale or mel filterbank. The formula for mel scale is:

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right)$$

In above equation ln has importance of its own because it is a channel normalization technique which allows cepstral mean subtraction [27].

- *Training and Classification*

The feature extracted from the method discussed above is further used in training. From the abovementioned procedure 13-dimensional feature vector has obtained for each class.

Traditional classification method such as SVM for multiclass classification along with k-NN with its variant is used for classification. Classification using k-NN is a lazy learning approach [31]. k-NN represents each sample point in V-dimensional space where V shows the number of features. In this experiment fine k-NN uses one neighbor with equal weight, while medium k-NN uses 10 neighbor, and coarse kNN takes 100 neighbor, equal weight, and Euclidian distance measurement for classification algorithm. In weighted k-NN instead of equal weight squared inverse weight is used with 10 neighbors and Euclidian distance calculated via MATLAB 2016b.

Principle of Nearest Neighbor Classification Approach

The kNN is a proximity-based lazy learner. This classifier is flexible to find training instances that are comparatively similar to test instances. A k-nearest neighbor classifier characterizes each instance as a data point in V-dimensional space where V is the number of features. By having test examples, proximity is computed to the rest of the instances in the training set by using proximity measures [32]. The algorithm for nearest neighbor classification is presented below.

Ensemble models are combination of multiple models created for improved results. Random subspace method (RSM) based on randomized search improves the accuracy but has drawback that randomly decided features cannot guarantee useful information. k-NN is very sensitive to subspace selection; hence a method reduces the error of non-discriminant information. The method subspace is not chosen randomly but selected as the points which reduce the weighted error rate in each boosting step [31]. Algorithm for ensemble subspace k-NN is shown in text box.

12.7 Results

This paper presents a method of speaker recognition and gender identification for bilingual speech signal. To examine the accuracy of speaker and gender recognition, various tests have been performed for classification with different classifiers. The classifiers used are SVM, coarse kNN, medium kNN, fine kNN, ensemble kNN, and weighted kNN. Table 12.1 shows the result in terms of classification accuracy for different classifiers.

From the above table, weighted k-NN performance is best among all the classifiers. Ensemble method is said to be more accurate but this gives less accuracy than others for current dataset. The area under the curve for gender classification and speaker recognition is given in Fig. 12.8. The AUC for gender recognition is 0.98 for weighted k-NN and it is 0.97 for speaker recognition with same classifier.

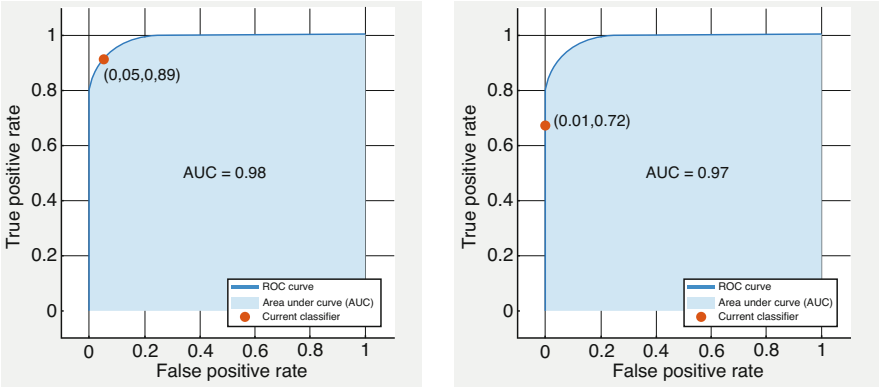
AUC is the area of ROC curve drawn between false-positive rate and true positive rate obtained from the classifier. The AUC is the measure of evaluation for the proposed model. If the model is perfect AUC would equal to 1 and if the model performs random guessing AUC would be 0.5. With reference to Fig. 12.8 it can be seen that the proposed model has AUC near to 1 which shows the better performance of the model using weighted kNN classifier.

Comparative analysis of different classifiers is shown in Fig. 12.9. The comparison can be discussed in following points:

- The bar chart represents higher classification accuracy for gender recognition than the speaker recognition problem.
- Recognition performance for Hindi utterances for both the cases showed maximum accuracy than classical recognition for English utterances.
- Mixed utterances gave lesser accuracy for both the cases that emphasize on the effect of language on classification.

Table 12.1 Classification accuracies with different algorithms

Classifiers / Data	SVM	Coarse kNN	Medium kNN	Fine kNN	Ensemble kNN	Weighted kNN
Gender classification English utterances	82.7	86.5	91.8	92	91.8	**92.6**
Gender classification Hindi utterances	87.1	89.7	93.5	93.9	93.4	**94.2**
Gender classification mixed utterances	84.3	86.4	91.4	91.6	90.7	**92.2**
Speaker recognition English utterances	78.9	71.3	81.7	82.2	82.9	**83.5**
Speaker recognition Hindi utterances	81.2	76.5	86.3	87.5	88	**88.1**
Speaker recognition mixed utterances	76.7	71.1	81.2	82	81.9	**83.2**



(a) AUC for gender recognition mixed utterances (b) AUC for speaker

Fig. 12.8 AUC obtained for weighted k-NN. (**a**) AUC for gender recognition mixed utterances. (**b**) AUC for speaker recognition mixed utterances

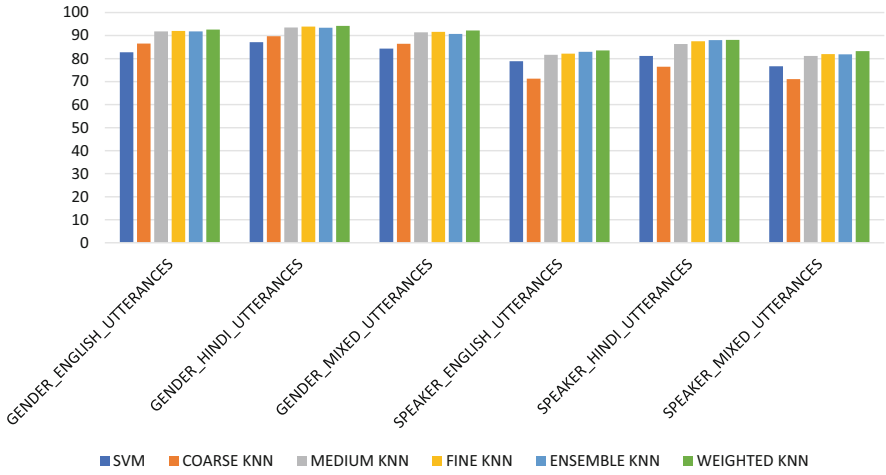


Fig. 12.9 Comparison between various classification methods

- Weighted kNN provides better results for all combination.
- Conventional SVM classifier for speaker recognition gives least accuracy for the present scenario of different languages while fine and coarse kNN have significant performance.

Comparison between various classification methods shown in bar graph (Fig. 12.9).

12.8 Conclusions

In this paper we have discussed the effect of language in the classification of gender and speaker. From the result shown in Sect. 4 weighted k-NN classification accuracy is maximum; this shows k-NN, itself a lazy classifier, better classifies than eager classifiers like SVM for the similar dataset described in the paper. Ensemble method of subspace k-NN also gives significant accuracy.

Language is another factor that has been considered in the current work to improve speech processing techniques. kNN-based bilingual speaker and gender recognition model has been proposed to increase the versatility of recognition systems. kNN-based model predicts the speaker and gender of the speaker from the list of present speakers. Extracted MFCC feature effectively creates envelope of vocal tract. Experimental results showed that weighted kNN outperforms than other classifiers for both speaker and gender recognition. Weighted kNN classification accuracy is maximum; this shows kNN, itself a lazy classifier, better classifies than eager classifiers like SVM for the similar dataset described in the current work.

From Table 12.1, it is seen that language affects the classification performance of both gender and speaker, speaker recognition performance more affected than gender recognition. Since the speakers and background environment are same for both Hindi and English recording sequence, classification performance degrades due to different handsets used for recording.

Future scope of the presented work can be in the direction of improving the performance of speaker and gender recognition. The approach can be extended to identify speakers with the presence of multiple factors such as noise, emotions, and language simultaneously. Multilingual dataset can also be taken into account for the current work's future direction.

References

1. B. Gold, N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (Wiley, New York, 1999)
2. J.P. Campbell, Speaker recognition: a tutorial. Proc. IEEE **85**(9), 1437–1462 (1997)
3. W. Yuan, B. Xia, A speech enhancement approach based on noise classification. Appl. Acoust. **96**, 11–19 (2015)
4. N. Singh, R.A. Khan, R. Shree, Applications of speaker recognition. Procedia Eng. **38**, 3122–3126 (2012)
5. P. Rose, Technical forensic speaker recognition: evaluation, types and testing of evidence. Comput. Speech Lang. **20**(2–3 Special issue), 159–191 (2006)
6. A. Khosravani, M.M. Homayounpour, A PLDA approach for language and text independent speaker recognition. Comput. Speech Lang. **45**, 457–474 (2017)
7. H. Beigi, *Fundamentals of Speaker Recognition* (Springer US, Boston, 2011)
8. M.F. Abdollah, M.N. Kamarudin, H.N.M. Shah, M.Z. Ab Rashid, C.K. Lin, Z. Kamis, Biometric voice recognition in security system. Indian Journal of Science and Technology **7**(2), 104–112 (2014)

9. C.S. Kumar, P.M. Rao, Design of an automatic speaker recognition system using MFCC, vector quantization and LBG algorithm. Int. J. Comput. Sci. Eng. **3**(8), 2942 (2011)
10. M.A. Grasso, Automated speech recognition in medical applications. M.D. Comput. Comput. Med. Pract. **12**(1), 16–23 (1995)
11. K. Bain, S.H. Basson, M. Wald, Speech recognition in university classrooms, in *Proceedings of the Fifth International ACM Conference on Assistive Technologies – Assets'02*, 2002, p. 192
12. J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications* (Academic Press, Amsterdam, 2015)
13. F. Saki, N. Kehtarnavaz, Background noise classification using random forest tree classifier for cochlear implant applications, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, 2014
14. N. Mohammadiha, P. Smaragdis, A. Leijon, Simultaneous noise classification and reduction using a priori learned models, in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2013
15. A.F. Martin, C.S. Greenberg, NIST 2008 speaker recognition evaluation: performance across telephone and room microphone channels, in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 2579–2582
16. N. Cvijanović, P. Kechichian, K. Janse, A. Kohlrausch, Effects of noise on arousal in a speech communication setting. Speech Commun. **88**, 127–136 (2017)
17. A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, C. Di Natale, Continuous estimation of emotions in speech by dynamic cooperative speaker models. IEEE Trans. Affect. Comput. **8**(3), 314–327 (2017)
18. S. Gold, Voice biometrics: real-world issues and solutions. Biometr. Technol. Today **2010**(5), 6–7 (2010)
19. O. Nehru, K. Kumar, Review paper of voice biometrics. Int. J. Eng. Sci. Comput. **6**(5), 5257–5260 (2016)
20. S. Chu, S. Narayanan, C.C.J. Kuo, Environmental sound recognition with time frequency audio features. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1142–1158 (2009)
21. F. Richardson, S. Member, D. Reynolds, N. Dehak, Deep neural network approaches to speaker and language recognition. IEEE Signal Process. Lett. **22**(10), 1671–1675 (2015)
22. R. González Hautamäki, M. Sahidullah, V. Hautamäki, T. Kinnunen, Acoustical and perceptual study of voice disguise by age modification in speaker verification. Speech Commun. **95**(March), 1–15 (2017)
23. G. Liu, J.H.L. Hansen, An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 1978–1992 (2014)
24. S. Safavi, M. Russell, P. Jančovič. Peter, Automatic speaker, age-group and gender identification from children's speech. Comput. Speech Lang. **50**, 141–156 (2018)
25. Z. Tang, L. Li, D. Wang, R. Vipperla, Collaborative joint training with multitask recurrent model for speech and speaker recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(3), 493–504 (2017)
26. E. Lleida, L.J. Rodriguez-fuentes, Speaker and language recognition and characterization: introduction to the CSL special issue. Comput. Speech Lang. **49**, 107–120 (2018)
27. A. Maurya, D. Kumar, R.K. Agarwal, Speaker recognition for Hindi speech signal using MFCC-GMM approach. Procedia Comput. Sci. **125**, 880–887 (2018)
28. N. Upadhyay, A. Karmakar, Speech enhancement using spectral subtraction-type algorithms: a comparison and simulation study. Procedia Comput. Sci. **54**, 574–584 (2015)
29. P. Nguyen, D. Tran, X. Huang, D. Sharma, Automatic speech-based classification of gender, age and accent, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6232 LNAI (2010), pp. 288–299

30. S. Sremath, S. Reza, A. Singh, R. Wang, Speaker identification features extraction methods: a systematic review. Expert Syst. Appl. **90**, 250–271 (2017)
31. N. García-Pedrajas, D. Ortiz-Boyer, Boosting k-nearest neighbor classifier by means of input space projection. Expert Syst. Appl. **36**(7), 10570–10582 (2009)
32. T. Pang-Ning, M. Steinbach, V. Kumar, Introduction to Data Mining (2006)
33. http://practicalcryptography.com/miscellaneous/machine-learning/tutorial-spectral-subraction/