International Journal of Speech Technology (2022) 25:773–782
https://doi.org/10.1007/s10772-022-09986-5

Timbre features with MEDIAN values for compensating intra-speaker variability in speaker identification of whispering sound

Vijay M. Sardar[1] · **Manisha L. Jadhav**[2] · **Saurabh H. Deshmukh**[3]

Received: 20 March 2020 / Accepted: 17 June 2022 / Published online: 3 August 2022

Abstract
Because voiced phonations are absent in the whisper, distinguishing among the speakers with whispered voice is a difficult task. The selection of audio descriptors appropriate for the type of database and application ensures the accuracy of the speaker identification system. The various audio descriptors are investigated here; the timbre features outperform others in identifying the whispering speaker. The Hybrid Selection Algorithm sorts only the best-performing and thus a limited timbre features. When tested on the CHAIN database, the timbre features combined in the form of vector (i.e. MFCC + Roll-off + Brightness + Roughness + Irregularity) increase the identification results by 7.72% compared to traditional MFCC features. Also, to avoid the spread among intra-speaker samples, MEDIAN values of feature vector are investigated, and it reported further enhancement of 2.23%.

Keywords Speaker identification · Whisper · MFCC · Timbre feature · K-means classifier · K-NN classifier · MEDIAN · FAR · FRR

1 Introduction

Speaker Identification (SID) technology is voice biometrics gaining popularity for voice-assisted devices and authentication applications. Traditionally, the MFCC is being used to extract the speaker-specific information (Maurya et al., 2018) represented as audio features or descriptors. However, existing noise in a real scenario, intra-speaker variability, etc. degrades performance in speaker identification. Many approaches are proposed in the literature for noise removal (Al-Allaf, 2015; Manasa & Rama, 2020); noise reduction is always a big issue for many well-known reasons. First, the noise signals in speech are non-stationary, so estimating the statistics for noise removal is tedious. Second, speech distortion is usually observed while speech enhancement.

Due to its hidden and perceptual characteristics, identification of the person from the whispered utterances is a complicated process. The rich phonation's identifiable separation ability is absent from a whisper (Bimbot et al., 2004; Singh & Joshi, 2020). The ADs that work well with the neutral database don't work efficiently with the whispered one. As a result, the customized strategy indicated in Fig. 2 is being utilized, in which the suitable ADs are selected at the beginning only. The identification rate will be improved by picking the proper ADs suitable for the type of database and application.

In the literature, many low-level audio descriptors are explored. The energy, speech bandwidth, spectral centroid, zero-crossing rate etc. are all important attributes of audio signal (Bhattacharjee et al., 2018; Fan et al., 2011). Mel-Frequency Cepstral Coefficients (MFCC), roll-off, and brightness are examples of second high-level and complex descriptors utilized for speech and speaker identification. The parametric analysis of the spectral envelope is used for such descriptors (Davis & Mermelstein, 1980). The different audio descriptors used for speech processing should be decorrelated from every other descriptor. For music and voice

✉ Vijay M. Sardar
vijaysardar@jspmjscoe.edu.in

Manisha L. Jadhav
manisha.shinde1@gmail.com

Saurabh H. Deshmukh
saurabh.h.deshmukh@gmail.com

[1] Jayawantrao Sawant College of Engineering, Pune, M.S., India

[2] MET's Institute of Engineering, Nasik, M.S., India

[3] Maharashtra Institute of Technology, Aurangabad, M.S., India

classification, which is a comparatively simple task, typical classification results employing improved non-correlated MFCC features reported accuracy of up to 95% (Hermansky & Malaya, 1998). These outcomes, however, are based on a clean audio signal setting. When we consider the problems of noise, inter-session variability and telephone speech etc. the performance degrades considerably (Dobrowohl et al., 2019; Toonen Dekkers & Aarts, 1995). The whisper is like a noise due to air turbulence by vocal efforts; hence the traditional audio descriptors are replaced by selected timbre descriptors features to maximize the accuracy.

The general strategy found in the literature for speaker identification research is to develop a statistical model to justify the applicability of audio features for the database in question and then utilize it (Foulkes & Sóskuthy, 2017; Karvanagh, 2519). It is shown in Fig. 1

However, the unknown and hidden reasons for the perceptual audio feature's good performance may not be justified. The whispered database lacks phonations, so the intangible timbre features are proposed; each researcher defined timbre differently. It must capture some additional concealed speaker-specific information (Failed, 2004).

While sorting the best-performing features, eight probable ADs are targeted here that are good mixes of various domains like time, frequency, cepstral, and wavelet. Any of the unknown attributes of these ADs may contribute to the enhancement of performance; therefore Hybrid Selection method is a good choice. As a result, the modified strategy is adopted like one illustrated in Fig. 2; where a selection of the suitable ADs is processed in the beginning only.

This paper is organized as follows. Audio features are classified in two different ways in Sect. 2, followed by a description of timbre features included in the MIR toolbox

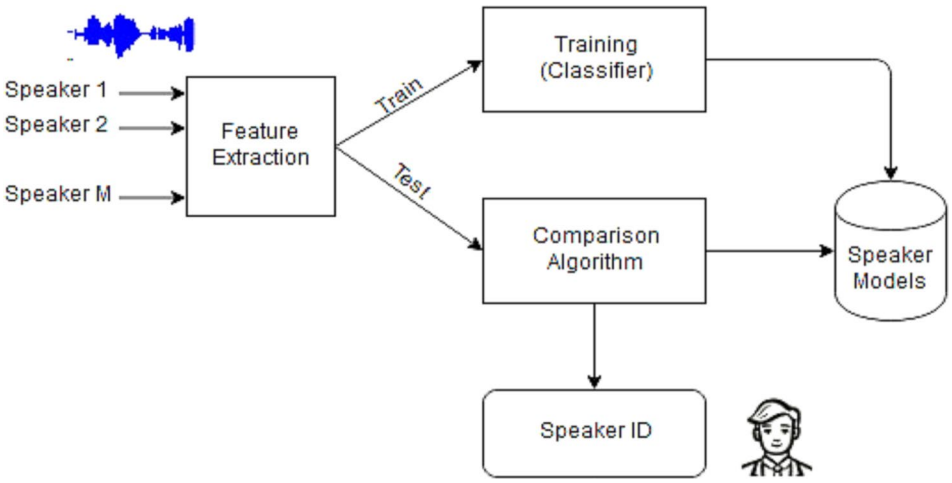

Fig. 1 Block diagram of Generalized Speaker identification System

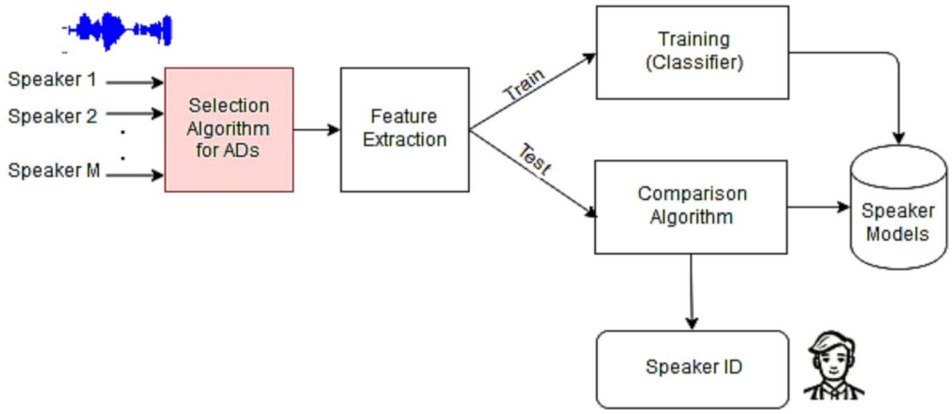

Fig. 2 Block diagram of Modified Speaker identification System

and concluded the impact of timbre feature selection on the identification accuracy in the whispered database. Section 3 is dedicated to the System Description. The methodologies and resources used for the study are explored in detail. It includes a database, a Hybrid selection algorithm, and Classifiers (K-means and K-NN). Section 4 emphasizes the use of Median Values of Timber Features for the whispered database. It also supports its impact on decreasing the intra-speaker spread by illustration. Results on the performance of k-means/ K-NN, different features as MFCC only, timbre features, and median values of timbre features are presented in Sect. 5. Few results on FAR (False Acceptance Rate) and FRR (False Rejection Rate) are also presented.

2 Audio Features

2.1 Review of audio features

Audio features can be divided into two categories. The global descriptor is a type of feature in which the computations are done on the entire signal. For example, the whole duration of an audio stream can be used to determine the attack time of a sound. Instantaneous descriptors are another class of descriptors that works with a single frame of audio data at an instance (40 ms). Because the spectral centroid in a audio signal can change over time, it is referred to as an instantaneous descriptor. As an instantaneous descriptor generates many values for a given number of frames, statistical processes (such as mean or median, standard-deviation, and inter-quartile range) are needed to provide a single value representation. A list of 166 audio features is offered in the CUIDADO project (Peeters, 2004).

Further differentiation can be made based on the method of extraction as shown in Figure 3:

Every individual method will be differently effective for the type of database.

2.2 MIR toolbox Matlab for timbre audio descriptors

The Musical Information Retrieval (MIR) toolbox is mainly designed to enable the study of the relation between musical attributes and music-tempted sensation. MIR toolbox uses a modular outline. It is well known that the common algorithms are used in audio processing like segmentation, filtering, framing etc. with an addition of one or more distinguished algorithms at some stage of processing. These algorithms are available in a modular form and the individual blocks can be integrated to capture some features (Albert-Ludwigs-Universität Freiburg, 2007).

We see MIRToolbox, an integrated set of functions written in Matlab, dedicated to the extraction of sound records related to timbre, tonality, rhythm or form of music. It offers the modular and craftsmanship of a computational approach for Music Information Retrieval (MIR). The different algorithms are decomposed into stages, formalized using a minimal set of elementary mechanisms, and integrated with different variants. We have formulated a piece strategy (Fig. 4) for this study. Before that, it is essential to define the timber features of concern that are used in the subsequent work.

Roll-off frequency Roll-off is assessed from the foremost energy (85% or 95% as a standard) contained below the predefined frequency.

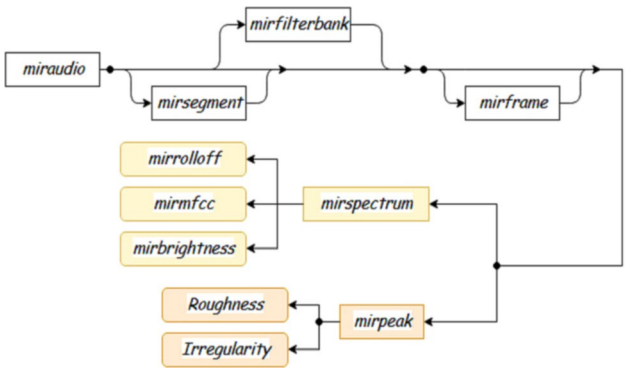

Fig. 4 Philosophical integration of modules for the timbre features of concern in MIR toolbox

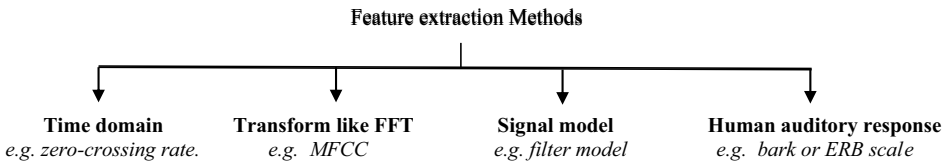

Fig. 3 Classification of Audio features based on the extraction method

Roughness It estimates the average disagreement between all peaks of the signal. It is also an indicator of the presence of harmonics generally higher than the 6th harmonic.

Brightness It is the measure of the percentage of energy spread above some cut-off frequency.

Irregularity It may be calculated as the sum of the square of the difference in amplitude between adjoining partials or the sum of the amplitude minus the mean of the past, the same component and subsequent amplitude.

After defining the algorithm of timbre features of concern, philosophical discussion on the integration of modules in the MIR toolbox resumes. For illustration to measure irregularity and brightness, we need the implementation of an algorithm like reading audio samples, segmentation, filtering, and framing as the common processes between them. Within the last arrangement, due to characteristic contrasts, irregularity needs a peaking algorithm and brightness is spectrum analysis. Even, the integration of different stages depends upon parameter variations. E.g. *mirregularity (…, 'Jensen'),* where the adjoining partials are taken into consideration and *mirregularity (…, 'Krimphoff')* which considers the mean of the preceding, same and next amplitude.

miraudio: This command loads the appropriate format of an audio file. *E.g. miraudio ('speaker.wav').*

mirsegment: This process splits a continuous audio signal into homogeneous segments.

Mirfilterbank: A set of filters are required which are useful to select neighboring narrow sub-bands that cover the entire frequency range. The effect like aliasing in the reconstruction process is avoided *e.g. mirfilterbank (..., 'Gammatone')* processes a Gammatone filterbank decomposition.

mirframe: The frame decomposition can be performed using the *mirframe* command. The frames can be specified as follows:

mirframe (x,…, 'Length', w, wu).

mirspectrum: Discrete Fourier Transform decomposes the energy of a signal (be it an audio waveform, an envelope, etc.) along with the frequencies.

Mathematically, for an audio signal x;

$$Xk = \sum_{n=0}^{N-1} xne^{\frac{-2\prod ikn}{N}} k = 0, \ldots, N-1 \quad (1)$$

This decomposition is performed using a Fast Fourier Transform by the *'mirspectrum'* function.

Mirpeaks: Many features like irregularity require the Peaks analysis. Peaks are calculated from any data x produced in the MIR toolbox using the command *'mirpeaks(x)'.*

In most of the studies, Timber features have been used for Music processing. Timbre feature covers almost all the domains of feature extractions that showed in the Fig. 3. There exists a variety of processing mechanisms available in the MIR toolbox. Hence, it is estimated to be helpful to capture hidden speaker-specific information in the whispered sound using the timbre class.

3 Description of proposed system

The major system components like the database and the role of the Hybrid Selection Algorithm in the selection process of ADs and classifiers are described in the subsequent discussion.

3.1 Database

The database utilized for the undertaking comprises 36 speakers with 33 tests each; with a good blend of male and female voice tests (20 guys and 16 females). It is the CHAIN database created at 'School of Computer Science and Informatics College Dublin' (Cummins et al., 2006). The duration of 2–3 s is recorded at 44.1 kHz. The sentences are chosen from CSLU and TIMIT database which guarantees the phonetic adjustment within the corpus. The database may be divided into different sub-databases (DB1, DB2, DB3, and DB4) to determine the contribution of individual ADs. Figure 5 presents the framework to analyze all the databases and automatically select the appropriate audio descriptors which can maximize the results. The processes can be divided into two parts, using the application software and the system software.

3.2 Hybrid selection algorithm

Hybrid selection is an iterative process that begins with the targeted timbre class of Audio descriptors and progresses to the ADs with the best identification result (Deshmukh & Bhirud, 2012). This technique was also utilized to classify abnormal images of liver tissue in Li et al. (2016).

After every iteration, the sorted AD which maximizes the classifier accuracy is appended by the remaining ADs for next iteration. The process continues until no further increase in accuracy is observed.

As shown in Fig. 6 below, all the eight targeted ADs are individually investigated for the accuracy in speaker identification experiment. (i) *Iteration I:* sorted three features offering the highest accuracy namely MFCC, Roll-off, and brightness. (ii) *Iteration II:* In this iteration, the sorted single ADs are combined with all remaining ADs and performance is evaluated for the combination of two ADs. The first three highest performances with the combination of two ADs

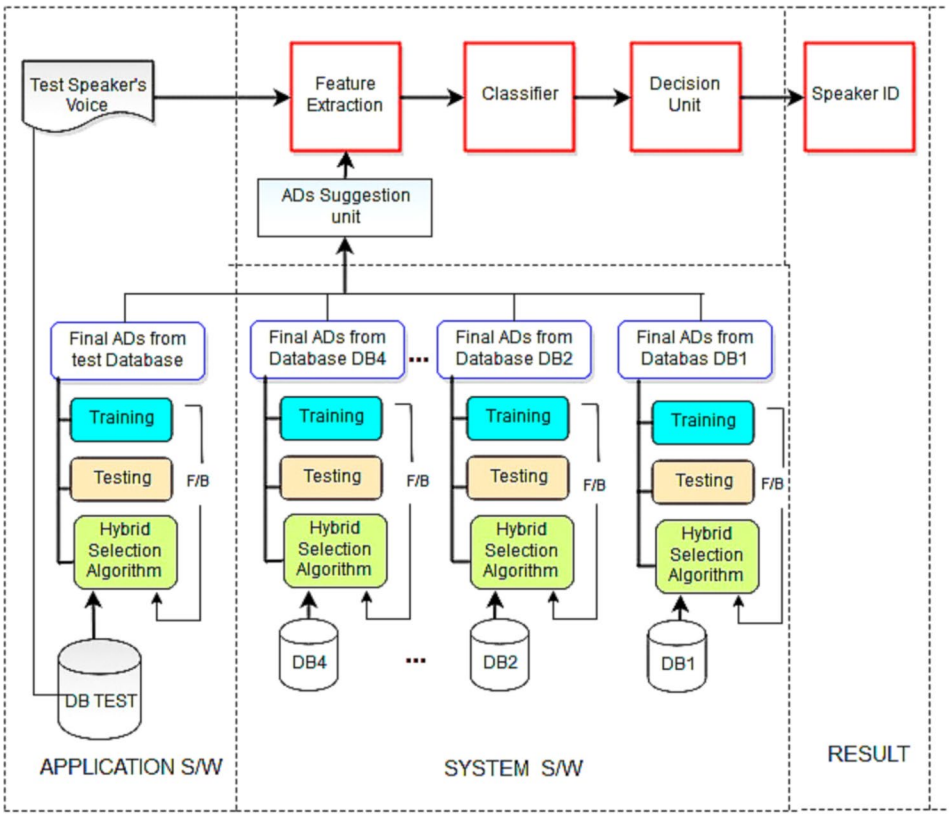

Fig. 5 Speaker Identification Architecture with feature selection algorithm

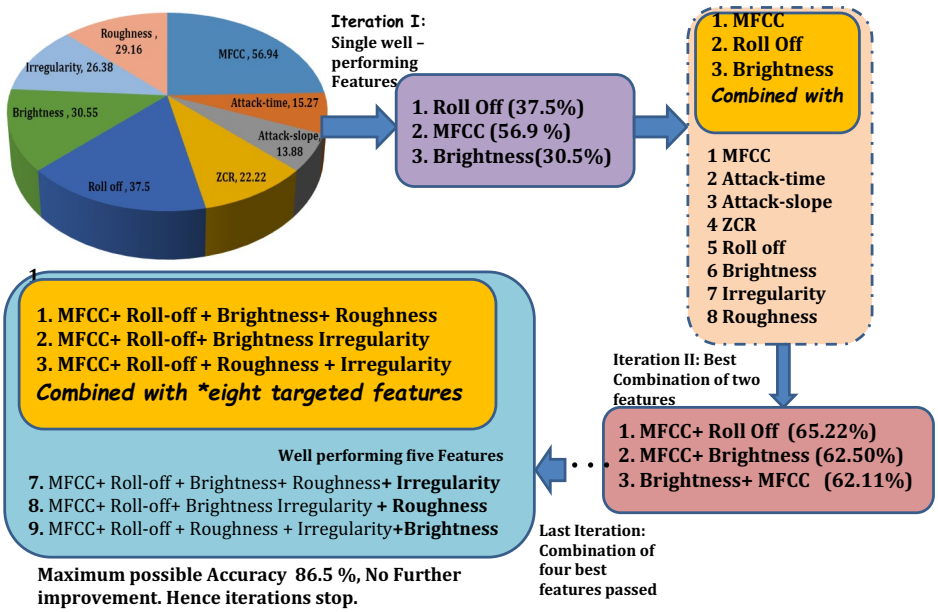

Fig. 6 Pictorial demonstration of Hybrid Selection Algorithm

are sorted MFCC + Roll Off, MFCC + Brightness, Brightness + MFCC (iii) Next iteration sorts the first three best performances with the combination of three ADs, four ADs and, (iv) the *Last iteration* sorts the best performances with the combination of best MFCC, Roll-off, Roughness, Brightness, and Irregularity. Now the process terminates as there is no further improvement by appending the ADs.

3.3 Classifiers

3.3.1 K-means classifier

A semi-supervised learning strategy of K-means clustering is adopted in this study. The audio feature samples are partitioned into clusters by the algorithm. To partition the data into clusters, 'k' number of centroids is assumed. Each feature is combined in a particular cluster based on the minimum distance from a particular centroid. K-means clustering aims to partition the 'n' observations into k ($\leq$n) set S = {S_1, S_2,.., S_k} so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance) (Ito et al., 2005). To be specific, the purpose is to find:

$$\arg\min \sum_{i=1}^{k} \sum_{x \in si} \|x - \mu\|^2 = \arg\min \sum_{i=1}^{k} |si| \text{ Var Si} \quad (2)$$

3.3.2 K- nearest neighbor (K-NN)

K-NN may be a straight forward and non-parametric algorithm which separates the data points into several classes. The points in the query samples are combined in the defined classes based on the distance metric. It is also called a lazy algorithm as this classification does not make any assumptions about the distribution of data. The real-world data does not comply with the usually assumed pattern (e.g. linear regression models). Hence, K-NN classifier is useful in general. While managing with this classifier, the following parameters are used: the number of nearest 'neighbors' (k), a distance function (d), decision rule and n labeled samples of audio files X_n. The query sample is assigned to one of the labels among the existing classes based on the minimum distance in the proximity of several neighbors (2-NN (nearest neighbors), 3-NN) from the training classes. In another word, K-NN calculates a posteriori class probabilities P (wi|x) $for P(w_i)$ outcome as below:

$$P(w_i|x) = \frac{k_i}{k}.P(w_i) \quad (3)$$

where k_i is the number of vectors which belongs to a class within the subset of k vectors (Shah et al., 2015).

KNN classifier allocates a class label to the query sample based on the closest distance from the training classes called the nearest neighbors. The selected five features like brightness, roll-off, irregularity, roughness and MFCC are extracted and rearranged in a vector form. The distances between the query feature vector and the feature vector of all other existing classes are calculated. The Euclidean distance is a popular distance metric, and the City-block distance is another for minimizing the effect of any the much-deviated feature/s; both distance metrics are exercised in the study (Sreelekshmi & Syama, 2017).

- Euclidean Distance: n- dimension Euclidean distance applies as:

$$D(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \quad (4)$$

  where x is the coordinates of the training feature vector and y is the coordinates of a query feature vector.
- City-block: The City -block (Manhattan) distance between a pair of points, x and y, with n dimensions is calculated as:

$$\sum_{j=1}^{n} \left| x_j - y_j \right| \quad (5)$$

The vector consists of multiple features; some features may have high intra-speaker variations (though undesirable) for some speech samples. The effect of such a high difference in a single dimension is diminished since the distances are not squared for City-block distance.

For our system, all the variants of the KNN classifier are verified to maximize the identification accuracy. The variations tested for the number of nearest neighbor are 1-NN, 2-NN, and 3-NN. Two distance functions Euclidean and City-block are investigated. The rules namely nearest and consensus are also tested. After a variety of experiments, it is concluded that a combination of *3-NN* neighbors, *City-block* distance and the *Nearest* rule give the maximum identification accuracy in Sardar and Shirbahadurkar (2018).

4 The role of median values of timber features

In the speaker identification task, inter-speaker variability is one of the reasons to degrade the performance. Standard Deviation (σ) is one of the statistical tools that are used to examine the variations among the same speaker and hence the corresponding feature values. The evaluated standard deviation value needs to be either added or subtracted from the feature value (i.e. feature value $\pm \sigma$). However, using the standard deviations for feature modifications is intricate for two reasons. It requires a complex decision algorithm for every feature value of every speaker sample. Second, modification of the feature values with standard deviation may exceed the normalization range. MEDIAN

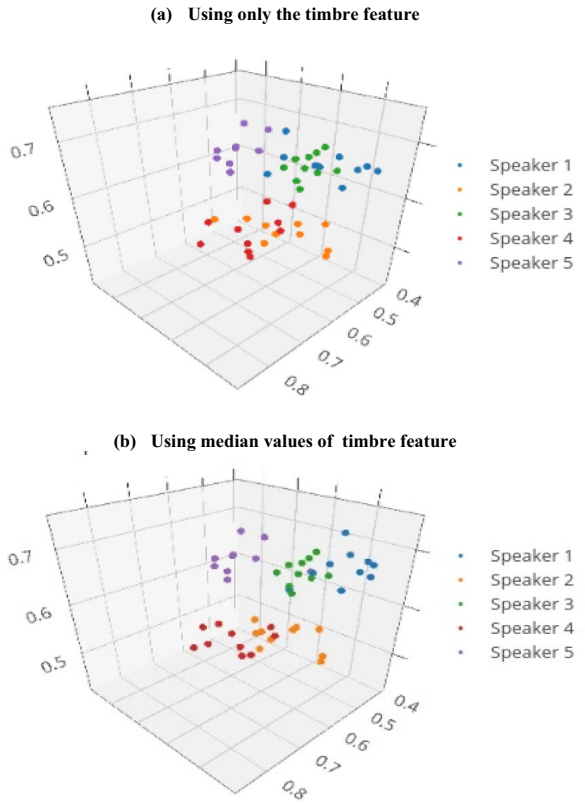

Fig. 7 Effect on the intra-speaker variability due to use of Median values of timbre features

formula considering even and odd number of samples is as below:

$$= \begin{cases} X\left[\frac{n}{2}\right] if\ n\ is\ even \\ \frac{X\left[\frac{n-1}{2}\right]+X\left[\frac{n+1}{2}\right]}{2} if\ n\ is\ odd \end{cases} \quad (6)$$

X = ordered list of values in a data set; n = number of values in the data set.

As a result, MEDIAN can be thought of as the fully sheared mid-range. The median values of the individual features are opted to minimize the intra-speaker spread. The following illustration uses a few samples of five speakers to represent the feature vector(MFCC + Roll-off + Brightness + Roughness + irregularity in the feature space by a single-valued dot. Part (a) of the figure shows the plot of the feature vector when the direct values of timbre features are utilized; while part (b) is the plot after using the median values of the timbre features (Fig. 7).

The illustration in Figure x proves that the feature samples of each class (speaker here) are closely spaced with minimum intra-speaker distance when median values of timbre features are used instead of absolute values.

5 Results and evaluation

5.1 Identification accuracy

Mel Frequency Cepstral Coefficient (MFCC) is a widely used feature for speaker identification tasks. Table 1 illustrates the comparative performance of K-means and K-NN classifier. A total of 35 speakers with 33 whispered samples each from a CHAIN database in a whisper train-whisper test scenario are tested. The samples of each speaker are selected with a choice of 70% samples for training and 30% for testing.

The results shown in Table 1 for the same feature (i.e. MFCC) proved that the K-NN classifier is most suited here.

The Table 2 shows results using the K-NN classifier with parameter settings as—Rule: nearest, Neighbor: 3, and distance Metric: City-Block distance. The selected Timbre features by the Hybrid selection Algorithm are MFCC, Roll-off, Brightness, Roughness, and irregularity. Also, the results are examined using the median values of timbre features (Fig. 8).

Compared to the conventional MFCC features, the identification accuracy utilizing chosen timbre descriptors is upgraded by 7.72%. Further, using median values of timbre features enhances the outcomes by about 2.23%. It is due to the compensation of intra-speaker spread by the advent of Mean values (Table 2).

The results are compared with a baseline speaker identification system. The baseline system also used the whispered data from the CHAINs database. The highest identification accuracy using NDMP Based Fusion System ($\alpha = 0.70$) + SVM whisper train-whisper test setting reported the highest results as 83.75% in Wang et al. (2015) which are reproduced in Table 3.

Compared to the highest result results given by baseline system (83.75%) shown in Table 3, results achieved by the

Table 1 Comparative accuracy using MFCC features and K-means/KNN classifiers

Classifier	K-means clustering	KNN
Audio feature	MFCC	MFCC
% Identification accuracy	67.04	78.81

Table 2 Comparative Identification accuracy by using features MFCC, Timbre only and Timbre (Median)

Speech mode Training-testing	% Accuracy		
	MFCC only	Timbre	Timbre (Median values)
Whisper-Whisper	78.81	86.53	88.76

CHAIN database of 35 speakers, MFCC only/timbre feature/timbre (median), and K-NN Classifier

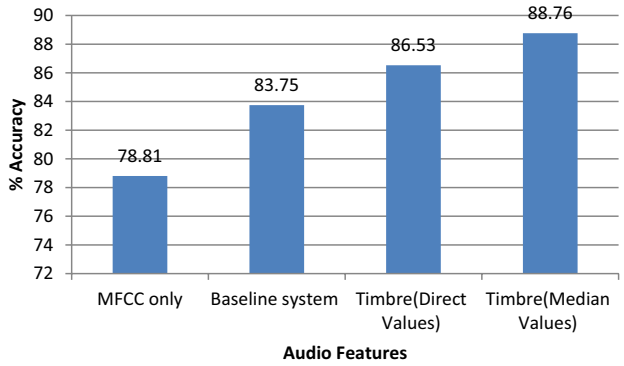

Fig. 8 Comparison of speaker identification accuracy by proposed study with MFCC and Baseline system

Table 3 Baseline results of speaker identification accuracy by Timbre features

Speech mode		% Accuracy
Training	Testing	
Neutral	Neutral	95.0
Whisper	*Whisper*	*83.75*
Neutral	Whisper	73.0

Whisper-train and whisper-test conditions in the base paper is emphasized and compared with this research

proposed system using median values of timbre features i.e. 88.76% (Table 2) report the increase by 5.01%.

5.2 False acceptance and false rejection rate

The false-positive rate (FPR) is the proportion of all negatives that still yield positive test outcomes while the False-negative rate(FNR) is the proportion of all outcomes which yield negative tests.

$$\mathrm{FPR} = \mathrm{FAR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN}) \tag{7}$$

$$\mathrm{FNR} = \mathrm{FRR} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TP}) \tag{8}$$

True Positive (TP), True Negative (TN), False positive (FP) and False Negative

Table 4 shows performance calculation on sample basis, i.e.FPR and FNR. Randomly five speakers 18 to 22 are considered for calculations of FAR and FRR.

Table 4 Sample calculations of performance parameters for speakers 18 to 22

Speaker	18	19	20	21	22
TP	7	10	8	9	9
TN	341	339	340	338	340
FP	1	1	0	2	0
FN	1	0	2	1	1
FPR	0.00292	0.00294	0	0.00588	0
FNR	0.125	0	0.2	0	0.1

FNR and FPR should be un-doubtfully low, but they are differently influencing the different applications.

6 Conclusion

A variety of sound descriptors are accessible that are selected to agree to the application. The drastic change in the characteristics of the whisper is observed compared to the neutral voice. Hence, multidimensional and the perceptually motivated timbre features are assumed to be most appropriate. However, it suggests utilizing constrained and well-performing for high speed and performance. The Hybrid Selection Algorithm sorted five features based on best performance using the CHAINs database. The selected timbre features MFCC, Brightness, Roll-off, Roughness, and irregularity) are used as a feature vector for the speaker identification. *It enhances t*he identification accuracy using the timbre features by 7.72 % compared to the most used MFCC features. The speaker identification task generates false positive outcomes due to intra-speaker variability. Hence, the MEDIAN values of timber features are utilized to reduce intra-speaker spread that further reported enhancement in the speaker identification by 2.23 %. The aggregate result is considering the complete database. This fact seeds the future scope to investigate the effectiveness of the selected features on unvoiced phonemes in whispered speech. It will put light on speaker identification and other speech processing applications.

Annexure A

CODE:

```
n=1;
for i=1:Speakers
for j =1:Samples

filename = sprintf('%d_%d.wav', i , j);
filename=[drct '\' filename];
            [x fs] = audioread(filename);
if (size(x,2) == 2)
                x= x(:,1);
end
            a = miraudio(x);

if op1==1
                MF=mfcc(x,fs,20);
                MF=MF';
                Training_Features(n,1:20)=MF(1,1:20);
end

if op2==1
                AB=1;

                BRIGH= mirbrightness(a);
                Training_Features(n,AB)= mirgetdata(BRIGH);
                AB=AB+1;

                ROUGH = mirroughness(a);
                Training_Features(n,AB)=mean(mirgetdata(ROUGH));
                AB=AB+1;

                IR = mirregularity(a);
                Training_Features(n,AB)=mirgetdata(IR);
                AB=AB+1;

                MF = mfcc(x,fs,20);
                MF=MF';
                Training_Features(n,AB:AB+19)=MF(1,1:20);
end
end
                n=n+1;
end
```

Data availability All relevant data are within the paper except the Database used in the manuscript is publically available at: (http://chains.ucd.ie/).

References

Al-Allaf, O. (2015). Removing noise from speech signals using different approaches of artificial neural networks. *International Journal of Information Technology and Computer Science, 7*, 8–18. https://doi.org/10.5815/ijitcs.2015.07.02

Albert-Ludwigs-Universität Freiburg. (2007). A Matlab toolbox for music information. In *Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V.* (pp. 261–268).

Bhattacharjee, M., Prasanna, S., & Guha, P. (2018) *Time-frequency audio features for speech-music classification*. Project: Broadcast Video Analytics.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., & Reynolds, D. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing*. https://doi.org/10.1155/S1110865704310024

Cummins, F., Grimaldi, M., Leonard, T., & Simko, J. (2006). *The CHAINS speech corpus: Characterizing individual speakers*. Dublin School of Computer Science and Informatics University College.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/TASSP.1980.1163420

Deshmukh, S., & Bhirud, S. G. (2012). A hybrid selection method of audio descriptors for singer identification in North Indian Classical Music. In *Fifth international conference on emerging trends in engineering and technology* (pp. 224–227).

Dobrowohl, F. A., Milne, A. J., & Dean, R. T. (2019). Timbre preferences in the context of mixing music. *Applied Sciences, 9*(8), 1695–1695.

Fan, X., Godin, K. W., & Hansen, H. L. (2011). Acoustic analysis of whispered speech for phoneme and speaker dependency. In *Proceedings of the annual conference of the international speech communication association, INTERSPEECH* (pp. 181–184).

Foulkes, P., & Sóskuthy, M. (2017). Speaker identification in whisper. *Letras de Hoje, 52*(1), 5–14.

Hermansky, H., & Malaya, N. (1998). Spectral basis functions from discriminant analysis. In *International conference on spoken language processing*.

Ito, T., Takeda, K., & Itakura, F. (2005). Analysis and recognition of whispered speech. *Speech Communication, 45*(2), 139–152.

Karvanagh, C. (2011). Intra- and inter-speaker variability in duration and spectral properties of English. *The Journal of the Acoustical Society of America, 130*, 2519. https://doi.org/10.1121/1.3655046

Li, H., Lai, L., Chen, L., Lu, C., & Cai, Q. (2016). *Computational and mathematical methods in medicine*. PB - Hindawi Publishing Corporation.

Manasa, Y., & Palaparthi, R. (2020). Minimization of noise in speech signal using mel-filter. IJEDR 2321-9939, Vol. 5, No. 2.

Maurya, A., Kumar, D., & Agarwal, R. K. (2018). Speaker recognition for Hindi speech signal using MFCC-GMM approach. *Procedia Computer Science, 125*, 880–887. https://doi.org/10.1016/j.procs.2017.12.112

Park, T. H. (2004). *Towards automatic musical instrument timbre recognition*, PhD thesis, the department of music. Princeton University.

Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification in the CUIDADO project)*.

Sardar, V. M., & Shirbahadurkar, S. D. (2018) Speaker identification of whispering speech: An investigation on selected timbrel features and KNN distance measures. *International Journal of Speech Technology*.

Shah, J. K., Smolenski, B. Y., Yantorno, R. E., & Iyer, A. N. (2015). Sequential k-nearest neighbor pattern recognition for usable speech classification. In *IEEE Xplore*.

Singh, A., & Joshi. A. M. (2020). Speaker identification through natural and whisper speech signal. In V. Janyani, G. Singh, M. Tiwari, & A. d'Alessandro (Eds.), *Optical and wireless technologies, Lecture Notes in Electrical Engineering* (Vol. 546). Springer. https://doi.org/10.1007/978-981-13-6159-3_24

Sreelekshmi, S. K., & Syama, R. (2017). Speaker identification using K-Nearest neighbors (k-NN) classifier employing MFCC and formants as features. *International Journal of Advanced Scientific Technologies, Engineering and Management Sciences*, *3*.

Toonen Dekkers, R. T. J., & Aarts, R. M. (1995) *On a very low-cost speech-music discriminato*. Technical Report124/95, Nat. Lab. Technical Note.

Wang, J.-C., Chin, Y.-H., Hsieh, W.-C., et al. (2015). Speaker identification with whispered speech for the access control system. *IEEE Transactions on Automation Science and Engineering, 12*, 1–9. https://doi.org/10.1109/TASE.2015.2467311