International Journal of Speech Technology
https://doi.org/10.1007/s10772-018-9527-4

Speaker identification of whispering speech: an investigation on selected timbrel features and KNN distance measures

V. M. Sardar[1] · **S. D. Shirbahadurkar**[2]

Received: 22 March 2018 / Accepted: 26 June 2018

Abstract
Speaker identification from the whispered speech is of great importance in the field of forensic science as well as in many other applications. Whispered speech shows many changes in the characteristics to its neutral counterpart. Hence the task of identification becomes difficult. This paper presents the use of only well-performing timbrel features selected by Hybrid selection method and effect of distance measures used in KNN classifier on the identification accuracy. The results using timbrel features are compared with MFCC features; the accuracy with the former is observed higher. KNN classifier with most probable distance function suitable for a whispered database like Euclidean and City-block are also compared. The combination of timbrel features and KNN classifiers with city block distance function have reported the highest identification accuracy.

Keywords Speaker identification · Timbrel audio descriptors · Whispering speech · Distance function · K-Nearest neighbor · Confusion matrix

1 Introduction

Speaker analysis includes applications like speaker identification/verification, gender and age group labeling, accent/dialect, etc. In any text-independent analysis of speaker, it is required to characterize the speaker's voice by some unique parameters called features. The normal voiced phonation is considered as the important source for characterization or modeling of a speaker; as a rich resonance information is available in a high-energy periodic signal. However, while whispering, an air turbulence without vibrating vocal chord changes the general condition of phonation (Beigi 2012). This is the most probable difficulty among all other reasons discussed in the literature for whispering speaker identification. Significant changes found between whisper and neutral speech in terms of periodicity, formants' location, and spectral slope boundaries of vowel regions. However, it is proved that vocal effort while whisper does not disturb unvoiced consonants as much (Fan and Hansen 2011). Hence, unvoiced part in neutral and whispered speech plays major role to identify speaker in neutral-whisper scenario. Secondly, speakers found it difficult to continue whispering for long duration (beyond 30 s). It is proved by good identification results for (i) long and whispered, and (ii) short and normal (non-whispered) compared to (iii) short and whispered (Foulkes and Sóskuthy 2017). So longer whisper (2–3 s) will consist of partial voiced phonation, thus increasing speaker identification accuracy.

The success of speaker identification in the whispered speech depends upon following factors mainly:

Quality of whispered recording (Signal to noise ratio) A SNR of 10 dB or higher is recommended for better speaker identification (Audio Engineering Society 2010). Hence it is required to record whisper in a noise-free environment. Whispered and neutral samples used in CHAIN database are above 15 dB. Also the duration of recording is 2–3 s for better identification results.
Selection of features MFCC is widely used in speaker identification experiment when database consists of neutral utterances. Here we have used limited well-perform-

✉ V. M. Sardar
vijay.sardar11@gmail.com

S. D. Shirbahadurkar
shirsd112@yahoo.in

[1] Department of E & TC, RSCoE Research Centre, S.P. Pune University, Pune, India

[2] Department of E & TC, Zeal College of Engineering, S.P. Pune University, Pune, India

Published online: 18 July 2018

ing timbrel features which are multidimensional and perceptually motivated (Deshmukh and Bhirud 2014).
Selection of classifier KNN classifier used here for classification. The choice of the number of nearest neighbors (k) and the distance measure are important factors. The optimum value of k is database dependent. The most outperforming distance metrics namely Euclidean and City-block are used here.

A large number of features are available to model or characterize the speaker. However, a limited number of features which contribute to accuracy are recommended to use for the speaker identification task. Linear discriminant analysis (LDA), is suggested in (Bistritz and Zilca 2001) for reducing the features but further reported that the performance is appropriate for clean speech but not for telephone speech due to noise. Whispered speech is also basically a noise like structure. Hence, using traditional LDA for whispered speaker database may not the appropriate choice.

While speaker identification, all useful features should be combined in a single vector. A comparison should be made on the basis of aggregate score of all features in the vector. As far the KNN classifier is concerned, various distance functions are available which are tested on various data. A review of the various distance function is compiled in (Surya Prasatha et al. 2017) which mentions that Euclidean and city-block (Manhattan) distance are the most common outperforming distance functions on various databases. One more distance function proposed by Hassanat for face recognition problem is found better but it is slower. Fast searching of the speaker in a large database is one of the challenge in the speaker identification which is overcome by cosine distance and i-vector. Cosine distance gives value -1 and $+1$ for i- Vectors of two speakers pointing in the same direction and opposite directions respectively (Schmidt et al. 2014).

2 System description

Figure 1 illustrates the experimental steps for speaker identification in whispered speech. A set of most probable 20 audio descriptors (timbrel and non-timbrel) are tested on a speaker database of whispering speech and only the best descriptors which give maximum accuracy are selected. The identification accuracy using only MFCC and the selected timbrel features is compared. Further, results with Euclidean and City-block distance functions of KNN are compared. This setup finally reports the best results with the alternatives used for experiments.

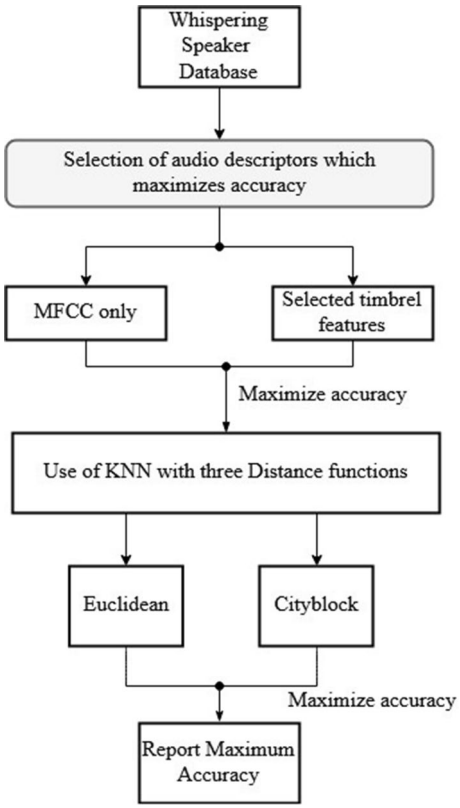

Fig. 1 Speaker identification system block diagram

2.1 Speaker database

This study employs the CHAIN corpus developed at School of Computer Science and Informatics University College Dublin (Cummins et al. 2006). The corpus consists of a total of 36 speakers, with 20 males and 16 females with a sampling frequency 44.1 KHz and 2–3 s. duration. For training the speaker identification system, phonetically rich and balanced sentences should be selected. This requirement is satisfied by CSLU's nine phrases and 24 sentences from TIMIT database. Corpus employs neutral and whispered speech from the speakers in the English language. The neutral recording session was carried out in a professional recording studio using a Neumann U87 condenser microphone. The whispered speech was recorded in a quiet office environment, using an AKG C420 headset condenser microphone.

2.2 Hybrid selection algorithm

It is recommended to use limited well-performing features for the given database. The purpose is to keep processing time and memory requirement minimum. It is also found

that the non-related features are simply messed which increases crowd in the feature space giving overlapping classes and hence defeating accuracy. Analysis of overlapping of the models due to a large number of features including non-relevant features are addressed in (Kwon and Narayanan 2007). Keeping in view, the common acoustic environment and similarities in many speakers, the need of selecting useful features only is emphasized. Speaker models are trained to separate data and select feature vectors that are estimated to add to discrimination (Kwon and Narayanan 2007). Principle Component Analysis is an attractive tool proposed in the literature to get rid of the dimensionality of features. But it should be noted that the features used here have diversified range of magnitude. Hence normalization is required before using all features in a vector. However normalizing will result in a need of more Principle Components to explain the same amount of variance in the data; otherwise, there will be a major loss of data.[1] Hence a well performing audio descriptors are selected by simple and reliable Hybrid selection algorithm (Deshmukh 2014).

MPEG7 standard consists of about 52 audio descriptors including low-level descriptors, which are broadly classified as Basic, Basics spectral, Signal parameters, and Temporal Timbral, Timbral and Spectral basis representations categories. Timbre is the complex attribute of sound as it can neither be mapped to a one-dimensional scale nor be uncoupled from the other one-dimensional components. Timbre cannot be expressed by physical quantity, rather it is a perceptual kind of nature. According to American National Standards Institute (ANSI 1960, 1973). Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.

Examples of timbrel audio descriptors are Zero-crossing rate, Roll-off, Brightness, MFCC etc. and examples of non-timbrel descriptors are pitch, energy, fundamental frequency etc. Moreover, a very large set of audio features for sound description are elaborated in depth along with classification, mathematical analysis, operation strategy etc in (Peeters 2004). It is a compressive study which covers almost all audio features available for sound description including temporal, energy, spectral, harmonic, harmonic spectral shape, perceptual, perceptual spectral envelope shape and, miscellaneous. A list of 166 audio features is also provided in (Peeters 2004).

While applying Hybrid section methods in the complex problem of North Indian Classical Music's singer identification (Deshmukh and Sunil 2012), total 20 audio descriptors have been considered as a primary set. The audio descriptors from a non-timbrel group are pitch, RMS energy, Low energy, fundamental frequency, Linear predictive coding coefficient(LPCC), Inharmonicity, Mode, Harmonic Change Detection Function (HCDF), Spectral centroid, spectral spread, Kurtosis, flatness, and Entropy. The primary timbrel features are Zero crossing rate, Roll-off, Brightness, MFCC, Roughhness, irregularity and Rhythm. They are also representing a good mix of various domains like time, frequency, cepstral and, wavelet domain. All these audio descriptors are used in this paper except two audio descriptors namely, pitch and fundamental frequency are replaced by the Attack-time and Attack-slope. The omission of two descriptors is due to the obvious reason of loss of periodicity in the whispered speech.

Consider a set of speaker data X, each having 'n' number of samples. ADt is an empty array of all probable features which are targeted. It is an iterative process where 20 timbrel and non timbrel features are tested for accuracy and arranged in highest to lowest accuracy order. As a thumb rule, top 50% of timbrel and 25% from non-timbrel features are selected. After every iteration, the feature which maximizes the classifier accuracy is appended in combination with previous feature/s. i.e. New well-performing features go on adding for every iteration till no further increase in accuracy is observed. m = iteration number, dm = dimension of ADt at iteration 'm'.

Algorithm steps for Hybrid Selection method (Deshmukh and Sunil 2012):

1. Training data of speakers = X←[x_1, x_2,x_3 … x_i…x_n]
2. Initialize ADt= { }, m=0, dm= 0.
3. ADm= { } Subset of Audio Descriptors Selected.
   while dm < ADt do
   k=1;
   while j < AD do { S_{final} = Sdm U (xi, k) where 1<i<n; }
   Calculate the classifier accuracy for all S'
   k=k+1;
   end while
4. k_0 ← Audio Descriptor(s) index that gives the highest accuracy.
   Sdr+1 = Sdr U(xi,jo) 1<i<n and by general thumb rule,
   (Sdm <= approximately 50% of total audio descriptors if Timbral group and 25% if Non timbral Group)
   dm+1 ←dm +1 ; m ←m+1; end while
   Ensure: SDs {where ADs total number of selected features}
   End

After applying Hybrid selection algorithm, the set of features maximizing accuracy found are Zero Cross Rate, Roll off, Brightness, Roughness, Irregularity, and MFCC (Mel frequency cepstral coefficient). A vector consisting of all these timbrel features is used for speaker identification.

2.2.1 Correlation: audio descriptors, intra-speaker and inter-speaker voice features

Correlation is used to know how much strong or weak relationship exists between two variables. Assuming two set of

[1] A Tutorial on Principal Component Analysis, Machine Learning, arXiv:1404.1100.

Table 1 Discrimination ability of selected audio descriptors evaluated by correlations

Correlation	ZCR	Roll-off	Roughness	Brightness	Irregularity	MFCC
ZCR	1	0.47780	0.73976	0.42834	−0.40951	0.10664
Roll-off	–	1	0.68775	0.71273	−0.49334	0.47900
Roughness	–	–	1	0.83338	−0.12968	0.58068
Brightness	–	–	–	1	−0.13331	0.57224
Irregularity	–	–	–	–	1	0.12129
MFCC1	–	–	–	–	–	1

variables array x and y, Pearson correlation coefficient for two variables is given by:

$$r = \frac{N \sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[N \sum x^2 - (\sum x)^2\right]\left[N \sum y^2 - (\sum y)^2\right]}} \quad (1)$$

where N is the Number of duos of scores, $\sum$xy is the sum of the products of paired scores, $\sum$x is the sum of x scores, $\sum$y is the sum of y scores, $\sum x^2$, $\sum y^2$ is the sum of squared x and y scores respectively (Ke et al. 2008). The coefficient value ranges from − 1.00 to 1.00. If the coefficient value is in the negative range, then that means the two variables are inversely correlated. i.e. if one variable increases, the other variable decreases. Positive correlation coefficient indicates that two variables increase or decrease together.

For better speaker identification accuracy: it is essential that the audio descriptors should be discriminated, inter-speaker variation should be least and inter-speaker variation should be maximum. The discrimination ability of four spectral audio descriptors namely centre of gravity, standard deviation, skewness, and kurtosis are investigated for speaker identification in (Karvanagh 2011). In our work, we have calculated Pearson correlation to observe the desired dissociation between six audio descriptors as shown in Table 1. All six feature values are extracted for ten speech samples of the same speaker. Each feature with ten values are arranged in an array. A correlation coefficient is calculated between an array of one feature and every other feature. E.g. an array of ten values of ZCR for the same speaker compared with an array of ten values of roll-off, roughness and so on.

It should be noted that threshold value of correlation coefficient for associated or dissociated data arrays can be set by analysis of complete data in consideration. Here we have judged association or dissociation relatively. Correlation of one feature with every other feature is small positive or negative value. Irregularity feature have negative and different correlation values. Negative correlation values prove that irregularity is dissociated from other features and different values of correlation with respect to other features implies that all features are sufficiently isolated from each other.

Table 2 Association among intra-speaker feature vectors

Correlation Speaker sample no.	1_1	1_2	1_3	1_4	1_5
1_1	1	0.93082	0.89871	0.89871	0.83527
1_2	–	1	0.61691	0.67767	0.69523
1_3	–	–	1	0.93173	0.89228
1_4	–	–	–	1	0.86676
1_5	–	–	–	–	1

Table 3 Discrimination among inter-speaker feature vectors

Correlation					
Speaker	Speaker_1	2	3	4	5
1	–	0.23947	0.78663	0.600239	0.392054
2	–	1	0.31248	0.428876	−0.27452
3	–	–	1	0.557393	0.667848
4	–	–	–	1	0.628116
5	–	–	–	–	1

The desired low intra-speaker deviation and high inter-speaker deviation is validated by subsequent observations. Five speech samples (1_1 to 1_5) of speaker_1 are considered for the experiment and a feature vector consisting values of six audio descriptors are listed. Afterward correlation among each speaker sample with every other sample on the basis of feature vector is calculated as shown in Table 2. E.g. A vector consisting of ZCR, roll-off, roughness, brightness, irregularity and MFCC for sample 1_1 compared with 1_2, 1_3, 1_4, and 1_5. The high positive values indicate that all samples of same speakers are highly correlated.

Correlations of feature vectors for five different speakers (Speaker 1–5) are listed in Table 3. The low positive or negative values in the table indicates that all five speaker features are sufficiently dissociated from each other.

2.3 Selected audio descriptors

The definitions and significance of selected audio descriptors are discussed below (MIR toolbox 1.3.3 (Matlab Central Version) 2011).

- *Zero Cross Rate* It measures the rate of crossing the X-axis by a signal in time domain. ZCR is high for unvoiced and low for voiced part of the audio.
- *Roll off* It is a spectral feature of audio which is defined as the frequency below which 85 or 95% of the total signal energy is present. While measuring roll-off, only the number of samples (R) for which 85% of the total energy is concentrated are considered. If a spectrum S_t consists of' N′ samples having total aggregated energy as $\sum_{i=1}^{N} S_t[n]$ then roll of is calculated as:

$$\sum_{i=1}^{R} S_t[n] = 0.85 \times \sum_{i=1}^{N} S_t[n] \tag{2}$$

Brightness It is a measure of audio energy above a certain cut of frequency f_c. Three values of f_c are recommended viz 1000, 1500, and 3000 Hz. An audio signal is processed frame-wise with a frame of length 25 ms and 50% overlap.

MFCC While extracting Mel-frequency-cepstral–coefficient (MFCC), the steps are followed as (i) framing (25 ms frame with 50% overlap), (ii) windowing (Hamming window), (iii) Fast Fourier Transform (size 512), Mel filtering and discrete cosine transform. Here 13 MFCC coefficients are sufficient to characterize the spectral shape of the audio signal. Total 13 'mel' filters will pick-up the required number of coefficients.

Roughness Related to the beating of sinusoids pair, close in frequency, an estimation of sensory disagreement is termed as roughness. While calculating the total roughness all the peaks of the spectrum are located, and the average of all the disagreement between all possible pairs of peaks is taken. The frequency components are assumed to be stationary over the small duration (25 ms) of each audio file.

Irregularity It is the degree of variation among the peaks of the spectrum of the signal. It can be calculated by considering successive particles as in Eq. (3) or previous, present, and, next particle components as in Eq. (4).

$$\sum_{k=1}^{n} (a_k - a_{k+1})^2 / \sum_{k=1}^{n} a_k^2 \tag{3}$$

$$\text{or} \sum_{k=2}^{N-1} \left| a_k - \frac{a_{k-1} + a_k + a_{k+1}}{3} \right| \tag{4}$$

2.4 K-nearest neighbor (KNN) classifier

The KNN classifier does not require the prior knowledge of data hence called as a non-parametric method for classification. The important parameter used in KNN are a number of nearest neighbors (k), a distance function (d), decision rule and n labeled samples of audio files X_n. A new query vector is labeled a class based on the minimum distance from the predefined classes. Mathematically, it is a matter of calculating a posteriori class probabilities P (wilx) as

$$P(w_i|x) = \frac{k_i}{k} \times P(w_i) \tag{5}$$

where ki is the number of vectors which belongs to class wi within the subset of k vectors (Jashmin et al. 2004). A large value of k is recommended, in general, to reduce the effect of noise on the accuracy. Also, the odd value of k is chosen for binary classification.[2] The results are also affected by the way of calculating distances between the training and testing vectors by various distance metrics available.

2.4.1 Distance metric

KNN classifier assigns a class label to the test sample on the basis of the nearest distance from the training classes which is called the nearest neighbor. Here, six features namely ZCR, brightness, roll-off, irregularity, roughness and, MFCC are arranged in a vector. A distance between query feature-vector and feature vector of existing classes is calculated. The method of calculating Euclidean and city-block distances are shown below (Jashmin et al. 2004):

- Euclidean distance: N-dimension Euclidean distance applies as:

$$D(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \tag{6}$$

where x is the coordinates of training feature vector and y is the coordinates of query feature vector.

- City-block: The City -block (Manhattan) distance between a pair of points, x and y, with n dimensions is calculated as:

[2] 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, Volume 3, Special Issue, April (2017).

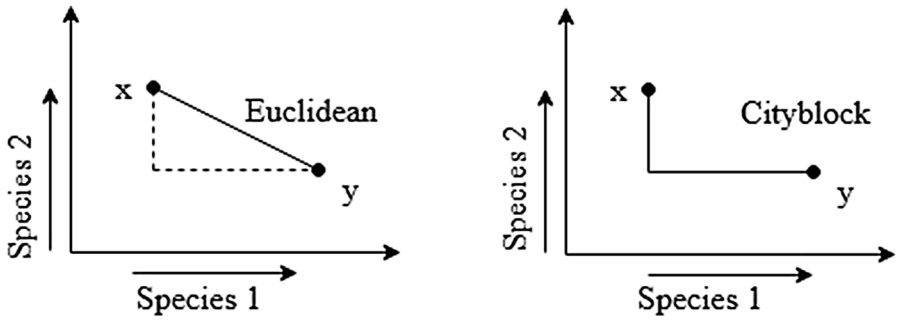

Fig. 2 Graphical representation for calculating Euclidean and City-block distances

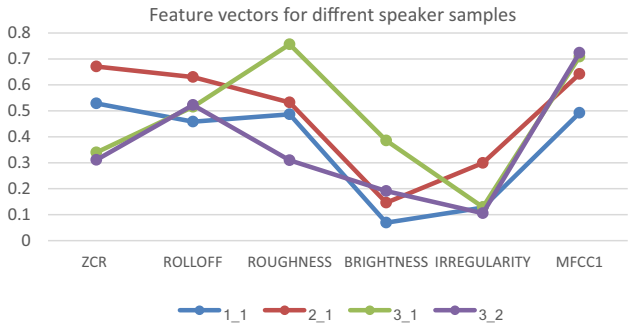

Fig. 3 Feature values for four speech samples of speakers

$$\sum_{j=1}^{n} \left| x_j - y_j \right| \tag{7}$$

The graphical representation to calculate Euclidean and City-block distance is as shown in Fig. 2. Vector consist of multiple features; some features may have high intra-speaker variations (though undesirable) for some speech samples. Effect of such a high difference in a single dimension is diminished since the distances are not squared for City-block distance.

Above graph (Fig. 3) shows the plot of feature vectors for three training samples and one test sample. Three speech samples used for training are 1_1, 2_1 and, 3_1 which are one sample each of the three different speakers. The speech sample 3_2 which is another sample of speaker-3 is used for testing. The large difference in roughness values can be observed for the intra-speaker samples (3_1 and 3_2). This difference will be further magnified by Euclidean distance function to cause to misclassify which is proved in subsequent discussion.

Table 4 shows the feature values namely zero crossing rate (ZCR), Roll-off, Roughness, brightness and also the first coefficient of MFCC. Three class labels 1, 2 and 3 are generated for the speakers 1_1, 2_1, 3_1 respectively. Now a different sample of speaker 3 (i.e. 3_2) is given for testing. The class label will be identified on the basis of the distance measured between the testing sample and every available class. Taking $k = 1$ (nearest three neighbors), sample 3_2 is measured at first minimum distance from class 3 (sample 3_1) by City-block distance which is correct. However, by using Euclidean distance, the test query is wrongly labeled as class 1. This misclassification is due to the effect of high difference in roughness value between 3_1 and 3_2; which is magnified further due to squaring as in equation to calculate Euclidean distance (7). This effect is avoided in the City block distance as it uses simple subtraction.

3 Results

The selected timbrel audio descriptors are adapted in this paper namely Zero Cross Rate, Roll off, Brightness, Roughness, Irregularity, and MFCC by using Hybrid Selection algorithm. These features are analyzed

Table 4 Comparison of Euclidean and City-block distances for deciding the class label

Speaker sample no.	Features						Class label	Distance	
	ZCR	Roll-off	Roughness	Brightness	Irregularity	MFCC		Euclidean	City Block
1_1	0.52852	0.45832	0.48671	0.06904	0.12718	0.49240	1	0.389327	0.833374
2_1	0.67084	0.63007	0.53246	0.14626	0.29901	0.64203	2	0.486589	1.009899
3_1	0.33991	0.51550	0.75632	0.38563	0.12912	0.70839	3	0.489019	0.716006
3_2	0.31099	0.52246	0.30960	0.19091	0.10566	0.72362	=>	Class1	Class 3

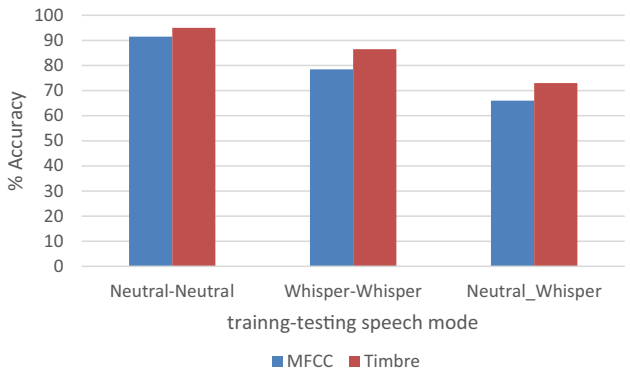

Fig. 4 Accuracy in three training–testing modes of speech using MFCC and timbre

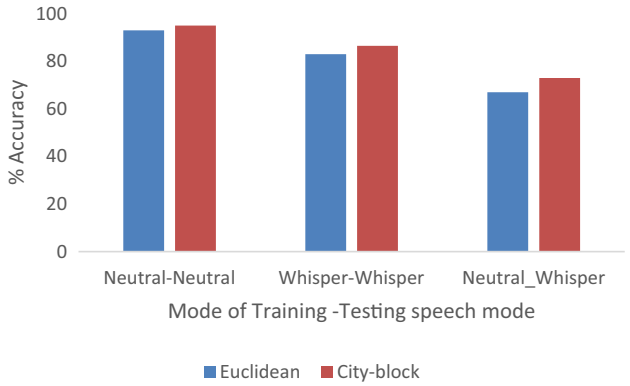

Fig. 5 Comparative accuracy by using Euclidean distance and City-block distance

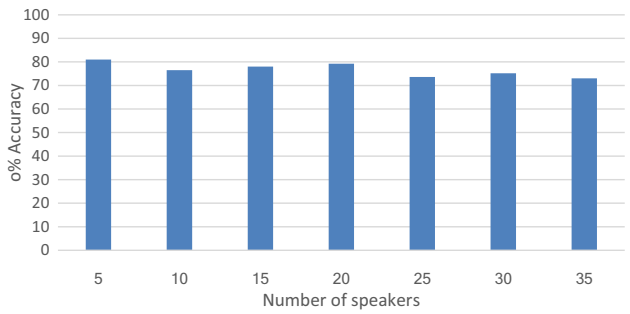

Fig. 6 % accuracy with increasing number of speakers

(Sect. 2.1.1) for the desired inter-feature discrimination for better identification. In addition, inter-speaker similarity and intra-speaker dissimilarity in the features are investigated. Also, the demonstration shows the possibility of correct class labelling by City-block distance compared to Euclidean distance. Hence, the proposed speaker identification system claims that the selected timbrel features and the K-NN classifier with city-block distance measure is expected to maximize the identification results. It is proven in the subsequent sections. The proposed features and classifier outperforms in all three different train-test scenario. Following results are reported on the CHAIN database described in Sect. 2.1. Figure 4 shows the comparison between accuracy using MFCC and timbre features. The comparison between results using Euclidean and City-block distance is represented in Fig. 5. The identification results are also observed with increasing number of speakers (Fig. 6). Table 5 lists the identification results for three train-test modes of speech (namely neutral–neutral, whisper–whisper and, neutral-whisper) comparing MFCC and timbrel features. Table 6 shows the comparative results using Euclidean and City-block distance function. The Identification accuracy with increasing number of speakers is given in Table 7.

Table 5 Comparative speaker identification accuracy for 35 speakers with MFCC and timbre features

Speech mode		% Accuracy	
Training	Testing	MFCC	Timbre
Neutral	Neutral	91.5	95.0
Whisper	Whisper	78.5	86.5
Neutral	Whisper	66.0	73.0

Table 6 Accuracy rate of baseline system using four different systems

System	Accuracy rate (%)
I. PDM based fusion system ($\alpha = 0.75$)	83.13
II. NDMP based fusion system ($\alpha = 0.70$)	83.75
III. Pyknogram based system	79.51
IV. MFCC-GMM	76.04

Table 7 Speaker identification accuracy with Euclidean and City-block distance

Speech mode		% Accuracy	
Training	Testing	Euclidean	City-block
Neutral	Neutral	93.0	95.0
Whisper	Whisper	83.0	86.5
Neutral	Whisper	67.0	73.0

K-NN with $k = 3$, and City-block distance which gives maximum accuracy is tested for the increasing number of speakers

3.1 Identification accuracy for different speech modes along with different features & distance functions, and for increasing number of speakers

CHAIN database consists of 36 speakers with 33 speech samples each in whispered and neutral mode. Following experiment used 20 speech samples each for 35 speakers and as a general rule, 70% samples used for the training and 30% for the testing. KNN classifier with 3- nearest neighbor and city-block distance is used for the experiment. By increasing k progressively; it is found that k = 3 is the optimum value which gives highest identification accuracy.

Results with three combinations of training and testing modes are used in the speaker identification as (i) neutral–neutral, (ii) whisper–whisper and, (iii) neutral-whisper are listed with two types of features MFCC and timbre.

A baseline speaker identification system with a whispered speech which is proposed for the access-control system uses the CHAIN database (Wang et al. 2015). Four different systems used in the system and results are reproduced here for the reference.

Above results are generated for the whispered training–testing utterances from CHAIN database. The system using nonparametric density model (NPDM) gives highest accuracy among four system discussed above. The identification accuracy of our proposed system in whisper train-whisper test scenario is 2.75% higher than the best results given by NDMP Based Fusion System in (Wang et al. 2015).

Further, the comparative results with two distance functions namely Euclidean and City-block are listed below.

It is generally seen that the identification accuracy is decreasing with increasing number of speakers.

3.2 Confusion matrix

Confusion matrix shown in Fig. 7 gives information about the parameters like true positive (TP), true negative (TN), False positive (FP) and false negative (FN) in a simple way for binary decision. However, our experiment includes multi-classes, hence overall values of these parameters are possible by using generalized formulae described in (Manliguez 2016). Further, parameters to evaluate the performance of the system like accuracy, precision, sensitivity and, specificity can be calculated.

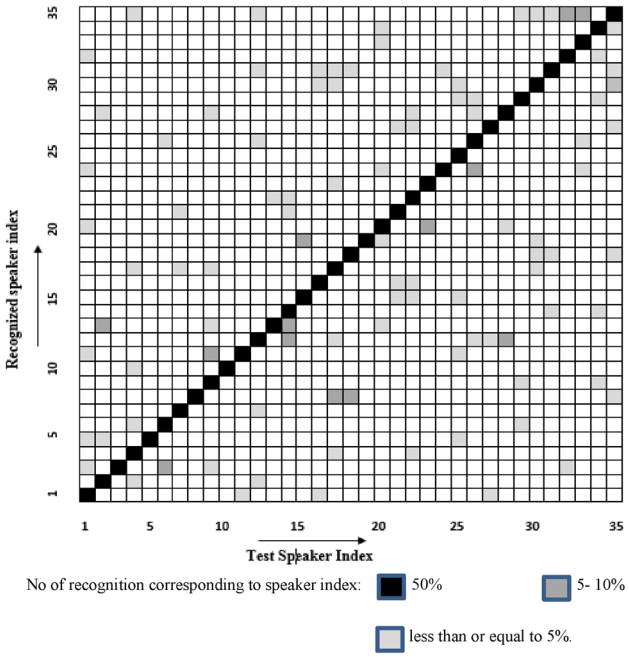

Fig. 7 Confusion matrix for speaker identification experiment of 35 speakers with Neutral training-whispered testing

4 Conclusion

We have selected well-performing audio descriptors only by using Hybrid selection algorithm for speaker identification with neutral and whispered speech. The selected features are found good in terms of the discrimination ability from their inter-correlation. The correlation analysis also confirmed the intra-speaker similarity and inter-speaker dissimilarity with respect to their feature vector values. The results with a combination of timbrel features namely zero-crossing rate, roll-off, irregularity, brightness, roughness, and Mel-frequency cepstral coefficient (MFCC) are compared in three training and testing modes of speech i.e. neutral–neutral, whisper–whisper, and, neutral whisper. However, neutral training and whispered testing for speaker identification is targeted. Timbrel features reported 6% increase in identification accuracy compared to MFCC features for 35 speakers with neutral-whisper condition.

The second set of experiments aimed to find better distance function among Euclidean and City-block. Accuracy by using City-block is found 6% more compared to Euclidean distance at similar conditions. Accuracy using k = 3, City-block distance and, timbrel features is observed with increasing number of speakers. Accuracy decreases with increasing database. However, decrease in accuracy is relatively lower which proves that timbrel features are robust enough (Table 8).

We propose the further scope to find common well- performing features to be used for any whispered database by experimenting with different whispered databases. Further, a system can be evaluated for the performance parameters like accuracy, precision, selectivity and, specificity using

Table 8 Speaker identification accuracy for 35 speakers with increasing number of speakers

No of speakers	5	10	15	20	25	30	35
% Accuracy	81.0	76.5	78.0	79.25	73.6	75.2	73.0

multi-class analysis. Similarly, an analysis for developing an optimized system for speed and other performance parameters which will be suitable for a huge speaker database.

References

Audio Engineering Society. (2010). Paper, C presented at the 129th convention, November 4–7, San Francisco, CA, USA.

Beigi, H. (2012). Speaker recognition: Advancements and challenges. In *New trends and developments in biometrics*. Rijeka: InTech.

Bistritz, Y., & Zilca, R. D. (2001). *Feature concatenation for speaker identification*.

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.

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.

Deshmukh, S. H. (2014). On the selection of audio descriptors and identification of singer in North Indian Classical Music, Ph.D. dissertation, Department of Computer Science Engineering, NMIMS Deemed-to-be University.

Deshmukh, S. H., & Bhirud, S. G. (2014). A novel method to identify audio descriptors, useful in gender identification from North Indian Classical Music Vocal. *International Journal of Computer Science and Information Technologies, 5*(2), 1139–1143.

Fan, X., & Hansen, J. H. (2011). Speaker identification within whispered speech audio streams. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1408–1421.

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

Jashmin, K., Shah, Brett, Y., Smolenski, R. E., Yantorno, & Iyer, A. N. (2004). Sequential k-nearest neighbor pattern recognition for usable speech classification. In *12th European, IEEE Xplore, signal processing conference, 2015*.

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.

Ke, Y., Cheng, J., & Ng, W. (2008). Efficient correlation search from graph databases. *IEEE Transactions on Knowledge and Data Engineering, 20*, 1601–1615.

Kwon, S., & Narayanan, S. (2007). Robust speaker identification based on selective use of feature vectors. *Pattern Recognition Letters, 28*(1), 85–89.

Manliguez, C. (2016). Generalized confusion matrix for multiple classes. https://doi.org/10.13140/RG.2.2.31150.51523.

MIR toolbox 1.3.3 (Matlab Central Version). (2011). *User's manual Olivier Lartillot, Finnish centre of excellence in interdisciplinary music research*. Jyväskylä: University of Jyväskylä.

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

Schmidt, L., Sharifi, M., & Moreno, I. L. (2014). Large-scale speaker identification. In *International conference on acoustic, speech and signal processing (ICASSP), IEEE*.

Surya Prasatha, V. B., Alfeilatb, H. A. A., Lasassmehb., O., Hassanatb, A. B. A. (2017). Distance and similarity measures effect on the performance of K-nearest neighbor classifier-a review. CoRR, abs/1708.04321.

Wang, J.-C., Chin, Y.-H., Hsieh, W.-C., Lin, C.-H., Chen, Y.-R., & Siahaan, E. (2015). Speaker identification with whispered speech for the access control system. *IEEE Transactions on Automation Science and Engineering, 12*, 1191–1199.