Speaker Identification of Whispering Sound: Effectiveness of Timbre Audio Descriptors

V.M. Sardar & Dr. S.D. Shirbahadurkar

Savitribai Phule Pune University

ABSTRACT

Identification of a person from the whispered voice is challenging task as many variations are observed in the speech attributes of the same the speaker in whispered and neutral mode. The success of the speaker identification system relies on the selection of good audio features and this paper mainly focus on the feature selection for the task. There are hundreds of audio features available for sound description, but their performance depends upon the type of the database. The motivation of this paper is to investigate the suitability of timbre features for whispered database. The choice of timbre features is due to their perceptual and multidimensional approach. However, all the features may not be contributing to the maximum speaker identification accuracy. Hence, a careful selection of limited audio descriptors from the available large set is essential to increase the speaker identification with low process time. The Hybrid Selection method is used to select the well-performing audio descriptors from all available descriptors in MPEG-7. Five timbre features namely roll-off, roughness, brightness, irregularity and MFCC are found outperforming for the database used. Here, a comparison of results is being done among traditional MFCC feature and timbre features where later reported an absolute accuracy of 10.4%. The database consist of about 480 utterances including neutral and whispered speech mode. K-NN classifier with three nearest neighbour and Euclidean distance is used.

Keywords: speaker identification, feature extraction, mpeg-7, musical information retrieval, timbre features, whispered speech.

Classification: H.5.1

Language: English

London Journals Press

LJP Copyright ID: 975734
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science and Technology

Volume 19 | Issue 2 | Compilation 1.0

Speaker Identification of Whispering Sound: Effectiveness of Timbre Audio Descriptors

V.M. Sardar[α] & Dr. S.D. Shirbahadurkar[σ]

ABSTRACT

Identification of a person from the whispered voice is challenging task as many variations are observed in the speech attributes of the same speaker in whispered and neutral mode. The success of the speaker identification system relies on the selection of good audio features and this paper mainly focus on the feature selection for the task. There are hundreds of audio features available for sound description, but their performance depends upon the type of the database. The motivation of this paper is to investigate the suitability of timbre features for whispered database. The choice of timbre features is due to their perceptual and multidimensional approach. However, all the features may not be contributing to the maximum speaker identification accuracy. Hence, a careful selection of limited audio descriptors from the available large set is essential to increase the speaker identification with low process time. The Hybrid Selection method is used to select the well-performing audio descriptors from all available descriptors in MPEG-7. Five timbre features namely roll-off, roughness, brightness, irregularity and MFCC are found outperforming for the database used. Here, a comparison of results is being done among traditional MFCC feature and timbre features where later reported an absolute accuracy of 10.4%. The database consist of about 480 utterances including neutral and whispered speech mode. K-NN classifier with three nearest neighbour and Euclidean distance is used.

Keywords: speaker identification, feature extraction, mpeg-7, musical information retrieval, timbre features, whispered speech.

Author α: Research Student, Department of E & T C, RSCoE, S.P. Pune University, Pune, India.
σ: Professor, Department of E & TC, Zeal College of Engineering, Pune, India.

I. INTRODUCTION

Due to basic differences in the vocal efforts while whispered and neutral mode, there exist differences in their characteristics in terms of structure, positions of formants, spectral slope and energy. The formants in whispered speech are shifted to the higher frequency which may not be captured by widely used Mel Frequency Cepstral Coefficient [1]. Speech modes are classified as neutral, loud, shouted, soft, and whispered, where whisper exhibits very low energy and minimum spectral slope. A low spectral slope indicates that energy contents in case of the whisper are concentrated in high frequency. Such variations degrade the performance of the speaker identification system developed for neutral speech when tested with whispered voice [1]. Hence, use of the perceptual feature which can accommodate the whisper-neutral variability is emphasized. Whispered speech has a low signal-to-noise ratio (SNR), hence using only high SNR whispers for identification are suggested. It is also mentioned that identification accuracy can be increased if a comparison is made by only unvoiced part in neutral and whispered samples of the same speaker as unvoiced part in both the modes remains almost the same [1-2]. Linear Frequency Cepstral Coefficient feature (LFCC) is also useful for a whispered sound, as formants are shifted to the higher frequency [3]. Using a feature mapping technique in [4], an additional enhancement of 10% in identification is found. But the delay is introduced as the feature mapping is being done in the testing phase. Another approach to feature transformation from neutral to whispered speech is adapted in [5], overcome the problem of system speed as the processing is done in training phase only.

London Journal of Research in Computer Science and Technology

In reference [6], the combined feature vector of Mel Frequency Cepstral Coefficient (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC) are used for speaker identification. Results are compared with three classifiers namely LBG-VQ, Gaussian mixture model (GMM) and back propagation neural network (BPNN) which are found to be 59.2%, 70.9% and 84.7% respectively. Also, it had demonstrated the use of the neural network with 14 inputs of the combined feature vectors of MFCC and GFCC and 90 neurons in its output layer to classify the speakers. The hidden layer uses 17 neurons. However, this number requires investigation through trial and error method. Here a Genetic algorithm is used for reduction in feature dimension of MFCC and GFCC. When first Principal Component investigation is applied, 39 feature vectors each of MFCC and GFCC is reduced to 24 feature vectors. Further GA (Genetic algorithm) is applied and feature vectors reduce to 24 features. Two similarity measures are used in speaker identification namely distance and the highest frequency [7]. In the distance measure, speaker samples assigned to codebook having a minimum distance. In the second method, along with the minimum distance, the codebook having the highest frequency pair of input voice sample is selected. Sound files are also processed with Least Mean Square (LMS) and reported the highest speaker identification accuracy as 82.35% with the distance measure and 90.72% with the frequency measure. In reference [8], the vectors in the low dimensional space named as i-vectors are addressed for speaker identification to accommodate both speaker and channel variability. I-vector method proposed a single space which is named the total variability space. The identification accuracy is investigated in three environments namely clean background, White noise and Babble noise environment. Improved Locality Preserving Projections (LPP) with i- vector reported improvements in Equal Error Rate (EER) and Min Detection Cost Function (MinDCF). EER and MinDCF with improved LPP are found 4.45 and 0.17 respectively compared to 7.32 and 0.53 with use of GMM. Linear Discriminant Analysis (LDA) dimension reduction method is also used which reduces the feature dimension to 200 compared to 512 in UBM.

MPEG-7 is rich with many perceptual based audio descriptors and the best performing audio descriptors can be found by Hybrid Selection Method [9]. This method is used for singer identification from North Indian classical music which is more complex due to a separation of vocal from this music. Hence, only limited and sufficient perceptual features combination (namely RMS energy, brightness and fundamental frequency) are selected by Hybrid Selection method and successfully used. The singer identification accuracy using K-means classifier offered the accuracy of about 70%. MPEG-7 descriptors are mainly found to be used for the singer and musical instrument identification with few efforts of a speaker, gender or age identification but in neutral mode of speech. To reduce the size of MPEG-7 descriptors, speech features are extracted with a spectrum basis projection like Principal Component Analysis (PCA). While an experiment of the speaker and gender recognition, use of Independent Component Analysis (ICA) had proven the best compared to Normalized Audio Spectrum Envelope (NASE) and PCA [10]. Timbre features such as vibrato and the attack-decay are used for singer identification and analyses harmonic content and the dynamic characteristics of sound. It is stated that spectral envelope is a base to distinguish between two different musical instruments even playing the same note and amplitude which is a perceptual work indeed. feature is found effective and accuracy reported as 87.8% in segment level singer identification [11].

II. ANALYSIS OF WHISPERED SPEECH

This section analyses the whispered speech to its neutral counterpart from the spectrogram. Further, it explains the limitations of a whispered speaker identification method on the basis of unvoiced part only.

2.1 Attributes of Whispered Speech

Speech characteristics in whispered mode drastically change compared to neutral speech. The whisper is generated due to air passing through vocal constriction without vibrations and hence, the periodic structure is not found. The formants in whispered speech shifts to higher frequencies and their bandwidth also expand compared to neutral speech [12].

Spectrogram shown below is generated using PRAAT for the sentence "If it doesn't matter who wins, why do we keep score?" both in neutral and whispered mode. From formant analysis, it is observed that the formants F1 and F2 are shifted to higher frequency compared to neutral speech and formant bandwidth also increases.

Spectrogram for: (a) Neutral speech (b) Whispered speech

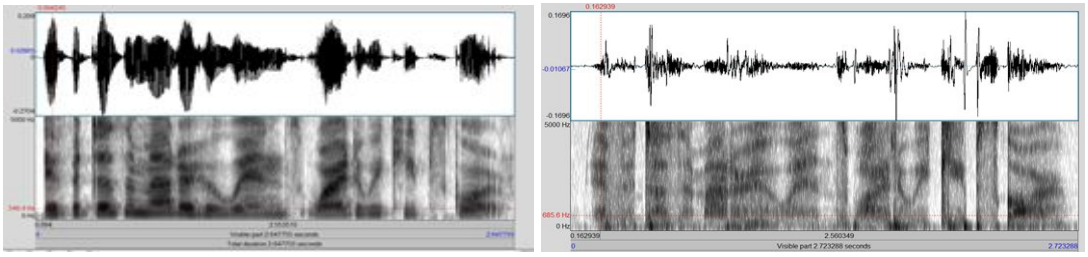

Fig. 1: Spectrogram for the same sentence and the same speaker in neutral and whispered Speech

The average formants values for all speakers in the database (described in section 4.a) are observed as below:

Table 1: Comparative average formant values and Bandwidth of F1 and F2 for neutral and whisper speech

Sr.	Type of sound	Formant Frequency (Hz)		Median of Bandwidth (Hz)	
		F1	F2	F1	F2
1.	Neutral	345-524	1814-2016	265.75	189.34
2.	Whisper	413-681	1895-2135	545.50	356.33

In addition to the formant and bandwidth shift, such other variations [13] in the whispered speech compared to neutral, makes the speaker identification task difficult.

2.2 Comparison between neutral an whispered speech by unvoiced utterances only

It is stated that the unvoiced part of speech in both whispered and neutral speech remains almost constant. Hence, it may be good choice to execute neutral-whisper speaker identification using unvoiced utterances only. A voiced or unvoiced decision is based on a range of zero-crossing rate (ZCR) and energy values together i.e. Low energy and high ZCR confirm an unvoiced decision [14].

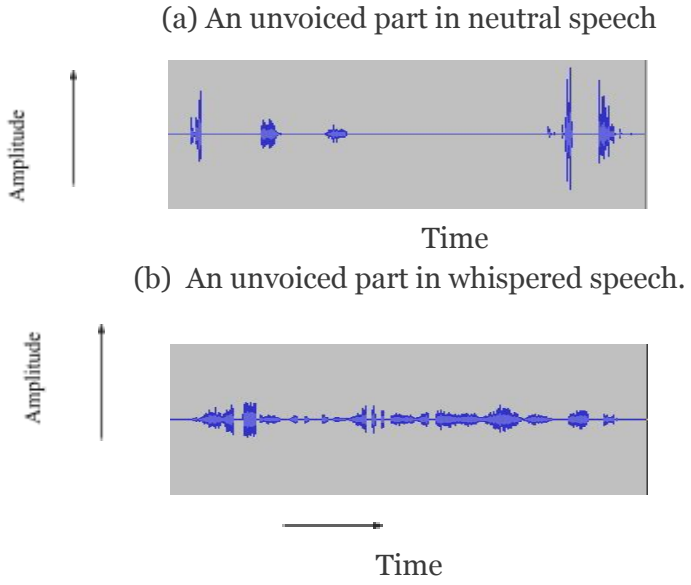

Fig.2: An unvoiced part in neutral and whispered speech

However, the unvoiced part in neutral voice is very small as shown in Fig. 2. It means a major part of the sound signal is not involved in the comparison process; hence there may be a loss of the intelligence. As fixed frame size is used while speech processing, the same frame may include voiced and unvoiced portion which confuses the decision.

III. AUDIO DESCRIPTORS AND TIMBRE

Audio Descriptors represent unique speaker dependent attributes. Specific attributes are identified and used as per suitability to the particular application. One can extract attributes in a time domain or frequency domain. Fourier transform is a widely used tool for visualizing the useful and sufficient characteristics of sound for classification in a most convenient way. MPEG-7 consists of rich and different classes of audio descriptors suitable for a variety of applications. The elements of the audio framework of a MPEG-7 standard are:

Descriptors (D): It includes timbrel temporal descriptor, timbrel spectral descriptors, basic spectral descriptors, spectral basis descriptors, signal parameter descriptors and silence descriptors.

Description schemes (DSs): It is way of integrating the relationships between the components of descriptors which supports other applications. It includes musical instrument, melody description audio signature, general sound recognition and indexing, and spoken content.

A description definition language (DDL): The DDL is syntax in schema language which is useful to express and combine audio descriptors and description schemes. It is a very powerful element in MPEG-7 which makes possible for user to create extension and modify descriptors, DSs and or even to define new ones [15].

Timbre is not understood clearly to the researchers and everyone has defined it in own way [16].

Definition 1: cannot be quantified like any other physical quantity but defined as a perceptual and subjective attribute of sound.

Definition 2: Timbre owe to many unknown dimensions, where the qualities and importance of feature are not very clear. Though cannot be mapped to a one-dimensional scale, it is even not uncoupled from the other one-dimensional components.

It is stated in [16] that trained musicians can "easily" identify instruments from a musical tone, but the error rates by machine are found higher. This fact pointed out the need of incorporating perceptual based method for extracting unknown useful information which cannot be done by traditional way.

IV. SYSTEM DESCRIPTION

The detailed process of the speaker identification of the whispered sound and components used in this paper are explained in Fig. 4.

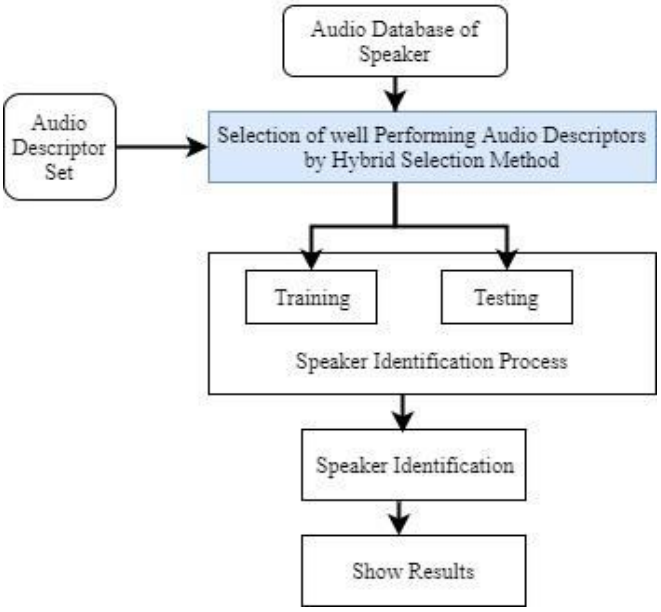

Fig. 3: System block diagram for Speaker Identification of whispered sound

The speaker identification mainly includes three steps: feature extraction, training and testing. Our system uses database of 24 speakers consisting audio samples both in neutral and whispered mode. Hybrid Selection method examines the performance of all probable audio descriptors and eliminates low performing descriptors. Selected timbre audio features are used for training the system. A whispered query given for testing is classified by K-NN classifier and identified speaker identity is declared.

4.1 Audio Database of Speakers

Speaker database consist of 24 speakers, about 10 samples each in both neutral and whispered mode. It includes 17 male and 7 female candidates. The database includes utterances in English language having duration of 2-3 seconds. Acer TravelMate 8000 series notebook was used for recordings, with a sampling frequency of 16 kHz and stored as 16-bit PCM-encoded WAV files. As the recorded speech (neutral and whispered both) is having bandwidth below 8 kHz, sampling frequency is selected according to the nyquist criteria. While recording the database, phonetically balanced sentences from TIMIT database are selected. The recording is done in a room with no acoustic conditioning to check the robustness of the proposed system in a bit noisy environment.

4.2 Hybrid selection method for finding the best Feature for speaker identification in whispers

MPEG-7 standards defined more than 52 audio descriptors including all low level descriptors. Using all of the descriptors will make the system complex; moreover all of them are not useful in the whispered database. The various dimension reducing algorithms like PCA, Sequential forward generation (SFG) or Sequential backward generation (SBG) are suggested. The SFG

algorithm starts with an empty set of features and adds new well performing feature at each level. The process continues adding new feature for identification task until no further improvement is found. The SBG process is similar to SFG but starts with a full set of features and eliminates those features giving the lowest results until no further fall in result is found [16]. Methods more of less similar to SFG or SBG can be adopted for its ease of implementation.

The Hybrid Selection Algorithm used in this paper is demonstrated below. The results confirmed our assumption that perceptually motivated timbre features are most suitable for the speaker identification within the whispered database. When the algorithm is executed for our database; the features namely roll-off, roughness, brightness, irregularity and MFCC are found to be giving the best identification accuracy. The performances of audio descriptors are also the database dependent [17]. Musical Information Retrieval (MIR) toolbox is easily incorporated in the Matlab environment. It makes use of functions available in public-domain tool boxes such as the auditory toolbox (Slaney, 1998), NetLab (Nabney, 2002), or SOM toolbox (Vesanto, 1999). MIR toolbox is adopted for easy use and computation simplicity. Among the selected timbre features, irregularity & roll-off are based on an envelope analysis while roughness & brightness are the spectrum based analysis and MFCC uses the Mel spectrum.

The algorithm described below is similar to SBG which starts with set of all probable features and goes about eliminating the low performing features at successive level.
The steps followed for the algorithm are as below:

1. Take all the probable Audio Descriptors (ADs).
2. *Level 1*: Test all of them separately and calculate classification accuracy for each.
3. Rearrange all ADs in the descending order of accuracy and choose only first three ADs giving highest accuracy.
4. *Level 2*: Now the selected ADs are appended with all ADs, one by one, and the best three feature vector as a combination of two ADs are passed to the next level.
5. *Level 3*: Feature vector of two ADs are appended by the third AD sequentially and the best vectors of three ADs are passed to the next level.
6. *Level n*: This is the last iteration before (n+1) level, where the further addition of AD does not increase the accuracy.

After the last iteration, the identification accuracy is maximum with the audio descriptors namely Roll-off, roughness, irregularity, brightness and, Mel Frequency Cepstral Coefficient (MFCC) which is proven in result section.

4.3 Timbre audio descriptor significance and Vector space

This section covers the definitions, significance [17-19], and vector space of timbre audio descriptors. It also compares the feature values for inter-speaker and intra-speaker voice samples.
Roll-off frequency: The roll-off is useful for distinguishing voiced from an unvoiced speech from envelope nature. Hence, roll-off is very important descriptor in whispered speech. It is estimated from the major energy (85% or 95%) concentrated below the frequency bins selected. Mathematically,

$$\sum_{i=1}^{R} S_t[n] = 0.85 \times \sum_{i=1}^{N} S_t[n] \qquad (1)$$

Roughness: Average roughness is found by *mir-roughness* function which estimates the average variance between all peaks of the spectrum of the signal. Another way, roughness can find the presence of harmonics generally higher than the 6th harmonic.

Brightness: It is the midpoint of the energy distribution of the frequency.

Irregularity: It analyses all peaks of the spectrum, and variations are calculated among successive peaks which is elaborated by,

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

For better speaker identification, it is desired that the feature vectors used for the different speakers should exhibit discriminative property. Three features namely brightness, irregularity and roughness (for the sake of example) are plotted as a three-dimensional vector in the 3-D space as shown in fig. 4. For each speaker sample, three features mentioned above are considered as three dimensions of a vector and every sample of similar speaker uses a unique colour. Fig. 4 shows that all the samples of the same speaker are located nearby which indicates intra-speaker similarity. While every speaker sample is sufficiently isolated from every other speaker.

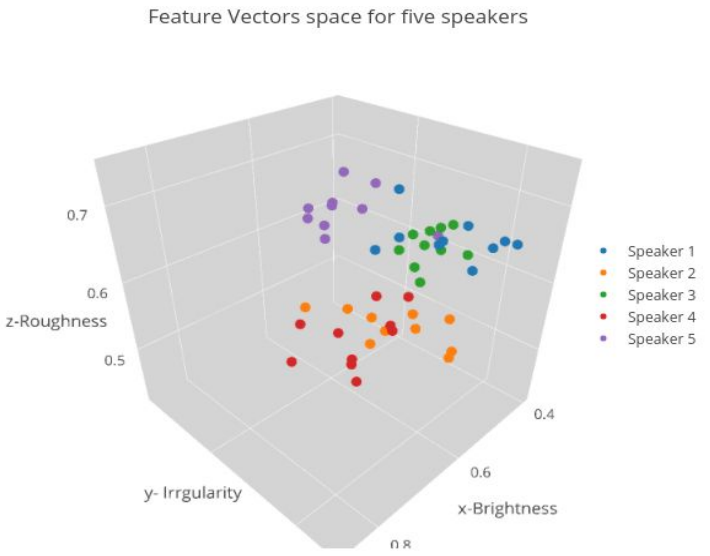

Fig. 4: Feature vectors in 3-D space for five speakers each with 10 samples

Fig.5 shows the acoustic distance for selected timbre features as Roll-off, Roughness, Brightness andregularity. Four samples of speakers are investigated w.r.t. above features. Here, two speech samples of two different speakers are used for illustration. SP1_1 and SP1_2 are samples of speaker 1 while SP2_1 and SP2_2 are samples of speaker 1. From below plot, it is depicted that there is an appreciable similarity in the acoustic distances of speaker samples of the same speaker i.e. acoustic distances for two samples of speaker 1 (SP1_1 and SP1_2) and speaker 2(SP2_1 and SP2_2) are nearby. On the other hand, acoustic distances of two different speakers (SP1_1 and SP2_1, SP1_1 and SP2_2, SP1_2 or SP2_1, SP1_2 or SP2_2) are discriminatory.

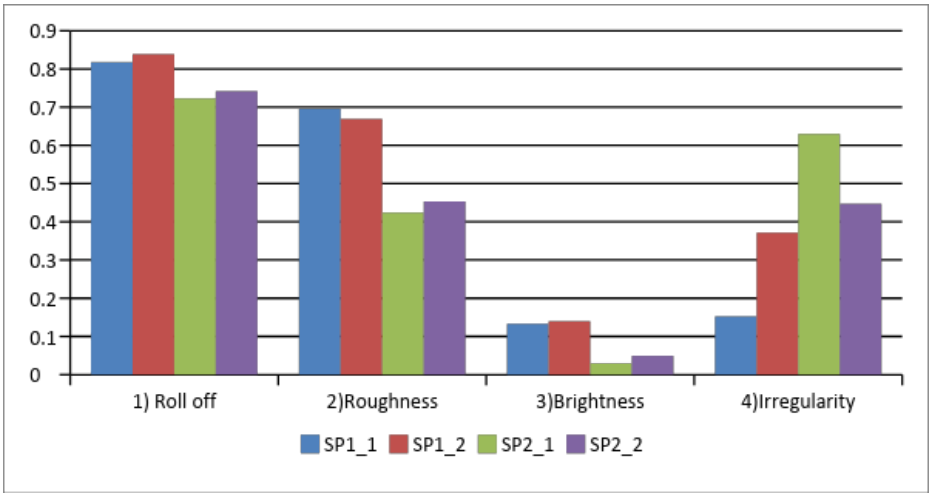

Fig. 5: Inter-speaker and intra-speaker feature comparison

The distance models are presented as a compact and easy way to understand raw data which can be useful the informational redundancy detaining of the audio descriptors [20]. Here, MFCC is not included in the feature vector due to large dimension (20 coefficient) which would make the representation somewhat redundant and unclear. While actual execution, Roll-off, Roughness, Brightness, irregularity and 20- MFCC coefficients form a vector which is used as a representation of unique characteristics of the individual speaker. The discrimination ability among selected features and different speakers is investigated by the standard deviation. The following table shows the standard deviation (σ) analysis. Standard deviation (σ) estimates the tendency of an individual elements of a group to deviate from the mean value (μ) for the group. A low standard deviation means that the group elements are distributed to be close to the mean, while a high value indicates that the data points are widely spread from the expected value (mean).

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (3)$$

Individual feature values of five speakers with 10 samples are tabulated in independent arrays first. Then the intra-speaker standard deviation for each audio descriptor is calculated and listed as below.

Table 2: Standard deviation for five Speakers pertaining to the selected timbre features

Selected Feature	Standard Deviation				
	SP_1	SP_2	SP_3	SP_4	SP_5
1) Roll-off	0.059773	0.067684	0.036448	0.041838	0.162628
2) Brightness	0.030004	0.039599	0.022873	0.026962	0.036932
3) Roughness	0.045886	0.126484	0.178834	0.024304	0.056731
4) Irregularity	0.082886	0.101791	0.141807	0.10608	0.117716
5) MFCC	0.083468	0.072469	0.046284	0.080897	0.035945

From the above table, the intra-speaker deviation for every feature is found to be very small which implies the small variations in the samples of the same speaker. Secondly, the standard deviation for every independent feature of the same speaker is different. Also, the standard deviation is different for every other speaker pertaining to all individual features. Hence, intra-speaker similarity and discrimination for inter-feature and inter-speaker acoustic properties can be predicted. As mentioned earlier, the identification process uses all selected features in the

form of vector. Hence, it is recommended to analyse inter-speaker discrimination on the basis of a feature vector. Hence, the standard deviation between all the samples of the same speaker for each feature are arranged in the form of an array (feature vector) and a box chart is represented for all 24 speakers in the database. Every speaker is represented by a range of values with min, max and,median. A major observation is noted here that each speaker shows a discriminative range of values w.r.t. feature vector.

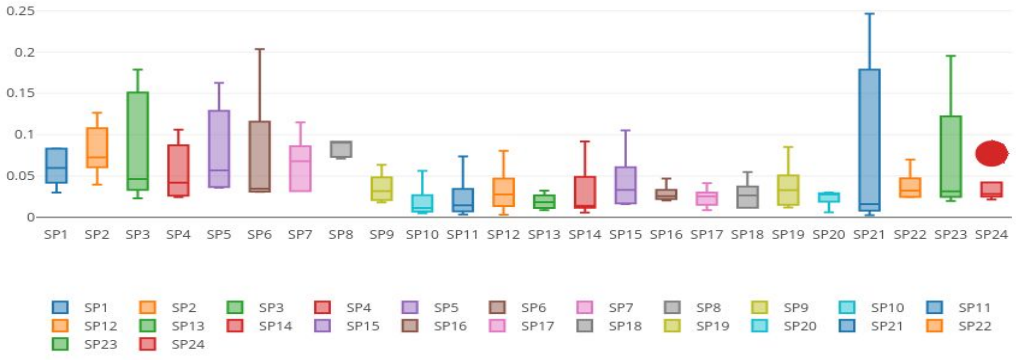

Fig. 5: A box chart for 24 speakers with all selected features as a vector

4.4 K-NN classifier K-Nearest Neighbor (KNN) Classifier

The KNN classifier lazy learning algorithm which does not require the prior knowledge of data. K-NN offers variations based on a number of nearest neighbors (k), a distance function (d), and a decision rule. For an audio file X_n with n samples, 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 (wi|x) as

$$P\,(w_i|x) = \frac{k_i}{k} \,.\, P(w_i) \qquad (4)$$

where ki is the number of vectors which belongs to class w_i within the subset of k vectors[21]. 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 [22]. The results are also affected by the way of calculating distances between the training and testing vectors by various distance metrics available.

V. RESULTS

The system is initially trained and tested with the whispered voice samples for the selection of well-performing features suitable for whispered voice. Here, 80% samples are used while training and 20% samples were used while testing. Results are compared with only MFCC features and timbre features (zero-crossing, brightness, roughness, roll-off and irregularity) including MFCC. The K-NN classifier is used for classification.

Table 3 elaborates selection of audio descriptors on the basis of the best classification results for the whispered database. The best audio descriptors found are namely MFCC, roll-off, brightness, irregularity and roughness.

Table 3: Classification accuracy using MFCC only and timbre with MFCC for five speakers

Sr. No.	Audio Descriptors	Combination*		
		MFCC+Roll-off+Brightness+Roughness	MFCC+Roll-off+Brightness+Irregularity	MFCC+Roll-off+Roughness +Irregularity
1.	MFCC	....	....	....
2.	Attack-time	48	32	42
3.	Attack-slope	46	18	20
4.	ZCR	34	36	30
5.	Roll-off	....	....	....
6.	Brightness	....	....	74
7.	Irregularity	74	....	....
8.	Roughness	....	74	....

It is seen that classification accuracy is increasing as we go about appending the audio descriptors and it is maximum as 74% when all five descriptors mentioned above are appended. However, when attack-time and attack-slope and zero-crossing are added in the feature vector, the accuracy suddenly drops, hence they are eliminated.

A baseline system using traditional MFCC features and k-NN classifier as described in [23] is used. We could have used some state of the art classifier, however our focus of study is on the performance of timbre features for whispered database and use of other good classifier is left for future scope. Again, K-NN is proven to be the simplest and efficient classifier invariety of applications. As a thumb rule, 80% neutral samples are used for training and 20% whispered samples are used for testing.

Comparative testing accuracy using MFCC-KNN and timbre with MFCC-KNN is shown in Table 4 below.

Table 4: % Testing accuracy of Speaker identification with MFCC only and Timbre with MFCC

No. of speakers	Training (neutral) samples	Testing (whispered) samples	MFCC only	Timbre with MFCC
4	32	8	62.5	75.0
8	64	16	68.7	81.2
12	96	24	66.6	75.0
16	128	32	65.6	71.8
20	160	40	47.5	65.0
24	192	48	41.6	52.0

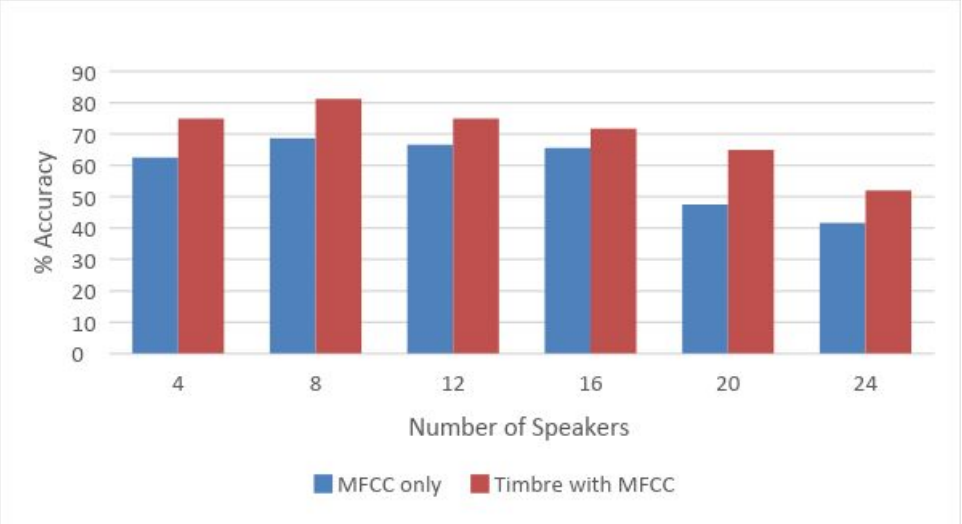

Fig. 7: Comparative identification results by Timbre features and MFCC features

Above results are observed with best performing timbre features and KNN classifier.

VI. CONCLUSIONS

The paper contributes (1) Creation of whispered-neutral speaker adaption database, (2) Analysis of whispered speech to know the limitations of a method like speaker identification on the basis unvoiced part only (3) Selection of the best audio descriptors useful for whispered speech database. (4) Analysis of timbre features on the basis of standard deviation for intra-speaker similarity and inter-speaker and inter-feature discrimination ability. (5) Further, experimental results proved that the perceptual based timbre features which are selected by Hybrid Selection Algorithm (brightness, roughness, roll-off, irregularity and MFCC) are found outperforming giving an enhancement in accuracy by 10.4% compared to traditional MFCC features only.

It is obvious that the set of well-performing feature depends upon the database used. By using a possible number of different databases of same kind e.g. whispered speech, list of most common and best-performing features can be known. Also, the robustness of timbre features can be investigated in depth by adding known and controlled noise in the noise-free (standard) speaker database.

K-NN classifier used here. However, use of any other advance classifier is recommended to enhance the results further, retaining the same feature extraction module used in the above work.

REFERENCES

1. Xing Fan and John H. L. Hansen, "Speaker Identification Within Whispered Speech Audio Streams", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No.5, pp. 1408 - 1421, 2011.
2. Chi Zhang and John H.L. Hansen, "Analysis and Classification of Speech Mode: Whispered through Shouted", in Proc. Interspeech, pp. 2289—2292, 2007.
3. Seiichi Nakagawa, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition", IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 4, May 2012.
4. Xing Fan and John H.L. Hansen, "Speaker Identification with Whispered Speech based on Modified LFCC Parameters and Feature Mapping", ICASSP, IEEE International conference on Acoustics , Speech and Signal Processing, pp. 2425-2428, 2009.
5. Xing Fan and John H.L. Hansen, "Speaker Identification for Whispered Speech Using a Training Feature Transformation from Neutral to Whisper", IEEE Transactions on

Audio, Speech, and Language Processing 2011, Volume: 19, Issue: 5, 2011, pp.1408 – 1421.
6. Abrham Debasu Mengistu, Dagnachew Melesew Alemayehu, "Text Independent Amharic Language Speaker Identification In Noisy Environments using speech Processing Techniques", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 5, No. 1, 2017, pp. 109 – 114.
7. Inggih Permana, "A Comparative Study on Similarity Measurement in Noisy Voice Speaker Identification", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 1, No. 3, 2016, pp. 590 -596.
8. Di Wu, Jie Cao, Jinhua Wang, "Speaker Recognition Based on i-vector and Improved Local Preserving Projection", TELKOMNIKA Indonesian Journal of Electrical Engineering, Vol.12, No.6, 2014, pp. 4299 – 4305.
9. Saurabh H. Deshmukh Dr. S.G. Bhirud, "A Novel Method to Identify Audio Descriptors, Useful in Gender Identification from North Indian Classical Music Vocal", IJCSIT, Vol. 5 (2) , 2014, pp 1139-1143.
10. Hyoung-Gook Kim, Edgar Berdahl, Nicolas Moreau, Thomas Sikora, "Speaker recognition using MPEG-7 descriptors", 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003, Geneva, Switzerland, 2003.
11. Swe Zin Kalayar Khine Tin Lay New Haizhou Li, "Exploring Perceptual Based Feature for Singer Identification", CMMR, Computer Music Modeling and Retrieval. Sense of Sounds, 2007, pp 159-171.
12. Xing Fan, Keith W. Godin, John H.L. Hansen, "Acoustic Analysis of Whispered Speech for Phoneme and Speaker Dependency", in proceedings, Acoustic AO, INTERSPEEH, 2011, pp 181-184.
13. T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," Speech Comm., vol. 45, pp. 139–152, 2005.
14. Mark Greenwood, Andrew Kinghorn, "SUVING: Automatic Silence/Unvoiced /Voiced Classification of Speech", Undergraduate Coursework, Department of Computer Science, The University of Sheffield, UK, 1999.
15. Hyoung-Gook Kim,Nicolas Moreau, Thomas Sikora.(2005). MPEG-7 - Audio and Beyond Audio Content Indexing and Retrieval, a textbook by John Viley & sons Publication [Online]. Available: http://www. wiley.com
16. Tae Hong Park, Towards Automatic Musical Instrument Recognition, Ph.D. thesis, the department of music, Princeton University, 2004.
17. Saurabh H. Deshmukh, On the Selection of Audio Descriptors and Identification of Singer in North Indian Classical Music, Ph.D. dissertation, Dept. Comp. Engg., NMIMS Deemed-to-be University, 2014.
18. MIR toolbox 1.3.3 (Matlab Central Version) User's Manual Olivier Lartillot, Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland, July, 12th, 2011.
19. A Matlab Toolbox for Music Information Retrieval Olivier Lartillot1, Petri Toiviainen1, and Tuomas Eerola1, Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert- Ludwigs- Universität Freiburg, March 7-9, 2007, pp.261-268.
20. Geoffroy Peetersa, Bruno L. Giordano Patrick Susini and Nicolas Misdariis, "The Toolbox: Extracting audio descriptors from musical signals", Journal of the Acoustical Society of America, 130(5), pp. 2902-2916.
21. Jashmin K Shah, Brett Y Smolenski, Robert E Yantorno and Ananth N Iyer, "Sequential k-Nearest Neighbor Pattern Recognition For Usable Speech Classification", Signal Processing Conference, 2004 12th European, ieee Xplore, 2015.
22. Sreelekshmi S Kumar, and Syama R," 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.

23. Fransiska, Ranny, "Voice Recognition Using k-Nearest Neighbor and Double Distance Method", 1-5. 10.1109/ ICIMSA. 2016. 7504045.

This page is intentionally left blank