International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018) pp. 6660-6666

Speaker Identification of Whispering Sound using Selected Audio Descriptors

V. M. Sardar[1], S. D. Shirbahadurkar[2]

[1]*Department of Electronics and Telecommunication, RSCoE and Research Centre, SPPU, Pune, India.*

[2]*Department of Electronics and Telecommunication, Zeal College of Engineering, Pune, India.*

Abstract

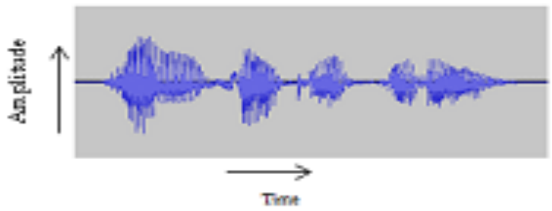
Whispering speech mode is adapted by speakers for any one of the reasons like secrecy of confidential data, avoiding being overheard in public places or hiding the identity intentionally. As acoustic properties of whispered speech are drastically changed compared to neutral speech; it makes difficult to identify the speaker from a whispered sound. This task requires the perceptual analysis of the whispering sound signal as do the humans. Many researchers presented various techniques for speaker identification of whispering sound but have some limitations. This paper describes the efficient method of identifying the speaker within the whispered speech using timbrel features which haven't been used so far in the whispered case. They are suitable here due to their multidimensional nature and perceptual ability. But all the timbrel audio descriptors are not well-performing for the whispered data. Hence, by using Hybrid Selection method, the most suitable timbrel audio features are selected and used. Timbrel features show an increase in the identification accuracy as 10.9 % compared to traditional MFCC features. A database containing 650 utterances (whispered and neutral) of 35 speakers is created and used for the experiments. K-means classifier with a random choice of the centroid is used for classification.

Keywords: Musical Information Retrieval (MIR); Speaker Identification; Timbre; Whispered Speech

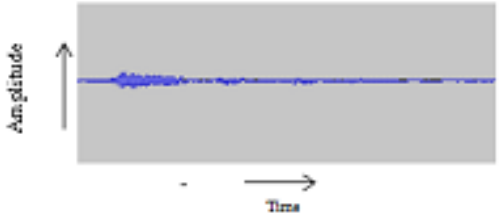
INTRODUCTION

The speaker identification mainly includes three steps as (i) Feature extraction: Compact and the unique speaker-specific information is extracted in this process. (ii) Training: It uses the classifier to model the speaker from its multiple voice samples to consolidate all the intra-speaker variations. (iii) Testing: In this step, the speaker query is tested for speaker identification. It is compared with the speaker models generated from the database while training. Speech modes are classified as shouted, loud, neutral, soft, and whispered with decreasing order of energy level. The spectral slope is found as the minimum for whispered speech with respect to neutral speech. This indicates that energy contents in case of whispered and soft speech are concentrated in higher frequency [1]. Widely used Mel Frequency Cepstral Coefficient (MFCC) works on a Mel scale, failing to capture some details in the high-frequency range. Hence the selection of feature suitable to the type of database is an essential task in speaker identification.Whispered speech properties drastically change compared to neutral speech due to variation in vocal efforts. While speaking in neutral mode, the vocal folds vibrate periodically. However, during a whispering speech, a continuous air stream without vibration and periodicity is propagated. The major changes in characteristics of whispered speech compared to neutral speech are summarized as: Loss of periodic excitation or harmonic structure, shifting of formants to the higher frequencies, flatter spectral slope and lower energy. Hence the performance of the system (neutral train - neutral test) with traditional MFCC features and Gaussian mixture model for classification (MFCC-GMM) degrades considerably when tested with whispered samples of utterances.

The signal-to-noise ratio (SNR) is a very important factor for achieving better speaker identification. The whisper speech can be classified qualitatively as high and low-performance whisper based on SNR and high-performance whisper reported better identification [1]-[2]. Figure 1(a) and (b) show the time domain waveforms of speech signals of neutral and whispered speech. The low magnitude and low SNR are major difficulties in whispers. When the whispered database files used in our experiment are investigated for the contrast ratios, they are found in the range 7 to 9dB. Hence, it needs to be processed for noise reduction.



(a) Neutral voice sample



(b) Whispered voice sample

Figure 1. Time domain waveforms of neutral and whispered voice

When KL divergence between whispered and neutral speech is investigated; it is found low for the unvoiced consonants compared to non-unvoiced consonants. It means unvoiced consonant does not deviate much in whispered and neutral mode [2]. So the identification can be increased if a comparison is made by only unvoiced part in neutral and whispered samples. Voiced and unvoiced parts are separated on the basis of energy and zero- crossing rate i.e. high energy & low ZCR for a voiced part and low energy and high ZCR for unvoiced part of utterance [3]. But the whispered speech is similar to noise having low SNR and high ZCR. This will misinterpret the decision as an "unvoiced" rather than "silence." Secondly, while processing audio files, generally fixed size of framing will lead to mixing voiced and unvoiced utterances in the same frame. Hence separation of a voiced and unvoiced part will be errorful.

Linear frequency cepstral coefficient (LFCC) is also the useful feature for the speech containing high- frequency formants. It is concluded that LFCC should be most recommended for the female trials [4]. This is due to the reason that linear scale captures details in higher frequency range equally and a female pitch is higher. Similarly, LFCC is useful for a whispered speaker case, as formants are shifted to the higher frequency. However, from the values of formant listing, the shift of formant in whispered speech is not found consistent when investigated for the speech files in our database. The formants shift in the whispered speech compared to neutral speech is speaker dependent to much extent.

Further, using feature mapping technique, further improvement in speaker identification compared to MFCC is observed [5]. But as the feature mapping of whispered audio with neutral audio is being done in a testing phase, it may slow down the speed of the system. Another approach to feature transformation from neutral to whispered speech is adapted in [6]. It offers two advantages compared to feature mapping technique: the pseudo-whispered generated data solved the problem of limited availability of whispered data. Secondly, pseudo-whisper is generated before testing phase; hence system speed will be better. This approach uses Vector Tylor Series (VTS)/ Constrained maximum likelihood linear regression (CMLLR) which contributed the increased accuracy of 46.26% compared to 79.29% of the baseline system. However, in the real scenario, the variation in whispered utterances compared to neutral may not be same as that generated by the pseudo whispered features. Again the variation in whispered speech compared to neutral speech for the individual speaker is also not the same.

Timbre based features such as vibrato and attack-decay are used for singer identification. They are found effective and accuracy is 87.8% in segment level singer identification [7]. The task of singer identification from North Indian classical music is addressed in [8]. This type of music is more complex structure and very difficult to separate vocal from music. Here only selected perceptual timbre features combination by Hybrid Selection method; namely RMS Energy, brightness and fundamental frequency are used. The accuracy is reported to be 70% when K-means clustering has been used for classification. When redundancy in the feature information removed, the best performance is observed. Particle Swarm Optimization technique (PSO) is used in the feature selection process and classification accuracy is reached to 98.15% for speaker identification with Arabic neutral speech [9].

Various applications may be extended by authorizing the speaker like stereo system operations or voice based door access [10]. Various ascents or languages can be used to develop the universal speaker identification system. A system to identify Quranic Reciter in the Arabic language is presented in [11].

SYSTEM DESCRIPTION

Following sections illustrate the processes and components used for the speaker identification in the

A. Steps in Speaker Identification Process

Three steps are used in the speaker identification process. They are listed below as shown in Figure 2.

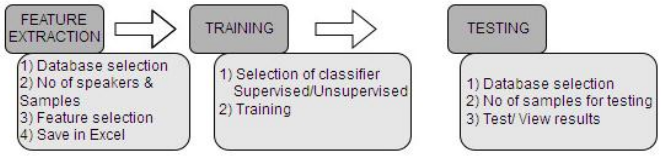


Figure 2. System setups for Whispered Speaker Identification

Feature extraction: Features are the compact representation of speaker-specific information. Features are extracted for every audio file in the speaker database while training as well as testing phase. The flexibility may be offered in this step like (i) Selection of the particular database out of the available sets (ii) Using number of speakers and number of samples per speaker for training (iii) Selection of desired features which include MFCC, zero crossing rate, roll-off, roughness, brightness and irregularity disjointedly or in combination as a features vector. Finally, the extracted features of all the speaker samples can be saved.

Training: This module may offer the options like (i) Selection of the desired classifier for training. The speaker models are generated by this process. Both supervised and unsupervised classifiers can be used for an experiment.

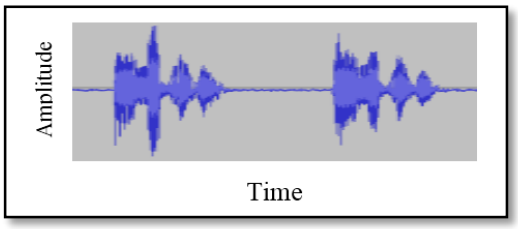
Testing: While testing, the facilities are incorporated into the system as (i) Selection of the database for testing (it is essential to have the same number of speakers to be tested as used in training.) (ii) Number of samples used while testing. The samples used while testing is more than that used while training. The test results are displayed which shows the identification accuracy in two ways: (a) Accuracy of identification with neutral samples which are selected while training, (b) Accuracy with the additional whispered samples used while testing. Flexibility proposed in the system makes us possible to investigate the various set of experiments with variable speaker numbers, a variety of features and classifiers.

B. Speaker Audio Database

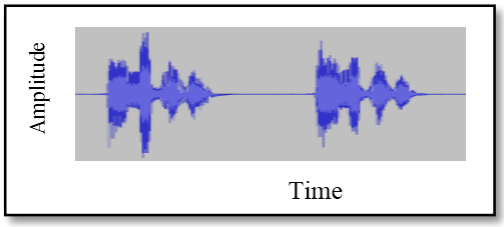
Speaker database consists of 650 samples of 35 speakers, about 10 samples each, in both neutral and whispered mode. The speaker recordings are the mix of male and female candidates i.e. 24 male and 11 female. The recording is done in a room without sound-proofing so that the effect of noise is also incorporated. Utterances are in English language having the duration of 2-3 seconds. Acer Travel Mate 8000 series notebook was used for recordings, with a sampling frequency of 16 kHz and stored as PCM-encoded WAV files. The database is further processed for removing noise. Two databases are ready: DB1 consists of all voice samples including the noise while database DB2 is processed for noise reduction.

C. Noise processing

While recording, noise is mixed with the speech recording from various sources. To differentiate noise from the speech is difficult by using any of the filters, as the mixed noise is random. Here, ambient noise is taken as reference and subtracted to get noise-free samples to some extent as shown in Figure 3. In speaker identification experiment, SNR as a one of the quality measure should be around 10 dB for better identification [12].



(a) Recorded speech with noise



(b) Speech sample after noise removal

Figure 3. Recorded speeches with noise and after noise removal

D. Selection of Audio Descriptors

Only well- performing timbrel features from the set of all audio descriptors are identified as below:

Algorithm for selection of Audio Descriptors

1. The algorithm starts with all probable Audio Descriptors (ADs) under consideration.
2. All the ADs are tested separately for classification accuracy.
3. All ADs are arranged in the ascending order of accuracy and choose only first three ADs. (Selecting ½ of them is a good choice).
4. Selected ADs are tested in combination with all single ADs and now the combination of two ADs are arranged in ascending order of accuracy. Again choose only first three combinations.
5. Best combinations of two ADs selected from the previous step are further combined with the third AD successively and best combinations of three ADs are passed to next level.
6. Iterations will stop when the further addition of new AD does not increase the accuracy.

It is known that in speaker identification experiment, the confusion among the classes' increases when the number of speakers increases. Hence, a need for feature choice based on high robustness along with high accuracy is evolved for larger database size. However, the same algorithm is used to validate the robustness of features as well with increasing data.

E. Definitions and significance of Timbral Descriptors

MPEG-7 includes efficient and diversified audio descriptors which are having widespread application areas like audio storage and retrieval applications, classification on the basis of contents, recognition, surfing or classification of data. The major descriptors include signal parameter descriptors (fundamental frequency, harmonicity), timbral temporal descriptor (timbre), timbral spectral descriptors (spectral features in a linear frequency space) which are primarily applicable to the musical timbre perception and Spectral Basis Descriptors (sound indexing and organization description tools) [13].

Roll-Off frequency: It is indicative of the frequency below which major magnitude of the spectrum (85% or 95%) is concentrated. The roll-off distinguishes voiced from an unvoiced speech by spectral shape and considered as a way to represent the extent of high-frequency content of the speaker's voice. Hence, it is useful while speaker identification with whispered speech.

Roughness: It is an estimate of sensual disagreement (difference) which exists due to beating phenomenon between close frequency peaks. It estimates the average variance between all peaks of the spectrum of the signal.

Brightness: It is the measure of energy above certain cut-off frequency.

Irregularity: It is the measure of variation on the successive peaks of the spectrum. Irregularity is a ratio given by the square of the sum of the difference between successive peaks for the entire signal by square of the number of signal points.

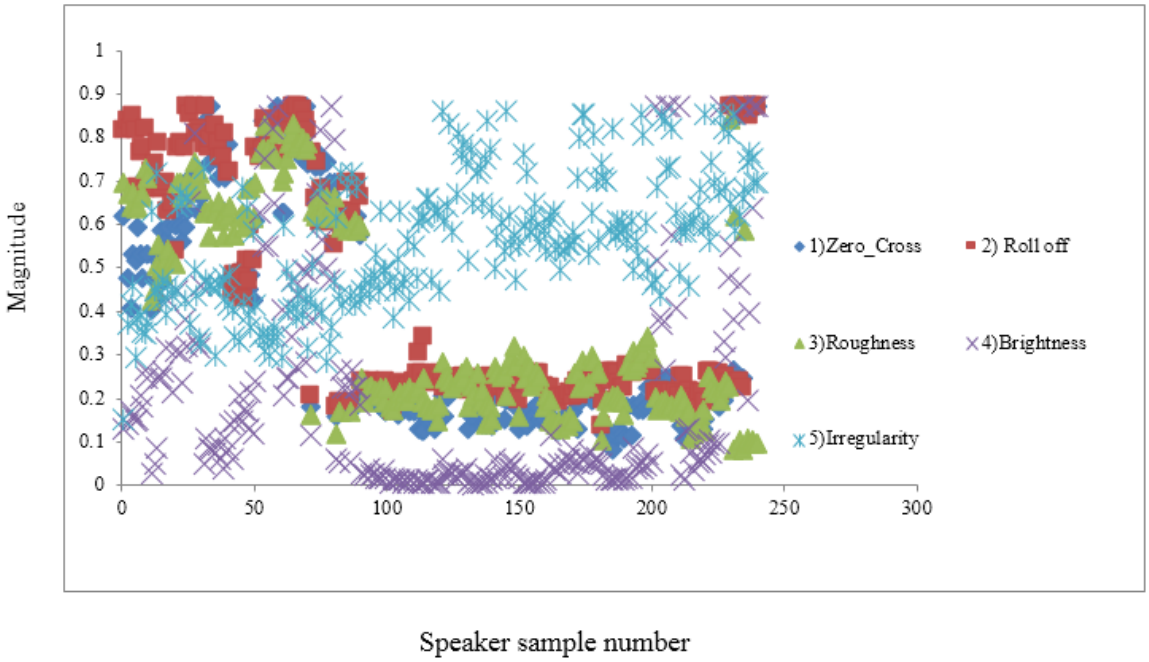


Figure 4. Feature space of Timbral descriptors extracted from the speaker audio samples

Zero crossing rate (ZCR): It is the rate of sign-alterations along a signal, i.e. the rate at which the signal toggles from positive to negative. ZCR is low in voiced part of speech and found high for unvoiced part [14].

F. K-means classifier

K-means algorithm partitions all the features into several clusters which are the representative for classification of the features of unknown identities. The clusters are generated by defining k centroids which are equal to the number of clusters. Each feature is grouped in a particular cluster on the basis of the nearest centroid. If a set of 'n' observations, each d-dimensional real vector (x_1, x_2,…, x_n) are available, k-means clustering aims to partition the 'n' observations into k (≤ n) set S = {S_1, S_2, .. , S_k } so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance) [15]. To be specific, the purpose is to find:

$$\arg\min \sum_{i=1}^{k} \sum_{x \in si} \|x - \mu\|^2$$

$$= arg\ min \sum_{i=1}^{k} |si|\ Var\ Si \qquad (1)$$

where μ_i is the mean of points in the given set S. This is the same as minimizing the squared deviations from the pairwise points in the same cluster. This algorithm uses a number of iterations for refinement of clusters. Generally, considering the random centroid, the algorithm assigns each observation to the cluster based on the nearest mean, in general, Euclidian distance. In the update step, the centroids are assigned to new values. This continues till convergence when the assignments no more change further. K-means offers the linear time for computation with increasing data.

RESULT AND ANALYSIS

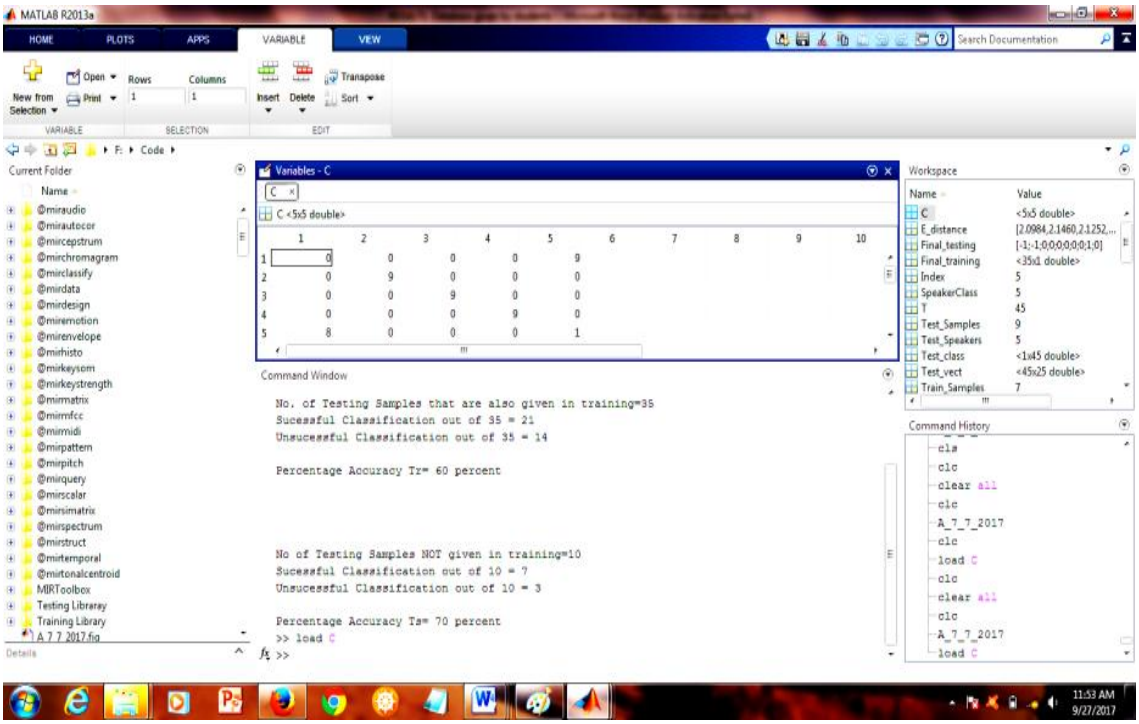
The database consists of 35 speakers having about 10 samples per speaker both in whispered and neutral mode. The speaker identification system is trained for the neutral speech samples and tested with the whispered voice samples. The best performing features viz. MFCC, roll-off, brightness, zero-crossing, irregularity, and roughness are used out of eight timbrel features. The performance of these features is validated by using the algorithm stated in section 2.4. The algorithm is tested for 10 speakers only. From the following Table 1, the maximum classification is found to be 68% when a combination of MFCC+ roll-off+ brightness+ zero-crossing+ irregularity +roughness is used. However, when feature vector is appended by Attack-time or Attack-slope, the accuracy is reduced to 43% and 36% respectively, hence they are discarded.

Table I. Validation of the best feature combinations giving highest accuracy

Sr. No.	Audio Descriptors	MFCC+Roll-off +Brightness +ZCR + Roughness	MFCC+Roll-off +Brightness +ZCR +Irregularity	MFCC+Roll-off+ Brightness +Roughness +Irregularity
1.	MFCC	….	….	….
2.	Attack-time	32	43	26
3.	Attack-slope	26	36	18
4.	ZCR	….	….	68
5.	Roll-off	….	….	….
6.	Brightness	….	….	….
7.	Irregularity	68	….	….
8.	Roughness	….	68	….

The K-means classifier is used for classification. The database selected is processed for noise removal, and same is used for all variety of experiments to compare the results. The average accuracy is noted after a number of trials till maximum accuracy is not achieved. The reason to go for multiple trials is that in K-means algorithm, the centroid is updated progressively till the convergence.

A snapshot as shown in Fig. 5 is an example of results to clear the considerations for calculation of accuracy.



No. of speakers = 5, Neutral samples while training = 35, whispered samples while testing= 10.

Figure 5. Snapshot of a result from Matlab

Here, five speakers with 35 samples of neutral speech are used while training. Out of 35 samples, 21 samples are correctly classified. Hence accuracy is 60%. Next, 10 samples of whispered speech are given for testing where five samples were correctly identified. Hence, testing accuracy is 50% which is only considered in the results. This is because the testing accuracy is the measure of a neutral-whisper mismatch which is important in this work. From the confusion matrix for 5 speakers with total 45 samples of neutral and whispered speech, the correctly identified samples are 26. Hence, overall accuracy is 57.7 % which is higher than testing accuracy but does not address a whisper-neutral mismatch.

Table 2. Accuracy using MFCC only and Timbre with MFCC

No. of speakers	Training (neutral) samples	Testing (whispered) samples	% Testing Accuracy			
			(for DB1) Low SNR		(for DB2) High SNR	
			MFCC only	Timbre with MFCC	MFCC only	Timbre with MFCC
5	35	10	50.0	50.0	60.0	70.0
10	70	20	45.0	55.0	58.0	68.0
15	105	30	43.3	46.6	52.0	65.3
20	140	40	37.5	42.5	47.5	61.0
25	175	50	34.0	36.0	42.4	55.2
30	210	60	28.3	25.0	39.0	51.3
35	245	70	22.8	24.2	33.1	44.0

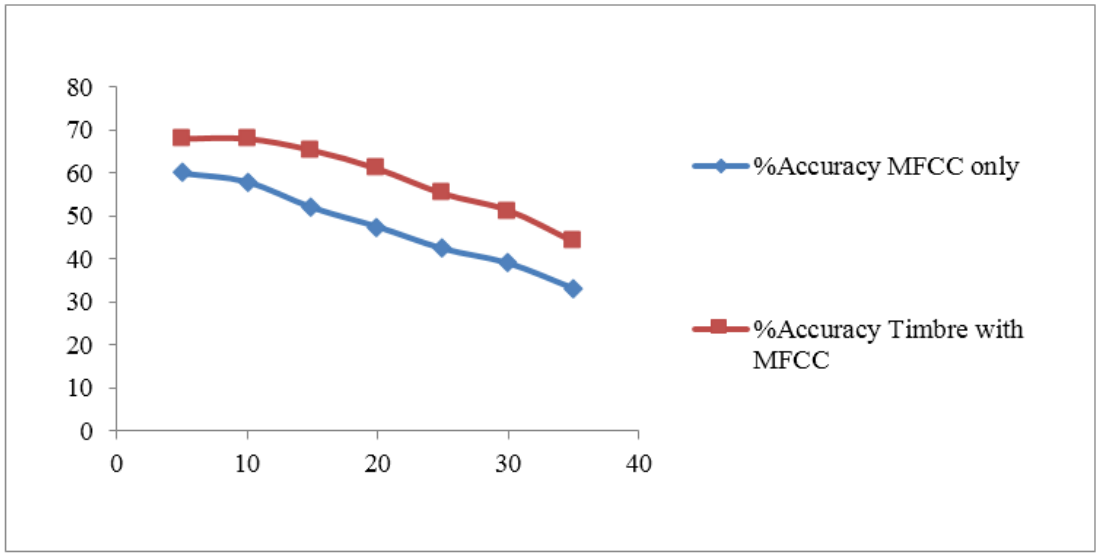


Figure 6. % Accuracy using only MFCC and using timbre including MFCC

Here, results are observed for reduction in accuracy with increasing number of speakers. Average reduction is found only 4.8% by using timbrel features along with MFCC whereas it is 6.58% using MFCC only. This indicates that timbrel features are more robust.

CONCLUSION

In the whispered speaker identification task, neutral trained automatic speaker identification system degrades significantly due to differences in vocal efforts between neutral and whispered voice. It is concluded that timbre features are effective to overcome whisper-neutral mismatch due to their perceptual ability. However, the feature vector should be compact containing only well-performing features. Best suitable features for any application may be found by the iterative algorithm discussed in this paper. Feature vector with limited timbrel features namely zero-crossing, brightness, roughness, roll-off, irregularity, and MFCC are found outperforming and robust. Timbre features reported an average accuracy enhancement of 11.83%. It is also seen that the identification results increases after noise processing.

Here, we have used K-means with a random selection of the centroid which is updated in successive iteration and the number of iterations for maximum accuracy is unknown. So the other approach of k-means clustering with fixed centroid can be used which may give the maximum value of accuracy without multiple trials. Also, the compensation techniques for whispered speech are to be used. The scope of this paper is limited to investigating the performance of most suitable timbrel features, selected by Hybrid Selection method, compared to traditional MFCC features. However, Timbrel features may be compared with other features in addition to MFCC. Further, selection of the well performing features can be also compared with using principle component analysis (PCA).

REFERENCES

[1] Chi Zhang and John H.L. Hansen, "Analysis and Classification of Speech Mode: Whispered through Shouted", INTERSPEECH 2007, Eighth Annual Conference of the International Speech Communication Association, 2007.

[2] Xing Fan and John H. L. Hansen, Fellow, IEEE, "Speaker Identification within Whispered Speech Audio Streams", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No.5, July 2011.

[3] Mark Greenwood, Andrew Kinghorn, "SUVING: Automatic Silence/Unvoiced /Voiced Classification of Speech", Undergraduate Coursework, Department of Computer Science, The University of Sheffield, UK, 1999.

[4] 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.

[5] Xing Fan and John H.L. Hansen, "Speaker Identification with Whispered Speech based on Modified LFCC Parameters and Feature Mapping", ICASSP 2009, IEEE International Conference on Acoustics, Speech and Signal Processing, 2009pp.

[6] 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 pp.1408 – 1421.

[7] Swe Zin Kalayar Khine Tin Lay New Haizhou Li, "Exploring Perceptual Based Timbre Feature for Singer Identification", CMMR 2007: Computer Music Modeling and Retrieval. Sense of Sounds pp 159-171.

[8] Saurabh H. Deshmukh Dr. S.G. Bhirud, "A Novel Method to Identify Audio Descriptors, Useful in Gender Identification from North Indian _ Classical Music Vocal", JCSIT, Vol. 5 (2), 2014.

[9] Ahmed Al-Hmouz, Khaled Daqrouq, Rami Al-Hmouz, Jaafar Alghazo, "Feature Reduction Method for Speaker Identification Systems Using Particle Swarm Optimization", International Journal of Engineering and Technology, Vol. 9,_No.3, 2017.

[10] Kayode Francis Akingbade, Okoko Mkpouto Umanna, Isiaka Ajewale Alimi. "Voice-Based Door Access Control System Using the Mel Frequency Cepstrum Coefficients and Gaussian Mixture Model", International Journal of Electrical and Computer Engineering (IJECE),_Vol. 4, No. 5, October 2014, pp. 643~64.

[11] Teddy Surya Gunawan, Nur Atikah Muhamat Saleh, Mira Kartiwi, "Development of Quranic Reciter Identification System using MFCC and GMM Classifier", International Journal of Electrical and Computer Engineering (IJECE) Vol. 8, No. 1, February 2018,pp372-378

[12] Barinov, Andrey, "Voice samples recording and speech quality assessment for forensic and automatic speaker identification. 129th Audio Engineering Society Convention 2010. 1.

[13] Hyoung-Gook Kim, Nicolas Moreau, Thomas Sikora MPEG-7, "Audio and Beyond Audio Content Indexing and Retrieval", a textbook by John Viley& sons Publication

[14] Olivier Lartillot , "MIR toolbox 1.3.3 (Matlab Central Version) User's Manual", Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland, July, 12th, 2011.

[15] Mahnoosh Mehrabani, John H.L. Hansen, "Singing speaker clustering based on subspace learning in the GMM mean super-vector space', Speech Communication 55 (2013) 653–666.