VOL. 13, NO. 13, JULY 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

SPEAKER IDENTIFICATION WITH WHISPERED SPEECH: DIFFERENT METHODS AND USE OF TIMBRAL AUDIO DESCRIPTORS

V. M. Sardar[1], S. D. Shirbahadurkar[2]
[1]Department of E and TC, R. S. CoE and Research, Pune, India
[2]Department of E and TC, Zeal College of Engineering, Pune, India
E-Mail: vijay.sardar11@gmail.com

ABSTRACT

Whispered mode of speech is preferred by people for secret conversations or avoiding to be overheard. E.g. sharing information like credit card number, bank account number or to hide the identity intentionally. This study focuses on various methods and techniques used for enhancing the accuracy in whispered speaker identification. MFCC is a most popular feature in the speaker identification experiment as the Mel scale is closer to the human hearing pattern. But the experiments with different feature-classifier combinations are tried by different researchers. However, considering the changes in vocal efforts while whispering, use of linear scale in feature extraction, separation of voiced and unvoiced part of utterances, whispered island detection, feature transformation from neutral to whisper, whispered to neutral efforts, contributes a lot. MIR toolbox has large number of feature sets suitable for representing the speaker-specific information efficiently, which may further increase the identification rate, especially with the timbral features.

Keywords: whispered, speaker identification, feature extraction, classifier, MIR, timbre.

1. INTRODUCTION

Among all biometric access applications, person identification based on his voice, is most simple and reliable means, relieving from the fear of forgetting or stealing passwords. Speaker recognition applications can be designed in two ways: text- dependent and text independent recognition [1]. In the former method, the same text (like customer number, passwords etc.) is used for training and testing phase both. Whereas in the later, speaker recognition is independent of the text spoken by speakers. The speaker identification may be defined as the process of finding the particular speaker from the stored speakers' database. It is used to confirm a speaker's identity and give access to confidential information areas. Whereas in the speaker verification process, identity claim of a speaker is accepted or rejected, which is mainly used for forensic applications [1].

All speaker recognition systems contain two main processes: Training (feature extraction and modelling) and Testing (feature matching and Identify)[2]. For both the phases, the feature extraction step is essential.

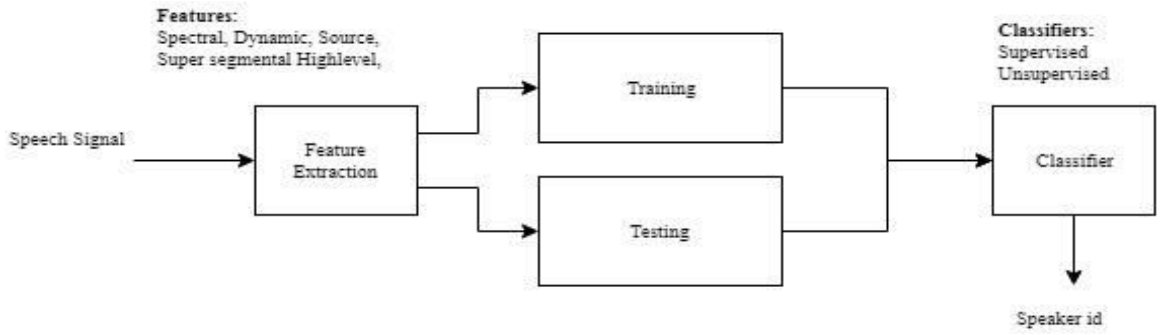

Figure-1. General speaker identification system.

The features represent the speech waveform by some kind of compact representation as, even for small duration speech, data is quite large. Hence, while selecting any type of the features, reduced data size while retaining speaker specific discriminative information is the major attribute. In training process, after feature extraction from multiple samples of the same speaker, the speaker model is generated which is stored in the database. The most widely used features are Mel-frequency cepstrum coefficients (MFCC), linear prediction coding (LPC), linear predictive cepstral coefficients (LPCC), Perceptual linear predictive coefficients (PLPC) etc. In the testing mode, speaker features are compared with the database of speaker models and identification results are declared (Figure-2).

Three types of models are adapted for classification: Stochastic model which includes e.g. Gaussian mixture model (GMM), Hidden Markov model (HMM)), Deterministic model (e.g. Support Vector Machines (SVM)),Template based model which are e.g. Dynamic Time Warping (DTW), Neural Networks (NN), Vector Quantization (VQ). Each of the model has its own pros and cons, hence selection of any classifier and features or even combination of feature- classifier depends

upon the type of application or type of speech used for a speaker identification process. (e.g. neutral, whispered, telephonic recording, singing, articulation disorders) Before extracting features, pre-emphasis and framing are essential steps. Simple 1st order High- pass filter is applied to the waveform for pre-emphasis. When continuous speech is divided into frames (15 to 20 ms), it is assumed to be stationary [3].

2. FEATURE EXTRACTION TECHNIQUES

The Mel Frequency Cepstrum (MFCC) is generated when the linear cosine transform of the log power spectrum on a non –linear Mel scale of frequency is taken. After pre-emphasis and framing, steps followed are windowing, Mel filter bank processing, discrete cosine transform. Different types of windows like the triangular, rectangular and hamming window may be used. The selection of particular windowing function should be selected according to the feature extraction method and the classifier type both. (For example MFCC- Hamming is the better combination; also the number of centroids used in VQ classifier also affects the result.) [4]

Linear predictive coding (LPC) is used in speaker identification along with other applications like speech synthesis and speech storage. LPC represents a current speech as a linear combination of the previous samples [5].

$$s(n) = \sum_{k=1}^{p} \alpha_k \, s(n-k) \quad (1)$$

Where s(n) is the predicted signal value, s(n-k) is the previous sample value and α_k is predictor coefficient. Coefficient (αk) is calculated by minimizing the sum of squared differences between the actual speech samples and the linearly predicted ones and to determine it from the speech frame is really difficult. Linear prediction (PLP) captures the short-term spectrum of the speech by several psychophysically based transformations and mostly similar to LPC [6]. Generally, the Speaker identification (especially in whispered mode) by a human is more accurate than a machine. The most important reason for this is human uses the perceptual ability which is more accurate rather than simply a template matching. Linear frequency cepstral coefficient (LFCC) is similar to the MFCC except frequency warping process is not required; LFCC uses the linearly spaced filter bank. Mel scale being the logarithmic scale, it captures the details in low-frequency range from the speaker data. MFCC is most popular features in speech as it represents logarithmic scale similar to human auditory perception. However, for the whispered case, LFCC is found outperforming in case of whispered speech compared to neutral speech [7].

3. CLASSIFIERS FOR SPEAKER IDENTIFICATION

Among two types of classifiers namely template based and stochastic approach, the former approach is the spectral template matching or spectrogram approach. A stochastic approach is based on a probabilistic model which focus on the probability rather than distances to feature vectors which may be more suitable for unpredictable utterances by a speaker.

Classifiers may also differentiate as supervised and unsupervised type. If you are training your machine learning task for every input with a corresponding target, it is called a supervised. It will be able to provide the target for any new input after sufficient training (e.g.SVM).Here, the learning algorithm seeks a function from inputs to the respective targets. Contrary, if you are training your machine learning task only with a set of inputs, it is called unsupervised learning (VQ, NN, SVM, GMM, HMM, DTW) which will be able to find the structure or relationships between different inputs.

Vector Quantization (VQ) is a basic classifier which provides the means of compressing the training data. A set of extracted short-term training feature vectors data would be large without the VQ. Therefore, the VQ method divides a large set of feature vectors into groups on the basis of the closest distance between them [8]. These may be called as codebooks. Each group is represented by its centroid point. The codebook is the compressed representative of large feature vector points. However, being the form of compression, the perceptual information in case of whispered speech may be lost.

Support Vector Machine (SVM) is used for discrimination of data points based on a separation with hyperplane [9]. Hence, major work in the SVM algorithm is to find the optimized hyperplane which will have the largest minimum distance (margin) to the training examples. i.e. larger margin will offer minimum confusion between the classes. SVM maps the given input to a high dimensional plane and it separate the classes with a hyper plane [9]. SVMs may not perform well for speech signals due to its restriction to work with fixed length vectors and binary decision. A binary decision is probably not very suitable in case of speaker identification due to variability in the sessions of same speaker. However, SVM can be used with prosodic, spectral and high level features.

A Gaussian mixture model is a generic probabilistic model for multivariate densities which can represent arbitrary densities, so it is well suited for unconstrained text-independent applications. It was first described in [10] for text-independent speaker identification. GMM computational time will dramatically increase when dealing with a large set of data.[11] Therefore, banking authentication systems often verify user identity instead of identifying a user voice with a full set of data.

Hidden Markov model represents a speech production system which is hidden or non-observable stochastic process. However, that can be observed through another stochastic process (i.e. speech signal). HMM is a modified version of a finite state machine having a set of hidden states Q, an output alphabet (observations) O, transition probabilities A, output (emission) probabilities B, and initial state probabilities Π . The current state is not observable. Instead, each state produces an output with a certain probability (B). However, HMM is not limited to characteristics or sequence only, but even probabilistic

models of these events also calculated from the feature. HMM is computationally more complex and needs more storage space, needs more training data to deal with inter-session issue [12], as even same utterances of the same speaker may vary trial to trial.

KNN is robust to noisy training data but distance-based learning is not clear. As whispered is a most noisy signal, KNN may be very useful especially when we use an inverse square of the weighted distance to classify.

4. LITERATURE SURVEY FOR SPEAKER IDENTIFICATION WITH WHISPERED SPEECH

Speaker Identification in whispered speech differs compared to neutral speech due to the facts: Loss of periodic excitation or harmonic structure in whispered speech, shifting of formants to higher frequencies, flatter spectral slope, lower energy. Due to these differences, traditional neutral-trained Mel-frequency cepstral coefficient–Gaussian mixture model (MFCC-GMM) speaker ID systems degrade significantly when tested with whispered speech. To investigate the reason, the KL divergence of unvoiced consonant and non-unvoiced consonants are compared. It is found that KL divergence is low for unvoiced consonant and high for the non-unvoiced consonants which indicates that unvoiced consonants do not vary as much as other phonemes between whispered and neutral speech [13]. The decision of voiced and unvoiced part in the utterance is made on the basis of energy and Zero- crossing rate i.e. Low ZCR &high energy is found for a voiced part and low energy & high ZCR for an unvoiced part of utterance [14]. However, this mechanism may suitable for clean and neutral speech. The difficulties found in whisper speech are: Whispered speech exhibits low SNR and the noise signal will have high ZCR which will mislead the decision as an “unvoiced” rather than “silence”. Fixed size of framing will create the possibility of mixing voiced and unvoiced utterances in the same frame. The whispered speech quality was also evaluated the quality of whisper speech as either high- or low-performance whisper based on SNR, spectral tilt in [13]. Here, high confidence part was tested on combined MFCC+LFCC/EFCC-GMM and this system achieved an absolute improvement of 8.85%-10.30% in speaker recognition, compared to the MFCC-GMM baseline. The various frequency scale used for the feature analysis like Mel, bark, exponential, linear will have their own advantages and disadvantages. So to combine the advantages of two or more, experiments carried on the exponential and linear frequency scales. This clue is useful to experiment with the combined scores of various features. i.e. even the multiple good performing features can be combined in the form of vectors while testing. In reference [15], it is also concluded that LFCC should be more widely used, at least for the female trials. The reason is that the linear scale captures details in higher frequency range and a female pitch is higher. So it can be related that LFCC is equally suitable to the whispered case as formant are shifted to the higher frequency.

In reference [16], the researcher used the modified linear frequency cepstral coefficient baseline system and also features mapping which involves fricative decision algorithm. Here MFCC was replaced by LFCC; increase in identification rate is reported. Thereafter, by using feature mapping and LFCC, an additional +10% improvement shown. But it may slow down the speed of the system as the mapping calculation being done in the testing phase.

An alternative approach to feature transformation from neutral to whispered speech is adapted in [17]. The method has two advantages: first, the non-availability or very less whispered data problem is solved. Secondly, it is the front end process on features hence requires no additional data transportation or calculation is required while testing phase. This model used Vector Tylor Series (VTS)/Constrained maximum likelihood linear regression (CMLLR) adaptation which offered the increased accuracy of 46.26% compared with the 79.29% accuracy of the baseline system. But probably this system has one threat that the whispered utterances in the real scenario may show a large deviation from the pseudo whispered features, as the variation of the whisper from the neutral speech is different among a set of speakers.

5. MIR TOOLBOX FOR WHISPERED SPEAKER IDENTIFICATION

Audio descriptors are the efficient and compact representation of discriminative information which can be extracted from digital audio waveforms in order to compare and classify the data. MPEG-7 consists of a strong set of audio descriptors which covers the characteristics of sound in diversified aspects. The major descriptors include Signal parameter descriptors (fundamental frequency, harmonicity), Timbral temporal descriptor (timbre), Timbral spectral descriptors (spectral features in a linear frequency space), especially applicable to the perception of musical timbre, Spectral Basis Descriptors (sound classification and indexing description tools) [18].

However, if all of the descriptors are used for classification, then the system becomes complex and messy [19]. MIR toolbox can be directly incorporated with Matlab software. Its strength lies, in particular, the computation of a large range of features from databases of audio files that can be applied to statistical analysis. Broadly, features in MIR toolbox can be classified as Energy related (RMS energy, low energy etc.), Timbral (MFCC, roll_off, brightness, roughness, irregularity, zero-cross etc.), statistical (centroid, spread, skewness, kurtosis,flatness, etc.), tonal (modality, HCDF etc.), Pitch related s (pitch, in-harmonicityetc).There exist more than 52 audio descriptors in literature including all low level descriptors specified in MPEG7 standards.

In this study, we propose the “timbral features” for Speaker Identification with whispered speech. Timber is a multidimensional concept which is not understood fully, nor its number of dimensions known. This might be due to the reason that it covers fields such as acoustics, music, engineering, and psychology.

As per "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."In the thesis written by Tae Hong Park,"Towards Automatic Musical Instrument Timbre Recognition" expressed Timber in terms of at least five major parameters like the range between tonal and noise-like character, spectral envelope, time envelope in terms of rise, duration and decay, the change both of spectral envelope (formant glide) or fundamental frequency (micro-intonation), The prefix, an onset of a sound quite dissimilar to the ensuing lasting vibration."

Next study explores several perceptually motivated features, namely harmonic, vibrato and timber features [20]. Here, harmonic content analysis to derive log energy of each band is done by passing each audio frame through harmonic filters for harmonic content. Thereafter, Octave Frequency Cepstral Coefficients (OFCChar) is computed using the discrete cosine transform. Then the harmonic filter is replaced with vibrato filters to get OFCC (vibrato) coefficients and similarly by harmonic and Mel-scale filters TimBrecepstral coefficients (TBCC) are obtained. With these three features, singer ID system performance using: a) vocal segments, b) non-vocal segments are studied and found that an average accuracy is 87.13% for vocal (so far Timbrel feature offers the highest accuracy 87.8%) and 38% for non-vocal.

In reference [21], the setup is elaborated to identify the gender of the singer in two steps: singer identification and then gender recognition. Here, the complex problem of "North Indian classical Music's singer identification" with the Timber feature had cracked. This validates the strength of this feature to catch the speaker specific perceptual and unknown characteristics and also motivates to use it for whispered speaker Identification. As far first part of our interest, researcher selected the timbral audio descriptors like Attack time and Attack slope, brightness, Mel-frequency cepstrum coefficients (MFCC), Zero-crossing rate (ZCR), roughness, roll -off and irregularity from MIR toolbox. The attack time and attack slope are eliminated from the experiment list by pre-processing audio files to have all the energy of the signal remains present throughout the recording. The K- means clustering is used as a classifier keeping in view, its simplicity and effectiveness for small data. Initially, individual audio descriptors are tested for efficiency in singer identification. The first four audio descriptors are considered giving highest identification rate. Further, the combinations of these audio descriptors are tested with each other remaining audio descriptors. An obvious advantage is to generate another record of best combinations of audio descriptors giving maximum singer identification accuracy. While selecting the best combinations of audio descriptors, efficiency from highest to lowest. This consideration probably will help as the sequence of an audio descriptor in the test vector should be highest to lowest efficiency. This process is repeated until we reach the maximum efficiency is attained [22].

From the above discussion, the most probable features which will be suitable for speaker identification in the whispered speech may be zero-crossing, brightness, roughness, roll-off, irregularity and MFCC.

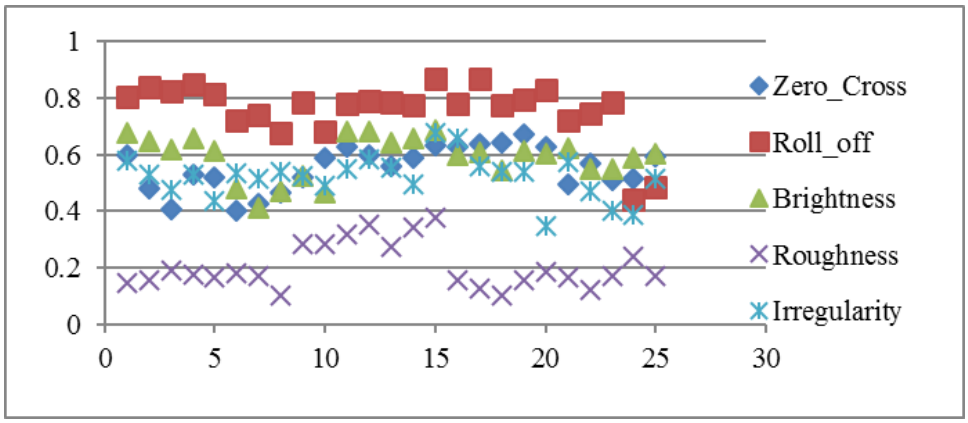

Figure-2. Selected audio descriptor values for five speakers, each five speech samples.

Here we are exploring the feature space for the probable features (shown in Figure-2) to be used for the experiment. For example, the feature values for 5 speakers, each having 5 samples are plotted. A limited number of speakers are selected and MFCC features are also not plotted to avoid a crowd of samples. All these features combined in the vector will be used for classification.

A database is created for speakers having multiple samples of the same speaker for both neutral and whispered mode. Speaker identification system is trained with neutral while whispered sample will be used for testing. The results will be compared with supervised as well as unsupervised classifier.

6. CONCLUSION AND PERFORMANCE EVALUATION

Speaker identification with a whispered speech by various SI system degrades performance compared to neutral speech due to basic differences in characteristics. Moreover, identification of speaker in whispered mode is mostly perceptual work as a human can do more

accurately. So a set of features among timbral features is recommended, which exhibits multidimensionality hence hidden or not well-defined parameters of whispered speech may be perceived. In the proposed approach, the independent feature performance is calculated. The best performing features are selected and then used as a vector with combinations of two, three or more till the maximum accuracy is achieved. The probable features suitable to our problem are zero-crossing, brightness, roughness, roll-off, irregularity [23]. So far, MIR toolbox is widely used for of musical instrument and singer identification. Rather than the pseudo- whisper efforts, it is recommended to include some realistic whispered samples in training process along with neutral database. Compared to GMM,HMM or any other classifier, KNN is robust to noisy training data hence will be more useful for a whispered case as the SNR is very much less.

The performance of the proposed system is to be measured by determining accuracy, precision, recall and selectivity. Precision is the proportion of matches found that are correct, while recall is the proportion of matches found for speakers that should match. Precision is more crucial requirement than recall as correct matches are more important than finding all matches. For sensitive applications like financial or secrecy, incorrect find is riskier than no identification.

REFERENCES

[1] Thesis by David Sierra Rodriguez. 8008. Text-Independent Speaker Identification. AGH University of Science and Technology, Krakow.

[2] Zhong-Xuan Yuan & Bo-LingXu & Chong-Zhi,Yu. 1999. Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification.in IEEE Transactions on Speech and Audio Processing. 7(1), IEEE, NewYork, NY, U.S.A.

[3] Roberto Togneriand Daniel Pullella. 1974. An Overview of Speaker Identification: Accuracy and Robustness Issues. J. Acoustic. Soc. Amer. 55: 1304.

[4] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman. 2004. Speaker Identification Using Mel Frequency Cepstral Coefficient. 3rdInternational Conference on Electrical & Computer Engineering ICECE 2004, pp. 28-30.

[5] Jeet Kumar, Om Prakash Prabhakar, Navneet Kumar Sahu. 2014. Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems: A review. International Journal of Innovative Research in Computer and Communication Engineering. 2(1).

[6] Florian Hong, Georgstemer. 2005. Revising Perceptual Linear Prediction. Interspeech.

[7] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, Shihab Shamma. 2011. Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition. IEEE Workshop on Automatic Speech Recognition & Understanding. pp. 559-564.

[8] R. M. Gray. 1984. Vector Quantization. IEEE ASSP Magazine. pp. 4-29.

[9] T. Hastie, R. Tibshirani and J. H. Friedman. 2009. Elements of Statistical Learning. (2nd edition), Springer series.

[10] Rose R. C. and Reynolds D. A. 1990. Text-independent speaker identification using automatic acoustic segmentation. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. pp. 293–296.

[11] WW. Chen, Q. Hong, X Li. 2012.GMM-UBM for Text-Dependent Speaker Recognition. IEEE International Conference on Audio, Language and Image Processing (ICALIP), Shanghai. pp. 432-435.

[12] Sylvain Meignier, Jean- Franqois Bonastre Evolutive. 2000. HMM For Multi-Speaker Tracking System.2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. 6: 3506-3509.

[13] Xing Fan and John H. L. Hansen, Fellow, IEEE. 2011. Speaker Identification within Whispered Speech Audio Streams. IEEE Transactions on Audio, Speech, and Language Processing. 19(5).

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

[15] Seiichi Nakagawa. 2012. Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition. IEEE Trans. on Audio, Speech, and Language Processing. 20(4).

[16] Xing Fan and John H.L. Hansen. 2009. Speaker Identification with Whispered Speech based on Modified LFCC Parameters and Feature Mapping.

ICASSP 2009, IEEE International conference on Acoustics, Speech and Signal Processing. p. 2009.

[17] Xing Fan and John H.L. Hansen. 2011. Speaker Identification for Whispered Speech Using a Training Feature Transformation from Neutral to Whisper. IEEE Transactions on Audio, Speech, and Language Processing. 19(5): 1408-1421.

[18] 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

[19] Saurabh H. Deshmukh, Dr. S. G. Bhirud. 2014. Analysis and application of audio features. Research gate article Feb.

[20] Swe Zin Kalayar Khine Tin Lay Nwe Haizhou Li. On Timbre Based Perceptual Feature For Singer Identification. International Symposium on Computer Music Modeling and Retrieval. pp. 159-171.

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

[22] Saurabh H. Deshmukh. 2014. 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.

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