International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 3, May-June 2015 **ISSN 2278-6856**

A Survey On: Continuous Voice Recognition Techniques

[1.]Swati Atame, [2.]Prof. Shanthi Therese S. , [3].Prof. Madhuri Gedam

[1.]Department of Computer Engineering, Mumbai University,
Shree L.R. Tiwari College of Engineering and Technology,

[2.]Department of Information Technology, Mumbai University,
Thadomal College of Engineering and Technology,

[3.]Department of Information Technology, Mumbai University,
Shree L.R. Tiwari College of Engineering and Technology

Abstract

There is a large amount of requirement for the digital data which is also called as digital multimedia which is used worldwide for almost all activities that carried out our day-to-day lives. To support these activities there are different formats of data for audio depending upon the requirement . It can be jpeg, mpeg, .wav etc. For efficient use and retrieval of audio signal different technologies are used by different users. One of the technologies used in these field is Automatic Singer identification which is used to recognize from features of the audio signal , the singer of the song or who is the one who is singing, the genuine singer[5].This same area of singer identification and recognition can be utilized in the bioinformatics where by the voice of the singer is used to gain access to a particular singer. The system would be very much useful to singers of the song who are actual singers and also to the common man who can store their speech and can later this input signal it to access the system. For the professional singers , there are many possibilities that the original singers voice is get mimicked which may in this situation lead to pirated copies of the voice of the singer which are then sold in to the market.

Keywords: Speech Recognition, Speaker Recognition , Singer Identification , Feature Extraction, MFCC, LPC, HMM,DTW, Pattern matching.

I. INTRODUCTION

Voice recognition and authentication consists of two parts: 1)Voice Identification 2)Voice Recognition/Verification .The voice is first captured using microphone with the support of some recording and editing software and the audio signal is stored with .wav extension in the database. Singer Identification and Recognition is identifying the voice of the singer from already existing database. The singer whose voice is recorded during the enrolment phase will only get identified which is also categorised as voice dependent singer identification. For the voice identification of the known singer, the features from the audio signal needs to be extracted. The continuous input audio signal which is first fed to the system gets converted into digital signal for processing purpose. Singers or speakers voice can be identified depending upon the characteristics stored in the database. The features vectors of a particular singer conveys lots of feature information about the said entity. This information can be proved to be very useful in the long run . As this information can be used to state whether the singer is male or a female , the pitch of the singer, ones mood characteristics which distinguish among individuals. Voice authentication is allowing access to the authorized and claimed entity. A technique based on the rhythm, so called as rhythm tracking method is used to analyze the continuous audio signal and segment them into beat space time frames. The are various techniques used inorder to perform feature extraction and modelling of classifiers.. Support Vector Machine(SVM) for vocal/instruments boundary detection and Gaussian Mixture Models(GMM) for modelling the singers voice[4].Speaker Identification or Singer Identification is considered as part of Speech recognition systems . Lot of work has been done in the area of Identification if speaker[1] ,Speaker recognition with isolated words[2], voice identification and recognition[3], Vocal and instrumental models are used in order to recognize and identify the Singer[4]

II. VOICE RECOGNITION SYSTEMS CATEGORIZATON

Voice recognition systems can be differentiated into various classes as per their categorization. Voice Identification and recognition is becoming more complex and a challenging task because of this variability and unstability and due to availability of noise in the input signal.

A.Voice Utterance Categories

A spoken word or words is termed as utterance that represents a single meaning to the computer system. Depending upon the requirement the system may vary. This utterances may vary depending upon how crucial the application is to the user. The Voice utterance categories are:

1) Isolated Words

A word spoken with its left and right side having empty or no signal as such is called as an isolated word. It requires isolated words to be accepted at a time , means it require a one word at a time. This works fine for scenarios where the user is asked to give only one word input or single word or commands, but it becomes very difficult for multiple word inputs or continuous voice

input. It is simple as well as it easy to implement one single utterance of a word because for a word its boundaries can easily be identified as it is pronounced clearly and hence the voice can be identified due to the fact the more emphasis is given on speaking a single word rather than speaking continuous sentence.

2)Connected words

Connected words are similar to isolated words, the only difference is that is allows separate utterances to be connected with less halts between them. We give less emphasis on each and every word in a sentence. The disadvantage of this type is choosing different boundaries affects the results. Eg. While saying just "Hello" ,we emphasize on specific word. But while saying "Hello how are u", we do not emphasize on each and every word in the sentence.

3)Continuous voice

Continuous voice system recognizer is the one which requires the user to speak in a very natural way. The continuous voice may or may not contain pauses. The computer determines the content that is contained in the input signal. Among all other systems, the most complex system to create and recognize is the Continuous speech recognition systems as it requires special techniques to determine utterance boundaries. The larger is the dataset more is the complexity and lesser becomes the accuracy and performance. Complexity also increases due to the presence of noise in the signal.

4)Spontaneous voice

Spontaneous voice is unplanned or unprepared voice in which rehearsal is not done and which is naturally said by the user. Spontaneous (unplanned/unprepared) voice may include mispronunciations, starting falsely etc.

B. Speaker Model Categories

Every individual who is the speaker for recognition systems is considered to have unique voice due to their distinguishing voice characteristics such as pitch, timbre, vibrato , glottal shape, vocal tract ,etc.

1)Speaker dependent models

Speaker dependent systems are basically designed for a particular speaker and depends on the speakers voice characteristics. Some of the features specified above can be used to distinguish and individually identify the particular speaker . The system is allowed to learn during the enrolment phase. During training phase specific user's voice is used to train the system. In the testing phase , pattern matching method searches for the best match between the test input and the voice that is already stored in the database. This model works very well with great accuracy for a particular singer or speaker whose voice is already stored in database , but much less accurate for rest of the singers.

2)Speaker independent models

Speaker independent systems are generally designed for live data or for variety of speakers. Enrolment phase is not required to take place priorly. Hence training is not possible since it contains huge and vast data. It recognizes the speech patterns of singer from huge amount datasets. It is very much difficult to generate such a system as it considers live data. But the advantage is, they are more flexible.

C. Vocabulary Categories :

i. Small vocabulary - Tens of words
ii. Medium vocabulary - Hundreds of words
iii. Large vocabulary - Thousands of words
iv. Very-large vocabulary - Tens of thousands of words
v. Out-of-Vocabulary- Mapping a word from the vocabulary with some new word .

III. WORKING PRINCIPLE OF SINGER IDENTIFICATION SYSTEM

Singer identification works in two phases

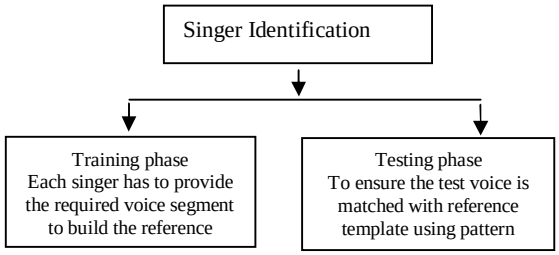

Figure 1. Singer identification system

B.TECHNIQUES FOR SINGER IDENTIFICATION

i. Voice Signal Analysis ii Extraction of Features iii. Model Training iv Testing/Pattern Matching

Once the audio signal are captured are converted into digital signal using some inbult or external hardware such as microphone, the voice signal is analysed and feature extraction is done by using different techniques such as Mel frequency cepsptral co-effeicient(MFCC) , Linear Prediction derived Cepstral Coefficients (LPCC). For proper classification of voice signal Support vector machine(SVM) can be used, and for modelling of singers voice can be done using pattern matching between two audio signal to identify the singer that is using Dynamic Time warp (DTW) or Hidden Markov Model(HMM).

A.Voice Signal Analysis

In voice analysis technique, input voice signal contains different types of information which can be helpful in identifying the speaker's identity. Various feature information listed which are specific to individual speaker. cochlea, glottal shape, critical bands, vocal tract, hearing limits in frequency, loudness , behaviour feature, pitch of the voice ,timbre. The physical shape of the glottal of a particular singer and vocal tract dimension as well as timbre, excitation source, vibrato are unique for each individual speaker which helps in distinguishing various individuals. The voice analysis deals dividing the continuous voice signal into frames and extracting the features from each frame which will be used for further processing[6]. The voice signal analysis is done with following three methods are given below.

1. Segment Analysis: In this method, voice signal is checked using the frame size and frame shift of the range

of 10-25 ms to extract features that represents specific speaker information. **2. Sub-segmental Analysis:** Voice is analysed using the frame size and frame shift in range 3-5 ms is termed as Sub segmental analysis. This analysis is performed to extract the feature from state of the excitation [7]. The transmitter of excitation which acts as a individual feature information is rapidly changing as compared to other information for eg. vocal tract information, so smaller frame sizes and shifts are required to capture particular singer information[8].

3. Supra-segmental Analysis: The extracted voice signal is analysed using frame size and frame shift of 100-300ms.

B.FEATURE EXTRACTION TECHNIQUES

Feature Extraction is very crucial part of any voice recognition and identification system. This techniques becomes very helpful when we want to distinguish one voice signal from the other voice signal or also to eliminate instruments noise from the song input. Since, there are many techniques that identifies the characteristics of individual speaker which further helps in speaker identification process.

1)Linear Predictive Coding (LPC)

One of the most powerful voice analysis techniques is the method of linear prediction. LPC [10] [11] of voice has become the most important method for estimating the features of voice signal. It provides both accuracy of the speech signal parameters and it is also an efficient computational model of speech signal. LPC is basically used to approximate the voice samples as a linear combination of past voice samples. By reducing the sum of squared differences for fixed interval between the actual speech samples and predicted values, a unique set of parameters or deterministic coefficient values can be estimated. These deterministic coefficients values are the basis for LPC of speech [12]. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients. The following figure shows the steps involved in LPC feature extraction.

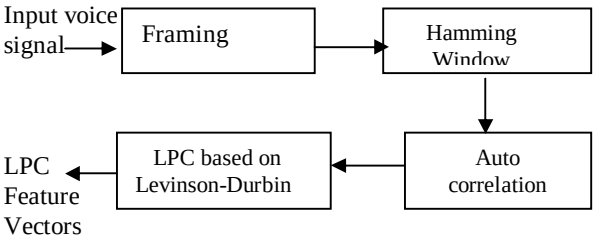

Figure 3. Steps involved in LPC Feature extraction

2)Frequency Cepstral Coefficients (MFCC)

The MFCC [10] [11] is the most evident example of a feature set that is extensively used in speech recognition. There are many researches done for MFCC feature extraction technique in which the placement of frequency bands are logarithmic positioned, In the diagram shown below the continuous audio signal with the frequency of 16000khz is given to the pre-emphasis block. The function of this block is to boost the high frequency components present in the audio signal which is then applied to next block that framing. Framing divides the continuous input signal into blocks of frames which contains N frames with adjacent frames separated by M(M<N). The frame size can of size 20-25 ms . The hamming window is applied on Framing to reduce the distortion at the beginning and ending of each frame. Discrete fourier transform is used to produce bank of Mel filters. The triangular filters are varied accordingly the total log energy in the critical band area which lies around the center frequency is included. The co-efficient numbers are obtained after performing warping.AS the last stage Inverse Discrete Fourier Transformer is used for the evaluation of cepstral coefficients values [8] [9]. It converts the time domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be evaluated by using the following formula

$$\textbf{Mel (f) = 2595*log10 (1+f/700)}$$

C MODELING TECHNIQUE

Modeling technique is to build a speaker models using speaker specific feature vector and train them . The singers identity is automatically identified on the basis of features extracted from the audio signal.

1. Dynamic Time Warping (DTW)

Dynamic time warping is an algorithm for calculating the similarity between two sequences which may change in time or speed. It is used to map 2 signals that may change in time and speed. It calculates distance between 2 audio signals which helps in matching the 2 signals that is between the training and testing samples . Linear form signal can be analyzed with DTW. DTW is a method that allows to find an best match between two given sequences with certain constrains. Dynamic programming is used to estimate the best optimal path. The warping of two audio signals is performed non linearly in the time domain which can be helpful in evaluating the amount of similarity even though there are non-linear changes in time domain. Some research has been done on isolated and connected words using DTW [13].

2.Statistical Based Approach

In this method, statistical model of speech variations is generated say for eg. Hidden Markov Model. There are many variations in speech which may be incomplete due to various reasons including multiple voice input, gummbling words, low voice etc. In this method random data is used which means non probable information is selected from the set of choices. The features that are applied as an input cannot be priory predicted[15]. The modelling assumptions in statistical models should be made before hand , which leads to inaccuracy , degrading the system's performance. In Voice recognition field , HMM is considered to have solution to problem as part of speech categorization [16]. The K-means algorithm is also used as a statistical tool and clustering group of siganls of voice based on the attribute of data. K means generates a number of separate clusters or groups having members that are close to each other. It requires same steps to be performed again and again , with no estimation of values, whereby no supervision is

required and takes cares of numerical values. The number of clusters is specified by K. Gaussian distributions is used to produce k means data to cluster the them in to vectors[17].

Table 2. Comparison of various Singer identification techniques considering their training data , features extraction technique and identification techniques

Author and Year	Title	Frame size and data sets	Feature Extraction Technique	Modeling Technique	Accuracy	Draw Back
Tong Zhang 2003	System and Method for Automatic Singer Identification [5]	15msecs And 45 songs	MFCC-characteristics of human voice LPC-harmonic components of audio signal.	GMM	80%	Accuracy drops with variation in GMM mixtures.
Namunu Chinthaka Maddage, Changsheng Xu, Ye Wang 2004	Singer Identification Based on Vocal and Instrumental Models[4]	Frame length based on inter-beat-interval and 100 songs	SVM with Octave scale Cepstral coefficient feature extraction	GMM (Gaussian mixture model)	87%	May not work for small data sets
Corneliu Octavian Dumitru, Inge Gavat 2006	A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language [10]	Speaker independent Continuous speech Considering large datasets.	PLP, MFCC, LPC	Hidden Markov Models (HMM)	MFCC-90,41%, LPC-63,55%. and PLP 75,78%	
Annamaria Mesaros, Tuomas Virtane, Anssi Klapuri 2011	Singer Identification in Polypnonic music using vocal separation and pattern recognition methods. [17]	Frame size of which is fixed 34 ms and 13 singers	MFCCs (Mel-frequency cepstral coefficients)	GMM-based maximum likelihood classifier. Kullback-Leibler deviation used in nearest-neighbour	70%	
Nadine Kroher, Emilia Gómez 2013	Automatic Singer Identification For Improvisational Styles Based On Vibrato, Timbre And Statistical Performance Descriptors [18]	Instead of frame wise extraction, the average for MFCC values for each song is calculated and Flamenco singing	Timbre feature extraction, Vibrato feature extraction	GMM(Gaussian mixture model)	Fl-Mono: 83.1% Fl-Poly: 86% Op-Poly: 76.5%	Less training done for model using timbre as a feature leads to less accuracy
R. Thangarajan, A.M. Natarajan and M. Selvam 2013	Phoneme Based Approach in Medium Vocabulary Continuous Speech Recognition in Tamil language [19]	25ms and Speaker independent Continuous Speech using medium datasets.	MFCCs (Mel-frequency cepstral coefficients)	Hidden Markov Model (HMM)	Good accuracy for known continuous voices	

VII. CONCLUSION

In this review, the basic techniques used in speaker identification recognition are discussed and its recent researches and work is highlighted . Some approaches available for developing an Singer Identification system are clearly explained with its merits and demerits , techniques of feature extraction , model training. The performance of the Singer Identification system based on the adapted feature extraction technique and the speaker identification and recognition approach for the particular individual is compared in this paper. In recent years, the there is huge requirement on singer identification and recognition on huge datasets. In researches it has been seen that HMM approach along with MFCC features is more suitable for these requirements and offers good recognition result. Where ever the above combination will be used to identification will help to generate large powerful systems.

REFERENCE

[1]. Abdul Syafiq Abdull Sukor, "Speaker identification system usinsg MFCC procedure and Noise reduction", University Tun Hussein Onn Malaysia, January 2012

[2]. Shivanker Dev Dhingra, Geeta Nijhawan , Poonam Pandit,"Isolated Speech recognition using MFCC and DTW", International Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 8, August 2013.

[3]. Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi,"Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques",

JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617.

[4]. Namunu Chinthaka Maddage1, 2, Changsheng Xu1, Ye Wang2,"Singer Identification Based on Vocal and Instrumental Models"1Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613 IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)1051-4651/04

[5]. Tong Zhang, "System and Method for Automatic Singer Identification", Imaging Systems Laboratory HP Laboratories Palo Alto,HPL-2003-8 January 15th , 2003.

[6]. GIN-DER WU AND YING LEI " A Register Array based Low power FFT Processor for speech recognition" Department of Electrical engineering national Chi Nan university Puli ,545 Taiwan

[7]. Nicolás Morales1, John H. L. Hansen2 and Doorstep T. Toledano1 "MFCC Compensation for improved recognition filtered and band limited speech" Center for Spoken Language Research, University of Colorado at Boulder, Boulder (CO), USA

[8]. B. Yegnanarayana, S.R.M. Prasanna, J. M. Zachariah, and C.S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed- text specker verification system," IEEE Trans. Speech Audio Process., vol. 13(4), pp. 575-82, July 2005.

[9]. N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010 1793-8201.

[10]. Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.

[11]. DOUGLAS O'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, VOL. 91, NO. 9, September 2003, 0018-9219/03$17.00 © 2003 IEEE.

[12]. A.P.Henry Charles & G.Devaraj, "Alaigal-A Tamil Speech Recognition", Tamil Internet 2004, Singapore

[13]. Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.

[14]. M.A.Anusuya, S.K.Katti "Speech Recognition by Machine: A Review" International journal of computer science and Information Security 2009.

[15]. Shigeru Katagiri et.al, A New hybrid algorithm for speech recognition based on HMM segmentation and learning Vector quantization, IEEE Transactions on Audio Speech and Language processing Vol.1, No.4

[16]. Alex weibel and Kai-Fu Lee, reading in Speech recognition,Morgan Kaufman Publisher,Inc.San Mateo,California,1990

[17]. Annamaria Mesaros, Tuomas Virtanen, Anssi Klapuri, " Singer Identification in polyphonic music using vocal separation and pattern recognition methods", 2011.

[18]. Nadine Kroher, Emilia Gómez," Automatic Singer Identification For Improvisational Styles Based On Vibrato, Timbre And Statistical Performance Descriptors"2013.

[19]. R. Thangarajan, A.M. Natarajan and M. Selvam, "Phoneme Based Approach in Medium Vocabulary Continuous Speech Recognition in Tamil language ", 2013.