Advances in Computational Research
ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, 2012, pp.-57-60.
Available online at http://www.bioinfo.in/contents.php?id=33

REVIEW OF DIFFERENT TECHNIQUES FOR SPEAKER RECOGNITION SYSTEM

BANSOD N.S.*, SEEMA KAWATHEKAR AND DABHADE S.B.

Dept. of C.S.& I.T. Dr.Babasaheb Ambedkar Marathwada University, Aurangabad, MS, India.
*Corresponding Author: Email- nagsenbansod@gmail.com

Received: February 21, 2012; Accepted: March 06, 2012

Abstract- Spoken language is the most natural way used by human to communicate information. Speech signal conveys linguistic information as well as speaker information (e.g. Emotional, regional and physiological characteristics). Such human ability has inspired many researches to understand production for developing the system that automatically process the richness of information in speech, this speech technology has many applications to find out "who is speaking" means speaker recognition system. This paper gives an overview of major techniques developed in each stage of speaker recognition. This paper has a list of techniques along with their results, merits and demerits. This also includes comparative study of different techniques, which are helping us to choose the technique for developing different language speaker recognition system like Marathi, Hindi and English. Time alignment of different utterances is a serious problem for distance measures and small shift would lead to incorrect identification. Dynamic time warping (DTW) is effective method. Vector quantization (VQ) is the classical quantization technique from signal processing. HMM is used for pattern matching. This paper shows the comparative result of PLC, PLPC and MFCC. MFCC have 73.62% is better result for Marathi language and 64.69% for Hindi language than PLC and PLPC techniques .
Keywords- MFCC, PLPC, PLC, HMM, GMM, SVM

Citation: Bansod N.S., Seema Kawathekar and Dabhade S.B. (2012) Review of Different Techniques for Speaker Recognition System. Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, pp.-57-60.

Introduction

Speaker communication medium means communication through speech. It is a most prominent & primary mode of communication therefore this is a popular choice for remote authentication due to the availability of devices for collecting speech samples (e.g. mobile phone, telephone, computer microphone and open space speech etc.) and its ease of integration. Speaker recognition is different from other methods, speech sample captured in a few second. Analysis occurs on a model in which changes over time are monitored, which is similar to other behavioral biometrics as a dynamic signature, keystroke recognition and emotion recognition.

Generally the speaker recognition can be divided into three specific tasks: identification, detection/ verification, and segmentation and clustering [1][5][3] speaker identification task is to determine which speaker out of a group of known speakers produces the input voice sample. There are two modes of operation closed-set mode means set of known voices and open-set mode mean unknown voices are referred to as impostors. The closed-set speaker identification can be considered as a multiple-class classification problem. Speaker verification, to determine whether a person (he or she) claims to be according to his/her voice sample. This task is also known as voice verification and speaker detection. Speaker segmentation and clustering techniques are used in multiple-speaker scenarios. In many speech recognition and speaker recognition applications, when the speech from the desired speaker is intermixed with other speakers, it is desired to segregate the speech into segments from the individuals before the recognition process commences. So the goal of this task is to divide the input audio into homogeneous segments and then label

them via speaker identity. Recently, this task has received more attention due to increased inclusion of multiple-speaker audio such as recorded news show or meetings in commonly used web searches and consumer electronic devices. Speaker segmentation and clustering is one way to index audio archives so that to make the retrieval easier [1][5].

General Structure of speaker recognition system

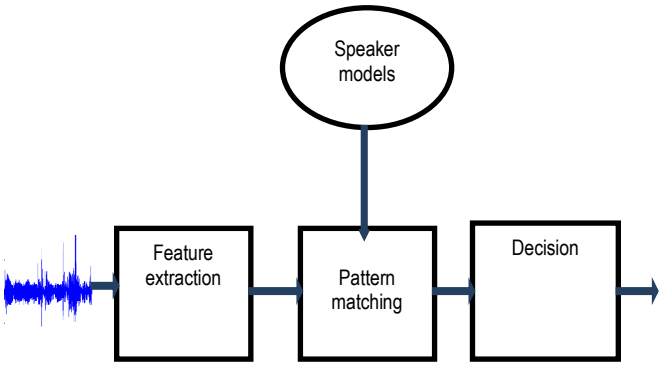

Fig.1- Classification

Feature extraction

Set of features of speaker speech production like semantic, phonologic, phonetic and acoustic, speaker-specific information. [6], [5]. The semantic level deals with transformation caused on the speech signal according to the communicative intent and dialog interaction of the speaker. For example, the vocabulary choice and the sentence formulation can be used to identify the socio-economic status and/or education background of the speaker [6]. The phonological level deals with the phonetic representation such as, duration and selection of phonemes, intonation of the sentence can be used to identify the native language and regional information. It is deals with the vibration of the vocal cords and the movements of articulators (lips, jaw, tongue, and velum) of the vocal tract [7]. For example, speaker can use a different set of articulator movements to produce the same phoneme [6]. The acoustic level deals with the spectral properties of the speech signal. For example, the dimensions of the vocal tract, or length and mass of vocal folds will define in some sense the fundamental and resonant frequencies, respectively [6] [10]

Pattern matching

The pattern matching is responsible for comparing the features to speaker models. There are various types of pattern matching methods. Some of the methods include Hidden Markov Models (HMM), Dynamic Time Warping (DTW), and Vector Quantization (VQ). In open-set applications (speaker verification and open-set speaker identification), the estimated features can also be compared to a model that represents the unknown speakers [3].

Decision

In verification module outputs a similarity score between the test sample and the claimed identity. In identification task, it outputs similarity scores for all stored voice models. The decision module analyzes the similarity score(s) (statistical or deterministic) to make a decision. The decision process depends on the system task [1][5].

Methods

MFCC

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency *t* measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' .The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as1000 Mels.

The extraction and selection of the parametric representation of acoustic signals is an important task in the design of any speech recognition system; it significantly affected the recognition performance. A compact representation would be provided by a set of mel-frequency cepsrtum coefficients (MFCC), which are the result of a cosine transform of the real logarithm of the short-term MFCCs are provide more efficient. It includes Mel-frequency wrapping and Cepstrum calculation.

A. Mel-frequency wrapping

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The Mel frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz .As a reference point, the pitch of a 1 KHz tone ,40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute them else for a given frequency *f* in Hz.

Mel (*f*) = 2595*log10 (1 + *f* / 700)

Ours approach is to simulate the subjective spectrum to use a filter bank, one filter for each desired Mel- frequency component. That filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant Mel-frequency interval. The Mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a Mel frequency scale.

B. Cepstrum

In this step, we convert the log Mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). In this final step log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).The discrete cosine transform is done for transforming the mel coefficients back to time domain.

$$Cn = \sum_{k-1}^{k} (\log\ S_k) \cos^{\,n\,k-\frac{1}{2}\bullet\frac{-}{k}}, \qquad n=1,2,.....k,$$

Where S_k k = 1,2,.....k are the outputs of last step. Complete process for the calculation of MFCC [12].

Comparison of different implementation of MFCC

The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by the number of filters. In this paper, several comparison experiments are done to find a best implementation. A. Effect of number of filters Results of the speaker recognition performance by varying the number of filters of MFCC to 12, 22, 32, and 42 are given for Marathi and Hindi language. The recognizer reaches the maximal performance at the filter number K = 32. Few or many filters and distance do not result in better accuracy. Hereafter, if not specifically stated, the number of filters is chosen to be K = 32.

MFCC with 12 filters

Table 1-

Speaker	No. Of Attempt	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	1
S3	4	0	2
S4	4	0	0
S5	4	0	2
total	20	0	5

Threshold value of distance = 130, Efficiency = 75

MFCC with 22 filters

Table 2-

Speaker	No. of Attempts	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	2
S3	4	0	2
S4	4	0	0
S5	4	0	3
total	20	0	7

Threshold value of distance = 150, Efficiency = 65

MFCC with 32 filters

Table 3-

Speaker	No. Of Attempt	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	0
S3	4	0	1
S4	4	0	0
S5	4	0	2
total	20	0	3

Threshold value of distance = 150, Efficiency = 85

MFCC with 42 filters

Table 4-

Speaker	No. Of Attempt	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	0
S3	4	0	2
S4	4	0	1
S5	4	0	1
total	20	0	4

Threshold value of distance = 85, Efficiency = 80%

The features extracted from the speech signal spectrum have shown to provide better performances in the speaker recognition system, specially the LPC-Cepstral, because these have proved to increase the robustness of the speaker recognition systems reducing the problem of speech signal distortion introduced by the communication channel [7]. The computation of the LPC-Cepstral is relatively simple since they can be obtained using a simple recursion after the Linear Prediction Coefficients (LPC) was estimated as follows [7]:

$$Cn = -a_n + \frac{1}{n}\sum_{i=k}^{n-1}(n-i)a_i c_{n-i} \qquad n>0 \quad (1)$$

Where Cn is the nth LPC-Cepstral coefficient, a_i are the linear prediction coefficients which are obtained by Levinson Durbin algorithm. In this application, 16 LPC-Cepstral coefficients were extracted in each frame.

This process is repeated until convergence is achieved. Here the Mixes order and the model parameters previous to the Maximization of the likelihood of GMM can be different depending on the application [3].

Mono-lingual experiments

Database of 60 speakers in each of two Indian languages, viz., Marathi and Hindi is considered for monolingual speaker identification experiments [12]. The population size is kept constant to make relative comparison of the performance of ASI for different languages. Training and testing was done with the same microphones for all the languages except for Oriya. Table-5 shows the results on monolingual speaker identification experiments for Marathi over different testing and training speech durations for LPC whereas Table-6 shows average success rates for monolingual speaker identification experiments in Marathi (M) and Hindi (H) for LPC. Finally, Table-7 shows the overall average success rates for Marathi, Hindi, Urdu and Oriya with LPC, LPCC and MFCC. Some of the observations from the results are as follows:

Table 5- (ASR for Marathi) Success rates (%) for LPC

TE	TR			
	30 s	60 s	90 s	120s
1 s	51.66	51.66	56.66	55
3 s	55	58.33	56.66	66.66
5 s	56.66	61.66	61.66	65
7 s	55	58.33	61.66	66.66
10 s	56.66	61.66	66.66	68.33
12 s	61.66	61.66	66.66	73.33
15 s	58.33	70	70	80
Av	56.42	60.47	62.85	67.85

TR = Training speech duration, TE = Testing speech duration

Cross-lingual experiments

In this section, ASR experiment is conducted for 17 pairs of identical twins (i.e., 34 speakers) in cross-lingual mode

Table 6- Average success rates (%) for LPC

L	TR			
	30 s	60 s	90 s	120 s
M	56.42	60.47	62.85	67.85
H	44.75	46.42	47.37	56.18

Table 7- Overall average success rates (%)

L	FS		
	LPC	LPCC	MFCC
M	61.89	66.72	73.62
H	48.68	50.17	64.69

FS = Feature sets, viz., LPC, LPCC, MFCC and L = language, M= Marathi, H =Hindi.

Segmentation and Clustering

The speaker segmentation and clustering system is to divide a speech signal into a sequence of speaker-homogeneous part. Segmentation and clustering may also be useful in information retrieval and as part of the indexing information of audio archives. Speaker segmentation is also called" speaker change detection" in the literature. Although speaker change detection also belongs to the family of pattern classification problems, and thus has a feature extraction module followed by classification/segmentation framework, no significant work has been reported on the feature extraction module. Various segmentation algorithms have been proposed in the literature, which can be categorized as follows:

Decoder-guided segmentation

The input stream is first decoded; then the desired segments are produced by cutting the input at the silence locations generated from the decoder ([13] [14]). Other information from the decoder, such as gender information could also be utilized in the segmentation.

Model-based segmentation

This involves making different models e.g. GMMs, for a fixed set of acoustic classes, such as telephone speech, pure music, etc. from a training corpus [7].

Metric-based segmentation

A distance-like metric is calculated between two neighboring.

Conclusion

In this paper various feature extraction techniques are used for speaker recognition. MFCC is well known techniques used in speaker recognition to describe the signals. MFCC results for Marathi language 73.62% and Hindi language 64.69%. LPCC results 66.72% and Hindi 50.17%. LPC results for Marathi 61.89 and Hindi 48.68%.

The performance of Linear Prediction Coefficients (LPC) better for 120s training speech duration with testing speech 15s, success rate is 80%.In feature I will find more accurate duration for speaker recognition.

The performance of the Mel-Frequency Cepstrum Coefficients (MFCC),affected by the filter and distance .In this paper various comparison experiment are done with various filter like 12, 22, 32 and 42 for Marathi and Hindi language. The accuracy is in increasing order up to 42 filter and threshold value of distance =85, efficiency=80% except 22 filter threshold value of distance =150, efficiency=65%.we can't increase quantity of filter to accuracy distance also affected.

Future Work

With the help of same algorithm and by using different number of filters and distance measures, increase the efficiency and accuracy in different languages.

References

[1] Furui S. (1997) *Pattern Recognition Letters*, 18, 859-872.

[2] Campbell J. P. (1997) *IEEE*, 85, 1437-1462.

[3] Kichul Kim and Moo Young Kim (2010) *IEEE Transactions on Consumer Electronics*, 56(3).

[4] B.S. (1976) *IEEE*, 64, 460-475.

[5] Nolan F. (1983) *Cambridge University Press.*

[6] Rabiner L.R. and Juang B.H. (1993) *Fundamentals of Speech Recognition*.

[7] Eric Simancas-Acevedo, Mariko Nakano- Miyatake, Hector Perez-Meana (2006) *Científica,* 10(3), 151-156, 1665-0654.

[8] Eric Simancas Acevedo, Akira Kurematsu, Mariko Nakano Miyatake and Perez Meana H. (2001) *Bio-Inspired Applications of Connectionism*, 287-294.

[9] Wolf J.J. (1972) *Journal of The American Statistical Association*, 51, 2044-2056.

[10] Vibha Tiwari (2010) *International journal on Emerging technology,* 1(1), 19-22.

[11] Hemant A. Patil, Basu T.K. (2009) Science+Business Media, LLC.

[12] Kubala F., Jin H., Matsoukas S., Nguyen L., Schwartz R. and Makhoul J. (1997) *DARPA Broadcast News Transcription and Understanding Workshop,* 90-93.

[13] Woodland P., Gales M., Pye D. and Young S. (1997) *DARPA Broadcast News Transcription and Understanding Workshop*.