Automatic Speaker Recognition System

Parul[1], R. B. Dubey[2]
ECE, Hindu College of Engineering, Sonepat, India[1,2]

Abstract

Spoken language is used by human to convey many types of information. Primarily, speech convey message via words. Owing to advanced speech technologies, people's interactions with remote machines, such as phone banking, internet browsing, and secured information retrieval by voice, is becoming popular today. Speaker verification and speaker identification are important for authentication and verification in security purpose. Speaker identification methods can be divided into text independent and text-dependent. Speaker recognition is the process of automatically recognizing speaker voice on the basis of individual information included in the input speech waves. It consists of comparing a speech signal from an unknown speaker to a set of stored data of known speakers. This process recognizes who has spoken by matching input signal with pre- stored samples. The work is focussed to improve the performance of the speaker verification under noisy conditions.

Keywords

Automatic speaker recognition system, speaker identification, speaker verification, MFCC, HMM, GMM, VQ

1. Introduction

Speech signals contain both language and speaker dependent information. The vocal tract characteristics of a speaker provide the main speaker-dependent information, which can be used to decide the speaker. Speech recognition aims at recognizing the word spoken in speech. The earliest methods of biometric identification included fingerprint and handwriting while more recent ones include iris/eye scan, face scan, voice print, and handprint. Biometric voice recognition and identification technology focuses on training the system to recognize an individual's unique voice characteristics. Speaker recognition is a process that enables machines to understand and interpret the human speech by making use of certain algorithms and verifies the authenticity of a speaker with the help of a database [1-4]. Speaker recognition can be divided into two specific tasks: identification, detection/ verification [5- 6]. If a speaker claims to be of a certain identity and their speech is used to verify this claim. This is called verification or authentication. There are two modes of operation for speaker identification namely, closed set and open set. 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. It can also be divided into three categories based on the type of spoken utterances: (I) text-independent, (2) text-dependent, and (3) text-prompted. In text-independent systems, the user can utter an arbitrary phrase, whereas in text-dependent systems, a fixed voice password is uttered, and in text-prompted systems, the user is asked to repeat a phrase [13, 14]. Speaker recognition technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, voice mail, telephone banking, telephone shopping, e-commerce transactions, internet-banking biometric security, voice operated vehicle/home door access by speaker recognition biometric security, forensic speaker identification/verification system, for corporate biometric time and attendance system, database access services, information services, security control for confidential information areas, and remote access to computers. The various algorithms used to process and store voice prints include frequency estimation, hidden Markov models, Gaussian mixture models, pattern matching algorithms, neural networks, matrix representation, vector quantization[16-17] and decision trees. The most important parts of a speaker recognition system are the feature extraction and the classification method. Mel Frequency Cepstral Coefficients (MFCC) [8, 12] algorithm is feature-extraction type speaker recognition method. Based on classification method, the speaker recognition algorithms are two types: text-dependant (constrained mode) and text-independent (unconstrained mode) speaker recognition systems. For text independent recognition, speaker specific vector quantization codebooks [18] or the more advanced Gaussian mixture models (GMM) [10-11] algorithm are used most often. For text dependent recognition, dynamic time warping (DTW) algorithm or hidden Markov models (HMM) algorithms are appropriate [9].

2. Proposed Methodology

In the verification phase, a speech sample or utterance is compared against a previously created voice print as shown in Fig. 1. For identification systems, the utterance is compared against multiple voice prints in order to determine the best matches while verification systems compare an utterance against a single voice print. Because of the process involved, verification is faster than identification.

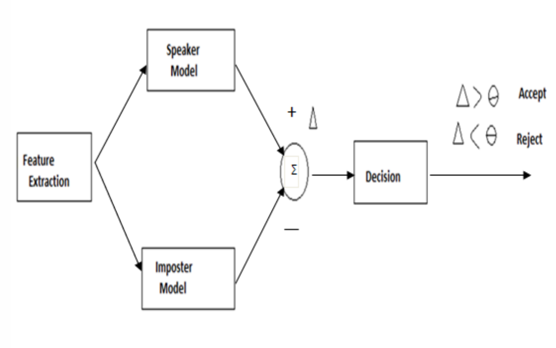

Figure 1: Speaker verification

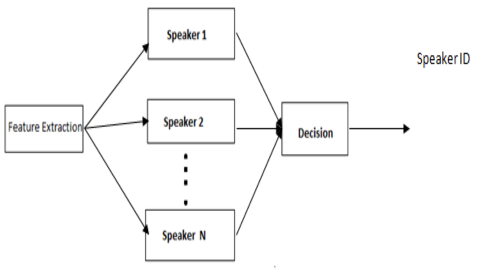

Figure 2: Speaker Identification

Speaker recognition system has two phases: Enrolment and verification. During enrolment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print, template, or model. Testing is the actual recognition task. In this phase, the input speech is matched with stored reference models and a recognition decision is made.

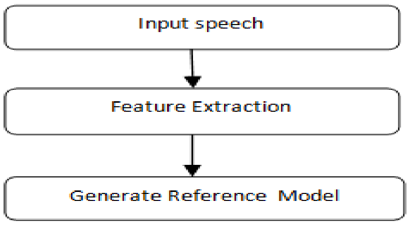

Fig. 3: Enrolment phase

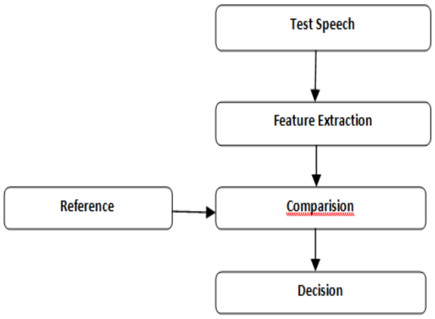

Fig. 4: Testing phase

2.1 Data collection

To implement this system, test data is required. The database was collected in a quiet office environment with little noise. The speech data was collected in one single session. Used data has been recorded in clips from 7 speakers. Each speaker data set consists of 5 speech clips of varying lengths. This data set is split into three categories – training, tuning and testing. To record the clips a microphone is used. One factor that affected the results was the distance between the microphone and the speaker's mouth. Too close or too far and the results were skewed.

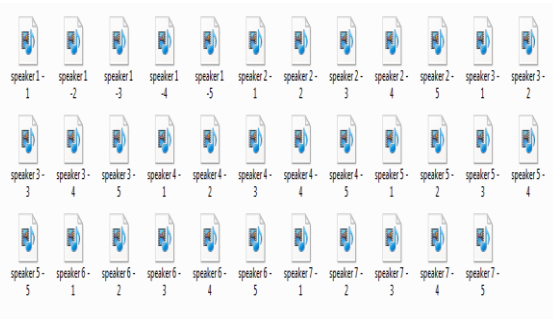

Figure 5: Recorded clips for database

2.2 Training phase

In the training phase, the background model is created. The background model is basically a large pool of all sample data. The wave files were converted into a different format so that they can be used for this analysis. The wave file is a continuous signal, which must be broken down in discrete parameter vectors. Each vector is about 10ms long; because we assume that in this duration the vector is stationary. This is not strictly true, but it is a reasonable approximation to make. Here we have used the MFCC format.

The conversion can be done as follows:
a) Divide signal into frames.
b) For each frame, obtain the amplitude spectrum.
c) Take the logarithm.
d) Convert to Mel (a perceptually-based) spectrum.
e) Take the discrete cosine transforms (DCT).

2.3 Testing phase

Once we have all the threshold values and speaker models, it is time to test the remaining files. This will help us determine if the analysis done above is accurate enough. Using the threshold values calculated in the tuning phase, remaining speaker files were tested. To ensure that the system is accurate while verifying users, we need to test the threshold values in two ways – for false alarms and false rejections. If the likelihood value of an imposter file is higher than the threshold for the speaker being tested, then the system will validate the imposter as the speaker. This is a false alarm. On the other hand, sometimes a speaker's own file may not have a likelihood value higher than the threshold and so the speaker is falsely identified as an imposter. This is a false rejection. An optimal threshold value would minimize both these values, keeping the error rate low.

2.4 Speech feature extraction

The purpose of this module is to convert the speech waveform to some type of parametric representation for further analysis and processing. This is often referred as the signal-processing front end. The speech signal is a slowly timed varying signal (it is called quasi-stationary). When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. MFCC's are based on the known variation of the human ear's critical logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

Mel-Frequency Cepstrum Coefficients (MFCC) Processing

A block diagram of the structure of an MFCC processor is given in Fig. 5. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behaviour of the human ears. In addition, rather than the speech waveforms themselves, MFFCs are shown to be less susceptible to mentioned variations.

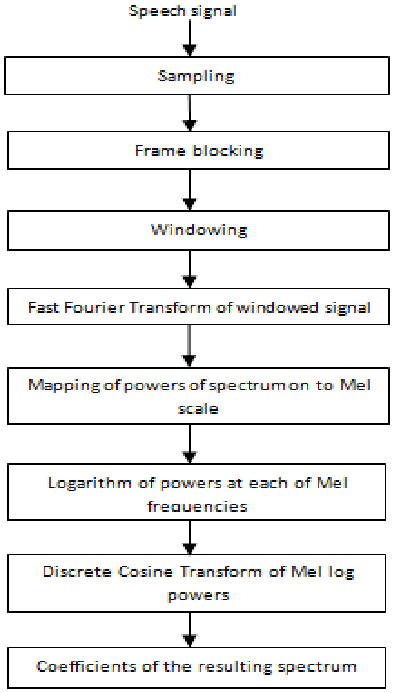
Speech signal
Sampling
Frame blocking
Windowing
Fast Fourier Transform of windowed signal
Mapping of powers of spectrum on to Mel scale
Logarithm of powers at each of Mel freauencies
Discrete Cosine Transform of Mel log powers
Coefficients of the resulting spectrum

Figure 6: MFCC processor

Frame blocking

In frame blocking, the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples. Similarly, the third frame begins 2M samples after the first frame (or M samples after the second frame) and overlaps it by N - 2M samples. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 100.

Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. In other words, when we perform Fourier Transform, it assumes that the signal repeats, and the end of one frame does not connect smoothly with the beginning of the next one. This introduces some glitches at regular intervals. So we have to make the ends of each frame smooth enough to connect with each other. This is possible by a processing called windowing. In this process, we multiply the given signal (frame in this case) by a so called window Function. Hamming window is used here, and is given by:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

Fast Fourier transform (FFT)

The next processing step is the fast Fourier transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the discrete Fourier transform (DFT) which is defined on the set of N samples {x , n**}**, as follow:

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N}, \qquad n = 0,1,2,\ldots,N-1$$

Mel-frequency wrapping

As mentioned above, psychophysical studies have shown that 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, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz:

$$mel\,(f) = 2595 * log_{10}(1 + \frac{f}{700})$$

Cepstrum

In this final 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).

2.5 Feature Matching

The problem of speaker recognition belongs to pattern recognition. The goal of pattern recognition is to classify objects of interest into one of categories or classes. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching. Furthermore, if there exist some set of patterns that the individual classes of which are already known, then one has a problem in supervised pattern recognition. This is exactly our case since during the training session, we label each input speech with the ID of the speaker. These patterns comprise the training set and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the test set. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm. VQ approach is used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called cluster and can be represented by its centre called a codeword. The collection of all code-words is called a codebook. Fig. 6 shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result code-words (centroids) are shown in Fig. 6 by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is vector-quantized using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

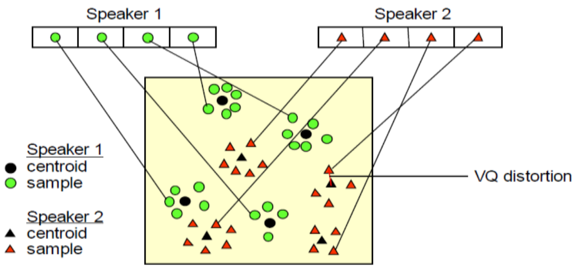

Figure 7: Conceptual diagram illustrating vector quantization codebook formation. One speaker can be discriminated from another based of the location of centroids.

3. Results and discussions

Experiments were performed on 7 different speakers. The speech data was collected in one single session. Total 7 clips have been recorded from 7 different speakers. Each speaker data set consists of 5 speech clips, of varying lengths. This data set splits into three categories – training, tuning and testing. These are the three phases of the project and the data will be required in each stage. A microphone was used for recording the clips. We apply test speech wave file to our speaker recognition system to find out who the speaker is. We run the program twice in order to get a more accurate result.

Recognition accuracy
We record the word "hello" of user1 for our test voice and store it in the database.
Test case1:
User1 speaks "hello"
Recognition accuracy: 80%
Rejection accuracy
Test case2: any other user, say user 2, speakers "hello"
Rejection accuracy: 85%
Recognition Speed
FFT calculation: 100 ms
MFCC feature extraction: 4~5s
DTW algorithm: 1~2s
Total recognition time: 5~7s

Table 1: Performance

Speaker ID	**Total Samples**	**Recognition accuracy (%)**	**Rejection accuracy (%)**
1	5	80%	85%
2	5	79%	80%
3	5	81%	84%
4	5	83%	81%
5	5	86%	80%
6	5	85%	83%
7	5	87%	82%

4. Conclusions

The goal of this work is to create a speaker recognition system and apply it to a speech of an unknown speaker. The feature extraction is done by using MFCC (Mel Frequency Cepstral Coefficients) and the speaker is modelled using vector quantization (VQ). A VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database. Although much care is taken to make an efficient speaker recognition system however, this task has been challenged by the highly variant input speech signals. The principle source of this variance is the speaker himself. Speech signals in training and testing sessions can be greatly different due to many facts such as people voice change with time, health conditions, speaking rates, the varied microphones and channels that people use can cause difficulties. Because of all these difficulties this technology is still an active area of research. Speaker verification is easy to use, has low computation requirements. The proposed approach could be used with conjunction with other ones, like face recognition, for better security and increasing the area of the particular application.

References

[1] R. Teunen, B. Shahshahani, and L. Heck, "A Model-based Transformational Approach to Robust Speaker Recognition'' ICSLP October 2000.

[2] S. Furui, "Recent advances in speaker recognition" AVBPA97, pp 237--251, 1997.

[3] J. P. Campbell, ``Speaker recognition: A tutorial," Proceedings of the IEEE, vol. 85, pp. 1437-1462, September 1997.

[4] D. A. Reynolds, "Special Issue on Speaker Recognition, Digital Signal Processing" vol. 10, January 2000.

[5] A. Higgins, L. Bahler and J. Porter, "Speaker Verification using Randomized Phrase Prompting'' Digital Signal Processing, vol. 1, pp. 89-106,1991.

[6] M. Carey, E. Parris, and J. Bridle, ``A Speaker Verification System using Alphanets," ICASSP, pp. 397-400, May 1991.

[7] A. D. Andrews, M. A. Kohler and J. P. Campbell, "Phonetic Speaker Recognition," Euro speech, 2001.

[8] Kumar, Ch Srinivasa, and P. Mallikarjuna Rao. "Design of an automatic speaker recognition system using MFCC, Vector quantization and LBG algorithm." Ch. Srinivasa Kumar et al./International Journal on Computer Science and Engineering (IJCSE) 3, no. 8 (2011).

[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications In Speech Recognition", Proceedings of the IEEE, 1989.

[10] Zhou, Yuhuan, Jinming Wang, and Xiongwei Zhang. "Research on Adaptive Speaker Identification Based on GMM." In Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on, vol. 2, pp. 330-332. IEEE, 2009.

[11] M. F. R. Chowdhury, S. A. Selouani, D. O'Shaughnessy, "Text Independent Distributed Speaker Identification And Verification Using Gmm-Ubm Speaker Models For Mobile Communications" 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010).

[12] J. Ramón, C. de Lara, "A Method of Automatic Speaker Recognition Using Cepstral Features and Vectorial Quantization" 10th Iberoamerican Conference on Pattern Recognition, CIARP 2005 Proceedings.

[13] Laxman, Srivatsan, and P. S. Sastry. "Text-dependent speaker recognition using speaker specific compensation." In TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region, vol. 1, pp. 384-387. IEEE, 2003.

[14] S.C. Yin, R. Rose(Senior Member, IEEE), and P. Kenny, "A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification" IEEE Transactions on audio, speech, and language processing, vol. 15, no. 7, september 2007 1999.

[15] Jin, Qin, Arthur R. Toth, Alan W. Black, and Tanja Schultz. "Is voice transformation a threat to speaker identification?." In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pp. 4845-4848. IEEE, 2008.

[16] Goto, Yuki, Tatsuya Akatsu, Masaharu Katoh, Tetsuo Kosaka, and Masaki Kohda. "An investigation on speaker vector-based speaker identification under noisy conditions." In Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on, pp. 1430-1435. IEEE, 2008.

[17] Tadokoro, Naoki, Tetsuo Kosaka, Masaharu Kato, and Masaki Kohda. "Improvement of Speaker Vector-Based Speaker Verification." In Information Assurance and Security, 2009. IAS'09. Fifth International Conference on, vol. 1, pp. 721-724. IEEE, 2009.

Parul was born on 1st January 1991. She received her B. Tech. degree in Electronics and Communication Engineering from M. D. U., Rohtak in 2011 and pursuing M. Tech. Degree in Electronics and Communication Engineering from D.C.R.U.S.T Murthal, Sonepat, India.

Rash Bihari Dubey was born in India on 10th November 1961. He received the M. Sc. degree in Physics with specialization in Electronics in 1984 from Agra University Agra, India, the M. Tech. degree in Instrumentation from R.E.C. Kurukshetra, India in 1989 and the Ph.D. degree in Electronics Enggg., from M. D. University, Rohtak, India in 2011. He is at present Professor and Head in the Department of Electronics and Communication Engineering at Hindu College of Engineering, Sonepat, India. He has well over 40 publications in both conferences and journals to his credit His research interest are in the areas of Medical Imaging, Digital Signal Processing, Digital Image Pprocessing, Biomedical Signal Analysis, and Industrial Real Time Applications.