Automatic Speaker Recognition

Debadatta Pati
Research Scholar
Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati

Abstract:

This report gives an overview of automatic speaker recognition. It discuses, terminologies, motivation and speaker information for automatic speaker recognition. It will then explain the various blocks present in the automatic speaker recognition system. And finally discusses some of the practical speaker recognition systems developed in the literature.

1. Introduction:

Speech is a natural way of communication between human beings. It conveys variety of information to listener. It includes Message, Characteristics of the speaker, language, accent, emotional condition of the speaker and physiological condition of the speaker.

Speech recognition is related to recognizing message present in the speech signal. So this process is related to the message information present in the speech signal. On the other hand speaker recognition is related to recognizing the speaker. So this process is related to the information present in the characteristics of the speaker.

Human recognition is a common experience. Human recognize people by hearing the speech signal. From this observation, people thought of training a machine for speaker recognition. This is the motivation behind ASR system. Humans recognize the people by perceiving the speech, where machine process the speech signal to recognize the people. The objective is to extract speaker specific information present in the speech signal. Speaker Specific information in the speech signal ,present in speaking rate, the quality of the speech signal ,pitch, Sound production organs such as vocal tract and excitation source. Speaker specific information should qualify certain characteristics to be a feature for recognition process .The characteristics of a speaker specific feature that can be used for speaker recognition are universality, permanence, distinctiveness and circumvention[].

We have some knowledge where speaker information is present and we know what should be the characteristics of a speaker specific feature but, how human is using these characteristics for recognizing the people is less understood. Since we do not have the knowledge, people are trying to exploit the characteristics of speaker with available tools from signal processing and pattern recognition .In this way automatic speaker recognition system(ASR) was developed in speech processing area. Depending upon the application ASR system is classified as automatic speaker verification (ASV) and automatic speaker identification (ASI) system. In verification, the machine validates the identity of a speaker. In identification, the machine searches the identity of the speaker. Identification process also classified as, open set and close set. In case of close set, the system is enrolled for N user and only those N users are allowed to use the system. If there is a chance of an outsider to use the system then, then it is an open set.

2. Description of ASR system

The input to the to the ASR system is the sequence of samples of the speech signal collected from either from microphone or from telephone line, sampled at 8000-16000 samples per second through A/D converter.

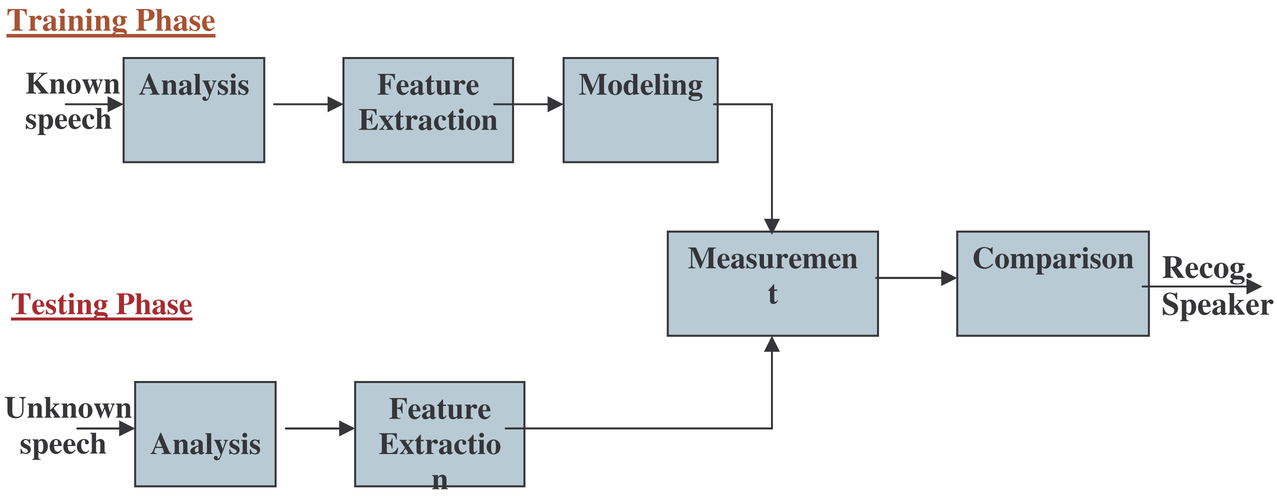


Fig-1: Block diagrams of ASR system

The general block diagrams of ASR system consist of two phase of operation, the training phase and testing phase. The training phase consists of analysis stage, feature extraction stage and modeling stage. The testing stage consists of analysis stage, feature extraction stage and measurement and comparison stage. The analysis and feature extraction stage in both the case are same and they differ in the last stage. So basically the block diagrams of ASR system consist of five blocks. Analysis stage, feature extraction stage, modeling stage, and measurement and comparison stage.

2.1 Analysis Stage

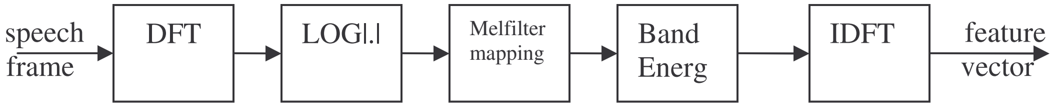
Speech is a non stationary signal. The usual concept to process the non stationary signal is the short term analysis. The function of this analysis block is to convert the speech signal into segments called as the analysis frames. Segmental analysis (10-30ms) for vocal tract characteristics, suprasegmental analysis (100-300ms) for high level features such as word duration and subsegmental analysis (1-5ms) for excitation source information is done. Overlapping frames also used to avoid end effects.

2.2 Feature Extraction Stage

The output of the analysis stage is series of frames with large number of samples. These large numbers of samples also contain some redundant information. The function of this feature extraction stage is to extract maximum amount of information from those samples with reduced data rate and convert it into vector form called as feature vector. These vectors are put in a k-dimensional space called as feature space, where k is the dimension of the feature vector. There are two standard feature extraction methods used in ASR system i.e Mel Frequency Cepstral Coefficients(MFCC) and Linear Prediction Cepstral Coefficients(LPCC).

Mel frequency cepstral Coefficient (MFCC)

- Based on perceptual frequency of the sound, mel scale.
- Log magnitude spectrum is mapped linearly on mel scale through overlapping band of filters .
- Energy in the bands are the cepstral coefficients in frequency domain
- Convert k-dimensional vector to n-dimensional vectors



$$Y(i,m) = \sum_{j=1}^{N_i} \log|S(k,m)| \times h(j,i) \quad \text{for } i = 1,2\ldots N$$

$$c_S(n,m) = F^{-1}\{\tilde{Y}(k,m)\}, \; m = \text{no. of frames}$$

$$\text{where, } \tilde{Y}(k,m) = Y(i,m), \; k = j_i, \text{ index of cen. freq. of } i^{th} \text{ filter}$$

$$= 0, \qquad k \neq j_i$$

$$n = \text{no of coefficients}$$

$$N_i = \text{no of samples in } i^{th} \text{ filter}$$

$$N = \text{no.of filters}$$

Linear Prediction Cepstral Coefficients (LPCC)

The LPCC computation is as follows:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k), \text{where } a_k \text{are the predictor coefficients}$$

The value of a_k are defined as

$$\sum_{k=1}^{p} a_k r_n(i-k) = r_n(i), p \text{ is the order of prediction,and}$$

$$r_n(i) = \sum_{m=0}^{N_w-1-i} s(m)s(m+i), 1 \leq i \leq p, 1 \leq k \leq p$$

$$c_1 = a_1,$$

$$c_n = \sum_{k=1}^{n-1} (1 - k/n) a_k c_{n-k} + a_n \;, \; 1 < n \leq p$$

$$c_n = \sum_{k=1}^{n-1} (1 - k/n) a_k c_{n-k} \;, \; n > p$$

2.3 Modeling

In the feature space these vectors are shared and overlapped each other. Ideally it is required that, these vectors should be distinct for speaker to speaker for better recognition. But that does not happen in practice. In the modeling stage a second level of compression is done, where a set of nearest feature vectors are clumped together and assigned with representative vector. A model is formed for a speaker , so that at the time of measurement, instead of comparing with all vectors, the comparison is done with the speaker model. One standard form of compression technique is vector quantization

Vector Quantization

- It divides large set of vectors in to clusters
- Each cluster is represented by a representative vector called as code vector.
- Set of all code vectors is called as code book.
- Codebook is created based on Binary split and K- means algorithm.

Other modeling technique used in ASR system are

Probabilistic Modeling Technique

- Gaussian Mixture Model
- Hidden Markov model
- Neural Network based
- Genetic algorithm based

2.4 Measurement and Comparision

In the measurement stage, the similarity between the incoming vector and the corresponding model is measured and decision is taken according to the requirement. The measurement criterion depends upon the modeling technique.

(A)Measurement

Euclidean distance

$$d(x,y) = [(x-y)^t (x-y)]^{\frac{1}{2}}$$

Mahalanobis distance

$$d(x,y) = [(x-y)^t w^{-1} (x-y)]^{\frac{1}{2}},$$

where, w is the intraspeaker covariance matrix

Bay's Decision Rule

Conditional probability density function generated by a speaker is the match score

$$P(w_j|x) > P(w_k|x),\ k \neq j,\ k = 1,...C$$
$$P(w_j) > P(w_k),\ k \neq j,\ k = 1,...C$$

Log likely hood Ratio

The Observation is a random vector.

$$\text{The like hood ratio for speaker A}$$
$$\lambda_A(z) = \frac{P_A(z|H_0)}{P_A(z|H_1)}$$

Also normalized log like hood ratio is used.

(B)Decision

Identification

- 1:N Comparison
- N+1 output

$$\text{Accuracy} = \frac{\text{No.of speakers Identified}}{\text{Total No.of speakers Tested}} \times 100$$

Verification

- -1:1 Comparison
- False Acceptance Ratio
- False Rejection Ratio
- Equal Error Rate

3. Development of ASR System

Two practical system developed in literature

Speaker recognition system – I

Text Independent speaker identification

Database

- 100 (50 male and 50 female) speakers, 10 single digits
- 5 sessions with 4 set of digit string per session in 2 months recorded from telephone.
- First 100 utterances are used as training set and remaining for testing

Analysis

- Sampling frequency 6.67 kHz
- Hamming window-45ms
- Frame size/ shift=45/30 ms

Feature Extraction

- 8th order LP coefficients are used as feature vectors

Modeling

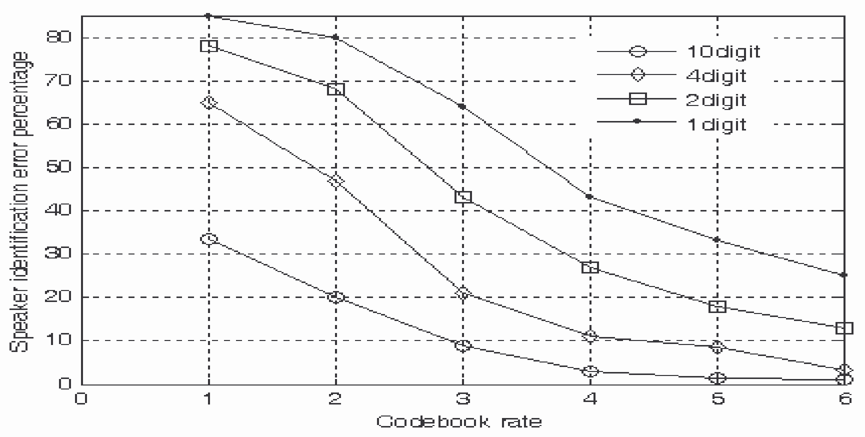
- Binary split algorithm is used to build 64 codebook vectors to model the speaker.

Measurement and decision

- Decision is taken on minimum Euclidean distance .

Performance

- 98% identification accuracy



Speaker recognition system - II

Text Independent speaker identification

Database

- TIMIT database, 30 speakers, 6 wave files for training, 2 wave files for testing

Analysis

- Sampling frequency 8 kHz
- Hamming window-20 ms
- Frame size/ shift=20/10 ms

Feature Extraction

- 1 to 13 MFCCs are used as the feature vectors

Modeling

- Binary split algorithm is used to build 512 codebook to model the speaker

Measurement and decision

- Decision is taken on minimum Euclidean distance .

Performance

- 95% identification accuracy

Codebook size/ Frame	**16**	**32**	**64**	**128**
160/80	**52**	**54**	**57**	**57**
	86.67	**90**	**95**	**95**

References

1]. J.R. Deller jr., J.H.L. Hansen, J G Proakis, Discrete-Time Processing of Speech Signal, IEEE press, 2000.

[2]. Campbell J.P., Speaker Recognition: A Tutorial. Proc. IEEE, vol. 85, no. 9, pp 1437-1462, Sept. 1997

[3]. O'Shaughnessy D, Speaker Recognition, IEEE ASSP Mag. vol. 3, pp 4-17, 1986

[4]. J.J.Wolf, "Efficient acoustic parameters for speaker recognition", J. Acoust. Soc. Amer. vol.51, pp.2044-2055, June, 1972.

[5]. S.R. Mahadeva Prasanna et al, " Extraction of speaker-specific excitation information from linear prediction residual of speech", Speech Communication 48 (2006),pp1243- 1261

[6]. Davis S. B, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans.Trans. on Acoust. Speech and signal processing, vol.28.pp.357-366.Aug.1980

[7]. B. S. Atal et al, "Speech analysis and synthesis by linear prediction of the speech wave", J.Acoust. Soc.Amer., vol.50, pp.637-655, Aug, 1971.

[8]. B.S.Atal, "Effectiveness of linear prediction characteristics of the speech for automatic speaker identification and verification", J.Acoust. Soc.Amer., vol.55, pp.1304-1312, June, 1974.

[9]. L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall Inc., 1993.

[10]. B.S.Atal, "Automatic Recognition of Speakers from their voices", Proc., IEEE, vol.64, no.4, April 1976

[11]. S.Pruzensky, " Pattern Matching Procedure for Automatic Talker Recognition ", J.Acoust.Soc.Amer., vol.35, pp.354-358, Mar, 1963.

[12]. S.Pruzensky, " Talker Recognition Procedure Based on analysis of Variance" , J. Acoust. Soc. Amer., vol.36, pp.2041-2047, Nov., 1964.

[13]. R. Schwartz et al, "The application of probability density estimation to speaker identification", Int. conf. on acoustic Speech and signal proc., vol.7, pp.1649-1652,May 1982.

[14]. S.Furui, “Cepstral Analysis Technique for Automatic Speaker Verification”, IEEE Trans. Acoust .,Speech, signal processing, vol.ASSP-29,no 2,pp.254-272, April.1981.

[15]. Marco Grimaldi, “Speaker Identification Using Instantaneous Frequencies”, IEEE Trans. on speech and audio processing vol.16, no.6, pp 1097-1111, Aug. 2008.

[16]. F K Soong et al, “A vector quantization approach to speaker recognition”, in Proc. Int.Conf. Acoustics, Speech and Signal processing, Tampa, FL, 1985, pp.387-390.

[17]. Douglas A Reynolds, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Trans. on speech and audio processing, vol.3.no.1,pp 72-83,Jan. 1995.

[18]. B.Yegnanarayana et al , “Combing Evidence From source ,Supra segmental and Spectral Features for a Fixed-text Speaker Verification System”, IEEE Trans. on speech and audio processing, vol.13, no.4, pp 575-582, July 2005.