A Study on Speaker Recognition System and Pattern classification Techniques

Dr E.Chandra[1] K.Manikandan[2] M.S.Kalaivani[3]

Dean, Department of Computer Application, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, India[1]

Assistant Professor, Department of Computer Science, P.S.G College of Arts and Science. Coimbatore, India[2]

Research Scholar, Department of Computer Science, P.S.G College of Arts and Science. Coimbatore, India[3]

Abstract Speaker Recognition is the process of identifying a person through his/her voice signals or speech waves. Pattern classification plays a vital role in speaker recognition. Pattern classification is the process of grouping the patterns, which are sharing the same set of properties. This paper deals with speaker recognition system and over view of Pattern classification techniques DTW, GMM and SVM.

Keywords Speaker Recognition System, Dynamic Time Warping (DTW), Gaussian Mixture Model (GMM), Support Vector Machine (SVM).

I. INTRODUCTION

Speaker Recognition is the process of identifying a person through his/her voice signals [1] or speech waves. It can be classified into two categories, speaker identification and speaker verification. In speaker identification task, a speech utterance of an unknown speaker is compared with set of valid users. The best match is used to identify the speaker. Similarly, in speaker verification the unknown speaker first claims identity, and the claimed model is then used for identification. If the match is above a predefined threshold, the identity claim is accepted The speech used for these task can be either text dependent or text independent. In text dependent application the system has the prior knowledge of the text to be spoken. The user will speak the same text as it is in the predefined text. In a text-independent application, there is no prior knowledge by the system of the text to be spoken.

Pattern classification plays a vital role in speaker recognition. The term Pattern defines the objects of interest. In this paper the sequence of acoustic vectors, extracted from input speech are taken as patterns. Pattern classification is the process of grouping the patterns, which are sharing the same set of properties. It plays a vital role in speaker recognition system. The result of pattern classification decides whether to accept or reject a speaker. Several research efforts have been done in pattern classification. Most of the works based on generative model. There are Dynamic Time Warping (DTW) [3], Hidden Markov Models (HMM) , Vector Quantization (VQ) [4], Gaussian mixture model (GMM) [5] and so forth.

Generative model is for randomly generating observed data, with some hidden parameters. Because of the randomly generating observed data functions, they are not able to provide a machine that can directly optimize discrimination.

Support vector machine was introducing as an alternative classifier for speaker verification. [6]. In machine learning SVM is a new tool, which is used for hard classification problems in several fields of application. This tool is capable to deal with the samples of higher dimensionality. In speaker verification binary decision is needed, since SVM is discriminative binary classifier it can classify a complete utterance in a single step.

This paper is planned as follows. In section 2: speaker recognition system, in section 3, Pattern Classification, AND overview of DTW, GMM, and SVM techniques .section 4: Conclusion.

II. SPEAKER RECOGNITION SYSTEM

Speaker recognition categorized into verification and identification. Speaker Recognition system consists of two stages .speaker verification and speaker identification. Speaker verification is 1:1 match, where the voice print is matched with one template. But speaker identification is 1:N match, where the input speech is matched with more than one templates. Speaker verification consists of five steps. 1. Input data acquisition 2.feature extraction 3.pattern matching 4.decision making 5.generate speaker models.

In the first step sample speech is acquired in a controlled manner from the user. The speaker recognition system will process the speech signals and extract the speaker discriminatory information. This information forms a speaker model. At the time of verification process, a sample voice print is acquired from the user. The speaker recognition system will extract the features from the input speech and compared with predefined model. This process is called pattern matching.

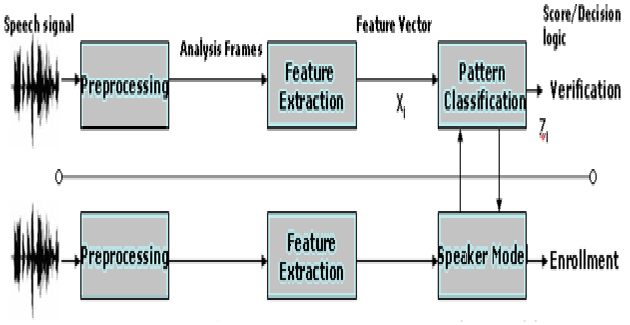


Fig 1. Speaker recognition system

A. DC Offset Removal and Silence Removal

Speech data are discrete-time speech signals, carry some redundant constant offset called DC offset [8].The values of DC offset affect the information, extracted from the speech signals. Silence frames are audio frames of background noise with low energy level .silence removal is the process of discarding the silence period from the speech. The signal energy in each speech frame is calculated by using equation (1).

$$E_i = \sqrt{\sum_{k=1}^{M} x_i(k)^2} \qquad i = 1, \ldots\ldots, N$$

M – Number of samples in a speech frames, N- Total number of speech frames. Threshold level is determined by using the equation (2)

Threshold = Emin + 0.1 (Emax – Emin) (2)

Emax and Emin are the lowest and greatest values of the N segments.

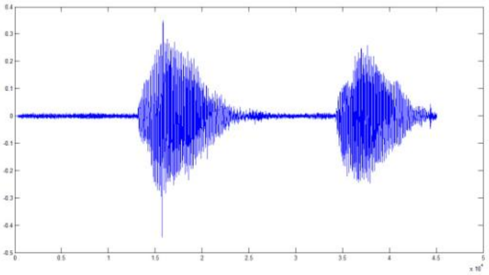


Fig 2. Speech Signal before Silence Removal

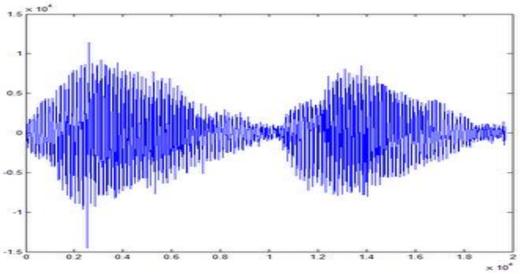


Fig 3. Speech Signal after Silence Removal

This technique is used to enhance the high frequencies of the speech signal. The aim of this technique is to spectrally flatten the speech signal that is to increase the relative energy of its high frequency spectrum. The following two factors decides the need of Pre-emphasis technique.1.Speech Signals generally contains more speaker specific information in higher frequencies [9]. 2. If the speech signal energy decreases the frequency increases .This made the feature extraction process to focus all the aspects of the voice signals. Pre-emphasis is implemented as first order finite Impulse Response filter, defined as

H (Z) = 1-0.95 Z-1 (3)

The below example represents speech signals before and after **Pre-emphasizing.**

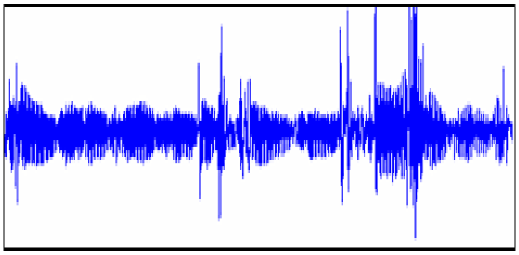


Fig 4. Speech Signal before Pre-emphasizing

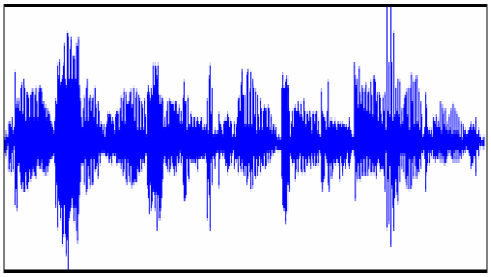


Fig 5. Speech Signal after Pre-emphasizing

A. Windowing and Feature Extraction:

The technique windowing is used to minimize the signal discontinuities at beginning and end of each frame. It is used to smooth the signal and makes the frame more flexible for spectral analysis. The following equation is used in windowing technique.

***y1(n) = x (n)w(n), 0 ≤□n ≤□N*-1 (4)** **N-** Number of samples in each frame.

The equation for Hamming window is(5)

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N\text{-}1 \quad (5)$$

There is large variability in the speech signal, which are taken for processing. to reduce this variability ,feature extraction technique is needed. MFCC has been widely used as the feature extraction technique for automatic speaker recognition. Davis and Mermelstein reported that Mel-frequency cepstral Coefficients (MFCC) provided better performance than other features in 1980 [10].

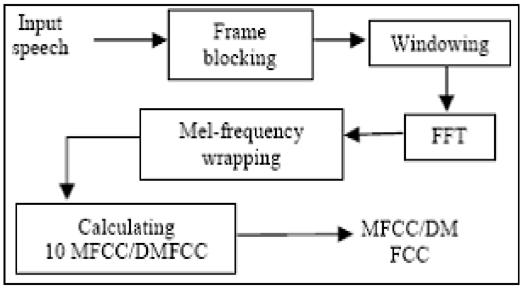


Fig 6. Feature Extraction

MFCC technique divides the input signal into short frames and applies the windowing techniques, to discard the discontinuities at edges of the frames. In fast Fourier transform (FFT) phase, it converts the signal to frequency domain and after that Mel scale filter bank is applied to the resultant frames. After that, Logarithm of the signal is passed to the inverse DFT function converting the signal back to time domain.

III. PATTERN CLASSIFICATION

Pattern classification involves in computing a match score in speaker recognition system. The term match score refers the similarity of the input feature vectors to some model. Speaker models are built from the features extracted from the speech signal. Based on the feature extraction a model of the voice is generated and stored in the speaker recognition system. To validate a user the matching algorithm compares the input voice signal with the model of the claimed user. In this paper three techniques in pattern classification have been compared. Those three major techniques are DTW, GMM and SVM.

A. Dynamic Time Warping:

This well known algorithm is used in many areas. It is currently used in Speech recognition, sign language recognition and gestures recognition, handwriting and online signature matching, data mining and time series clustering, surveillance, protein sequence alignment and chemical engineering, music and signal processing. Dynamic Time Warping algorithm is proposed by Sadaoki Furui in 1981.This algorithm measures the similarity between two series which may vary in time and speed. This algorithm finds an optimal match between two given sequences. The average of the two patterns is taken to form a new template. This process is repeated until all the training utterances have been combined into a single template. This technique matches a test input from a multi-dimensional feature vector T= [t1, t2…tI] with a reference template R= [r1, r2…rj]. It finds the function w (i) as shown in the below figure. In Speaker Recognition system every input speech is compared with the utterance in the database .For each comparison, the distance measure is calculated .In the measurements lower distance indicates higher similarity.

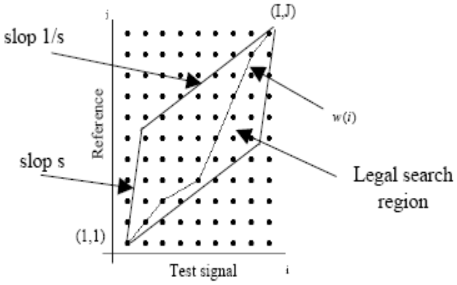


Fig 7. Dynamic Time Warping

B. Gaussian mixture model:

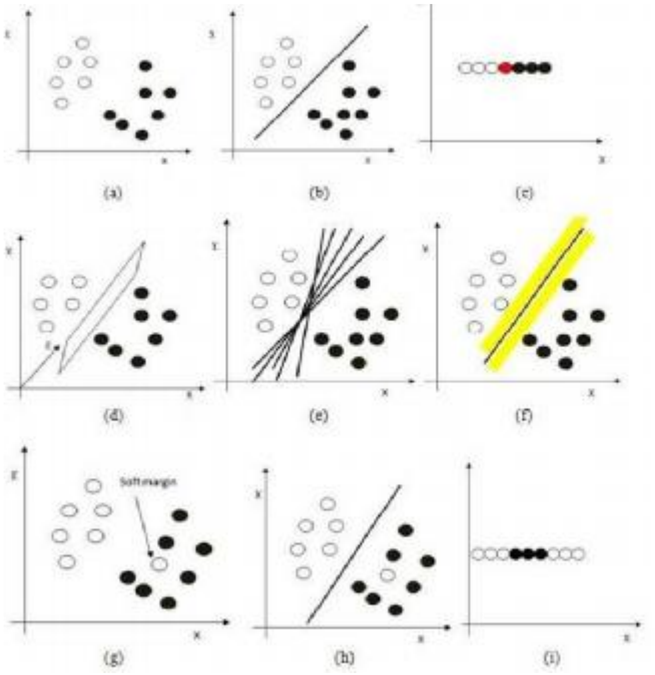
Gaussian mixture model is the most commonly used classifier in speaker recognition system. It is a type of density model which comprises a number of component functions. These functions are combined to provide a multimodal density. This model is often used for data clustering. It uses an alternative algorithm that converges to a local optimum. In this method the distribution of the feature vector x is modeled clearly using mixture of M Gaussians.

$$P(\mathrm{x}|M) = \sum_{i=1}^{m} a_i \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathrm{x}-\mu_i)^T \Sigma_i^{-1}(\mathrm{x}-\mu_i)\right)$$

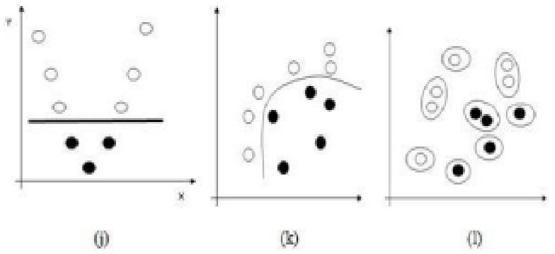
*mu*i- represent the mean and covariance of the i th mixture. x1, x2…xn, - Training data ,M-number of mixture. The task is parameter estimation which best matches the distribution of the training feature vectors given in the input speech. The well known method is maximum likehood estimation. It finds the model parameters which maximize the likehood of GMM. Therefore, the testing data which gain a maximum score will recognize as speaker.

C. Support Vector Machine:

Support machine was proposed in 1990 and it is one of the best machine learning algorithms. This is used in many pattern classification problems. Such as image recognition, speech recognition, text categorization, faces detection and faulty card detection, etc. The basic idea of support vector machine is to find the optimal linear decision surface based on the concept of structural risk minimization. It is a binary classification method. The decision surface refers the weighted combination of elements in a training dataset. These elements are called support vectors. These vectors define the boundary between two classes. In a binary problem +1 and -1 are taken as two classes. The size of the margin should be maximized to characterize the boundary between two classes. The below example explains pattern classification by using SVM. In the fig 3(a), there are two different kinds of patterns taken for process. A line is drawn to separate these two patterns. In the fig 3(b), by using a single line the patterns are separated, the patterns are presented in two dimensional spaces. The similar representation in one dimensional space in the fig 3(c), a point can be used to separate patterns in one dimensional space. a plane that separates these patterns in 3-D space ,represented in the fig 3(d),is called separating hyper plane.. The next task a plane should be selected from the set of planes whose margin is maximum. The plane with the maximum margin i.e. perpendicular distance from the marginal line is known as optimal hyper plane or maximum margin hyper plane as shown in fig 3(f). The patterns that lie on the edges of the plane are called support vectors



While classify the patterns, there may exist some errors in the representation, as shown in the fig 3(g), such types of errors are called soft margin. Sometimes, these errors can be ignored to some threshold value. The patterns that can be easily separated using line or Plane are called linearly Separable patterns .Non-linear separable patterns (fig-j, k,l)are difficult to classify. These patterns are classified by using kernel functions.



In order to classify non-linear separable patterns the original data's are mapped to higher dimensional space using kernel function.

IV. CONCLUSION

In this paper we have explained about speaker recognition system and discussed about three major pattern classification techniques, Dynamic Time Warping, Gaussian mixture model and Support Vector Machine. SVM will work efficiently on fixed length vectors. To implement SVM the input data should be normalized for better performance. In future, we have planned to implement these techniques in speaker recognition system and evaluate the performance. The performance of the models will also be evaluated by incrementing the amounts of training data.

REFERENCES

[1] Campbell, J.P., "Speaker Recognition: A Tutorial", Proc. Of the IEEE, vol. 85,no. 9, 1997, pp. 1437-1462.

[2] Sadaoki Furui., "Recent advances in speaker recognition",Pattern Recognition Letters. 1997,18 (9): 859-72.

[3] Sakoe, H.and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", Acoustics,Speech, and Signal Processing, IEEE Transactions on Volume 26, Issue 1, Feb 1978 Page 43 - 49.

[4] Lubkin, J. and Cauwenberghs, G., "VLSI Implementation of Fuzzy Adaptive Resonance and Learning Vector Quantization", Int. J. Analog Integrated Circuits and Signal Processing, vol. 30 (2), 2002,pp. 149-157.

[5] Reynolds, D. A. and Rose, R. C. "Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3, 1995, pp 72–83.

[6] Solera, U.R., Martín-Iglesias, D., Gallardo-Antolín, A., Peláez-Moreno, C. and Díaz-de-María, F, "Robust ASR using Support Vector Machines", Speech Communication, Volume 49 Issue 4, 2007.

[7] Temko, A.; Monte, E.; Nadeu, C., "Comparison of Sequence Discriminant Support Vector Machines for Acoustic Event Classification", ICASSP 2006 Proceedings, 2006 IEEE International Conference on Volume 5, Issue , 14-19 May 2006

[8] Shang, S.; Mirabbasi, S.; Saleh, R., "A technique for DCoffset removal and carrier phase error compensation in integrated wireless receivers Circuits and Systems", ISCAS apos;03. Proceedings of the 2003

International Symposium onVolume 1, Issue , 25-28 May 2003 Page I-173 - I-176 vol.1

[9] Vergin, R.; Oapos;Shaughnessy, D., "Pre-emphasis and speech recognition lectrical and Computer Engineering",Canadian Conference on Volume 2, Issue , 5-8 Sep 1995

[10] Davis, S. B. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustic, Speech and Signal Processing, ASSP-28, 1980, No. 4.

[11] Sadaoki Furui., "Cepstral analysis technique for automatic speaker verification", IEEE Trans. ASSP 29, 1981,pages 254-272.

BIOGRAPHIES

Dr.E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University, Coimbatore in 1994. She obtained her M.Phil. In the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007. She has totally 15 yrs of experience in teaching including 6 months in the industry. Presently she is working as Director, Department of Computer Applications in D. J. Academy for Managerial Excellence, Coimbatore. She has published more than 30 research papers in National, International Journals and Conferences in India and abroad. She has guided more than 20 M.Phil. Research Scholars. Currently 3 M.Phil Scholars and 8 PhD Scholars are working under her guidance. She has delivered lectures to various Colleges. She is a Board of studies member of various Institutions. Her research interest lies in the area of Data Mining, Artificial Intelligence, Neural Networks, Speech Recognition Systems, Fuzzy Logic and Machine Learning Techniques. She is an active and Life member of CSI, Society of Statistics and Computer Applications. Currently she is Management Committee member of CSI Coimbatore Chapter.

K. Manikandan received his Bsc from Bharathidhasan University, Tiruchirappalli in1998 and received his MCA from Bharathiadsan University, Tiruchirappalli in 2001. He received M.Phil in the area of soft computing from Bharathiyar university, Coimbatore in 2004. He has 12 years of experience in teaching. Currently, he is working as a Assistant Professor, Department Of Computer Science, PSG College of arts and Science, Coimbatore and pursuing PhD in Bharathiar University, Coimbatore.He has presented research papers in National and International Conferences and published a paper in International Journal. His Research Interest is Soft Computing. He is Life a member of IAENG. He has guided more than 4 M.Phil Research Scholars. Currently 3 M.Phil Scholars are working under his guidance. He has delivered lectures to various Colleges.

M.S.Kalaivani received her BCA from P.S.G College of Arts and Science, Coimbatore, in 2005 and received her MCA from National Institute of Technology, Tiruchirappalli in 2008.She has 4 years of working experience at software industry. Presently, she is working as a Research Scholar, Department of Computer Science, P.S.G. College of Arts and Science, Coimbatore. Her research interests are Machine Learning and Fuzzy logic.