ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

A Review on Speech Analysis and Automatic Speaker Recognition

Alahari.Neelima[1], Gattupalli.Deepti[2]

M. Tech Student, Department of ECE, Bapatla Engineering College, Bapatla, India[1,2]

Abstract*:* Speech recognition technology has been a multi-disciplinary research area since many decades and it found wide range of applications in many areas. Speech analysis can be done in time and frequency domain. In this paper, different techniques for speech analysis (feature extraction) such as short time Fourier transform, linear predictive coding, Mel Frequency cepstral analysis etc. are discussed. Also different methods for automatic speaker recognition also have been discussed. Thus this paper gives a technological review on methods involved in feature extraction and automatic speaker recognition implemented so far.

Keywords: Linear Predictive Coding, Mel Frequency, Wavelet, Vector Quantization, Dynamic time warping, Hidden Markov Model, Artificial Neural Network, Support Vector Machine, Gaussian Mixture Model.

I.INTRODUCTION TO SPEECH ANALYSIS

Speech Analysis refers to extraction of the properties or features of a speech signal s(n).Mostly speech analysis is extracted features are used for different aspects while suppressing other aspects of speech.

Short-Time Speech Analysis

Speech analysis assumes that signal properties mostly vary with time. A short time window of speech is used to extract parameters presumed to be fixed for the duration of the window. To model the dynamic parameters, we must divide the signal into successive frames, so that the parameters can be calculated to follow relevant changes.

Windowing

Windowing refers to multiplication of speech signal s(n) by a window w(n).Most of the applications prefer speech averaging to yield output parameter contour which represents slowly varying physiological aspects of vocal tract movements. A set of speech samples are weighted by the shape of the window w(n). w(n) may have infinite length, but mostly w(n) have finite length to simplify computation.

Time Domain Parameters

Time domain analysis transforms a speech signal into a set of parameter signals, which usually vary much more than the original signal.

Frequency Domain Spectral Parameters

Frequency Domain provides most useful parameters for speech processing. Parameters are analysed easily in spectral domain than t time domain.

Filter Bank Analysis

This method uses a bank of band-pass filters each analysing a different range of frequencies of the input speech. Filters often follow bark scale i.e. equally spaced; fixed bandwidth filters up to 1 kHz and then logarithmically increasing bandwidth. The amplitude outputs from a bank of filters provide a more efficient spectral representation than DFT.

Speech enhancement techniques have been evolved from the very basic techniques based on short and long term spectral averages, predictive coefficients (LPCC,PLPC) to widely used filter coefficients(MFCC,LFCC) . The widely used techniques for feature extraction are characterized into various categories. One is filter bank coefficients (MFCC,LFCC) and the other one is predictive coefficients modelled by all pole model (LPC,LPCC).

II.FEATURE EXTRACTION METHODS

A) Linear Predictive Coefficients

In this feature extraction technique the speech reduction model is represented by these coefficients. The linear predictive coefficients are the coefficients of this all pole filter. The speech signal is represented as

$$S(n)=-\sum_{k=1}^{p} a(k)s(n-k)+Gu(n)$$

Where u(n) is the excitation signal, s(n) is the speech signal . This technique is modelled by all pole model which is based on the principle of sound production. Andcoming to the disadvantage there will be performance degradation in presence of noise.

B) Linear Predictive Cepstral Coefficients

When the LPC coefficients are represented in cepstrum domain then the obtained coefficients are linear predictive cepstral coefficients. They are robust and reliable than LPC. Cepstrum is obtained by taking inverse DFT of logarithm of the magnitude of the DFT of the speech signal. Cepstral coefficients are totally FFT based which are not consistent with human hearing due to representation by linearly spaced filters.

C) Mel-frequency Cepstral Coefficients

This technique is based on a linear cosine transform of a log power spectrum on a non-linearMel-scale of frequency.This is called as short-term power spectrum of a sound. So these coefficients collectively make up the MFCC.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

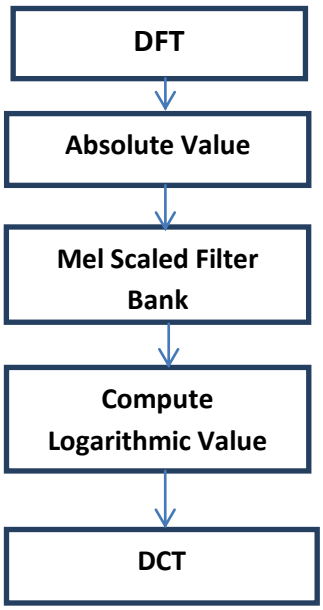

Fig.1. Flow chart for MFCC technique

For each frame of the speech signal, DFT is calculated as

$$\mathbf{X(k)}= \sum_{n=0}^{n-1} \mathbf{x(n)} e^{-j2*pi*k*n/N} \quad ; \text{N-1<=k<=0}$$

After that the absolute values obtained by powers of DFT is determined and mapped by Mel-scaled filter bank. After calculating logarithmic values of these mapped power spectrums, cepstral coefficients are obtained by taking DCT as

$$\mathbf{X(k)}= \sum_{n=0}^{n-1} \mathbf{x(n)} \cos[\frac{\mathbf{pi}}{\mathbf{N}\left(\mathbf{n}+\frac{\mathbf{1}}{\mathbf{2}}\right)\mathbf{k}}];\text{N-1<=k<=0}$$

This technique is totally based on filter bank coefficients, so this involves more information about lower frequencies than higher frequencies due to Mel spaced filter banks hence behaves like a human ear. And as this technique is based on STFT so it has fixed time-frequency resolution

D)Wavelet Decomposition

In this frequency extraction technique the speech signal is decomposed into different frequency sub-bands. At each level signal is partitioned into low level and high level frequency bands the .And the coefficients obtained can be used for further decomposition. The main advantage of this technique is that only the low frequency bands can be used in fusion with MFCC by extracting features from selective frequency sub-bands and can be used for further processing. This technique involves decomposition to sub-bands. The main advantage of this technique is that the characteristics of time frequency localization and the multi-resolution analysis are suitable for non-stationary speech signal.

III.AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is the process by which a computer converts the speech signal into corresponding text with the use of algorithms. By theory, ASR could be as simple as a large dictionary in which each entry (input utterance) is stored as a digitized waveform labelled with text pronunciation. In case of any conversion from speech to text, the system would pick up a closest text match to its input utterance. Thus ASR search is vast and expensive.

Basic Pattern Recognition Method

The basic pattern recognition approach involves different stages such as normalization, parameterization and feature estimation, similarity evaluation and decision. It can be explained by the figure as follows:

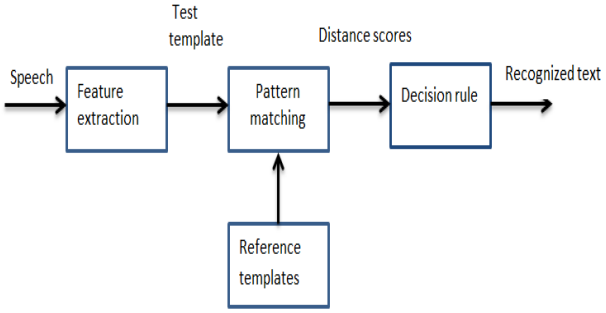

Fig. 2. Pattern recognition method

The first two steps in this process constitute ASR front end. In the process of pattern recognition method, the input speech utterance is subjected to feature extraction which allows us to extract the required parameters for further processing, using different methods. Later the test template is subjected to match with the reference templates in the database using distance measures for evaluating the similarities between them. Based on this similarity measure, final decision is made and thus the input speech is recognized as corresponding text.

A. Dynamic Time Warping

This method deals with the variant speaking rate and the variable length of input vectors. Dynamic time warping specifically addresses the problem of time alignment by non-linearly stretching (warping) of one template in the process of synchronizing similar acoustic segments in test and reference templates. The time normalized distance is calculated between the two patterns and finally an optimal path is selected through a network of possibilities which gives the minimum distance measure. Linear time alignment is a special case of DTW.

B. Vector Quantization:

Vector quantization is often used in ASR because of its efficient data reduction same as in speech coders. The utility of vector quantization lies in the efficient use of code books which consists of code vectors, as reference models and also in the use of code book searches instead of using costly evaluation methods. Each vocabulary word has been provided with its own vector quantization code book during training phase. During the matching phase, the test speech is evaluated by all the code books and then ASR chooses the word with its code book having lowest distance measure. In this way vector quantization method of speech recognition is carried out. Mostly vector

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

quantization in ASR involve full search algorithm in order to find optimal code word. Computations could be reduced by using binary tree or look up tables for distances. One of the examples of vector quantization is clustering. Although it consumes more time in training phase, yields good performance in recognition phase. Hence this is about vector quantization.

C. Hidden Markov model:

It is the stochastic approach and most commonly used network in automatic speech recognition. In the first order Markov chain, the likelihood of being in a present state depends only on the prior state to it and is independent on any other previous states. In this model, only left to right transitions are allowed.

Thus there exist three states in this model i.e. initial, middle and final states corresponding to beginning, middle and end of the utterance which is being modelled. The parameters here to be considered are the number of distinct observations, probability distribution of observation symbol, probability of changing from one state to another and probability of lying in particular state initially.

The training features are represented in terms of probability measures which train the HMM speaker model. The HMM model which yields the highest probability is thus selected.

D. Support Vector Machine:

Support Vector Machine (SVM)is a binary non-linear classifier in which the input vector can be classified into either class 1 or class 2 by constructing a decision hyper plane. It works on fixed length vectors.

The decision hyper plane can be defined by weight vectors w, data points x and bias b. the equation for a decision hyper plane can be given by

$$W.X + b = 0$$

Points lying on the hyper plane satisfy the equation:

$$W.X + b = 0$$

These points are known as support vectors.

Points lying on one side of this hyper plane are denoted by class C1 as positive examples satisfying:

$$W.X + b > 0,\ d(i) = +1$$

Points lying on the other side of this hyper plane are denoted by class C2 as negative examples satisfying:

$$W.X + b < 0,\ d(i) = -1$$

SVM can also be used for multiclass problem in which the input utterance taken as one class and rest as the other.

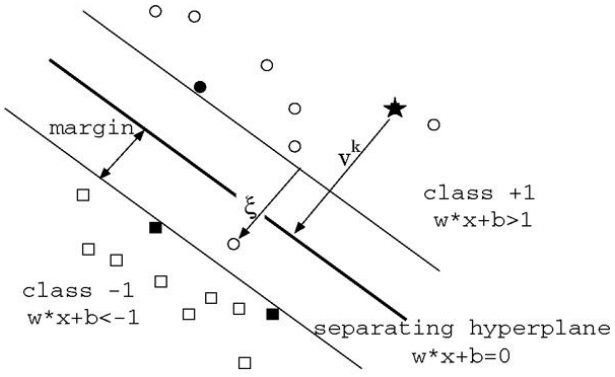

Fig 3: Hyperplane dividing input data into two classes

E. Artificial Neural network:

Artificial neural network(ANN) can accomplish useful pattern recognition tasks including ASR. An individual node simulating a neuron can make a binary classification (via the presence or absence of a firing), based on a set (frame) of M input data samples (e.g., M spectral parameters from a frame of speech). A group of N such nodes all receiving the same M inputs can thus classify speech frames into one of L ~ 2N classes. Usually, such a basic ANN is configured so that only one of the N nodes should emit a firing for each frame (i.e., L = N). Such a single-layer ANN or perceptron (one set of nodes, all receiving the same inputs) is a linear classifier that partitions an acoustic space of M dimensions into N hyper ellipsoids or classes. Such simple classifiers do not solve most ASR problems; e.g., deciding to which of N phonemes an M-dimensional speech frame belongs. In the M-space for most speech parameters (e.g., LPC, cepstra), phonemes often have very complex shapes, requiring complicated decision surfaces (and not the simple hyper planes of perceptron's). Allowing one layer of M} neurons to feed into a second layer of M2neurons create a two-layer connectionist ANN, capable of distinguishing among many convex surfaces in acoustic space. In practice, ASR often uses a third layer, so that the original set of M1 speech parameters feeds into a hidden layer of M2 nodes, in tum feeding another hidden layer of M) nodes, finally yielding a set of N outputs . While M1 and N correspond to the number of parameters in each input frame and the number of output classes (e.g., phonemes), respectively, the sizes of the hidden layers (M2 and M) are empirically chosen in a compromise between complexity and classification accuracy. Each node in an ANN emits a value of 1 (i.e., a neural firing) when a weighted sum of its inputs Xi exceeds a threshold .Many ANNs go beyond the original biological idea of a binary output, and allow a sigmoid output, which monotonically ranges from 0 to 1 as the sum in increases.

F. Gaussian Mixture Model:

It is a probability density estimator in which the distribution of an input feature vector can be modelled using a mixture of N models. It uses expectation minimization algorithm which is popularly known as EM

DOI 10.17148/IJARCCE.2015.4289

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

algorithm in order to estimate the mean and variance of input signal. The distance of given sequence of features is computed by log likelihood measure and thus the one with highest measure is then identified as the speaker identity.

IV.CONCLUSION

Thus this paper reviewed different the research that have been done in automatic speaker recognition. The techniques which have been used for analysing speech are discussed such as linear predictive coding, Mel frequency cepstral analysis etc. Each method has its own merits and demerits and among all those Mel frequency cepstral coefficients method can be viewed as the best one for feature extraction. Automatic speaker recognition classification was also explained with different methods such as support vector machine, dynamic time warping, vector quantization, artificial neural networks etc. and Gaussian mixture model is of more significance of all those. Hence, this paper aimed at bringing out the progress in speech recognition which may be useful for further research.

REFERENCES

[1] D.A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification", IEEE Trans. on Speech and Audio Processing, vol. 2, issue-4, pp. 639-643, October, 1994.

[2] X. Zhou, D. G. Romero, R. Duraiswami, C.E. Wilson, S. Shamma, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition", IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Waikoloa, HI, pp. 559 – 564, 11-15 December, 2011.

[3] Anil K.Jain, et.al., Statistical Pattern Recognition: A Review , IEEE Transactions on Pattern Analysis and Machine Intelligence ,Vol.22, No.1, January 2000.

[4] MdSahidullah, G.Saha, "A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition", IEEE Signal Processing Letters, vol.20, issue-2, pp. 149-152, February, 2013.

[5] SPEECH COMMUNICATIONS ,Human and Machine,SecondEdition,DouglasO'Shaughnessy,Universite du Quebec, Institut National de la RechercheScientifique,INRS-TELECOMMUNICATIONS.

[6] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery (Springer), vol. 2, issue-2, pp. 121-167, June, 1998.

[7] WW. Chen, Q. Hong, X. Li, "GMM-UBM for Text-Dependent Speaker Recognition", IEEE International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, pp. 432-435, 16-18 July, 2012.

[8] R. P. Ramachandran, K.R. Farrell, R. Ramachandran, R. J. Mammone, "Speaker Recognition—General Classifier Approaches and Data Fusion Methods", Pattern Recognition in Information Systems, vol. 35, issue-12, pp. 2801-2821, December, 2002.

[9] Ghahramani Z., "An Introduction to Hidden Markov Models and Bayesian Networks", International Journal of Pattern Recognition and Artificial Intelligence, vol. 5, issue-1, pp. 9-42, 2001.

[10] P.A.Devijver and J.Kittler, Pattern Recognition:A Statistical Approach , London, Prentice Hall, 1982.

[11] E.Oja, Subspace Methods of Pattern Recognition, Letchworth, HeHertfordshire,England:Research Studies Press, 1983.

[12] K.Fukunaga, Introduction to Statistical Pattern Recognition , second, Newyork:Academic Press, 1990.

[13] J.P. Campbell, "*Speaker Recognition: A Tutorial*", Proceedings of the IEEE, vol.85, issue-9, pp. 1437-1462, September, 1997.

[14] G.R. Doddington, "*Speaker Recognition – Identifying People by their Voices*", Proceedings of the IEEE, vol. 73, issue-11, pp. 1651-1664, November, 1985.