International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-5)

Research Article | **May 2015**

A Survey on Speech Recognition Algorithms

Gaganpreet Kaur
Research Scholar, Dept. of CSE,
P.T.U, (Punjab) India

Dr. Dheerendra Singh
Prof. & Head, Dept. of CSE
SUSCET, Tangori,(Punjab) India

Gagandeep Kaur
Student, Dept. ofCSE,
S.G.G.S.W.U., FGS. (Punjab) India

Abstract-

Speaker recognition is a process where a person is recognized on the basis of his/ her voice signals. Human voice is aunique characteristic for any individual.Speaker recognition is being applied in biometric identification, security related areas, remote access to computers etc.This paper delivers an overview of different techniques that can be used in application of speaker recognition such as MFCC, LPC, and LPCC for feature extraction and VQ, SVM, HMM; GMM for feature classification.It also helps in choosing the better technique based on the comparison done.

Keywords-Feature extraction, MFCC, LPC, LPCC, GMM, VQ, SVM, HMM.

I. INTRODUCTION

Human always identify speaker while they are talking to one another .The speaker may present in the same place or in different places. In this way a blind person can identify a speaker based soley on his/her vocal characteristics.Animals also use these characteristics to identify their familiar one [1].The development of efficient-Speaker Identification system has been a topic of active research during last two decades because they have a large number of potential applications in many fields that require accurate user identification such as shopping by telephone, bank transaction,accesses control and voicemail etc [2].

Speaker recognition is a generic term used for two related problems: Speaker identification and verification. In the identification task the goal is to recognize the unknown speaker from a set of N known speakers.In verification, an identity claim (e.g.,a username) is given to the recognizer and the goal is to accept or reject the given identity claim.In this paper we concentrate on the identification task.

Speaker identification can be further divided into two branches:

1. Open-set speaker identification (Speaker from outside the training set may be examined).
2. Closed-set speaker identification (The speaker is always one of a closed set used for training).

Depending on the algorithm used for the identification, this task can also be divided into text-dependent (The speaker must utter one of a closed set of words) and text-independent identification (The speaker may utter any type of words).

II. SPEAKER IDENTIFICATIONSYSTEMS: PHASES

A general speaker identification system consists of an enrollment phase and identification phase.

A. Enrollment Phase

In this phase,speech samples are collected from the speaker to train their models.The collection of enrolled models is also called a speaker database.

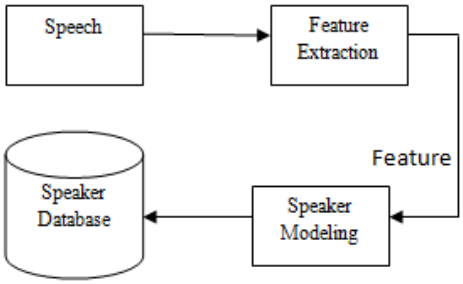

Fig. 1 Enrollment Phase

B. Identification Phase

In this phase,a test sample from unknown speaker is compared against the speaker database.Both phases include the feature extraction step to extract the speaker dependent characteristics from speech.

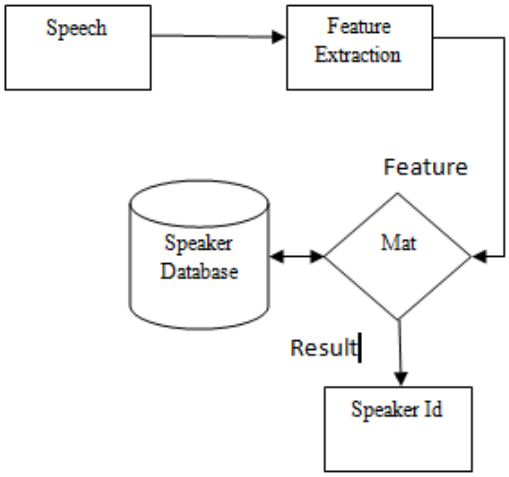

Fig. 2 Identification Phase

III. FEATUREEXTRACTION

The main aim of speaker identification is to determine speaker identity from his/her speech utterances. While speaking of every speaker have his own characteristics. These characteristics are called speaker features which can be extracted from speech utterances. A variety of choices can be used for feature extraction i.e..Digital filter bank,Fourier Transform,LPC, LPCC, MFCC etc.Some commonly used methods for speaker identification are LPC, LPCC and MFCC.

A. Linear Predictive Coding Analysis (LPC)

LPC is one of the good analysis techniques for extracting features and hence encoding the speech at low bit rate .LPC has capability for speech compression, synthesis and as well as identification .LPC is spectral estimation technique because it provides an estimate of the poles of the vocal tract transfer function.The LPC algorithm is a Path signal is stationary within and zero outside,the analysis window.

B. Linear Predictive Cepstral Coefficients (LPCC)

This technique is just an extension to the above mentionedLPC technique.When linear predictive coefficient is represented in cestrum domain then the obtained coefficients are linear predictive cepstral coefficients.Cestrum is obtained by taking inverse DFTof logarithm of the magnitude of the DFT of the speech signal.They are more robust and reliable then LPC.

C. Mel- Frequency Cepstral Coefficients (MFCC's)

The cepstrum coefficient is the result of a cosine transformation of the real logarithm of the short time energy spectrum expressed on a Mel-frequency scale.This is a more robust, reliable feature set for speech recognition then the LPC coefficients.The sensitivity of the low order cepstrum coefficient to overall spectral slope,and the sensitivity of the high-order cepstrum coefficient to noise, has made it a standard technique.It weights the cepstrum coefficient by a tapered window so as to minimize these sensitivities,frame and these are used as the feature vector.In MFCC's,the main advantage is that it uses Mel frequency scaling which is very approximate to the human auditory system.The coefficients generated by algorithm are fine representation of signal spectra with great data compression.The process of extracting MFCC's from continuous speech is illustrated in Fig [3].

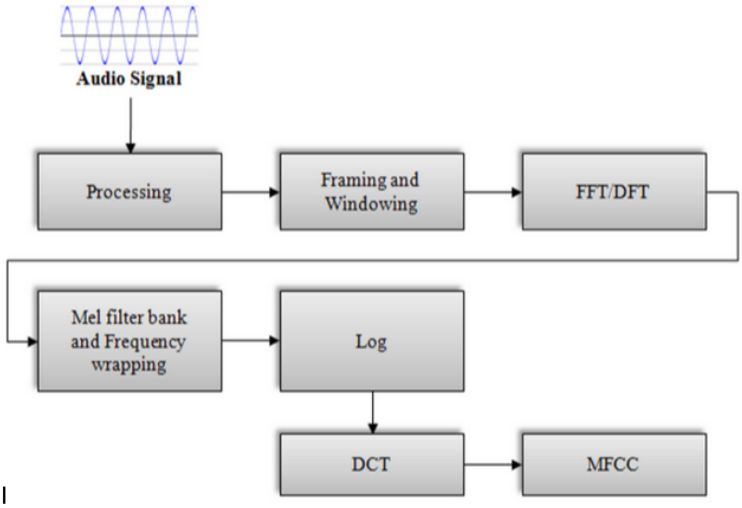

Fig.3Flow Chart for MFCC Technique [2].

TABLE I COMPARISON OF DIFFERENT FEATURE EXTRACTION TECHNIQUES [10]

S.No.	Technique	Principle	Merits and De-merits
I	Linear Predictive Coding	Modeled by all pole model	(a)Based on basic principle of sound production. (b)Performance degradation in presence of noise.
II	Cepstral Coefficients	FFT based	Not much consistent with human hearing due to representation by linearly spaced filters.
III	Linear Predictive Cepstral Coefficient	Modeled by all pole model	(a) Gives smoother spectral envelope and stable representation as compared to LPC. (b) Drawback due to linearly spaced frequency bands.
IV	Mel-Frequency Cepstral Coefficient	Filter bank coefficients	More information about lower frequencies than higher frequencies due to Mel spaced filter banks hence behaves more like a human ear as compared to other techniques.

IV. FEATURECLASSIFICATIONS

Using feature extraction, a speaker voice can be represented with feature vectors.To recognize these feature vectors various classifier can be used.The classifier consists of the various speaker models and the decision logic.Its operation constitutes two important steps:

1. In the training phase,feature vectors are used to obtain M speaker models.For each speaker,a different model is obtained from his/her speech.
2. In the testing phase,feature vectors from unknown speaker are first computed.For speaker identification,the features vectors are compared with each of M speaker models to get the scores Score (1) to Score (M).These scores are used to render a decision.The different classifiers are discussed and compared here:

A. Support Vector Machines (SVM)

Support Vector Machine is a supervised learning algorithm.It needs training of the tool before classification procedure gets started.This is best tool for binary classification of the data.The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes form the input. The hyper plane is constructed defined by set of weights W, data points X and a bias or offset b, such that:

$$W.X + b = 0 \qquad [10]$$

Where W.X denotes the dot product of the data and the normal vector to the hyper plane.The parameter b determines the offset of the hyper plane from the origin along the normal vector [10].

B. Hidden Markov Models (HMM)

HMM is popular statistical tool for modeling a huge range of the time series data.In the context of natural language processing (NLP),HMMs are usually applied with great success to problems such as part-of-speech tagging and noun-phrase chunking.This powerful statistical tool is also used for modeling generative sequences.

C. Vector quantization (VQ)

VQ is a classical quantization technique of signal processing which permits the modeling of probability density functions by the dividing of prototype vectors.The process starts by splitting a large set of points into clusters or groups having approximately the same number of points nearest to them. Centroid point represents each group.The density matching property of this quantization technique is very powerful,mainly in the density of large and high dimensional data identification.Data points are shown by their closest Centroid indexing,frequently occurring data have low error,rare data high error.Hence,this method is also appropriate for lossy data compression.

D. Gaussian mixture model (GMM)

The GMM is a density estimator. The distribution of the feature vector x is modeled clearly using a mixture of M Gaussians.Expectation maximization algorithm is used to estimate mean,covariance parameters.During recognition, a sequence of features is extracted from the input signal. Then the distance of the given sequence from the model is obtained by computing the log likelihood of given sequence. The model that provides the highest likelihood score is verified as the identity of the speaker.

TABLE II COMPARISONS OF DIFFERENTCLASSIFIERS [10]

S.No.	Classifier	Type of algorithm	Merits and De-merits
I	Support Vector Machines	Supervised	(a)Beneficial in case of binary classification. (b)Poor performance in speaker recognition due to its restriction to work with fixed length vectors.
II	Hidden Markov Model	Unsupervised	(a)Computationally more complex and needs more storage space. (b)Needs more training data to deal with intersession issue.
III	Vector Quantization	Unsupervised	(a)Memory requirement is feasible for real-time application. (b) Computationally less complex.
IV	Gaussian Mixture Model	Unsupervised	(a) Needs less training and test data. (b)Compromise between DTW and HMM.

V. CONCLUSIONS

This paper has reviewed the research done in the area ofspeaker recognition.The different methods used for feature extraction and feature classification have been discussed.Some techniques preferred over others such as MFCC for feature extraction has better performance rather than LPC or LPCC, because MFCC is more consistent with human hearing due to Mel scale representation.Otherwise choice can be made depending upon certain parameters such as number of system users,storage space,classification time etc.For feature extraction, GMM performs better as it requires fewer amounts of data to train the classifier hence memory usage also decreases for the system.

REFERENCES

[1] Md. Monirul Islam, Fahim Hassan Khan , Abul Ahsan, Md. Mahmudul Haque, "*ANovel Approach for Text-Independent Speaker Identification UsingArtificial Neural Network*", International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, 2013.

[2] V Sailaja, K.Srinivasa Rao, K.V.V.S. Reddy, "*Text Independent Speaker Identification with Finite Multivariate Generalized Gaussian Mixture Model andHierarchical Clustering Algorithm*", International Journal of Computer Applications (0975-8887), vol. 11, 2010.

[3] P. Sivakumaran, A. Ariyaeeinia and M. Loomes, "*Sub-band Based Text-dependent Speaker Verifcation,"Speech Communication,* vol. 41, pp. 485-509,2003.

[4] D.A. Reynolds, "*Speaker Identification and Verification Using Gaussian Mixture Speaker Models*," *Speech Commun.* vol. 17, pp. 91-108, 1995.

[5] Wu Junqin, Yu Junjun, "*An Improved Arithmetic of MFCC in Speech Recognition System*," *IEEE Transaction on Audio Speech processing, and Language, pp.719-722, 2011*.

[6] Jamal Amini, Abdoreza Sabzi Shahrebabaki, Navid Shokouhi, Hamid Sheikhzadeh, "*Speech Analysis/Synthesis by Gaussian Mixture Approximation of theSpeech Spectrum for Voice Conversion*", *IEEE Transaction on Audio Speech processing, and Language, pp.000428-000433, 2013.*

[7] Anjali Jain,"*Evaluation of MFCC for Speaker Verification on Various Windows*", IEEE International Conference on Recent Advances and Innovations in Engineering, pp.1-6, 2014.

[8] Gurpreet Kaur, Harjeet Kaur, "*Multi Lingual Speaker Identification on Foreign Languages Using Artificial Neural Network with Clustering*",International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 2013.

[9] Ning Wang, P. C. Ching, "*Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features speaker verification*," IEEE Transaction onAudio Speech processing, and Language,vol. 19, pp. 196-205,2011.

[10] Kirandeep Kaur, Neelu Jain, "*Feature Extraction and Classification for Automatic Speaker Recognition System-A Review*", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, 2015.

[11] Jun Tao, Xiaoxiao Jiang, "*A Domestic Speech Recognition Based on Hidden Markov Model*", IEEE CCIS, pp.606-609, 2011.

[12] Mansour Alsulaiman, Ghulam Muhammad, Zulfiqar Ali, "*Comparison of Voice Features for Arabic Speech Recognition*", IEEE, pp.90-95, 2011.

[13] Eleonora D'Arca, Neil M. Robertson and James Hopgood, "*Using the Voice Spectrum for Improved Tracking of People in a Joint Audio-Video Scheme*",IEEE, 2013.

[14] Joanna Grzybowska, Maciej Klaczynski,"*Computer-assisted HFCC-based learning system for people with speech sound disorders*", IEEE, pp.1-5, 2014.

[15] Peng Dai, Ing Yann Soon, "*An Adaptive Soft Voice Activity Detector for Automatic Speech Recognition System*", IEEE, pp.1-5, 2011.

