ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Review of Automatic Speech Recognition For Recognition of Speech and Speaker

Prof. Trupti K. Wable

Assistant Professor, Department of Electronics & Telecommunication Engineering, SVIT, Chincholi, Nasik, India

ABSTRACT: Speech is fundamental mode of communication among the human being. Every individual has distinct characteristics called speech. Due to his different characteristics speech becomes important attribute in aural biometrics. If this attributes further use to recognize particular word or phrase then it is termed as speech recognition. And if it is used to recognize individual human being i.e. speaker then it is termed as speaker recognition. A lot much research has been done in the speech and speaker recognition. Both speech and speaker recognition processes requires databases, speech processing techniques such feature extraction, feature classification and feature matching techniques. In this paper the review of speech and speaker recognition is taken with these basic processes

KEYWORDS: ASR (Automatic Speech Recognition), Speaker Recognition, Speech Recognition, Feature Extraction, Feature Classification, Speech Database.

I. INTRODUCTION

Speech and speaker recognition methods have made it possible for computer to follow human speech commands. The main goal of speech recognition area is to develop systems for speech input to machine to convert it into text. Automatic speech recognition systems has numerous application in human machine interface, such as automatic call processing in telephone networks, Voice command control, Access security systems, in super markets and in many such fields. Now days field like access to information: travel, banking, Avionics, Automobile portal, speech transcription, and Handicapped people (Blind people), railway reservations etc. [1, 2]

Speech and speaker recognition is art of device of identifying the words or phrases and individual person based on human speech. The process utilizes the characteristics of speech called as speech features. If such features are classified on the basis of uttered sound i.e. phoneme then it belongs to speech recognition. Speech and speaker recognition are both processes are affected by how the speech is produced and how it is perceived by human beings [5, 6]. The speech generation is a result of vocal tract mechanism with nasal cavity, tongue, lips, jaw & velum movement. The speech signal is slowly time varying signal in the sense that, when examined over short period it characteristics are fairly stationary. As a speech production is non stationary random process, speech is usually segmented into small parts or set of small data samples. This is for purpose of suitable processing, for data reduction. Data reduction is done through different signal processing steps such as framing and windowing. If long data segments are processed without these operations then ASR results are degraded due to non speech part processing (does not contain phoneme). There are some challenges in speech and speaker recognition which are addressed by different researchers recently i.e. environment or working condition (noise, signal to noise ratio etc), speaker age, sex, physical condition, vocabulary, speech style continuous or isolated

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

II. SPEECH AND SPEAKER RECOGNITION METHODS

A. *BLOCK DIAGRAM OF SPEECH RECOGNITION:*

A general speech or speaker recognition system is shown Figure 1.

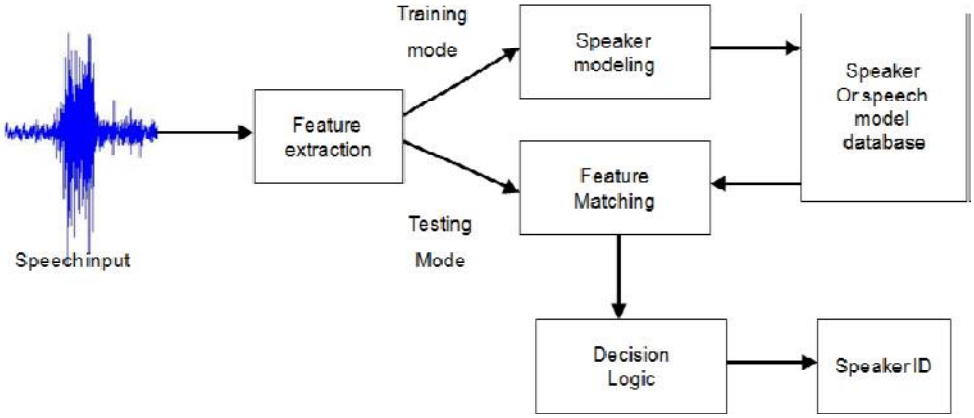


Fig.1 Speech or Speaker recognition system

The major part of any speech or speaker recognition system is feature extraction, feature classification and feature matching with database developed for speakers or speech [1].

B. *TYPES OF SPEECH AND SPEAKER RECOGNITION SYSTEMS*

Classification of speech recognition system is done in different ways. Figure 2 shows common classification of speech and speaker recognition systems.

Speech recognition systems can be separated in several different classes what type of speech/utterance they have ability to recognize such as isolated, connected word, continuous speech or spontaneous speech. Speaker recognition can be classified into identification and verification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker.

All speaker recognition systems have to serve two distinguished steps. The first one is the enrolment or training phase, while the second one is referred to as the testing phase. In the training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. In case of speaker verification systems, in addition, a speaker-specific threshold is also computed from the training samples. In the testing phase, the input speech is matched with stored reference model(s) and a recognition decision is made.

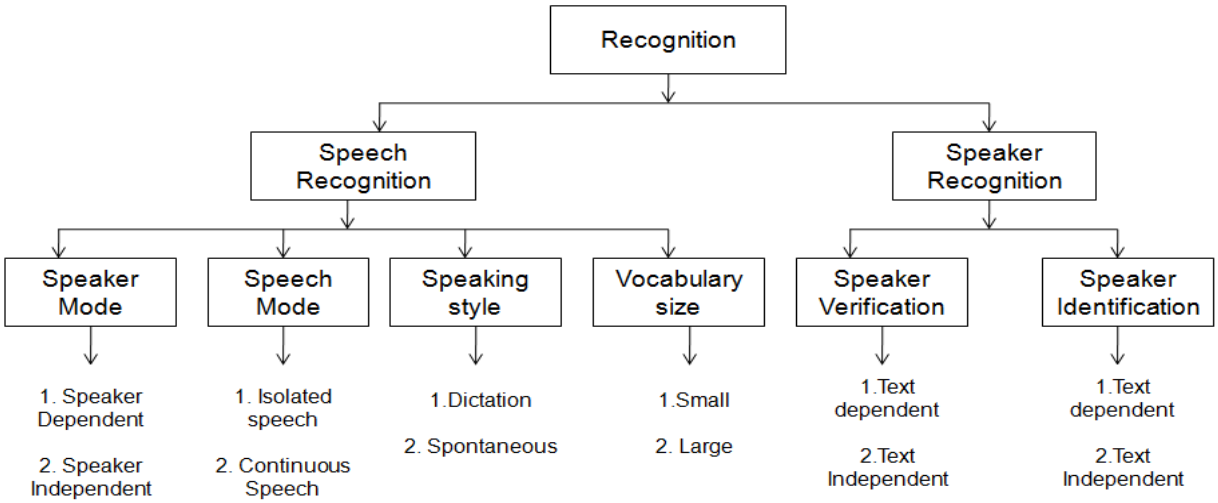
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017



C. *DIFFERENT FEATURE EXTRACTION METHODS:*

The purpose of feature extraction is to convert the speech waveform to some type of parametric representation for useful analysis and processing. This is termed as the signal-processing front end. **Table 1.** Feature Extraction Techniques.

Feature Extraction Method	**Characteristics**
Principal Component Analysis(PCA)	Basically it is Non linear feature extraction method, uses the Eigen vector.
Linear Predictive Coding.(LPC)	Static feature extraction Method, lower order linear coefficient used as features. 1st and 2nd Derivative of features is used. An important aspect of LPC is that it models well the spectral details around spectral peaks
Mel-Frequency Cepstrum Coefficient.(MFFCs) Wavelet	Cepstrum calculation and Mel scaling Bandwidth of FT with one proportional to unique frequency which has better time resolution at high frequencies than FT. 1st and 2nd Derivative of feature is used.
Spectral Subtraction (SS) and Cepstral Mean Normalization (CMN)	It uses the noise added features for robustness of system.

Different methods are used for feature extraction purpose, selection of suitable extraction algorithms are based on spectral analysis, available speech or speaker data and environment. Table 1 shows some feature extraction techniques commonly used [5, 6].

D. *FEATURE CLASSIFIERS AND MATCHING*

In speech and speaker recognition systems a supervised pattern classification system is trained with labelled examples; that is, each input pattern has a class label associated with it.

It is commonly seen as data reduction techniques. Pattern classifiers can also be trained in an unsupervised fashion. For example in a technique known as vector quantization classifier is commonly used for this purpose. Neural Network classifier is based on nearest neighbour rules also used. NN classifier not requires user specified parameters. Some other classifier which are based on distance metrics are multivariate Gaussian distributions, binomial distributions,

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

maximum likelihood estimator. Gaussian Mixture Model (GMM) is used to classify speech spurts into slow, medium and fast speech. The output likelihoods of these GMMs are used as input to a neural network whose targets are the actual phonemes [4, 5]. In the 1980s hidden Markov model (HMM) is one of the technologies developed can also be trained in an unsupervised fashion use for modelling.HMM technique is depend on high dimension feature vectors for small dimension short time speech.HMM is generally used for continuous speech with moderate vocabulary size[3].

III. SPEECH DATABASE

Variety of speech and speaker databases has a wider use in ASR. They are also used for speech synthesis, coding and analysis including speaker, language identification and verification. All these applications require large vocabulary amounts of recoded database. Following are some commonly used for Speech and speaker recognition systems.

A. *TIMIT AND DERIVATIVES*

TIMIT is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. DARPA commissioned TIMIT was designed to acoustic-phonetic knowledge and automatic speech recognition systems. It was worked on many sites including Texas Instrument (TI).

B. *TI46 DATABASE*

At Texas Instruments (TI) TI46 corpus was designed and collected. The speech was produced by 16 speakers, 8 females and 8 males, labelled f1-f8 and m1-m8 respectively, consisting of two vocabularies TI-20 and TI-alphabet. The TI-20 vocabulary contains the English digits from 0 to 9 and ten command words: enter, erase, go, help, no, robot, stop, start, and yes. The TI alphabet vocabulary contains the names of the 26 alphabets. For each vocabulary item each speaker produced 10 samples in a single training session and another two samples in each of 8 testing sessions.

C. *SWITCHBOARD*

SWITCHBOARD is a large multi-speaker corpus of telephone conversations. Although designed to support several types of speech and language research, its variety of speakers, speech data, telephone handsets, and recording conditions make SWITCHBOARD a rich source for speaker verification experiments of several kinds. Collected at Texas Instruments with funding from ARPA, SWITCHBOARD includes about 2430 conversations averaging 6 minutes in length; in other terms, over 240 hours of recorded and transcribed speech, about 3 million words, spoken by over 500 speakers of both sexes from every major dialect of American English. The data is 8 kHz, 8-bit mu-law encoded, with the two channels interleaved in each audio.

D. *YOHO*

The *YOHO* corpus was designed for evaluating speaker verification in text-dependent situation for secure access applications. This database consists of 138 speakers, out of which 106are male and 32 are female. It was recorded in different sessions over 90 days of period by fixed high-quality handset in the office environment. The text read was prompted digit phrases.

IV. CONCLUSION

Speech and speaker recognition are providing man-machine communication. Many developments are taken place in speech and speaker recognition. As speech behavioural attribute of human being, it is mostly used for access control, command control. Since last two decades researcher developed many noise robust algorithm to sustain the recognition system in any type of environment.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

REFERENCES

[1]John R. Deller, Jr., John H. L. Hansen, John G. Proakis, "Discrete-Time Processing of Speech Signals", John Wiley & Sons, inc., publication, IEEE Press.

[2] Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of speech Recognition", Prentice Hall, Englewood Cliffs, N.J., 1993.

[3] Mikael Nilsson, Marcus Ejnarsson. "Speech Recognition using Hidden Markov Model". Department of Telecommunications and Speech Processing, Blekinge Institute of Technology. 2002.

[4] L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals". Englewood Cliffs, NJ: Prentice-Hall, (1978).

[5] Douglas O'Shaughnessy "Acoustic Analysis for Automatic Speech Recognition", Proceedings of The IEEE, Vol.101, No.5,May 2013.

[6] Jr. John R. Deller, John G. Proakis, and John H. L. Hansen. Discrete-Time Processing od Speech Signals. Macmillian Publishing Company, 866 Third Avenue, New York, New York 10022, 1993

[7] G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors. EuroSpeech 97 Proceedings: ESCA 5th European Conference on Speech Communication and Technology, volume 1–5, 1998.

[8] Kjell Elenius and Johan Lindberg. FIXED1SV-FDB1000: A 1000 Speaker Swedish Database for the Fixed Telephone Network. Department of Speech Music and Hearing KTH, SE 100 44, Stockholm, Sweden, mar 1998.

.