A Study About Robust Speech Recognition And Speech Enhancements Basic Perception in Speech Processing

[1]P. R. Saranya, [2]N. Shanmugapriya
[1]Dept. of Computer science, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore
[2]Dept. of Computer Applications, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore

Abstract
Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to user's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers.

Keywords
Significances, Applications, Challenges, Speech Enhancement, Objectives, Techniques

I. Significance of Speech Recognition

The following definitions are the basics needed for understanding speech recognition technology [2].

A. Utterance

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences.

B. Speaker Dependence

Speaker dependent systems are designed around a specific speaker. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy.

C. Vocabularies/ Dictionaries

Vocabularies (or dictionaries) are lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. They can be as long as a sentence or two. Smaller vocabularies can have as few as 1 or 2 recognized utterances (e.g."Wake Up"), while very large vocabularies can have a hundred thousand or more!

D. Accuracy

The ability of a recognizer can be examined by measuring its accuracy − or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more! The acceptable accuracy of a system really depends on the application.

E. Training

Some speech recognizers have the ability to adapt to a speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy.

Training can also be used by speakers that have difficulty speaking, or pronouncing certain words. As long as the speaker can consistently repeat an utterance, ASR systems with training should be able to adapt.

II. Speech Recognition Applications

In any task that involves interfacing with a computer can potentially use Automatic Speech Recognition(ASR), the following applications are the most common they are,

A. Dictation

Dictation is the most common use for ASR systems. This includes medical transcriptions, legal and business dictation, as well as general word processing. In some cases special vocabularies are used to increase the accuracy of the system.

B. Command and Control

ASR systems that are designed to perform functions and actions on the system are defined as Command and Control systems. Utterances like "Open Netscape" and "Start a new xterm" will do just that.

C. Telephony

Some PBX/Voice Mail systems allow callers to speak commands instead of pressing buttons to send specific tones.

D. Wearable's

Because inputs are limited for wearable devices, speaking is a natural possibility.

E. Medical/Disabilities

Many people have difficulty typing due to physical limitations such as repetitive strain injuries (RSI), muscular dystrophy, and many others. For example, people with difficulty hearing could use a system connected to their telephone to convert the caller's speech to text.

F. Embedded Applications

Some newer cellular phones include C&C speech recognition that allow utterances such as "Call Home". This could be a major factor in the future of ASR and Linux..

F. Speaker Recognition Methods

Speaker recognition methods can also be divided into text-independent and text-dependent methods

- In a text-independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying.
- In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her speaking one

or more specific phrases, like passwords

III. Speech Processing Classification

The following tree structure emphasizes the speech processing applications. In speech processing, speech recognition, speaker recognition and language identification comes under recognition task. Speech recognition can be further classified according to the size of the vocabulary/ speaker mode/ speech mode/ speaking style as small, medium and large vocabulary continuous/ isolated and speaker independent/ speaker dependent/ speaker adaptive speech recognition systems. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in fig.

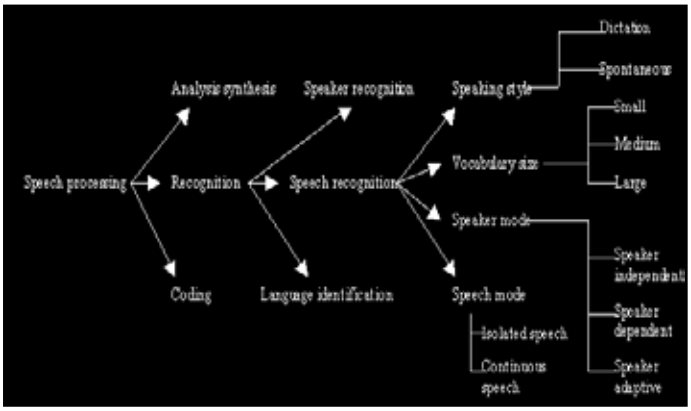

Fig. 1:

A. Speech Processing Classification

The main methodologies that made significant change in the speech recognition area are elaborated below.

B. Acoustic Phonetic Approach

This is the basis of the acoustic phonetic approach, which postulates that there exist finite, distinctive phonetic units in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time. The first step in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word from the phonetic label sequences produced by the segmentation to labeling.

C. Pattern Recognition Approach

The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns.

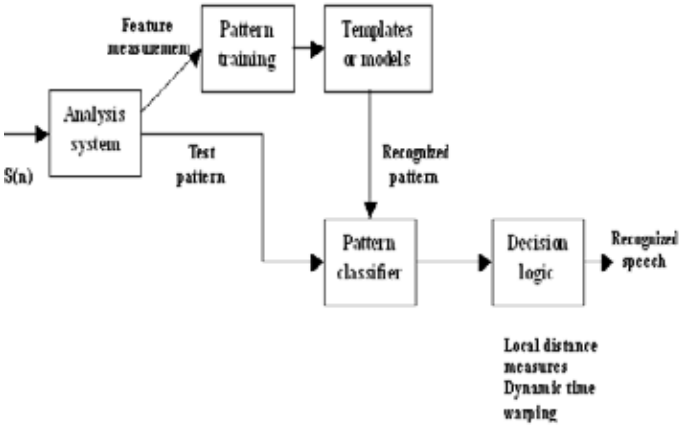

Fig. 2: Block Diagram of Pattern Recognition Approach

D. Template Based Approach

The term template is often used for two fundamentally different concepts: either for the representation of a single segment of speech with a known transcription, or for some sort of average of a number of different segments of speech. Both types of templates can be used in the DTW algorithm to compare them with a segment of input speech [3]. It has a sequence of consecutive acoustic feature vectors, a transcription of the sounds or words it represents, knowledge of neighboring templates, a tag with meta-information. Template based approaches, in which unknown speech is compared against a set of prerecorded words (template) in order to find the best match. This has the advantage of using perfectly accurate word models; but it also has the disadvantage that the prerecorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical. When considering the concrete implementation of template- based recognition, it quickly becomes apparent that the classical DTW algorithm with the Euclidean distance used as local distance metric, combined with a simple beam search will not do the job, neither from a performance nor from a computational point of view. Development of isolated word speech recognition system is based on a use of dynamic time warping (DTW) for speech pattern matching. The DTW process nonlinearly expands or contracts the time axis to match the same phoneme positions between the input speech and reference templates

E. Support Vector Machine (SVM)

Support Vector Machines are a comparatively new approach to the problems of classification, regression, ranking, etc While Artificial Neural Networks (ANNs) are widely used, Support Vector Machines (SVMs) are a comparatively new and efficient pattern recognition tool. SVMs are fast in training and guarantee a global optimum if the kernel satisfies Mercer's condition but require an appropriate choice of kernel function. ANNs are slow in training and can only guarantee local optima; the most successful solution seems to be using larger databases, trying to embed in the training set all the variability of speech and speakers. In particular, some alternative approaches, most of them based on Artificial Neural Networks . Most implementations of SVM algorithm require computing and storing in memory the complete kernel matrix of all the input samples.

F. Artificial Neural Network

Recent work on neural networks raises the possibility of new approaches to the speech recognition problem. Their use of many

processors operating in parallel may provide the computational power required for continuous-speech recognition. New neural net algorithms self-organize and build an internal speech model that maximizes performance. Auto Associative Neural Network (AANN) models for the task of speaker verification and speech recognition which produce comparable performance with that of GMM based speaker verification and speech recognition. There exists a relationship between principal component analysis and weights learned by a 3-layer AANN model. AANN model has been mostly used in applications involving dimensionality reduction.

G. Vector Quantization

Vector Quantization is a clustering technique that neglects the temporal information contained in a word in order to avoid the need for time alignment. The design of a vector quantization (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration. Given a vector source with its statistical properties known, given a distortion measure and given the number of code vectors, find a codebook and a partition, which result in the smallest average distortion. During the recognition phase the feature vectors extracted from the test word are compared to all reference codebooks. The codebook that produces the minimum distortion determines the spoken word .

H. Hidden Morkov Model

HMM is very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of application. The introduction of Hidden Morkov Models (HMMs) in the early 1980 provided much more powerful tool for speech recognition. The elements of HMM is characterized by following:

1. Number of state N
2. Number of distinct observation symbol per state
3. State transition probability,
4. Observation symbol probability distribution in state
5. The initial state distribution

IV. Challenges in Speech Recognition

Accurately and efficiently convert a speech signal into a text message independent of the device, speaker or the environment.

- Automatic generation of word lexicons.
- Automatic generation of language models for new tasks.
- Finding the theoretical limit for implementation of automatic speech recognition.
- Optimal utterance verification-rejection algorithm.
- Achieving or surpassing human performance on ASR tasks.

A. Heart of the Recognition Systems

Recognition Systems can be broken down into two main types [2].

1. Pattern Recognition Systems
2. Acoustic Phonetic Systems

B. Pattern Recognition Systems

Pattern Recognition Systems compares the patterns to known or trained patterns to determine a match.

C. Acoustic Phonetic Systems

Acoustic Phonetic Systems use knowledge of the human body to compare speech features. Most modern systems focus on the pattern recognition approach because it combines techniques and tends to have higher accuracy.

Recognizers had the following steps:

- Audio recording and Utterance detection
- Pre-Filtering
- Framing or Windowing
- Filtering
- Comparison and Matching
- Action

1. Audio Recording and Utterance Detection

It can be competent in a number of ways. Starting points can be found by comparing ambient audio levels with the sample just recorded. End point detection is harder because speakers tend to leave "artifacts" including breathing, teeth chatters, echoes.

2. Pre-Filtering

Pre-Filtering had a variety of ways, depending on other features of the recognition system.The most common methods are the "Bank-of-Filters" method which utilizes a series of audio filters to prepare the sample and the linear predictive coding method which uses a prediction function to calculate different errors. Different forms of spectral analysis are also used.

3. Framing or Windowing

It involves separating the sample data into specific sizes. This is often rolled into steps. This step involves preparing the sample boundaries for analysis.

4. Filtering

It is not always present. It is final preparation for each window before comparison and matching. It also consists of Time Alignment and Normalization.

D. Speech Enhancement

Speech enhancement aims to improve speech quality by using various Algorithms. The objective of enhancement is improvement in intelligibility or overall perceptual quality of degraded speech signal using Audio signal processing techniques.

Speech enhancement improves the quality of signals corrupted by the adverse noise, channel distortion such as competing speakers, background noise, car noise, room reverberations and low-quality microphones. A broad range of applications includes mobile communications, robust speech recognition, low-quality audio devices and aids for the hearing impaired.

1. Speech enhancement depends on good signal processing technique.
2. It also depends on human perceptual factor
3. Speech quality and intelligibility are dependent on short term spectral amplitude and unfeeling to spectral phase.

E. Objective of Speech Enhancement

1. Improve the performance of the voice communication device
2. To boost overall speech quality
3. To reduce listener fatigue
4. To increase intelligibility
5. Improving the speech quality and minimizing any speech intelligibility loss.

F. Speech Enhancement Techniques

There are four classes of Speech enhancement methods each with its own advantages and limitations subtraction of interfering sounds, filtering out such sounds, suppression of non harmonic

frequencies and resynthesis using a vocoder.
The first and most popular method simply estimates

1. The important speech related components of the distorted input speech signal.
2. The corrupting portions of the signal.

G. Spectral Subtraction and Filtering

If an interfering sound can also be captured apart from the desired speech and later is usually enhanced by subtracting out a version of the former. Best results usually require a second microphone placed closer to the noise source other than the primary microphone recording the desired speech. The second recording provides the noise reference, after processing from the primary recordings.

H. Harmonic Filtering

The Harmonic Speech enhancement attempts to identify the entire desired speech. If the desired sound is the strongest component in the signal, its frequencies can be identified and other frequencies can be suppressed. Such simple Wiener filtering improves SNR but has little effect on intelligibility

I. Parametric Resynthesis

The last Speech enhancement method improves speech signal by parametric estimation and speech resynthesis. Speech synthesizers generate noise-free parametric representation of either a vocal tract model or previously analysed speech.

V. Conclusion

In this Speech Recognition and Speech enhancement were studied and intricate clearly it can be useful for signal processing and recognition of speech and it can be converted into text based on these basic perceptions as a base for every actions. In Future Work would be based on comparing enhancement and recognition perceptions.

References

[1] Douglas O'Shanghunessy,"Speech communications, Human and machine".
[2] L.R.Rabiner, R.W. Schaffer,"Digital Processing of Speech signals", Prentice.
[3] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang "Springer handbook of speech".

P.R,Saranya had received B.COM(CA) in Commerce with Computer Applications in the year 2007 and also received Master of Computer Applications(MCA) in the year 2011. Her research interests include Digital signal Processing, Speech Recognition and Speech Enhancement. At present she is a research scholar in Speech Processing.

N.Shanmugapriya is an Assistant Professor in MCA Department in Dr.SNS Rajalakshmi college of Arts and Science. She had finished his UG in the year 1999 and PG of M.Sc (CT) in the year 2001 done at Bharathiar University and M.Phil in the year 2007. She had presented many papers and attended conferences in the field of Speech Recognition and Image processing areas. She had interested in Speech Processing, Speech Recognition and Speech Enhancements. She is currently researching Doctorate (Ph.d) in the field of Speech Recognition.