INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

A REVIEW ON AUTOMATIC SPEECH RECOGNITION USING NEURALNETWORKS

NITIN BHAVESH(178R1A0466), AMARENDRA GANTA(178R1A0480),
R. RAMGOPAL(178R1A04B1), R. VAISHNAVI(178R1A04B2)
Department of ECE, C M R Engineering college,
Hyderabad, Telangana, INDIA .

Abstract- Automatic speech recognition acknowledges the spoken words and converts them to a machine-readable format of text. By converting spoken audio into text, this technology allows users to control digital devices by speaking instead of using conventional tools like keystrokes and buttons. The challenges in speech recognition are the improvisation of the accuracy, varying user responsiveness, performance, reliability and fault tolerance. Since the world is moving at a rapid pace towards digitization, new technologies are being developed to make lives easy. Interactive Voice Response System is an example. Speech recognition (SR) is the use of voice inputs into a computing device. The device converts the speech signal to a format that computers can process. Speech recognition allows use of verbal commands, which may be a necessity if the user has physical disabilities. Essentially, it works by storing a human voice and training an automatic speech recognition system to recognize vocabulary and speech patterns in that voice. Feature extraction is accomplished by changing the speech waveform to a form of parametric representation at a relatively lesser data rate for subsequent processing and analysis. This is usually called the front end signal-processing Feature extraction is the most relevant portion of speaker recognition.

An acoustic model is used in automatic speech recognition to represent the relationship between an audio signal and the phonemes or other linguistic units that make up speech. The model is learned from a set of audio recordings and their corresponding transcripts. The language model provides context to distinguish between words and phrases that sound similar. In speech recognition, sounds are matched with word sequences. Ambiguities are easier to resolve when evidence from the language model is integrated with a pronunciation model and an acoustic model. Frequency Domain Multi-channel Acoustic Modeling for distant Speech Recognition. Conventional far field automatic speech recognition (ASR) systems typically employ microphone array techniques for speech enhancement in order to improve robustness against noise or reverberation. Speech recognition (SR) is the use of voice inputs into a computing device. The device converts the speech signal to a format that computers can process. This whole thing is done using MATLAB software.

Key words- Speech recognition, MATLAB software, automatic speech recognition

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

I. INTRODUCTION

In Information processing machines have become ubiquitous. However, the current modes of human machine communication are geared more towards living with the limitations of computer input/output devices rather than the convenience of humans. Speech is the primary mode of communication among human beings. On the other hand, prevalent means of input to computers is through a keyboard or a mouse. It would be nice if computers could listen to human speech and carry out their commands.

Automatic Speech Recognition (ASR) is the process of deriving the transcription (word sequence) of an utterance, given the speech waveform. Speech understanding goes one step further, and gleans the meaning of the utterance in order to carry out the speaker's command. The field of automatic speech processing has extensively been investigated over the past fifty years or so. Automatic speech recognition (ASR) is one of the different topics which have attracted particular attention and research effort. In spite of this and not withstanding important progress and cumulated improvements, the performance of ASR systems, under natural and realistic conditions, is still far inferior to human capabilities.

ASR is still among the most challenging areas of artificialintelligence and pattern recognition. The search for solution to the problem of automatic speech recognition has led to the development of a large number of techniques and models. Artificial Neural Networks (ANNs) have been applied over the past decade with some success. There is no surprise about this. Indeed for a long time, ANNs were successfully used to solve complex problems of pattern classification and recognition. Review of literature on speech recognition systems genuinely demands the very first attention towards the discovery of Alexander Graham Bell about the process of converting sound waves into electrical impulses and the first speech recognition system developed by Davis et al. for recognizing telephone quality digits spoken at normal speech rate. This effort for automatic recognition of speech was basically centered on the building up of an electronic circuit for recognizing ten digits of telephone quality. Spoken utterances were analyzed to get a 2-dimensional plot of formant 1 vs. formant 2.

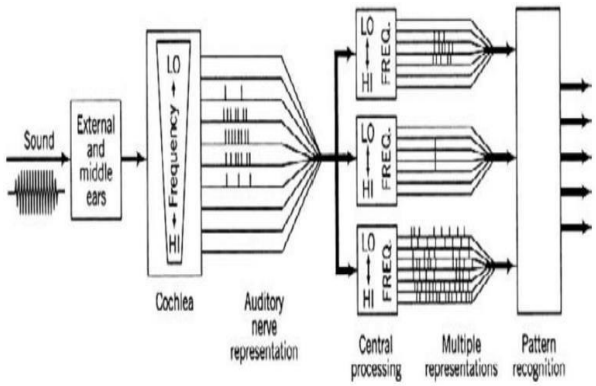
For pattern matching, a circuit was designed for determining the highest relative correlation coefficient between a set of new incoming data and each of the reference digit patterns. It was also observed that circuit adjustment helps the recognition system to perform well

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

for the speech of different speakers. An indication circuit was built to display the recognized spoken digit. The approaches to speech recognition, evolved thereafter, had a major stress on finding speech sounds and providing appropriate labels to these sounds. Various approaches and types of speech recognition systems came into existence in last five decades gradually. This evolution has lead to a remarkable impact on the development of speech recognition systems for various languages worldwide. Automatic speech recognition has been viewed as successive transformations of acoustic micro structure of speech signal into its implicit phonetic macro-structure. In other words, a speech recognition system is a speech-to-text conversion wherein the output of the system displays text corresponding to the recognized speech. Languages, on which so far automatic speech recognition systems have been developed, are just a fraction of total around 7300 existing languages. Russian, Portuguese, Chinese, Vietnamese, Japan, Spanish, Filipino, Arabic, English, Bengali, Tamil, Malayalam, Sinhala, Hindi are prominent among them. English is the language for which maximum work for recognition is done.Speech is probably

II. INTRODUCTION FUNCTION OF HUMAN RECONGITION

Here, Schematic view of the human ear showing the three distinct sound processing sections, namely: the outer ear consisting of the pinna, which gathers sound and conducts it through the external canal to the middle ear; the middle ear beginning at the tympanic membrane, or eardrum, and including three small bones, the malleus (also called the hammer), the incus (also called the anvil) and the stapes (also called the stirrup), which perform a transduction from acoustic waves to mechanical pressure waves; and finally, the inner ear, which consists of the cochlea and the set of neural connections to the auditory nerve, which conducts the neural signals to the brain



Depicts a block diagram abstraction of the auditory system. The acoustic wave is transmitted from the outer

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

ear to the inner ear where the ear drum and bone structures convert the sound wave to mechanical vibrations which ultimately are transferred to the basilar membrane inside the cochlea. The basilar membrane vibrates in a frequency- analysis of the sound selective manner along its extent and thereby performs a rough spectral analysis of the sound.

Distributed along the basilar membrane are a set of inner hair cells that serve to convert motion along the basilar membrane to neural activity. This produces an auditory nerve representation in both time and frequency. The processing at higher levels in the brain, shown in Figure

3.2 as a sequence of central processing with multiple representations followed by some type of pattern recognition, is not well understood and we can only postulate the mechanisms used by the human brain to perceive sound or speech.

Even so, a wealth of knowledge about how sounds are perceived has been discovered by careful experiments that use tones and noise signals to stimulate the auditory system of human observers in very specific and controlled ways. These experiments have yielded much valuable knowledge about the sensitivity of the human auditory system to acoustic properties such as intensity and frequency

and consist of an acoustic processor which analyze the speech signal and converts it into a set of acoustic (spectral, temporal) features, X, which efficiently characterize the speech sounds, followed by a linguistic decoding process which makes a best (maximum likelihood) estimate of the words of the spoken sentence,resulting in the recognized sentence w^ .

Once the words are chosen, the speaker sends appropriate control signals to the articulator speech organs which form a speech utterance whose sounds are those requiredto speak the desired sentence, resulting in the speech waveform s[n]. We refer to the process of creating the speech waveform from the speaker's intention as the Speaker Model since it reflects the speaker's accent and choice of words to express a given thought or request.

The processing steps of the Speech Recognizer areshown at the right side of Figure

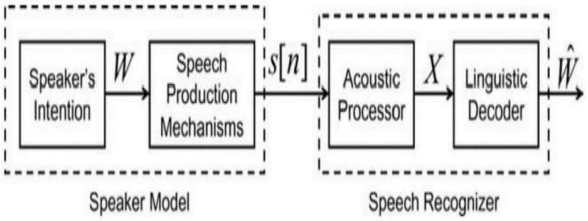


Fig 4.2 Process or procedure for Speech Recognition

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

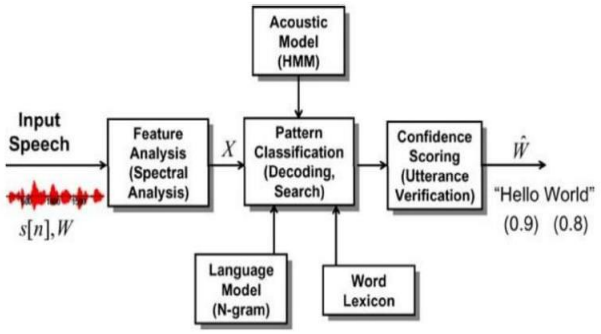


Fig 4.3| Block Diagram or layout of Speech Recognition

Figure-4.3 shows a more detailed block diagram of the overall speech recognition system. The input speech signal, s[n], is converted to the sequence of feature vectors, X ={x1,x2,...,xT }, by the feature analysis block (also denoted spectral analysis). The feature vectors are computed on a frame-by-frame basis using the techniques discussed in the earlier chapters. In particular, the MEL FREQUENCY COEFFICIENTS(MFCC)are widely used to represent the short-time spectral characteristics.

The pattern classification block (also denoted as the decoding and search block) decodes the sequence of feature vectors into a symbolic representation that is the maximum likelihood string, Wˆ that could have produced the input sequence of feature vectors. The pattern recognition system uses a set of acoustic models (represented as hidden Markov models) and a word lexicon to provide the acoustic match score for each proposed string. Also, an N-gram language model is used to compute a language model score for each proposed word string. The final block in the process is a confidence scoring process (also denoted as an utterance verification block), which is used to provide a confidence score for each individual word in the recognized string.

Each of the operations in Figure 4.4 involves many details and, in some cases, extensive digital computation. The remainder of this chapter is an attempt to give the flavour of what is involved in each part of Figure 4.4.

FEATURE ANALYSIS

Mel-Frequency Cestrum (MFFCs): Mel Frequency Cepstral Coefficients are based on the known variations of the human ear's critical bandwidths with frequencies which are below a 1000 Hz. The main purpose of the MFCC processor is to copy the behavior of human ears. The derivation of MFCCs is done by the following Figure 4.5

An audio signal is constantly changing, so to simplify things we assume that on short time scales the audio signal doesn't change much (when we say it doesn't change, we mean statistically i.e. statistically stationary, obviously the samples are constantly changing on even short time scales). This is why we frame the signal into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

it is longer the signal changes too much throughout the frame.

The next step is to calculate the power spectrum of each frame. This is motivated by the human cochlea (an organ in the ear) which vibrates at different spots depending on the frequency of the incoming sounds. Depending on the location in the cochlea that vibrates (which wobbles small hairs), different nerves fire informing the brain that certain frequencies are present. Our periodogram estimate performs a similar job for us, identifying which frequencies are present in the frame.
The periodogram spectral estimate still contains a lot of information not required for Automatic Speech Recognition (ASR). In particular the cochlea cannot discern the difference frequencies increase. For this reason we take clumps of periodogram bins and sum them up to get an idea of how much energy exists in various frequency regions. This is performed by our Mel filter bank the first filter is very narrow and gives an indication of how much energy exists near 0 Hertz. As the frequencies get higher our filters get wider as we become less concerned about variations. We are only interested in roughly how much energy occurs at each spot. The Mel scale tells us exactly how to space our filter banks and how wide to make them.
Once we have the filter bank energies, we take the logarithm of them. This is also motivated by human hearing: we don't hear loudness on a linear scale. Generally to double the perceived volume of a sound we need to put 8 times as much energy into it. This means that large variations in energy may not sound all that different if the sound is loud to begin with. This compression operation makes our features match more closely what humans actually hear. Why the logarithm and not a cube root? The logarithm allows us to use cepstral mean subtraction, which is a channel normalization technique.
The final step is to compute the DCT of the log filter bank energies. There are 2 main reasons this is performed. Because our filter banks are all overlapping, the filter bank energies are quite correlated with each other. The DCT decor relates the energies which means diagonal covariance matrices can be used to model the features in

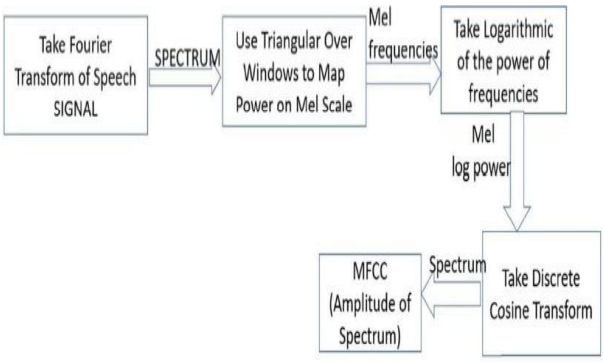


Fig 4.5 Derivates of Mel Frequency cepstral Coefficients

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

e.g. a HMM classifier. But notice that only 12 of the 26 DCT coefficients are kept. This is because the higher DCT coefficients represent fast changes in the filter bank energies and it turns out that these fast changes actually degrade ASR performance, so we get a small improvement by dropping them.

CONCLUSION

In conclusion, human speech recognition is a way of enabling a computer to decode human voice. In Speech Recognition a person must perform activities such as filtering and identifying specific noise from certain signal so as to recognize the speech. Speech does not have natural pauses between the boundaries of the word. Another feature I would want the synthetic speech to have is the ability to identify the origin of the sound, for instance, to be able to identify the specific object.
Speech recognition is a way of enabling a computer to decode human voice in more precise and perfect way. The computer translates the analog waves of voice into digital by analyzing the sound.

FUTURE SCOPE

More applications are being made specifically to be compatible with smart devices, such as smart household appliances. They can now help humans with tasks around the home such as controlling the heating, turning lights on and off and using entertainment systems. The future of voice recognition is looking bright.
Voice recognition devices are a growing trend among customers.

REFERENCES

[1]. Automatic Speech Recognition by k.Samudravijaya Tata Institute of FundamentalResearch
[2].Speech Recognition using deep Neural Networks by A.B.NASSIF in IEEE ACCESSS.
[3].MATLAB Deep Learning textbook by Phim Kim.
[4]. Introduction to Digital Speech Processing text book by Lawrence R. Rabiner and Ronald
[5]. Deep Learning for NLP and Speech Recognition Textbook by James Whitaker, John Liu [6]. Automatic Speech Recognition: A Deep Learning Approach by Dong Yu and Li Deng
[7]. Fundamentals of speech recognition by Lawrence Rabiner
[8]. Speech Recognition and Processing: Algorithms and Applied Principles by Marcus Hintz
[9]. Automatic speech recognition an IEEE Journel published at CHILEON by J Meng
[10]. Using Speech Recognition Software
Book by Calais J. Ingel
[11]. Advanced Speech Recognition: Concepts and Case Studies by Marcus Hintz
[12]. Designing Voice User Interfaces: Principles of Conversational Experiences Book by Cathy Pearl [13]. Audio Processing and Speech Recognition: Concepts, Techniques and Overviews Book by AnjanDutta, Nilanjan Dey, and Soumya Sen

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

[14]. How to Build a Speech Recognition Application: A Style Guide for Telephony Book by Bruce Balentine andDavid P. Morgan
[15]. Prosody and speech recognitionBook by Alex Waibel