International Journal of Engineering Research & Management Technology ISSN: 2348-4039

January-2017 Volume- 4, Issue-1

Email: editor@ijermt.org www.ijermt.org

# MANIPULATIVE AN AUTOMATED SPEECH RECOGNITION SYSTEM USING NEURAL NETWORKS

**[1]Vennu Santosh Kumar, [2]Dr. K.P. Yadav**

[1]Research Scholar,[2] Research Supervisor

[1,2]Department of Computer Science & Engineering

[1,2]Sunrise University, Rajasthan, Rajasthan

**ABSTRACT:**

We learn all the relevant skills during early childhood, without instruction, and we continue to rely on voice communication throughout our lives. It comes so naturally to us that we don’t realize how complex a phenomenon voice is. The human vocal tract and articulators are biological organs with nonlinear a property, whose operation is not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state. As a result, vocalizations can vary widely in terms of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed; moreover, during transmission, our irregular speech patterns can be further distorted by background noise and echoes, as well as electrical characteristics.

**KEYWORDS:** Voice, ranging, electrical

## INTRODUCTION:

Broadly speaking, speech recognition systems are usually built upon three common approaches, namely, the acoustic-phonetic approach, the pattern recognition approach and the artificial intelligence approach. The acoustic-phonetic approach attempts to decide the speech signal in a sequential manner based on the knowledge of the acoustic features and the relations between the acoustic features with phonetic symbols. The pattern recognition approach, on the other hand, classifies the speech patterns without explicit feature determination and segmentation such as in the formal approach. The artificial intelligence approach forms a hybrid system between the acoustic-phonetic approach and the pattern-recognition approach.

The artificial intelligence approach becomes the field of interest after seeing the success of this approach in solving problems (especially classification problems). The application of artificial neural networks is proposed to meet the needs of an accurate speech recognizer. For example, the neural network approach to phoneme recognition is proposed in Japanese vowel recognition. Besides, the combination of neural networks and linear dynamic models is proven in achieving a high level of accuracy in automatic speech recognition systems. Another problem in speech recognition is the increase of error in the presence of noise such as in a typical office environment. Some researchers propose the use of visual information such as the lip movement. In this case, image processing techniques and neural networks are applied to capture and analyze lip movement. Digit recognition is one of the common applications in this field, for example, mandarin digit recognition systems have been actively developed by researchers in China. Different systems have been proposed to recognize digits of different languages.

The application of neural networks in the pattern-recognition approach is discussed. We propose the use of a multilayer perception (MLP), which is trained using the back-propagation technique to be the engine of an automated digit recognition system. Firstly, the features of the training datasets are extracted automatically using the end-point detection function. The features are then used to train the neural network. The same function is used to extract the features of signals during the recognition stage. Several networks with different structures (different numbers of neurons) were trained with different numbers of samples and the performance in recognizing the unknown input patterns were compared. The system was built using MATLAB and accuracy greater than 95% was achieved for the unknown patterns.

The following section discusses the stages in designing the automatic speech recognition system. Firstly, the speech signal properties are discussed, followed by the end point detection method in finding the region of interest from the raw speech data. After the start point and the end point of a speech signal have been detected, it is then analyzed by various methods. The LPC method is used to represent the features of the speech signal which has been blocked into frames. Besides, by referring to the start point and end point of the signals, a finite number of frames is selected to become the input for the neural network. Finally, a comparison of the performance for various networks with different numbers of training datasets and different numbers of neurons was done.

## REVIEW OF LITERATURE

There are two basic approaches of using neural networks in speech classification, which are the static approach and the dynamic approach. In the static approach, the neural network accepts all input speech data at once, and makes a single decision. On the other hand, for the dynamic approach, the neural network processes a small window of the speech at one time, and this window slides over the input speech data while the network makes a series of local decisions, which have to be integrated into a global decision at a later time. Both approaches are being applied in phoneme recognition as well as word level recognition. In this paper, a neural network will be used to recognize digits at the word level.

Peeling and Moore (1987) applied Multilayer Perceptrons to digit recognition with excellent results. A static input buffer of 60 frames of spectral coefficients is applied in which the briefer words were padded with zeros and positioned randomly in the 60-frame buffer. By evaluating different network topologies, a single hidden layer with 50 units was found to perform efficiently. A performance of 99.8% was found in speaker-dependent experiments and 99.4% was found for multi-speaker experiments. Kammerer and Kupper (1988) found that single-layer perceptrons outperformed both multi-layer perceptrons and a dynamic time warping (DTW) template-based recognizer in many cases. A static input buffer of 16 frames was applied in which each word was linearly normalized, with sixteen 2-bit coefficients per frame. The system achieved the performance of 99.6% in speaker-dependent experiments and 97.3% for speaker-independent experiments.

Burr (1988) applied Multilayer Perceptrons in a more difficult task, alphabet recognition. A static input buffer of 20 frames was applied, in which each spoken letter was linearly normalised, with 8 spectral coefficients per frame. Training on three sets of the 26 spoken letters and testing on a fourth set, the performance achieved was 85% in speaker dependent experiments, matching the accuracy of a dynamic time warping (DTW) template-based approach.

## SPEECH SIGNAL REPRESENTATION

A speech signal is usually classified into three states. The first state is silence, where no speech is produced.

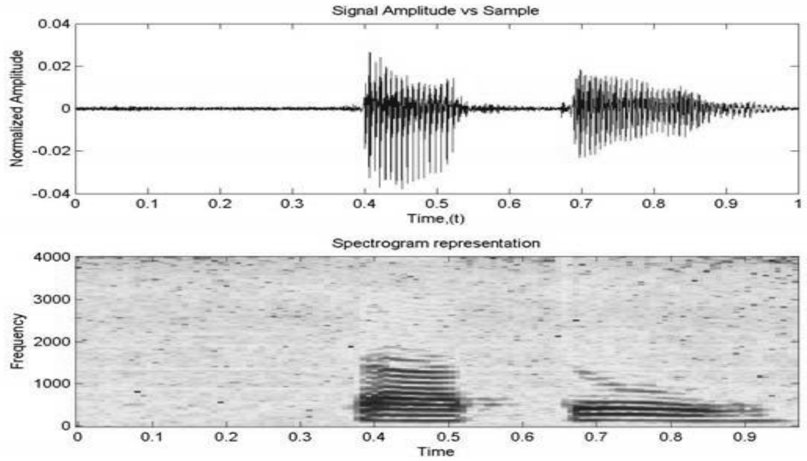
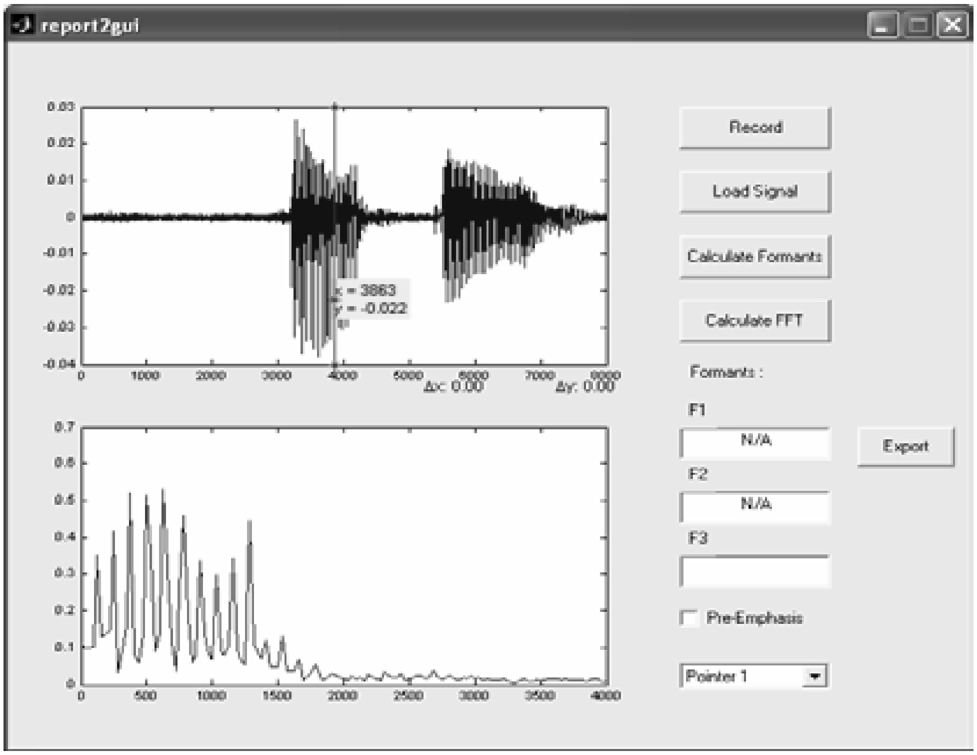


*Figure1: Spectrogram analysis by using FFT*

The second state is unvoiced, in which the vocal cords are not vibrating and the resulting signal is random in nature. The last state is voices, in which the vocal cords vibrate and produce a quasi-periodic signal. The silence state is usually the unwanted state and has to be removed in order to save the processing time of the speech recognition system as well as to improve the accuracy of the system. In the time domain, the amplitude of the speech signal at each sampling time is plotted over time. This representation gives the picture on how a speech varies over time, and requires large storages. Spectral representations illustrate the nature of speech signals in terms of their frequency contents. Figure 1 shows the spectrogram of a speech signal which corresponds to performing a fast Fourier transform on every 256 samples (32ms) with the analysis advancing in intervals of 64 samples (8ms).

For the sake of analysis, the speech signal is usually broken into frames. This has been applied in the spectrogram shown above in which the frequency contents of all frames are arranged one next to the other to form a three dimensional representation (the colors represent the third dimension). The frequency information of a specific frame can also be obtained by taking the fast Fourier transform of the specified frame. An analysis tool is built using MATLAB to perform this task. Figure 5.2 shows the frequency contents of a frame with 256 samples.



Formants :

F1 N/A

F2 N/A

F3

□ Pre-Emphasis

Pointer 1 ▾

Figure 2: Fast Fourier Transforms of a frame with 256 samples (from n=3863–127 to n=3863+128)

## CONCLUSION:

In this paper, the approach of using neural networks for speaker independent isolated word recognition has been studied. Besides, an automatic speech recognition system has been designed using MATLAB programming. By the fully automated training and recognition process without the interference of manual cropping, an accuracy of almost 85% is achieved for unknown pattern (spoken by unknown speakers). This opens the door to the implementation in embedded systems, which requires small programs and simple algorithms for certain applications.

The results show that the performance of a network improves when more training datasets are used to train the network. Besides, the networks that use the LPC coefficients as inputs also perform better than the networks that use the first three formants as the networks' inputs.

For large vocabulary systems, this approach can also work together with other models to achieve higher accuracy. For example, it can be modified in order to recognize the phonemes in the speech signal and work with Hidden Marker Models (HMM) to recognize mandarin monosyllables.

## REFERENCES:

1. Rumelhart D. E., Hinton G. E. & William R. J., "Learning Internal Representation by Error Propagation", In Rumelhart D. E. & McClelland J. L. (Eds.), Parallel Distributed Processing, 1, MIT Press, 318 – 362, (1986).
2. D. O. Hebb, "The organization of Behaviour: A Neuropsychological Theory", New York: Wiley, (1949).
3. S. Grossberg, "Some networks that can learn, remember, and reproduce any number of complicated space-time patterns", J. Math. Mech., **1 (19)** (1969).
4. J. J. Hopfield, "Neural networks and physical systems with emergent collective computational capabilities" in Proceedings National Academy of Sciences (USA) **79** p2554-2558 (1982).
5. Pandey A., Bansal K.K. "Performance Evaluation of TORA Protocol Using Random Waypoint Mobility Model" International Journal of Education and Science Research Review Vol.1 (2) pp 193-199
6. Yadav R.K., Bansal K.K. "Analysis of Sliding Window Protocol for Connected Node "International Journal of Soft Computing and Engineering (IJSCE) Vol.2 (5) PP. 292-294
7. Tiwari S.P.,Kumar S.,Bansal K.K.(2014) "A Survey of Metaheuristic Algorithms for Travelling Salesman Problem " International Journal Of Engineering Research & Management Technology Vol.1(5)
8. Kumar N.,Bansal K.K.,(2012) "Different Compression Techniques and Their Execution In Database Systems To Improve Performance" International Journal of Advanced Research in Computer Science and Software Engineering Vol.2(6) pp. 293-296
9. Gene Bylinsky, *Computers That Learn By Doing*, Fortune, September 6, 1993.
10. J. M. Zurada (1992), "Introduction to Artificial Neural Systems". Singapore: Info Access and Distribution.
11. Yao X., "Evolving Artificial Neural Networks", Proceedings of the IEEE, 87, 1423 – 1447, (1999).
12. Nolfi S., & Patisi D., "Learning to adapt to changing environments in evolving neural networks", Adaptive Behaviour, 5(1), 75 – 98, (2010).