PalArch's Journal of Archaeology of Egypt / Egyptology

SPEECH RECOGNITION WITH STATIC AND DYNAMIC APPROACHES IN ARTIFICIAL NEURAL NETWORK

[1]**Dr. Shobha Lal , Professor and Dean,** Deprtment of Computer science and Engineering, *Jayoti Vidyapeeth Women's University*, Jaipur, India

[2]**Dr. Kavita,** Research Supervisor, Deprtment of Computer science and Engineering, *Jayoti Vidyapeeth Women's University*, Jaipur, India

[3]**Jitendra Joshi,** Research Scholar, Deprtment of Computer science and Engineering, *Jayoti Vidyapeeth Women's University*, Jaipur, India

*Corresponding author. E-mail: lect.jitendra29@gmail.com

Dr. Shobha Lal, Dr. Kavita, Jitendra Joshi, Speech Recognition with Static and Dynamic approaches in Artificial Neural Network-Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(12), ISSN 1567-214x

Abstract

In this research paper, speech recognition refers to the identification of utterances through the movements of lips, tongue, teeth, and other facial muscles of the speaker without using the acoustic signal. This work shows the relative benefits of both static and dynamic approaches with speech features for improved visual speech recognition system. Artificial Neural Networks (ANN) is biologically inspired computer programs that simulate the way the human brain processes information. Artificial Neural Networks (ANN) gathers their knowledge by recognizing patterns and relationships in the data, speech and learning or training them through experience of speech.

Keywords: Speech, Static, Dynamic, Frame, Time Alignment**,** Signal Analysis.

1. Introduction

Speech recognition is become very complex with its large number of characteristics; these characteristics should be considered to find solution related speech recognition.

1.1. Vocabulary size and confusability

Efficiency of speech recognition comes with the small size of vocabulary. If we increase number of words that need to be recognized than confusion and error will raise proportionally.

1.2. Speaker dependence v/s independence

If we make a speech recognition system dependent upon single speaker than it will be much easier to recognize the speech. While our system is dealing with multiple speakers, it becomes very complex due to different voices. Different voices have different values on several factors.

1.3. Isolated, discontinuous, or continuous speech

Complexity of a speech recognition system also depends how we are providing input to the system. Here isolated speech means single word input to the system. Discontinuous speech is a full sentence with silence separated words. Continuous speech is naturally spoken sentences. It is easier to understand that the speech recognition system will give best output with isolated words over a word that is part of continuous speech.

1.4. Read v/s spontaneous speech

Read speech will be prepared before when preparing speech to be recognized by a system so it consists no error. Spontaneous speech has possibility to contain mispronounced words or incomplete sentences and it is more difficult to recognize.

2. Fundamentals of Speech Recognition

Multilevel pattern recognition applies upon speech recognition. In which acoustical signals are examined and structured into a hierarchy of sub-word units, words, phrases and sentences. Each level may provide temporal constraints based upon known pronunciations or allowable word sequences.

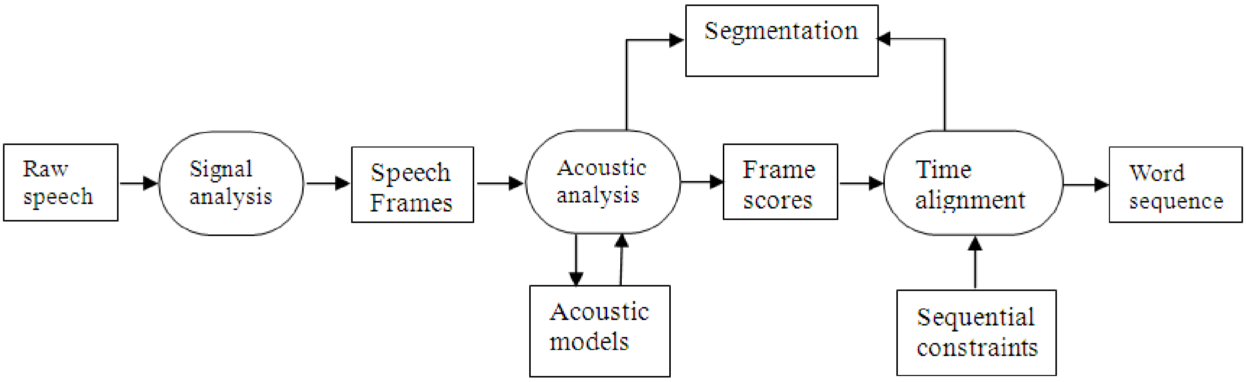

Figure 1: Structure of a standard speech recognition system

2.1. Raw Speech

Speech is sampled at a standard frequency of 16 KHz over a microphone. We sample this upon a sequence of amplitude values over time.

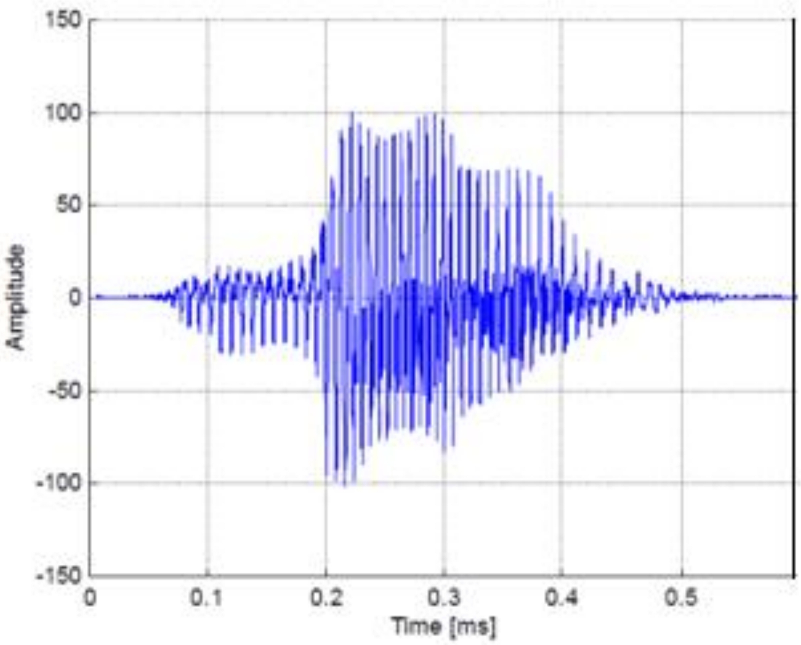

Figure 2: Speech Signal for the word ‘ZERO’.

2.1. Signal Analysis

Transformation and compression process applied on sampled raw speech to simplify the recognition process. Fourier analysis, Perceptual Linear Prediction, Linear Predictive Coding and Cepstral analysis all properly process the raw speech into a more usable state.

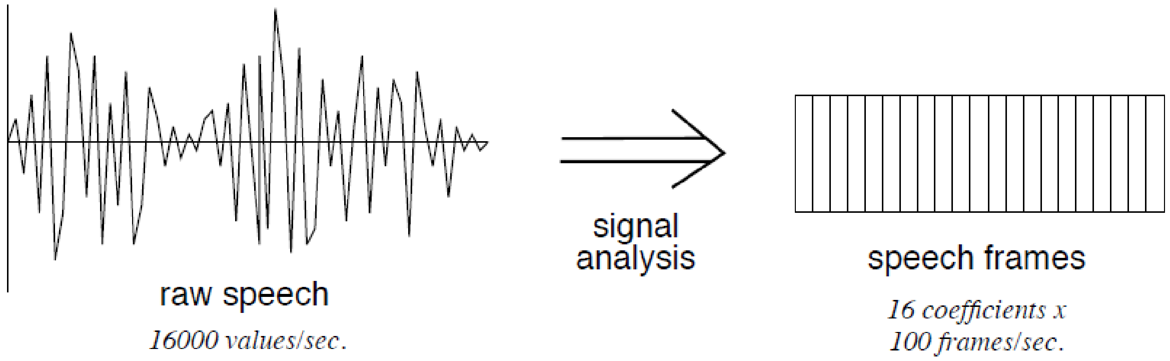

Figure 3: Signal analysis converts raw speech to speech frames

3. Static and Dynamic Classification

In neural networks there are two approaches for speech recognition. One is static classification and second is dynamic classification. In static classification neural network can access all the input speech at the same time and can make decision based on this input. In dynamic classification, neural network can see only a small window of the input speech. The small view window directs to a series of decisions that can be combined over the entire speech input. If neural network is being applied over isolated words or phonemes than static classification works well. However, it is proven that dynamic classification much better at classifying words or sentences that are spoken together.

3.1. Static Classification Approach

Dr. Richard Lippmann and Dr. William Huang demonstrated that neural networks are able to form complex decisions from speech inputs. Using a multilayer perceptron with 2 inputs, 50 hidden neurons and 10 outputs, the network was able to process spoken vowels and form decision regions. After 50,000 iterations of training the neural network, the decision regions became optimal and yielded very promising classification results.

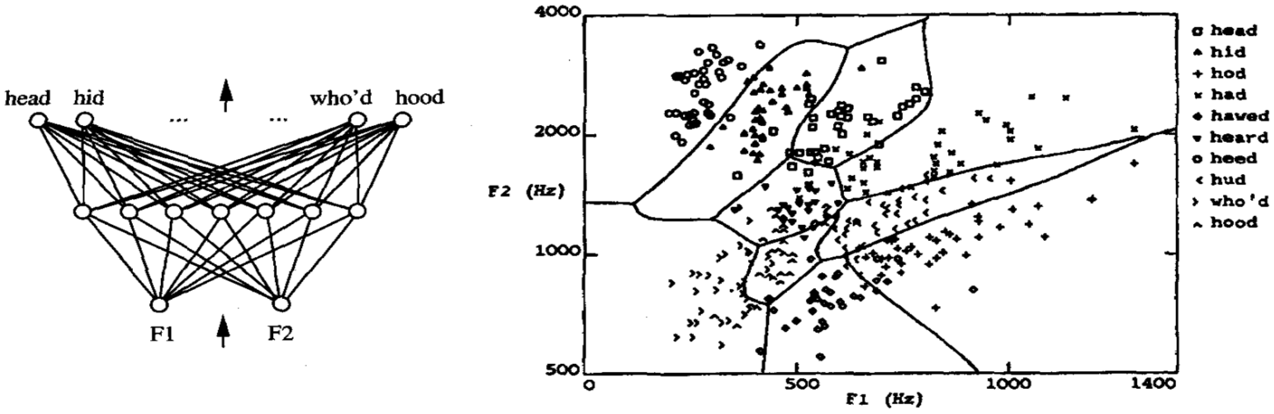

Figure 4 : The multilayer Perceptron and Decision Regions

3.2. Dynamic Approach

In English alphabet E-set of English Letters, “B,C,D,E,G,P,T,V and Z” are most difficult to implement in speech recognition. While in processing with neural network these letters with same characteristic, an 8% error rate would be consider as good result.

Dr. Alexander Waibel experimented with Time Delay Neural Network and had excellent results in phoneme recognition. He created a structure consisted of one input layer with three delays and hidden layer of five delays. The final output was computed by integrating over 9 frames of phoneme activations in the second hidden layer.

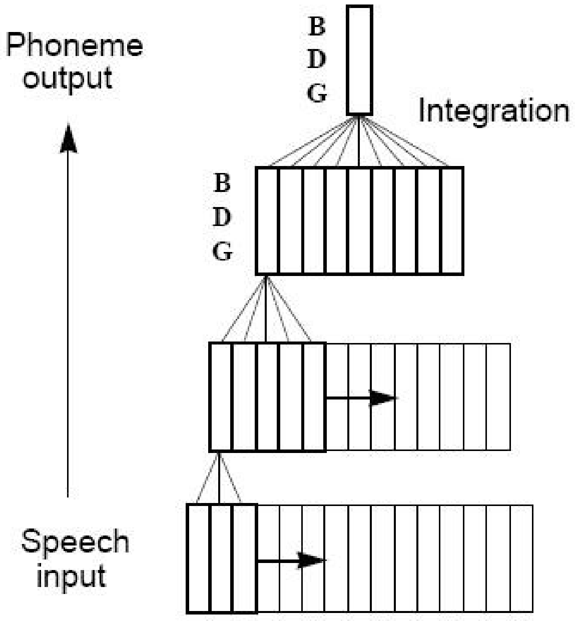

Figure 5: Time Delay Neural Network Structure

2000 samples of phonemes /b, d, g/ were tested and trained the network by manually excised from a database containing 5260 words. In the

Waibel achieved an error rate of only 1.5%, compare with 6.5% achieved by a simple Hidden Markov Model recognition system.

4. Conclusion

The main objective of this research work is to provide a comparative analysis and implementation of the most popular speech feature extraction techniques for Hidden Markov Model recognition system. Neural Network is considered as the most dominant pattern recognition techniques used in the field of speech recognition.

5. References

1. A. Waible, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, "Phoneme Recognition Using Time-Delay Neural Networks", *IEEE Trans. on ASSP*, vol. 37, no. 3, pp. 328-339.
2. G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Trans, on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30-2, 2012.
3. K. Tokuda, T. Kobayashi, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter generation algorithms for HMM-based speech synthesis", InProc. ICASSP, pp. 1315-1318,2000.
4. G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82-97,2012.
5. Bengio, Y. and Senecal, J.-S. Adaptive Importance Sampling toAccelerate Training of a Neural Probabilistic Language Model.IEEE Transactions on Neural Networks.
6. G. Melis, C. Dyer, and P. Blunsom. On the state of the art of evaluation in neural language models.arXivpreprint arXiv:1707.05589, 2017
7. Audhkhasi Kartik, Osoba Osonde, Kosko Bart (2013) Noisy hidden Markov models for speech recognition. In: Neural Networks (IJCNN), The 2013 International Joint Conference on, IEEE, p 1–6.
8. Zarrouk E, Ayed YB, Gargouri F (2014) Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study. International Journal of Speech Technology 17(3):223–233.

9. Bietti Alberto, Francis Bach, and Arshia Cont (2015). An online EM algorithm in hidden (semi-) Markov models for audio segmentation and clustering. In: 2015 I.E. International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, p 1881–1885.
10. G. Ororbia II, T. Mikolov, and D. Reitter. Learning simpler language models with the differential stateframework.Neural Computation, 2017.