Journal of Computer Networks, 2013, Vol. 1, No. 2, 32-37
Available online at http://pubs.sciepub.com/jcn/1/2/3

DOI:10.12691/jcn-1-2-3

Recognition of Nigerian Major Languages Using Neural Networks

Ibikunle Frank*, Katende James

Botswana Int'l University of Science and Technology, Botswana
*Corresponding author: faibikunle2@yahoo.co.uk

Received July 27, 2013; Revised August 08, 2013; Accepted August 10, 2013

Abstract Speech Recognition is the technology by which sounds, words or phrases spoken by humans are converted into electrical signals and these signals are transformed into coding patterns to which meanings are assigned. It has two main types: discrete word and continuous speech recognition systems. Each type can be further sub-divided into two categories as Speaker Dependent and Speaker Independent recognition systems. Speaker dependent system operates only on the speech of a particular speaker for which the system is trained, while the Speaker Independent systems can be operated on the speech of any speaker. The speech recognition system proposed here digitizes the isolated words spoken by a speaker and performs Mel Frequency ceptral analysis and other signal processing techniques on the digitized data. The processed speech signal is then passed on to a pattern recognition which takes action based on the type of command pattern received. Artificial Neural Network (ANN) is used as speech recognition engine. Two different corpora were collected of audio recordings of Yoruba, Igbo and Hausa language speakers, in which subjects read aloud different words. One of the collected corpora contained data with background noise and the other without background noise. The results obtained from simulation can be generalized to cater for larger vocabularies and for continuous speech recognition.

Keywords: *speech recognition, Artificial Neural Networks, MFCC, Yoruba, Igbo and Hausa languages*

Cite This Article: Ibikunle Frank, and Katende James, "Recognition of Nigerian Major Languages Using Neural Networks." *Journal of Computer Networks* 1, no. 2 (2013): 32-37. doi: 10.12691/jcn-1-2-3.

1. Introduction

Speech is one of the oldest and most natural means of information exchange between human beings. We learn all the relevant skills during early childhood, without instruction, and we continue to rely on speech communication throughout our lives. It comes so naturally to us that we don't realize how complex the phenomenon of speech is [1]. The human vocal tract and articulators are biological organs with nonlinear properties, whose operation are 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 (if telephones or other electronic equipment are used). Even with these irregularities we are still able to interpret the speech most of the time as long as the spoken language is the language that we are familiar with. For centuries people have tried to develop machines that can understand and recognize speech as humans do so naturally. The human brain is known to be wired differently than a conventional computer. In fact, it operates under a totally different computational pattern. While conventional computers use a very fast and complex central processor with specific program instructions and locally addressable memory, by contrast the human brain uses a massively parallel collection of slow and simple processing elements (neurons), densely connected by weights (synapses) whose strengths are modified with experience, directly supporting the integration of multiple constraints, and providing a distributed form of associative memory [2].

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results for applications such as commands and control, data entry, and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding [3]. Speech recognition system depends on several different factors that can influence the accuracy of the speech that include: *Vocabulary size and confusability:* Generally it is easy to distinguish among a small set of words, but error rate naturally increase as the vocabulary size grows. On the other hand even a small vocabulary can be hard to recognize if it contains confusable words; *Speaker dependence vs. independence:* By definition, a speaker dependent system is intended for use by a single speaker, but a speaker independent system is intended for use by any speaker. Speaker independence is problematic to achieve because a system's parameters become tuned to the speaker(s) that it was trained on, and these parameters tend to be highly speaker-specific; *Isolated discontinuous,*

or continuous speech: Isolated speech means single words. Discontinuous speech means full sentences in which words are artificially separated by silence. Continuous speech means naturally spoken sentences. Isolated and discontinuous speech recognition is relatively easy because word boundaries are detectable and the words tend to be cleanly pronounced. Continuous speech is more difficult, because word boundaries are unclear and their pronunciations are more corrupted by slurring of speech sounds; *Read vs. Spontaneous speech: Systems* can be evaluated on speech that is either read from prepared scripts, or speech that is uttered spontaneously. Spontaneous speech is vastly more difficult, because it tends to be peppered with disfluencies like "uh" and "um", false starts, incomplete sentences, stuttering, coughing, and laughter; and moreover, the vocabulary is essentially unlimited, so the system must be able to deal intelligently with unknown words; *Adverse conditions:* A system's performance can also be reduced by a range of adverse conditions. These include environmental noise (e.g., noise in a car or a factory); acoustical distortions (e.g., echoes, room acoustics); different microphones (e.g., close-speaking, unidirectional); limited frequency bandwidth (in telephone transmission); and altered speaking manner (shouting, whining, speaking quickly).

The speech recognition process can generally be divided into different approaches as shown in Figure 1.

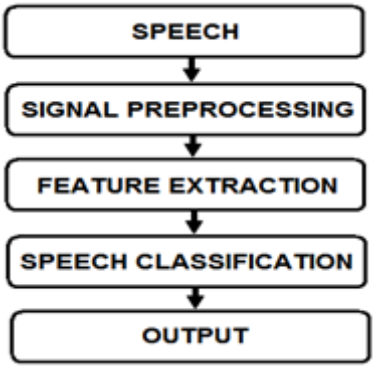

Figure 1. Speech recognition process

The first block consists of the acoustic environment plus the transduction equipment (microphone, preamplifier and AD converter) that have a strong effect on the generated speech representations. For instance we can have additional impact generated from additive noise or room reverberation. The second block is intended to deal with these problems, as well as deriving acoustic representations that are both good at separating classes of speech sounds and effective at suppressing irrelevant sources of variation. The third block must be capable of extracting speech specific features of the pre-processed signal. This can be done with a variety of techniques like cepstrum analysis and the spectrogram. The fourth block tries to classify the extracted features and relates the input sound to the best fitting sound in a known 'vocabulary set' and represents this as output. The commonly used approaches for speech classification include:

1) Template-based approaches: in which unknown speech is compared against a set of prerecorded words (templates), 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.

2) Knowledge-based approaches: in which "expert" knowledge about variations in speech is hand-coded into a system. This has the advantage of precisely modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully, so this approach was judged to be impractical, and automatic learning procedures were sought instead.

3) Statistical-based approaches: in which variations in speech are modeled statistically by Hidden Markov Models (HMMs), using automatic learning procedures. This approach represents the current state of the art. The main disadvantage of statistical models is that they must make a priori modeling assumptions, which are liable to be inaccurate, handicap the system's performance.

4) Learning based approaches: To overcome the disadvantage of the HMMs, machine learning methods could be introduced such as neural networks and genetic algorithm programming. In these machine learning models, explicit rules or other domain expert knowledge do not need to be given, they can be learned automatically through emulations or evolutionary process.

5) The Artificial intelligence approach: attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features. Expert systems are used widely in this approach. A good example of the artificial intelligence approach is the ANN. ANN offers an approach to computation that mimics biological nervous systems. The concept of ANNs is rooted deep into the recognition that though the human brain performs the functions about a million times slower than the digital computers, yet the human brain is more efficient when it comes to performing a complex set of the tasks such as speech synthesis, visual information processing, handwriting analysis etc. This is partially attributed to the fact that human brain is massively a parallel structure of biological neurons. ANNs are physical cellular system that can acquire, store and utilize experimental knowledge. ANNs have been applied to an increasingly number of real world problems of considerable complexity.

The paper is organized as follows. Section II is the Literature Review that critically reviews previous research similar to the speech recognition system identifying their limitations and shortcomings and the improvement this work gives to the different reviewed projects. Section III gives a detailed design of the system and its specifications. It explained speech signal processing, Speech Recognition and its Algorithms; Artificial Neural Networks and back-propagation, giving their key features and the feature extraction method used. Section IV present the implementation and testing of the speech recognition system. The results obtained are also discussed there. Section V concludes the work and presents the achievements of the project, it shortcomings and future improvements.

2. Literature Review

There are two basic approaches of using Neural Networks in speech recognition, 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. A few researches related to this method are discussed.

In [4], a Multi-layered Perception and the Recurrent Neural Network (RNN) models are applied for Standard Yoruba (SY) isolated tone recognition. The neural networks were trained on SY tone data extracted from recorded speech files. Their results led to three major conclusions that SY tone recognition problem can be implemented with MLP and RNN; the RNN training converges slower than the MLP using the same training data; the accuracy rates achieved using the RNN was found to be higher (although not significantly) than that of the MLP on the inside and outside test data sets; and that the SY mid tone has highest recognition accuracy. However, the efforts required for building the RNN is relatively more than those required for building the MLP. The MLP produced an accuracy of 87.50% for the tests carried out inside for H tone, it also produced 71.30% for the outside test. Similarly, the RNN model produced an inside test of 89.50% for the H tone while it produced 76.10% for outside test. The general conclusion from these results is that the RNN tone has the best recognition rate. A speech recognition component which would implement a set of reading lessons to assist adult illiterates in developing better reading capabilities was develop in [5]. The first stage involved the identification of the different alternatives for the different components of a speech recognition system, such as using linear predictive coding, Hidden Markov Models, Neural Networks or K-Nearest Neighbor Classifier for the pattern recognition block. The NN classifier trained using the Al-Alaoui Algorithm overcomes the HMM in the prediction of both words and sentences. They also examined the KNN classifier which gave better results than the NN in the prediction of sentences. The segmentation of Arabic sentences was also considered in their work, and they proved the problems in applying it to Arabic speech. In the project they implemented several classifiers for the Arabic speech recognition problem.

In [6], a Multilayer Perceptrons was applied in a more difficult task for alphabet recognition. A static input buffer of 20 frames was applied, in which each spoken letter was linearly normalized, 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.

The performance of a recurrent network and a feedforward network on a digit recognition task was compared in [7]. The feedforward network was an MLP with a 500 msec input window, while the recurrent network had a shorter 70 msec input window but a 500 msec state buffer. They found no significant difference in the recognition accuracy of these systems, suggesting that it's important only that a network have some form of memory, regardless of whether it's represented as a feedforward input buffer or a recurrent state layer.

3. System Design and Implementaton

There are a number of different approaches to the implementation of a speech recognition system, but this work considered the four major processing steps suggested in [2], namely: data preparation; training; testing and analysis; and implementation.

3.1. Data Preparation

The first stage of any recognizer development is data preparation. Speech data is needed both for training and for testing. In the system built, two different corpora were collected of audio recordings. One of the collected corpora is an open source audio recording of Yoruba, Igbo and Hausa Speeches, recorded in a silent environment and the other is a collection of Yoruba Igbo and Hausa recorded from scratch by different speakers in a noisy environment. It follows from above that before the data can be recorded, a phone set must be defined to cover both training and testing and a task grammar must be defined. The task grammar defines constraints on what the recognizer can expect as input. The system built provides an interface to receive input speech and recognizes it either as Yoruba, Igbo or Hausa. Each speaker repeated each word at least 5 times in different tones to provide accuracy of the system.

The speakers were given a list of sentences which they had to read aloud. Each word was pronounced at least five times by each speaker. The training corpus consisting of about 1000 words were recorded and labeled using AVS Audio Editing Software.

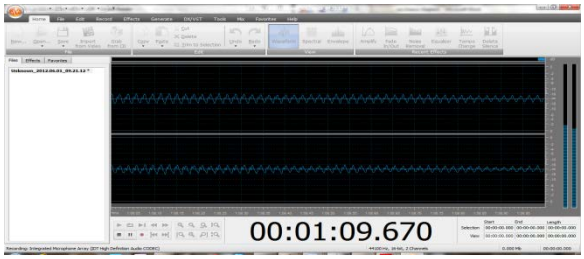

Figure 2. Screenshot of the data recording process using AVS audio editor

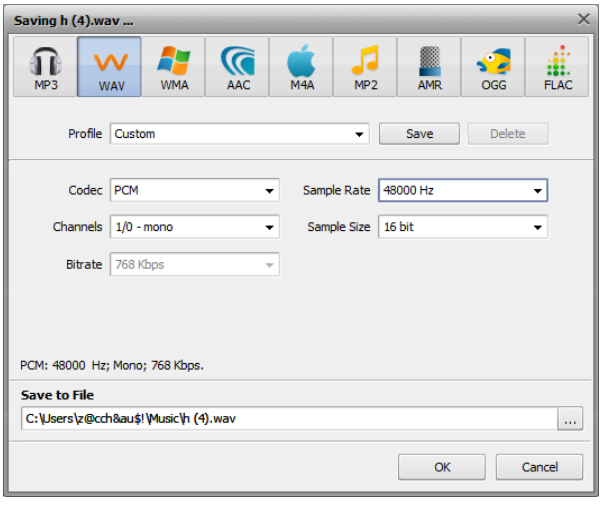

Figure 3. Speech file conversion interface

Silence was removed from each speech file using the AVS audio editing software which enables the user to select the portion of speech file to save.

All the speech data recorded were converted in order for it to be processed in MATLAB workspace as shown in Figure 3. Each file was converted to 4800Hz 16bits wav file using the AVS audio editing software.

3.2. Training

Back propagation is the best known training algorithm for multiple layer neural networks and still one of the most useful. It has lower memory requirements than most algorithms, and usually reaches an acceptable error level quite quickly, although it can then be very slow to converge properly on an error minimum. The Neural Network was designed in the MATLAB workspace using the following parameters.

Number of output: 2
Number of input: 13
Number of hidden neurons: 20
Number of hidden layer: 1
Number of training samples: 300
Number of validation samples: 15
Number of testing samples: 15
Performance: Mean Squared Error (MSE)
Data division: Random (DIVIDERAND)

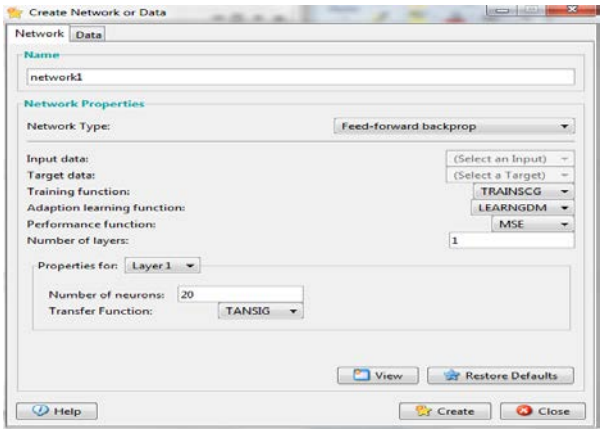

Figure 4. Creating the neural network data

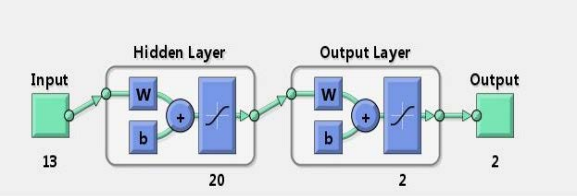

Figure 5. Realized Neural Network

For the feature extraction, the Mel Frequency Ceptral Coefficients (MFCC) features were computed with the following parameters.

Pre-emphasis coefficients(apre) = -0.95
Frame size = 256 samples (16ms)
Frame overlap = 85 samples (5.3ms)
Number of triangular bandpass filters = 20
Number of MFCCs =13

Thirteen (13) MFCCs are extracted from each speech file. Figure 6 depicts bar graph of one of the speech extracted.

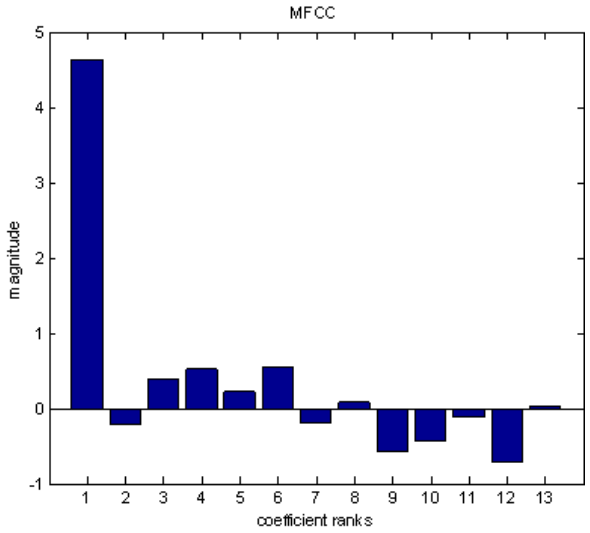

Figure 6. Plot of magnitude over coefficient ranks

4. Simulation Results

4.1. Testing Results

The recognition performance evaluation of the speech recognition system is measured on a corpus of data different from the training corpus. A separate test corpus, with new speech records was created as it was previously done with the training corpus. The test corpus consists of 15 different speech samples for each language and the labeled data were converted to MFCC's. In order to test for speaker independency of the system, some of the subjects who participated in creation of the testing corpus had not participated in the creation of the training corpus. A total of 6 speakers participated in this testing.

Table 1. Testing Results

Subject	Words correctly Recognized	Substitution Errors	Recognition Rate (%)
Subject1 (Yoruba)	14	1	92.8
Subject2 (Yoruba)	15	0	100
Subject 3 (Igbo)	13	2	86.6
Subject 4 (Igbo)	14	1	92.8
Subject 5 (Hausa)	12	3	80
Subject 6 (Hausa)	10	5	66.7

4.2. Training Results

The training results were automatically generated using the MATLAB "trainers" function as shown in Figure 7.

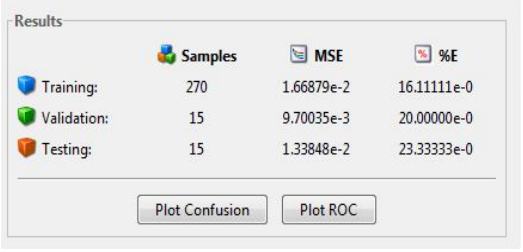

Figure 7. Screenshot of Training/Error Results

Figure 8. Screenshot of the Confusion Matrix

The confusion matrix of the network was generated as shown in Figure 8. This result shows that Yoruba and Hausa speech samples are more easily confused by the neural network model.

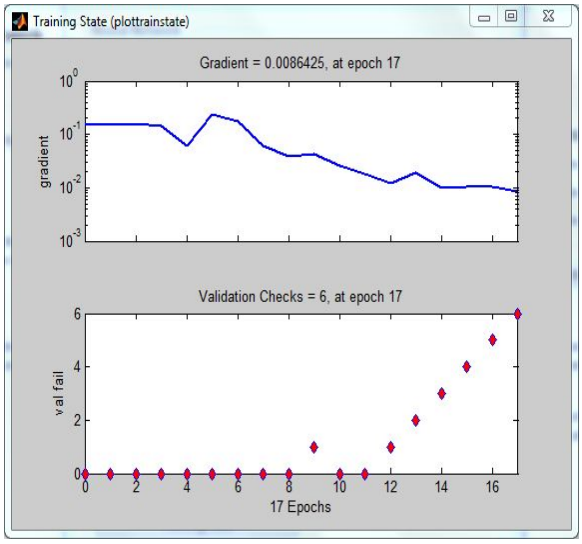

Figure 9. Screenshot of training state at 17 epochs

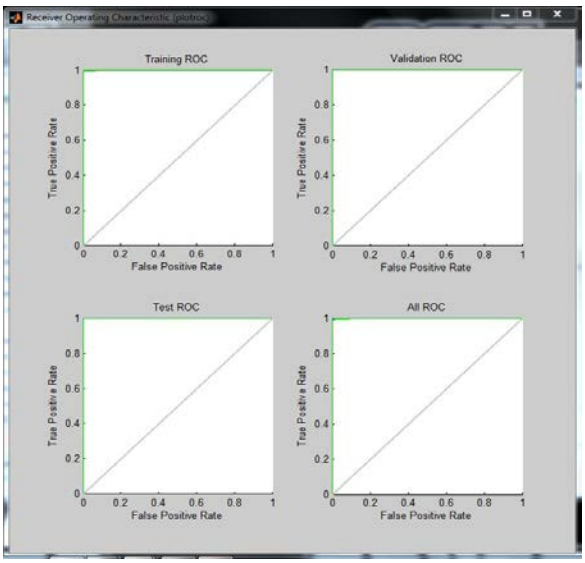

Figure 10. Screenshot of the Receiver Operating Characteristics

5. Performance Evaluation

From the simulation, results show that adequate functioning of neural networks depends on the sizes of the training and test set. The Percentage recognition rate used is computed as the total number of recognized speech sample to that of the total speech sample multiplied by 100 as shown in Figure 11.

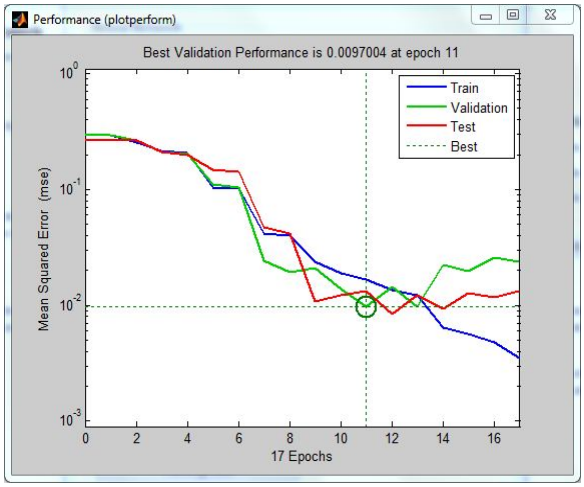

Figure 11. Performance analysis after Training

6. Conclusion

The aim and objective of this research is to simulate a speech recognition prototype that can recognize some words in Yoruba, Igbo and Hausa languages (the three major Nigerian native languages) using Artificial Neural Network. In order to meet this objective a limited word grammar was constructed, a dictionary created and data from different Nigerian language speakers was recorded and trained thereafter. The system was tested using testing corpus data and live data and the system has a relatively high recognition rate. This implies that the objective of creating a system that can recognize spoken Nigerian native languages was achieved. The Nigerian major languages speech recognition recipe accompanying this paper can be used by any researcher desiring to join language processing research. The research is however not all conclusive as it has catered for only some amount of words in each language. As much as it has created a basis for research, the work can be expanded to cater for more extensive language models and larger vocabularies.

References

[1] Rudnicky, A., Lee K., and Hauptmann A., (1992):"Survey of current speech technology". *Communications of the ACM*, 37(3): pp52-57.

[2] Picheny M., (2002). Large vocabulary speech recognition, *IEEE Computer*, 35(4):42-50.

[3] Reddy D.R., (1976). Speech Recognition by Machine: a Review. *Proceeding of IEEE*, 64(4):501-531.

[4] Odetunji Ajadi Odelobi, "Recognition of Tones in Yoruba Speech: experiments with Artificial Neural Network, Studies in Computational Intelligence 83, 23-47 (2008).

[5] Mohamad Adnan etal., "Speech Recognition using Artificial NeuralNetworks and Hidden Markov Models".

[6] Burr, D. (1988). “Experiments on Neural Net Recognition of Spoken and Written Text”, IEEE Trans. on Acoustics, Speech, and Signal Processing, 36, pp 1162-1168.

[7] Franzini, M., Witbrock, M., and Lee K., “Speaker-Independent Recognition of Connected Utterances using Recurrent and Non-Recurrent Neural Networks”, In Proc. International Joint Conference on Neural Networks, 1989.

[8] Rabiner, L, and Wilpon, J., (1979). “Considerations in applying clustering techniques to speaker-independent word recognition” *Journal of Acoustic Society of America*, pp.663-673.

[9] Haykin S, "Neural Networks: a comprehensive foundation", 2nd Edition, Prentice Hall, 1999.

[10] LawenceRabinar& Bine-hwangJuang, “Fundamentals of speech recognition’. Prentice Hall, Englewood Cliffs, 1999.

[11] Jurafsky, Daniel and Martin, James H, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition” (1st ed.). Prentice Hall, 1996.

[12] Daoudi, K. (2002): “Automatic Speech Recognition: The New Millennium”, Proceedings of the 15th International Conference on Industrial and Engineering, Applications of Artificial Intelligence and Expert Systems: Developments in Applied Artificial Intelligence, pp 253-263.