Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 4992-5002
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.512

Speech To Speech (S2s) User Interface Design For Emerging Artificial Intelligence Markets In Education

Santosh Gaikwad Department of Computer Science, Dr. Babasaheb Ambedkar Marathwada University Constituent Model College Ghansawangi, Jalna, India.

Bharti Gawali Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.

Sandeep Thorat santosh.gaikwadcsit@gmail.com[1], bharti_rokade@yahoo.co.in [2], sandipthorat16@gmail.com

Abstract

Socio-economic development has represented a variety of challenges, such as providing access to information, bridging communication gaps and technological and functional illiteracy. Significant advances have been made in the field of spoken language technology development for under – resourced languages, such as text-to-speech (TTS) and Speech to text (STT) systems. Speech recognition is expanding capabilities to support application in emerging markets.

More people around the world will gain the ability to use speech enable application in day-to-day activity. This experiment designs and develops speech to speech (S2S) interface for the calculator appliance. The MFCC is used for the feature extraction mode for the system. The system is tested based on a neural network approach.

The possibility for the designing of a software system of speech-to-speech recognition using one of the techniques of artificial intelligence applications neuron networks where this system is used for real time education-based speech UI design. The UI is created utilizing speech as input information and result turn into a speech. The people groups in developing business sector require a speech empower machine for them without hands free communication. The system is tested over the speaker independent mode. Performance of the system achieved 96.75% with minimum real time factor.

Keywords: Speech Recognition, Neural Networks, Artificial Networks, Signals Processing, MFCC, feature extraction.

1. Introduction

In the current era of research, the artificial Intelligence playing an important role for the technological development. The AI based neural network is proved as a performance-based speech recognition techniques, which improves the accuracy of the traditional speech recognition and its relevant UI

design. This Speech is one of the most important tools for communication between human and his environment. Therefore, building of automatic Speech Recognition System (ASR) is desire for him all the time [1]. In a Speech Recognition system, many parameters affect the accuracy of speech recognition system such as Vocabulary size, speaker dependant, speaker independent, time for recognition, type of speech (continuous, isolated) and recognition environment condition. A speech recognition algorithm is consisted of several stages that the most significant of them are feature extraction and Classification. In feature extraction the category, best presented algorithm are Mel Frequency Cepstral Coefficients (MFCC), linear discriminant analysis (LDA), Linear Prediction Cepstral Coefficient (LPCC), and Linear Prediction Coefficients. The enriched literature available on speech recognition hence reported research paper from 1960 to till today MFCC is most popular and robust technique for feature extraction [2]. The aim of a continuous speech recognizer is, to provide an efficient and accurate mechanism to transcribe speech into text.
The main contribution of paper is as…

1. Database of Continuous Marathi speech database using standard protocol with 1800 samples in which 10 male and 10 female's speakers with 16000 Hz sampling frequency.
2. A fusion of MFCC and AI technology for the S2S UI Design and implementation.
3. The implementation of application of UI design for the education purpose such as teaching and learning.

The research article is structured in five sections. Feature extraction is described in section II. CNN approach is explained in section III. Section IV deals with experimental design and Section V with Result and Discussion followed by conclusion is section VI.

II. Feature extraction

Feature extraction involves analysis of speech signal to reduce data size before classification. Using feature extraction techniques used for extracting a specific feature from the speech, these features carry the characteristics of the specific speech which are useful for differentiate the different speech, so these features will play the major role in speech recognition. Feature extraction technique is classified into temporal analysis and spectral analysis technique [3]. The design of descriptive feature for a specific application was the main challenge in building speech recognition system.

A. Mel-frequency cepstral coefficients

The cepstral coefficient is set of features, reported to be robust in different pattern recognition task concerning Speech [4]. The frequency bands of Mel-frequency cestrum are equally spaced on the Mel-Scale. It was approximating human auditory systems are response more closely than the linearly spaced frequency bands used in normal cepstrum. As a result, MFCC has been widely used for speech as well as speaker recognition in the recent years [3, 4, 5, 6]. Figure 2 describes the block diagram of MFCC extraction algorithm.

B. Speech Database

The database was created based on socio survey of various colleges in Marathwada region. The voice sample are collected using the connected dictionary which was prepared based on day-to-day sentences used in the teaching and learning. The database was collected onsite environment which was impacted by the background noise. The age group of speakers selected for the collection of database ranges from 20-year 3 month to 35-year 6 month. Mother tongue of all the speakers was Marathi. The total number of speakers was 20 out of which 10 were Females and 10 were Males.

The vocabulary size of the database consists of:
Sentences: 30 samples.
Utterances: 03
Speaker: 20
Total Vocabulary: 1800 Sentences

The following speech sample such as " आज तासिका होणार आहेत का " were consider in this work as a sample. The figure 1 describe a visual representation of Marathi continuous speech sample.

C. Environment Setup for Database Recording

To achieve a high audio quality the recording took place in the onsite environment with and without noisy sound and effect of echo. The sampling frequency for all recordings was 16000 Hz at the room temperature and normal humidity. The speaker was standing and talking in natural manner in front of the direction of the microphone with the distance of about 12-15 cm [3]. The speech data was collected with the help of PRAAT open-source software using the single channel.

The digitization of Continuous Speech Signal speech signal is shown in figure 1. This is continuous sentence. This speech signal is slowly varying over time, and it is called quasi stationery.

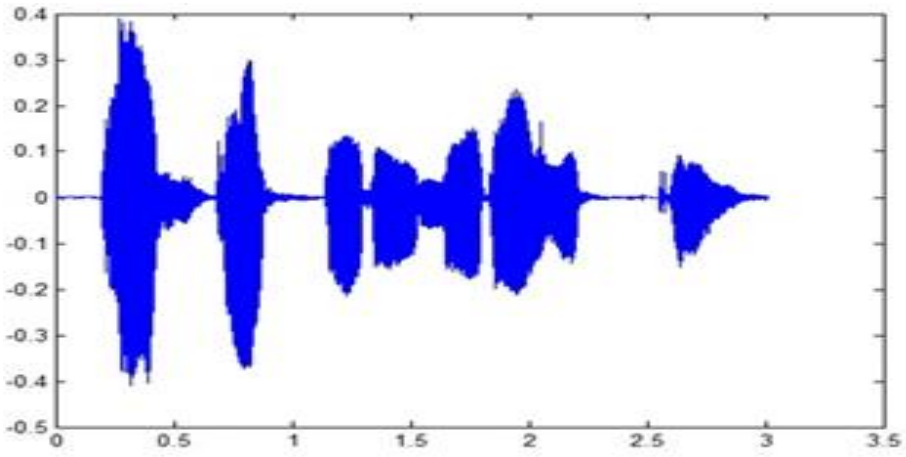

Figure 1: The representation of the voice sample A.

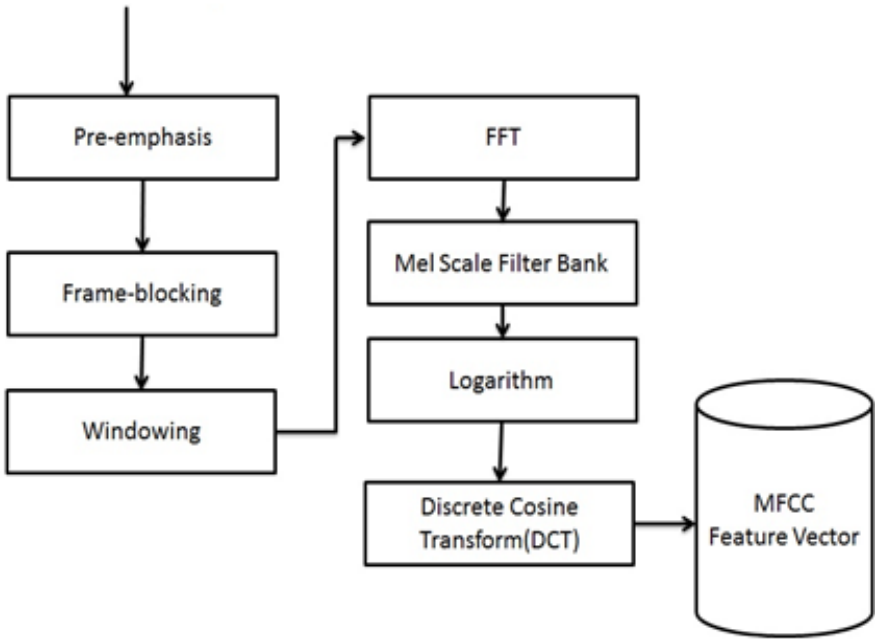

Figure 2: Block diagram of MFCC

In the Pre-emphasis each of the speech samples is down sampled to 16000 Hz for analysis purpose. The sample speech signal was pre-emphasized with filter. The output is called Mel Frequency Cepstrum Coefficients (MFCC). The MFCC is real numbers and can be converted into time domain using the Discrete Cosine Transform (DCT).The MFCC is used to discriminate the repetitions and prolongations in natural speech [6]. The figure 3 describes the original speech signal with MFCC feature.

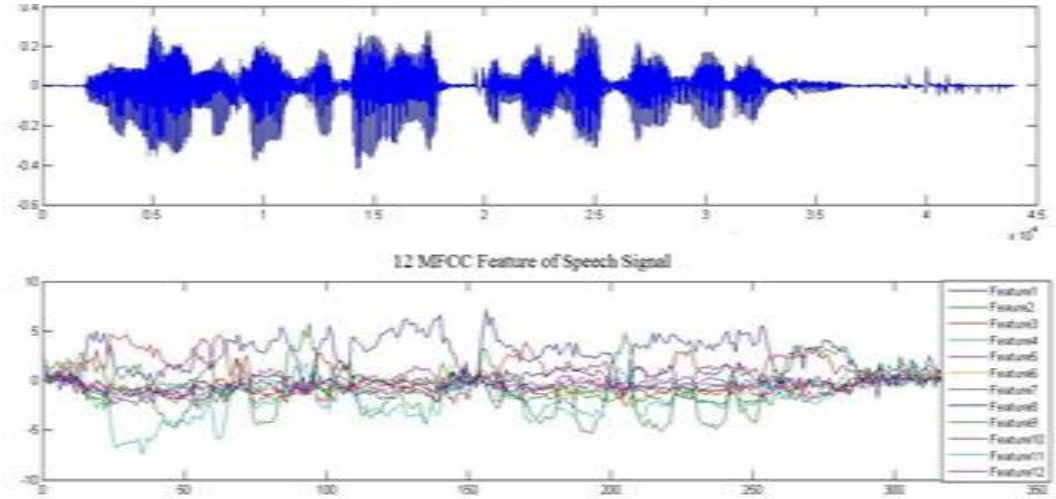

Figure 3: Original Speech Signal with MFCC feature

III. CNN Approcah for Speech Recognition

Neural network training is carried out through the consistent presentation of the training set, with simultaneous tuning scales in accordance with a specific procedure, until around the variety of

configuration error reaches an acceptable level[7,8]. A prototype of a neuron is nerve cell biology. A neuron consists of a cell body, or soma, and two types of external wood-like branches: Axon and dendrites. The cell body contains the nucleus, which holds information on hereditary characteristics and plasma with molecular tools for the production and transmission of elements of the neuron of the necessary materials. A neuron receives signals from other neurons through the dendrites and transmits signals generated by the cells of the body, along the axon, which at the end of branches into the fibre, the endings of synapses [9,10] The working architecture of

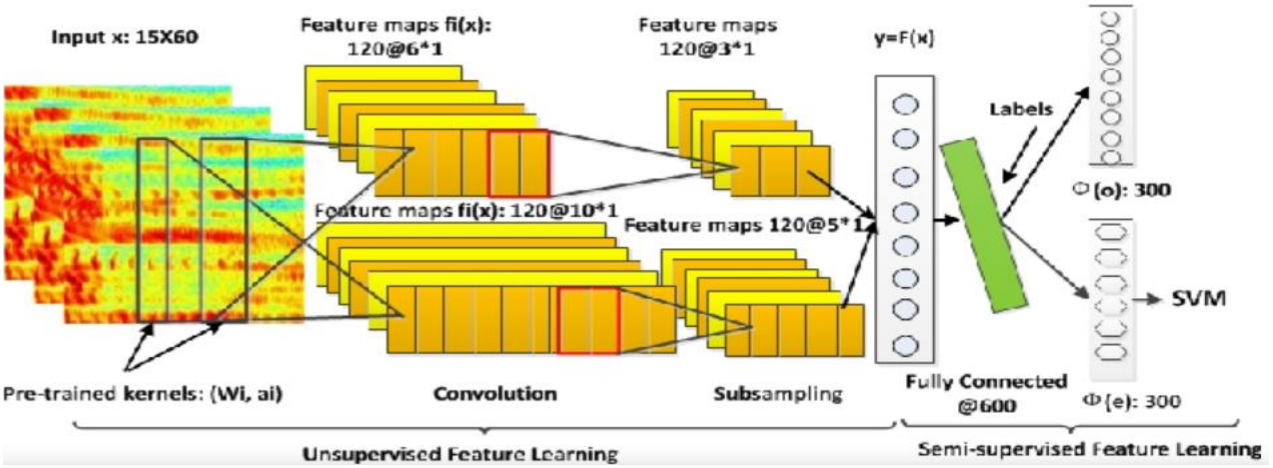

Figure 4: CNN architecture for speech recognition

The recognition system uses an acoustic model to recognize speech. It is the most important component of any ASR scheme. It provides mathematical representations of the sound for each phrase using audio recordings of the voice and their text transcriptions. Neural network refers to mathematical representations. - phoneme in the term has its own hidden layers, and each mathematical representation is given a mark called a phoneme[11,12,13]. Figure 5 depicts the acoustic model's teaching diagram.

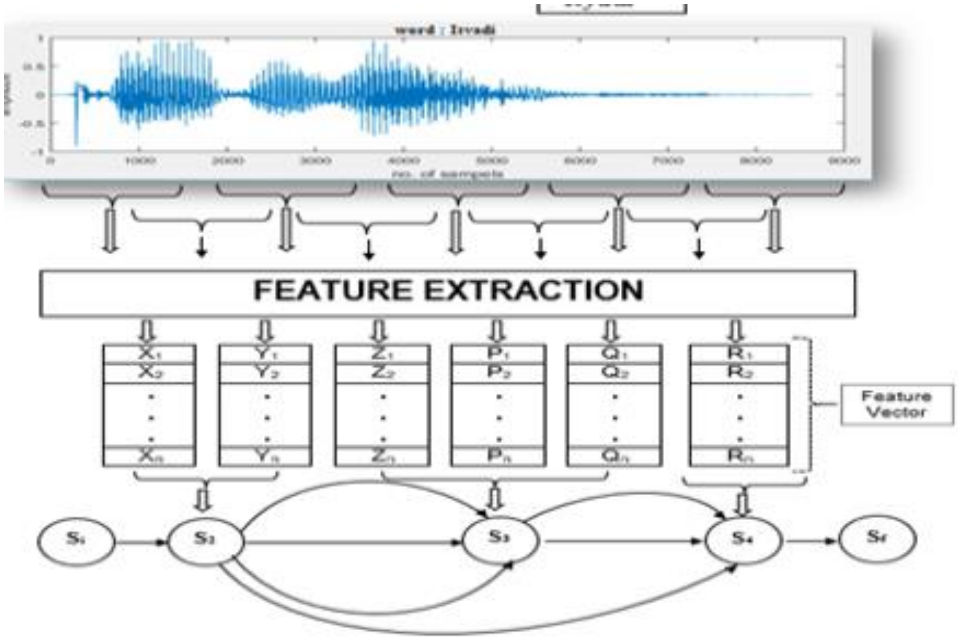

Figure 5: Acoustic model for speech recognition using MFCC and CNN

IV. Experiment Analysis

The experiment design for the said proposed S2S system is explain the below manner. The database was recorded in an onsite real time environment, so it required a preprocessing with the noisy echo. The collected dataset was tested on real time training and testing environments.

A. Data

The data collection for the said experiment were done based on LDCIL standard dataset. This dataset was collected onsite environment and real time applications for the education. The database was maintaining the standard guideline provided by the LDCIL speech community which was published in the COCOSDA conference[3,4].

B. Pre-processing

The pre-processing of the collected speech sample was done using the spectrogram, energy contour, frequency contour, FFT waterfall techniques. The voiced period of speech signal was identified and applied to all experiment used for pre-processing of speech signal [11].

Table 1: Extracted speech features for the speech sample A

	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6	Frame 7
Feature 1	-99.2611	-99.4898	-99.2722	-101.788	-98.56	-98.0701	-99.2988
Feature 2	-1.49357	0.029736	-1.11548	-1.50401	-2.42051	0.407593	-1.20999
Feature 3	3.957262	5.873194	3.317755	3.760828	4.293447	4.281029	4.695836
Feature 4	3.028113	3.551687	3.732127	2.988079	3.005539	1.835096	2.112846
Feature 5	0.221089	-0.03585	0.673257	0.56653	0.352773	-0.24075	1.664033
Feature 6	-3.70178	-2.59263	-0.45272	-2.98227	-2.01282	-0.55236	0.160449
Feature 7	-2.40732	0.812234	-2.19574	-2.04926	-4.78145	-2.3998	-1.31035
Feature 8	-3.14345	-1.96177	-3.86076	-4.50504	-3.45373	-1.71338	-3.74567
Feature 9	-2.60152	-1.94813	-3.0404	-3.51443	-4.07733	-2.47365	-4.03897
Feature 10	-2.45382	-2.71887	-2.42262	-3.00455	-1.31358	-1.03477	-2.01364
Feature 11	-2.58438	-2.47419	-2.05613	-1.64615	-1.79146	-0.76962	-0.70951
Feature 12	-2.61877	-1.32345	-1.83451	-2.31923	-2.47309	-2.56884	-2.68682
Feature 13	-0.09343	-1.08313	-0.07689	-0.18294	-1.13966	-1.73903	-1.56617

C. Training

For the training of the system based on MFCC and CNN network selected as 1000 sentences. The Training of the system was performed with the different combination of technique as mentioned below. The system was trained using MFCC individually as well as using MFCC and CNN approach. Result obtained by all technique for training time on same dataset were collected is shown in Table 2. The performance results of confusion matrix for MFCC and CNN is described in Table 3. The comparative performance of the MFCC and MFCC+ CNN approach is shown in Table 4. The training of the proposed model using CNN and MFCC is shown in figure 6.

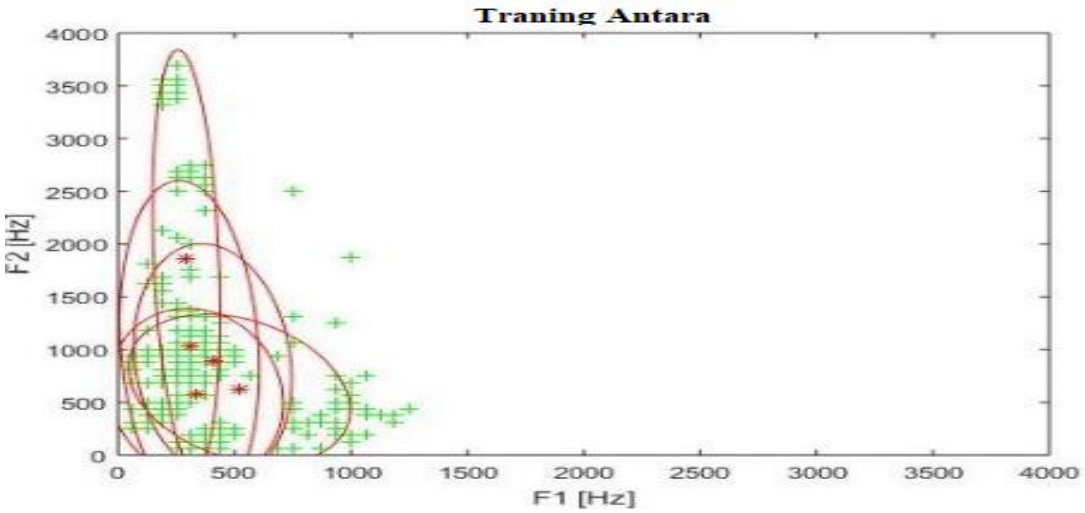

Figure 6: Training of speech samples using MFCC and CNN approach

Table 2 : Confusion Matrix for traditional MFCC

ex rd → A	B	C	D	E	F	G	H	I	J	K	No. Of Token Passed	No. Of token Confuse	Accuracy (%)
A	25	-	-	1	-	2	-	1	-	-	1	30	05	83.25
B	2	20	1	-	2	-	-	2	-	1	2	30	10	66.60
C	2	-	24	1	-	1	-	-	2	-	-	30	06	79.92
D	2	-	-	22	1	1	-	1	2	-	1	30	08	73.26
E	1	-	-	-	23	-	2	-	2	1	1	30	07	76.59
F	-	1	-	2	-	25	-	-	1	-	1	30	05	83.25
G	1	-	1	-	2	-	23	-	-	2	1	30	07	76.59
H	2	-	-	1	-	-	-	26	-	-	1	30	04	86.58
I	1	1	-	1	2	2	-	-	21	-	2	30	09	69.93
J	1	-	-	1	-	1	2	-	-	24	1	30	06	79.92
K	1	-	02	-	-	1	-	-	-	1	24	30	06	79.92
Overall Accuracy									1019.04/11= 92.64%					

Table 3: Confusion Matrix for MFCC and CNN

ex → / rd ↓	A	B	C	D	E	F	G	H	I	J	K	No. Of Token Passed	No. Of token Confuse	Accuracy (%)
A	28	1	-	-	-	1	-		-	-	-	30	02	93.33
B	-	26	-	-	1	-	-	1	-	2	-	30	04	86.58
C	-	-	27	1	-	1	-	-	-	-	1	30	03	89.91
D	1	-	-	28	-	-	-	1	-	-	-	30	02	93.33
E	-	-	1	-	27	-	-	-	-	1	1	30	03	89.91
F	-	1	-	1	-	26	1	-	1	-	-	30	04	86.58
G	-	-	1	-	-	-	29	-	-	-	-	30	01	96.57
H	1	-	1	-	-	-	-	27	-	-	1	30	03	89.91
I	-	1	2	1	1	1	-	1	23	-	-	30	07	76.59
J	1	-	-	2	-	1	-	-	-	25	1	30	05	86.25
K	1	-	-	1	-	-	-	-	-	1	28	30	02	93.33
Overall Accuracy									1073.93/11=**97.63%**					

Table 4: Performance of the system based on MFCC and proposed approach

SR.NO	NAME OF TECHNIQUE	ACCURACY	RTF
1	MFCC	92.34	4.35
2	MFCC AND CNN	97.68	3.39

IV: Results and Discussion

Table 2 and Table 3 shows detailed information about accuracy in the form of confusion Matrix .In the Confusion Matrix row and column correspond to the Marathi Continuous

sentences index A to K which was used as an index for the experiment . Number of token was passed randomly selected sentences passed for testing . A sentences at some exceptional case makes system confused about speech From Confusion Matrix we derived overall accuracy from following Method[15].

$$Accuracy = \frac{N - C}{N} * 100$$

The result obtained in this paper motivated to use MFCC and CNN approach for the modelling of S2S system for real time applications.

V. Conclusion

Model of speech recognition was based on artificial neural networks. In the current era of technology, the AI placed as an important role in various application. This research developed and tested the S2S system for the educational domain. The system is tested using the traditional MFCC approach which proved as a 92.34 % accuracy. The proposed modelling algorithm using MFCC and CNN approach proved as 97.68% accuracy. The prominent results obtained based on accuracy and real time factor. The author recommended the current AI based CNN approach with MFCC is proven the best for S2S educational applications.

References

[1] Mohammad Shahin Mahanta, Konstantinos N.Plataniotis "Linear Feature extraction using sufficient statistics",978-1-4244-4296-6/10 IEEE (ICASSP), 2010.

[2] Bharti W. Gawali, Santosh Gaikwad, Pravin Yannawar,Suresh C.Mehrotra "Marathi Isolated Word Recognition System using MFCC and DTW Features" ACEEE 2010, Source[online] http://www.searchdl.org/conference/ACS2010/73.PDF.

[3] M.Kesarkar, "Feature Extraction for Speech Recognition" Indian Institute of Technology, Bombay 2003.

[4] K.Ravikumar, B.Reddy, R.Rajagopal, and H.Nagaraj,"Automatic Detection of syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies,"in proceeding of Word Academy Science,Engineering and Technology,2008,pp.270-273.

[5] S. Memon, M.Lech, and H.Ling, "Using Information Theoretic Vector Quantization for Inverted MFCC Based Speaker Verification," in Computer ,Control and Communication,2009.IC42009,2nd International Conference on,2009,pp.1-5.

[6] Lim Sin chee, Ooi Chia Ai, M.Hariharan and Sazali Yaacob, "MFCC based Recognition of Repetition and Prolongations in Stuttered speech using k-NN and LDA", Proceedings of 2009 IEEE Student Conferences on Research and Development (SCOReD 2009),16-18 Nov,2009,UPM Serdang,Malaysia.

[7] Childer, D.G. (2004) The Matlab Speech Processing and Synthesis Toolbox. Photocopy Edition, Tsinghua University Press, Beijing, 45-51.
[8] Chien, J.T. (2005) Predictive Hidden Markov Model Selection for Speech Recognition. IEEE Transaction on Speech and Audio Processing, 13.
[9] Luger, G. and Stubblefield, W. (2004) Artificial Intelligence: Structures and Strategies for Complex Problem Solving. 5th Edition, The Benjamin/Cummings Publishing Company, Inc. http://www.cs.unm.edu/~luger/ai-final/tocfull.htm
[10] Choudhary, A. and Kshirsagar, R. (2012) Process Speech Recognition System Using Artificial Intelligence Technique. International Journal of Soft Computing and Engineering (IJSCE), 2.
[11] Ovchinnikov, P.E. (2005) Multilayer Perceptron Training without Word Segmentation for Phoneme Recognition. Optical Memory & Neural Networks (Information Optics), 14, 245-248.
[12] Guo, X.Y., Liang, X. and Li, X. (2007) A Stock Pattern Recognition Algorithm Based on Neural Networks. Third International Conference on Natural Computation, 2.
[13] Dai, W.J. and Wang, P. (2007) Application of Pattern Recognition and Artificial Neural Network to Load Forecasting in Electric Power System. Third International Conference on Natural Computation, 1.
[14] Shahrin, A.N., Omar, N., Jumari, K.F. and Khalid, M. (2007) Face Detecting Using Artificial Neural Networks Approach. First Asia International Conference on Modelling & Simulation.
[15] Lin, H., Hou, W.S., Zhen, X.L. and Peng, C.L. (2006) Recognition of ECG Patterns Using Artificial Neural Network. Sixth International Conference on Intelligent Systems Design and Applications, 2006.