ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Isolated Telugu Speech Recognition using MFCC and Gamma tone features by Radial Basis Networks in Noisy Environment

Shaik Shafee [1], Prof.B.Anuradha[2]

Research Student, Dept of ECE, S.V.University College of Engineering, Tirupati, Andhra Pradesh, India[1]

Professor, Dept of ECE, S.V.University College of Engineering, Tirupati, Andhra Pradesh, India[2]

ABSTRACT: In this paper, Radial basis neural networks[1][12][17] have been examined for speech recognition using speech features MFCC (Mel frequency Coefficients) and Gamma tone frequency coefficients for isolated Telugu words in noisy environment. Speech feature vectors are used to train, validate and test the Radial basis neural networks.Experiments conducted in Office environment under the presence of light/fans/Air conditioning noises, and the results are analyzed. MFCC (Mel frequency Cepstrum Coefficients) are preferably used features for Speech recognition in ASR Systems. In recent trends another speech features extraction technique called Gamma tone frequency Coefficients are being experimented for ASR. Both MFCC and Gamma tone features [2] [3] [4] have been analyzed under the same noisy conditions for isolated telugu words. The design and development of the neural network, features extraction are implemented in MATLAB environment [12][17] and analyzed the results.

KEYWORDS: Speech recognition, Telugu isolated words, Radial basis neural networks, ANN, MFCC, Gamma tone frequency Coefficients, K-means algorithm.

I. INTRODUCTION

The advancement of speech or voice recognition automation process improves the interface between man and machine in numerous applications. Isolated words speech recognition [10] [11] system can be used as voice controlled interface between human beings and artificial machines. Since many years HMM models with MFCC features have been extensively used for Speech recognition systems and gives good results for the system designed in clean speech environment (under no noise conditions). But MFCC features still not showing good results in noisy conditions. In this paper an attempt has been made to design Radial basis neural network system for Telugu isolated speech words recognition under noisy environment. Speech samples are collected under normal noisy conditions in office environment under the presence of lights, A.C (Air Conditioning), fans and Computer keyboard noises and examine the recognition success rate in Radial basis neural networks for MFCC and GFCC features.

II. RELATED WORK

Generally Speech recognition means converting speech into text or automatic speech recognition (ASR). Speech recognition system is a machine which understands the human speech and act accordingly. Speech Recognition [5] [6] [9] may be of isolated word to continuous speech recognition, speaker-dependent to speaker-independent recognition, and from a small vocabulary to a large vocabulary. The major steps involved in speech recognition are features extraction [8], train the system and validation/testing. Different types of features extraction techniques are LPC (Linear predictive coefficients), PLP (Perceptual Linear Predictive Analysis), MFCC (Mel Frequency Cepstral Coefficients) and GFCC (Gammatone frequency cepstral coefficients) and other different processes can be used to extract the features from the windowed overlap frames of speech signal. The supervised pattern recognition models: HMM (Hidden Markov Models) [21], back-propagation neural network (BPNN), Dynamic Time warping (DTW), Support Vector Machine (SVM) and Gaussian mixture model (GMM) , unsupervised models : fuzzy k-means algorithm and Kohonen self-organizing map (KSOM) , and dimensionality reduction techniques: principal component analysis (PCA), linear discriminant analysis (LDA), kernel LDA, and independent component analysis (ICA) are generally used

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

models to train the speech recognition system. MFCC and Hidden Markov Models combination is widely used techniques in speech recognition and it works well under clean environment from the past research studies.

Though the recognition success rate shows good results for the above feature extraction and training models in clean environment (noise less conditions), the performance of speech recognition is not showing good results in noisy environment such as machines noise in factories, vehicle noise in traffic, lights/A.C/ computer key board noises in Office environment and many different types of noises along with weather conditions affect the speech recognition performance. Recent research experiments [2][3][4] of speech recognition techniques using Gamma tone frequency Coefficients showing better performance over MFCC under noisy environment, Artificial neural networks(ANN) [9] such as feed forward neural networks, feed forward back propagation techniques , and other ANN techniques are being examined in recent researches [1]. In this paper an attempt has been made to examine the speech recognition using Radial Basis neural networks for MFCC and Gamma tone features.

III. SPEECH FEATURES EXTRACTION

A. *Mel Cepstral Coefficients:*

Mel Frequency Cepstral Coefficents (MFCCs) are widely used features in automatic speech and speaker recognition. This method mainly consists of two parts namely cepstrum calculation and mel scaling. Cepstrum method is the homomorphic process to find vocal tract filter and mel scaling is the human perception of the frequency content of the sound. MFCC feature extraction process includes framing and windowing the 16KHz speech signal into short frames of size 320 samples (20 msec duration) with 50% overlapping i.e 160 samples (10 msec duration), computing FFT (N=512 for 320 sample blocks) and then applying the mel filter bank, sum the energy in each filter , calculate the logarithm of all filter bank energies, then take the DCT (Discrete Cosine Transform) of the log filter bank energies. The frequency content of sounds is not liner so the pitch is measured on non linear scale called mel scale and mel frequencies can be calculated as:

$$F_{mel} = 2595 \log_{10}(1+F_{HZ}/700) \quad \text{------ eq. (1)}$$

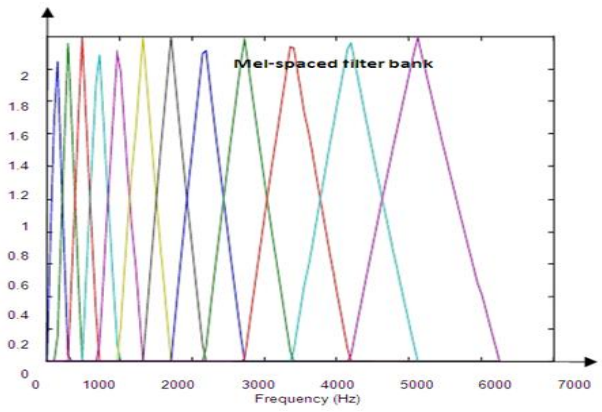


Fig.1: Mel spaced filter bank of order 12

Filter	Centre Frequency(Hz)	Lower limit(Hz)	Upper limit(Hz)
K=1	163.5	0	327
K=2	364.5	163.5	566
K=3	613.5	365	862
K=4	919.5	613	1226
K=5	1297.5	919	1675
K=6	1763	1297	2229
K=7	2338	1763	2913
K=8	3047	2338	3756
K=9	3921	3047	4795
K=10	5000	3921	6077

Table.1: Frequency limits for 10 filter banks

Let us assume 10 MFCC coefficients needs to be extracted from each frame for frequency range of (0-5000Hz) .Then 10 filters required to extract 10 MFCC coefficients from each speech frame. So 10 triangular filters need to be generated using mel sclae to get 10 MFCC coefficients.

The first 12 to 14 coefficients are enough to extract the speech information of the frame. Different count of filter banks has been analysed in this paper.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

B. *Gammatone Coefficients:*

Gamma tone function is defined in time domain by its impulse response as:

$$G(t)= a.t^{n-1}. \text{Cos}\,(2\prod f t + \varphi) \times e^{-2\prod bt} \quad \text{------ eq. (2)}$$

Where
n is the order of the filter;
b is the bandwidth of the filter;
a is the amplitude;
f is the filter centre frequency;
φ is the phase;

Patterson and Moore shown that the gammatone function of impulse response would best fit to the human auditory system and the filter shapes derived by Patterson and Moore in the year 1986. The Equivalent Rectangular bandwidth (ERB) of the auditory filter with the function has been proposed as:

$$\text{ERB}= 24.7\,(4.37\times10^{-3} f +1) \quad \text{------- eq.(3)}$$

In order to simulate human auditory behaviour, the central frequencies of the filter bank are often equally distributed on the Bark scale, the temporal-frequency presentation will be similar to that of the FFT-based short-time spectral analysis. The gammatone filterbank is commonly used to simulate the motion of the basilar membrane within the cochlea as a function of time, in which the output of each filter models the frequency response of the basilar membrane at a single place. The filterbank is normally defined in such a way that the filter centre frequencies are distributed across frequency in proportion to their bandwidth, known as the ERB scale (Glasberg and Moore, 1990). The ERB scale is approximately logarithmic, on which the filter centre frequencies are equally spaced.

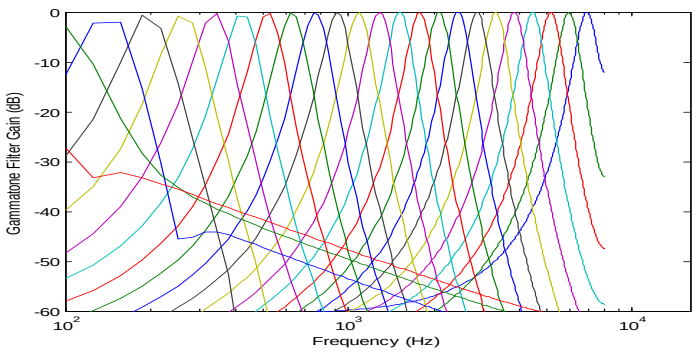


Fig. 2: Frequency responses of a gamma tone filter bank of order 24 on the ERB-rate scale.

GFCC features also computes in the same way as MFCCs are calculated and the only difference is the filterbank function. Each speech frame is processed through all the gammatone filters and energies are summed up and each filter will generate one coefficient. Different count of filter banks of order 12, 16 and 20 are examined for recognition success rate.

IV. K-MEANS ALGORITHM

K-means algorithm is used to classify the set of data to K number of groups based on attributes or features, where K is a positive integer number. The grouping is done by minimizing the Euclidian distance i.e. sum of squares of distances between data and the corresponding cluster centroid.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

The recorded speech samples may not be having of same length though the same words are collected from the same speaker. So there is a need to process the speech waveforms to a fixed set of feature vectors (same number of feature vectors to all the speech waves) to apply to the training system. By using End point detection algorithm the unvoiced samples will be removed at both the ends of speech waveforms. Then speech wave will be segmented to overlapped frames and then compute the feature vectors (MFCC or GFCC) frame wise. By using k-means algorithm with 'K' centroids, all the speech wave forms feature vectors are processed to a fixed set of k-feature vectors to each of the speech wave form.

Let the speech wave form is segmented to 'n' number of overlapped frames as 'Fi' with i = 1,...,n and the number of clusters are chosen as K. Let K centroids (vectors) will be initialized as S1, S2,...,SK ,Then Euclidean distance D_{ij} be calculated between the i^{th} vector and j^{th} centroid , Now calculate arg(min $\sum D_{ij}$) for i=1 to n and j=1 to K, and this value to be assigned to i^{th} vector, then calculate the mean of all the vectors assigned to the j^{th} cluster to obtain i^{th} centroid. The above process is repeated for finite number of iterations and obtained centroid vectors .

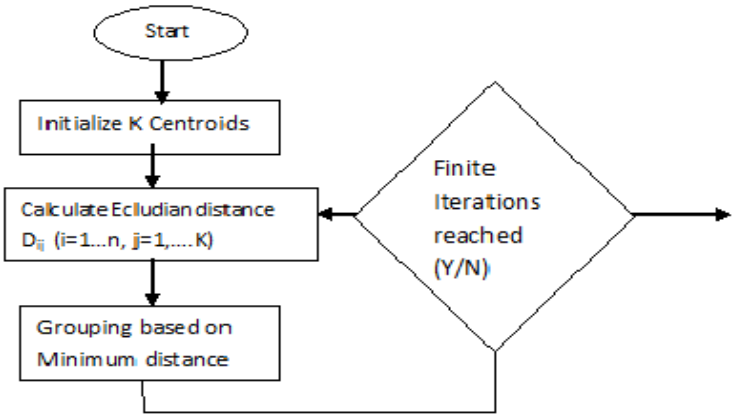


Fig.3: Flow chart of K-Means Algorithm

In this paper, the number of cluster centroids chosen as the no. of filter banks used to extract the features in MFCC and GFCC extraction process.

V. RADIAL BASIS NEURAL NETWORKS

In this project Radial basis networks have been chosen for analysis over the standard feed forward back propagation networks as the former will take less time for training and designs the network with zero error on the design/training vectors. The basic Radial basis networks are NEWRBE and NEWRB. Other variants of Radial basis networks are GRNNs and PNNs. NEWRBE is chosen to model the system and the basic neuron model will be as shown in below figure:

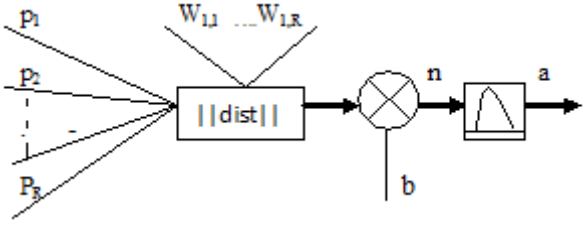


Fig.4: Neuron model of Radial basis function with R inputs.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

The final output 'a' is a radial basis function of 'n'. The net input to the radbas transfer function is the vector distance between its weight vector 'w' and the input vector 'p' multiplied by the bias 'b'.

$$n = \| w-p \| b \quad \text{-------eq.(4)}$$

$$a = \text{radbas}(n) = \exp(-n2) \quad \text{-------eq.(5)}$$

The radial basis function will have maximum value equal to '1' at the input is '0' i.e. if the distance between 'w' and 'p' decreases then the output increases. Thus, a radial basis neuron acts as a detector that produces 1 whenever the input 'p' is identical to its weight vector 'w'. The bias 'b' allows the sensitivity of the radbas neuron to be adjusted. For example, if a neuron had a bias of 0.1 it would output 0.5 for any input vector 'p' at vector distance of 0.8326 (0.8326/b) from its weight vector 'w'. NEWRBE is the basic function of Radial basis networks which will give zero error for the trained vectors.

net = newrbe (P,T,SPREAD) ----------eq.(6)

Where P is R by Q matrix of Q- R element vectors

T is S by Q matrix of S-element target class vector

SPREAD: spread of radial basis function (default=1)

The function newrbe takes matrices of input vectors P and target vectors T, and a spread constant SPREAD for the radial basis layer, and returns a network with weights and biases such that the outputs are exactly T when the inputs are P. The SPREAD should be chosen that it is large enough so that the active input regions of the radbas neurons overlap enough so that several radbas neurons always have fairly large outputs at any given moment. This makes the network function smoother and results in better generalization for new input vectors occurring between input vectors used in the design. However, SPREAD should not be so large that each neuron is effectively responding in the same, large, area of the input. Increase in SPREAD may also generate processing and numerical errors.

VI. EXPERIMENTAL SETUP

Total 8 isolated telugu words of .wav speech samples have been collected from 6 male speakers and 6 female speakers of the age group 15 to 40 years under office/Home environment in the presence of normal noisy conditions such as light/A.C /computer keyboard noise etc., each word has been recorded for 15 times from each speaker and so total of 8×12×15=1440 sample through MATLAB audio recorder object with 16 KHz sampling frequency .The words collected are general telugu commands as mentioned below .

Telugu word	Telugu Script	English Meaning
aagu	ఆగు	STOP
kadhulu	కదులు	START
kudi	కుడి	RIGHT SIDE
edama	ఎడమ	LEFT SIDE
muMdhuki	ముందుకి	FORWARD
venakki	వెనక్కి	BACKWARD
paiki	పైకి	UPWARD
kiMdhaki	కిందకి	DOWNWARD

Table.2: Telugu Isolated words collected in Office Environment(Noisy) .

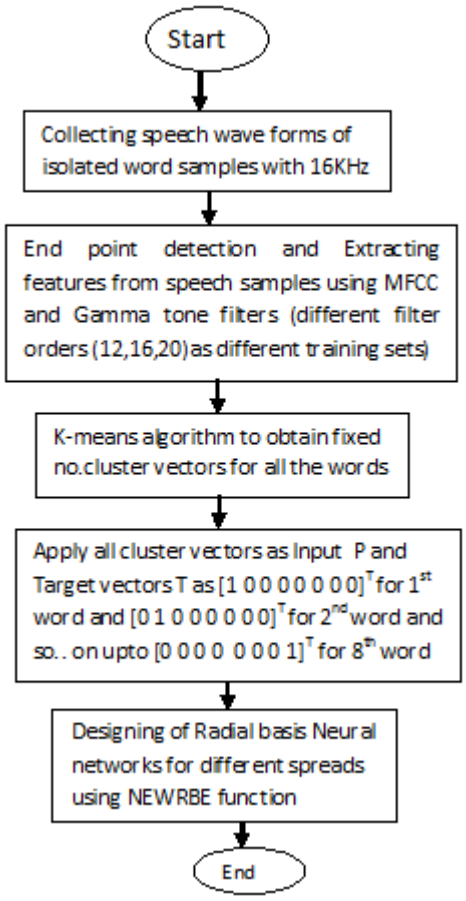
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

The above chosen set of telugu words are basic commands to direct any route or path .These commands can be used to direct any artificial machine through telugu voice commands under noisy conditions. Two different types of features MFCC (Mel Frequency Cepstral Coefficients) and Gamma tone coefficients (GFCC) have been extracted from all the collected samples. All the collected waveforms are first processed by end point detection algorithm to discard unvoiced part at both the ends of speech wave forms. Then the speech waveforms are segmented and windowed into frames of 20 msec (320 samples) with overlapping of 10 msec (160 samples) i.e 50% overlap. Each and every frame of the speech then processed and MFCC and Gamma tone coefficients are extracted .Both MFCC and GFCC filterbanks are designed with the help of MATLAB voice tool box, and then K-means algorithm is imposed on feature vectors and obtains the fixed set of cluster centroid vectors to all the speech waveforms. Now 50% of these MFCC and Gamma tone coefficients centroid vectors applied separately to NEWRBE function as training vectors and designed Radial basis neural networks .the rest of the 50% feature vectors used for testing the recognition success rate of the networks. Different experiments conducted for 3 set of Filterbanks (12,16 and 20 channels) and 3 different SPREADs of Radial basis networks. A separate set of testing features (not used in training) used in analyzing the recognition success rate.



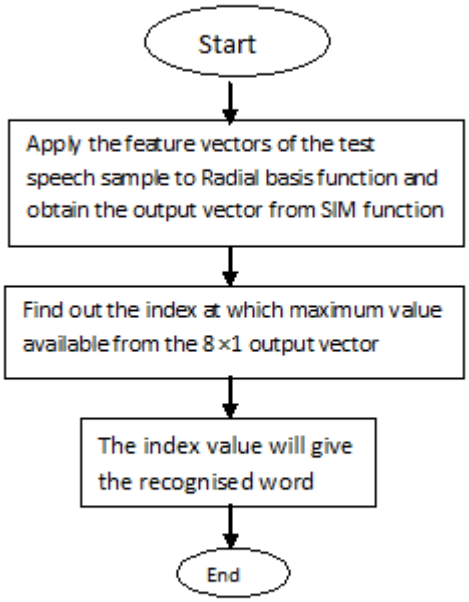


Fig 5(a): Training of Radial Basis Networks 5(b): Testing of Radial Basis Networks

VII. SIMULATION RESULTS

Two different types of experiments have been conducted. In case1 the training data from 12 different speakers with 10 samples of same word (total 8words*10 samples of each word*12 speakers (6 Male and 6 Female)) have been processed and MFCC and GFCC features vectors are given as input vectors and Radial basis neural networks designed separately for both MFCC and GFCC, In testing another set of samples which are not used in training phase but

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

collected from same speakers have been applied for testing and analyzed the results. Both MFCC and GFCC have given around 85% success rate.

In case2, the results are analyzed for different speakers which are not used in training. In this case, total 480 samples used for training (8 words*10 samples for each word*6 speakers) and the testing set consists of another 480 samples (8 words*10 samples for each word*6 another set of speakers), as neural networks perform inconsistently at different times, the same experiment have been conducted 10 times and the averages are calculated. At lower spreads MFCC shows better results and at higher spreads GFCC shows better. So spread should be chosen such that it fits for the feature vectors and smoothening the function. Overall GFCC features shows better results over MFCC for different channel counts (no. of filter banks) and the SPREAD of Radial basis Networks and the averaged results are tabulated:

Success Rate (%)	Spread-3		Spread-4		Spraed-5	
	MFCC	GFCC	MFCC	GFCC	MFCC	GFCC
Ch12	54.2	53	60.8	56.2	57.4	61.2
Ch16	54.8	53.6	61.4	58.1	58.8	62.4
Ch20	55.4	54.1	61	57.4	56.9	63.6

Table.3: Recognition success rate for different spreads and filter banks.

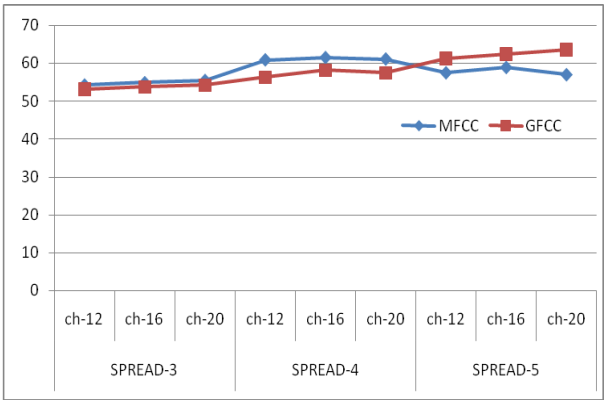


Fig.6: Percentage of success rate for different spreads and filter banks for both MFCC and GFCC.

VIII. CONCLUSION AND FUTURE WORK

From the results, it is observed that speech recognition success rate using GFCC features giving better results compared to MFCC in noisy conditions (Office/Home environment). The same experiment has been conducted 10 times and the average results are computed and considered.

It is also observed from the experiments that for low and very high Spread values the NEWRBE function is not giving promising results .For higher values of spread GFCC shows better results than MFCC. High Spread values may lead to mathematical complexities. So moderate SPREAD values have to be chosen such that the function will be smoothening the feature vectors and not giving any numerical and processing issues.

The work may be extended by examining the different noise conditions for different feature extraction procedures, training models such as ANNs (Artificial Neural Networks) and HMM models and examine for the improvement of recognition success rate in noisy conditions.

REFERENCES

[1] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, Senior Member, IEEE- JOURNAL OF AUTOMATIC CONTROL, UNIVERSITY OF BELGRADE, - Neural Networks used for Speech Recognition- VOL. 20:1-7, 2010

[2] Hari Krishna Maganti and Marco Matassoni - Auditory processing-based features for improving speech recognition in adverse acoustic conditions- - Magantiand Matassoni EURASIP Journal on Audio, Speech, and Music Processing May-2014

[3] Anirudha adiga and Chandrasekhar seemanthuls IISC Bangalore - Gammatone Wavelet Cepstral Coefficients for Robust Speech Recognition- - TENCON-2013

[4] JunQi, Dong Wang, Yi Jiang, Runsheng Liu- Tsinghua University, Beijing, China –AUDITORY FEATURES BASED ON GAMMATONE FILTERS FOR ROBUST SPEECH RECOGNITION-

[5] Md Jahangir Alam, Patrick Kenny, Douglas O'Shaughnessy- University of Quebec, Montreal, Quebec, Canad - Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum- Interspeech-2012

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

[6] Chanwoo Kim, Language Technologies Institute School of Computer Science Carnegie Mellon University Signal Processing for Robust Speech Recognition Motivated by Auditory Processing- -2010
[7] Xiaojia Zhao, Yang Shao and DeLiang Wang, "CASA-based Robust Speaker Identification," IEEE Trans. on Audio, Speech and Language Processing, vol.20, no.5, pp.1608-1616, 2012.
[8] Evaluation of Different Feature Extraction Techniques for Continuous Speech Recognition- International Journal of Information and Communication Technology Research- Volume 2 No. 12, December 2012
[9] Speech Recognition using Artificial Neural Networks and Hidden Markov Models- IEEE MULTIDISCIPLINARY ENGINEERING EDUCATION MAGAZINE, VOL. 3, NO. 3, SEPTEMBER 2008
[10] Prasad D Polur, Ruobing Zhou, Jun Yang, Fedra Adnani, Rosalyn S. Hobson - ISOLATED SPEECH RECOGNITION USING ARTIFICIAL NEURAL NETWORKS: - 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey.
[11] Philip Felber Illinois Institute of Technology - SPEECH RECOGNITION Report of an Isolated Word experiment.- April 25, 2001
[12] Heikki N. Koivo - NEURAL NETWORKS: Basics using MATLAB Neural Network Toolbox - 2008
[13] Automatic Noise Recognition Based on Neural Network Using LPC and MFCC Feature Parameters- Proceedings of the Federated Conference on Computer Science and Information Systems pp. 69–73 ISBN 978-83-60810-51-4
[14] THE OPTIMAL PERFORMANCE OF MULTI-LAYER NEURAL NETWORK FOR SPEAKER-INDEPENDENT ISOLATED SPOKEN MALAY PARLIAMENTARY SPEECH - ISSN 2231-7473 2010 Faculty of Computer and Mathematical Sciences, Universiti Teknolgi MARA, (UiTM), Malaysia
[15] Role of neural network models for developing speech systems- S¯adhan¯a Vol. 36, Part 5, October 2011, pp. 783–836. c Indian Academy of Sciences
[16] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA - RECENT ADVANCES IN DEEP LEARNING FOR SPEECH RESEARCH AT MICROSOFT- - ICASSP 2013
[17] Neural Network Toolbox For Use with MATLAB® Howard Demuth Mark Beale User's Guide Version 4 COPYRIGHT 1992 - 2004 by The MathWorks, Inc.
[18] Ganesh K Venayagamoorthy, Viresh Moonasar and Kumbes Sandrasegaran* Electronics Engineering Department, M L Sultan Technikon, Durban, South Africa gkumar@saiee. org. co.za *Institute for Information Sciences and Technologv (IIST), Massey Universiv, New Zealand K. Sandrasegaran@massey. ac. nz - VOICE RECOGNITION USING NEURAL NETWORKS- 0-7803-5054-5.0029 1998 IEEE
[19] Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques- JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617
[20] Development of Isolated Word Speech Recognition System- INFORMATICA, 2002, Vol. 13, No. 1, 37–46 2002 Institute of Mathematics and Informatics, Vilnius
[21] Rafik Djemili, Mouldi Bedda, and Hocine Bourouba- Recognition of Spoken Arabic Digits Using Neural Predictive Hidden Markov Models- The International Arab Journal of Information Technology, Vol. 1, No. 2, July 2004