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. 4, Issue 5, May 2016

Free Model Speech Recognition System Using MFCC Model

Bhavna Bhardwaj, Dr. Dinesh Kumar

M. Tech Student, Dept. of CSE, SRCEM, Palwal, MD University, Haryana, India

Prof. [HOD], Dept. of CSE, SRCEM, Palwal, MD University, Haryana, India

ABSTRACT : Being visually impaired or a laptop illiterate have invariably being a barrier for folks to use laptop to perform their tasks in a straightforward, economical and fast method. However, computer/laptops are solely a necessity of everybody nowadays however is convenient to be operated by anyone and everybody. This paper discus a system whereby the user are able to kind text on laptop by providing a voice input through his movable.

KEYWORDS: Speech Recognition System, MFCC, HMM, N-Gram Dataset.

I. INTRODUCTION

Irrespective of age, gender, academic background or physical impairment (excluding being dumb), speech is that the primary mode of communication employed by individuals to specific themselves to others. it's this human speech that forms the premise of 1 of the most well liked field of contemporary science i.e. Speech Recognition. The analysis work conferred during this paper uses a movable to produce speech input to a speech recognition system that successively gets written as text on the pc screen This paper is split into 2 major elements, within the initial half associate introductory summary of speech recognition system is provided at the side of a glimpse of a number of the recent analysises within the field and therefore the second half discuses the research work into consideration during this paper.

II. SPEECH RECOGNITION SYSTEM

Popularly called Automatic Speech Recognition (ASR), speech recognition could be a sub-category of pattern recognition whereby the speech input is 1st understood by the pc to perform the user desired task victimisation it. It's preliminary task isn't solely to retort instantly however additionally to figure effectively in noise or complete silence surroundings which too on heterogeneous inputs.

Types of Speech Input [1] [8]

Different ASR systems settle for speech input in several forms.

• Isolated Words: Isolated i.e. single utterances square measure fed as input to the recognizer however selecting word boundaries affects result obtained.

- Connected Words: Separate utterances together with minimum pause are input requirement of this system.
- Continuous Speech: A dictation by computer to the speaker, it is the most difficult recognizers to create.
- Spontaneous Speech: Speaker's natural speech acts as the input for the system.
- Connected Words: Separate utterances together with minimum pause are input requirement of this system.
- Continuous Speech: A dictation by computer to the speaker, it is the most difficult recognizers to create.
- Spontaneous Speech: Speaker's natural speech acts as the input for the system.

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. 4, Issue 5, May 2016

Stages of Speech Recognition Technique [2] [3]

1. Analysis: Vocal tract, excitation state characteristics and behavior characteristic of the speaker are identified.
2. Feature Extraction: Spectral features along with excitation source are identified. This stage is further divided into two steps. The first is the training step, shown in Fig. 1 below, wherein the system is familiarized with the speaker's voice characteristics and these act as reference models.

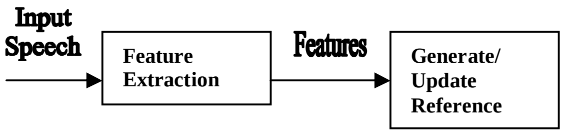

Fig.1 Training [7].

The second step is testing, shown in Fig. 2, where unknown utterances are matched with the reference model to find their best possible match.

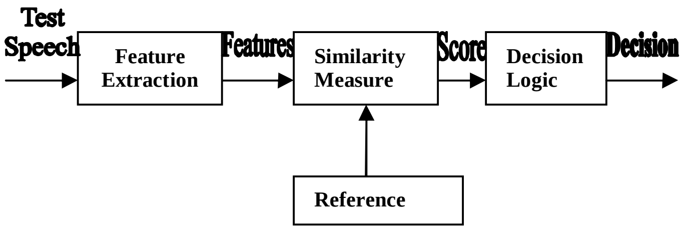

Fig. 2 Testing [7].

Modeling Technique: Here speaker models are generated which are categorized into speaker identification (where input speech signal helps to identify the speaker) and speaker recognition. Speaker recognition is further divided into speaker dependent (where the extracted characteristics identify speaker) and speaker independent (where only the content matters and not who is speaking it). The various modeling techniques available are :

- *Acoustic-Phonetic Approach:* Based on acoustic properties labels are allotted to speech sounds.
- *Pattern Recognition Approach:* First the system is trained with utterances which act as reference patterns and then unknown utterances are compared to these references to know their identity.
- *Template Based Approach:* In this there is a candidate word dictionary which acts as templates. Input unknown utterances are matched with these templates and the one that matches the best is selected. However, the production and storage of template per word is an impractical task.
- *Dynamic Time Warping (DTR):* It measures the similarities between two utterances that vary in speed and are "warped" non-linearly in time dimension, on a frame-by-frame basis.
- *Knowledge Based Approach:* The knowledge of experts related to variation in speech is hand coded in the system, though it is not easy to obtain such knowledge and use it effectively.

Matching Technique:

- *Whole Word Matching:* There is a large storage requirement in this approach as all incoming speech signals are compared to these pre-recorded word templates. Results are obtained quite quickly.
- *Sub-word Matching:* Recognition is done with the help of phonemes. Though there is less storage requirement for templates but processing time is more.

III. HIDDEN MARKOV MODEL: SPEECH RECOGNITION TECHNIQUE

As discussed above training and testing are the two phases of speech recognition. Hidden Markov Model i.e. HMM is a successful and flexible technique used in speech recognition. In this research work also HMM is used so a brief overview of it is given in this section.

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. 4, Issue 5, May 2016

Introduction of HMM

An extension of Markov model, HMM [6] [31] is a model in which the intermediate states that are responsible to transform the initial state to final output state are hidden. The triplet which represents HMM is given as:

$$(A, B, \pi) \tag{1}$$

where A is state-transition probability
B is observation/output probability
π is initial state
HMM is a probabilistic model wherein the model can go from one state to another within a time shift but it is only probabilistic.

Elements of HMM

- N number of states in the model - Though the principle of HMM is that it's states are hidden but still they have physical significance.
- M number of distinct observation symbols per state – It represent the physical output emitted by the model under consideration.
- S represent the individual states

Steps in HMM

1. Evaluation: In this step the probability of a model to generate an observation sequence is judged so as to find the best model available. For the HMM model λ, the observation sequence O is given as :

$$P(O \mid \lambda) \tag{2}$$

2. Decoding: It is the process wherein the best state sequence, Q, is obtained for the observation sequence, O.
3. Training (Learning): The most tedious step of HMM where the model parameters (A, B, π) are adjusted to maximize the observation sequence probability.

Basic Problems of HMM

Problem 1: For the given observation sequence O and a model $\lambda = (A, B, \pi)$, how to efficiently find the best model i.e. $P(O \mid \lambda)$.

Problem 2: For the given observation sequence O and a model λ, how to choose the state sequence Q which best explains the observation.

Problem 3: How the model parameters (A, B, π) should be adjusted to maximize $P(O \mid \lambda)$.

IV. MEL FREQUENCY CEPSTRUM COEFFICIENT (MFCC): FEATURE EXTRACTION TECHNIQUE

MFCC is the feature extraction technique employed in this research work which not only extracts but also selects the parametric representation which is best for acoustic signal. The steps involved in MFCC calculation are:

1. *Mel Frequency Wrapping*: The pitch under consideration is measured using a 'mel' scale as human speech does not follow linear scale for measuring frequency of speech signal. 'Mel' scale spaces linearly below 1000 Hz and logarithmically above 1000 Hz. Mathematically it is formulated as :

$$Mel(f) = 2595 * \log 10(1 + f / 700)$$

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. 4, Issue 5, May 2016

2. *Cepstrum*: The log mel spectrum obtained from above step, are converted to time from real numbers using discrete cosine transform, which represent local spectral properties of the speech signal.

Fig. 2.

$$C_n = \sum_{k=1}^{k} (\log S_k) \cos \{n(k-(1/2)*\pi /k)\}$$

PERFORMANCE OF SYSTEMS

Rate of accuracy and speed are the two parameters that measure the effectiveness of any speech recognition system [7]. Word Error Rate (WER) measures accuracy and is given in equation (1) :

$$WER = (S + D + I) / N \qquad (5)$$

where S are the number of words that are substituted
D are the number of words that are deleted
I are the number of words that are inserted
N are the total number of words under consideration

Speed is measured in terms of Real Time Factor (RTF) which is given in equation (2) :

$$RTF = P / I \qquad (6)$$

where P is the time that will be taken to process input of I duration

Speech recognizer is widely used in the field of voice authentication where the recognizer on the basis of input speech signal judges the authenticity of the speaker. The speaker should neither be authenticated by the recognizer when he should not be nor should the speaker be unauthenticated when he should i.e. a balance needs to be maintained between false acceptance rate (FAR) and false rejection rate (FRR). On plotting a graph between FAR and FRR, as shown in Fig. 3, the point of intersection is known as crossover error rate (CER) which must be low for a better system performance.

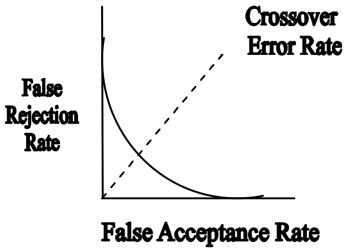

Fig. 3 FAR, FER, CER [4].

Analysis of Existing Systems

The table 1 below brings into picture some of the existing systems whose accuracy varies on the basis of which feature extraction technique they choose out of LPC (Linear Predictive Coding), MFCC (Mel Frequency Cepstral Coefficient) and PLP (Perceptual Linear Prediction) and which recognition technique they choose out of HMM (Hidden Markov Model), GA (Genetic Algorithm) and VQ (Vector Quantization).

Research Work Name	Feature Extraction Technique	Recognition Technique	Accuracy
Alaigal-A Tamil Speech Recognition [9]	PLP	HMM	70% - 80%
Hindi Speech Recognition System Using HTK [11]	MFCC	HMM	Word-accuracy and word-error rate of the system are 94.63% and 5.37% respectively
Speech Emotion Recognition System based on Integrating Feature and Improved HMM [12]	MFCC	HMM + GA	More than 77%
Continuous Speech Recognition System for Tamil Using Monophone-based Hidden Markov Model [15]	MFCC	HMM	92% accuracy in word level and 81% accuracy in sentence level
Automatic Speech Recognition for Bangla Digits [17]	MFCC	HMM	More than 95% for digits (0-5) and less than 90% for digits (6-9)
Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK) [19]	MFCC	HMM	97.99%
English Digits Speech Recognition System Based on Hidden Markov Models [27]	MFCC	HMM	56.25% - 72.5%
Human Computer Interaction Using Isolated Words Speech Recognition Technology [30]	MFCC	VQ	88%
Segment-Based Stochastic Modelings for Speech Recognition [32]	LPC	HMM + VQ	62% - 96%
Automatic Speech Recognition:Human Computer Interface for Kinyarwanda Language [33]	MFCC	HMM	92%

Table 1: Analysis of Existing Speech Recognition Systems.

PROPOSED SYSTEM

Initially the input signal is transformed from analog to digital form and then it is divided into frames each having their individual frequency. The spectral features extracted from the frames help in identifying the phones of the speech sample which are nothing but sounds that distinguish two words. MFCC is applied to the phones which identify unique discrete acoustic phone of each individual input speech sample. Then HMM is applied to each phone which identifies the likelihood of occurrence of a word W within the given acoustic observation P(W/A). This can be represented using Baye's rule format:

$$P(W/A) = P(A/W)P(W)/P(A) \qquad (7)$$

where P(A/W) represents the acoustic model

P(W) represents the language model giving the probability of sequence of words

P(W) is calculated using the Microsoft N-Gram dataset which has the ability to predict the next phrase, letter or word in a given sequence of input.

The N-Gram dataset is an error correction mechanism which corrects the error produce by the ASR system. The proposed method works on a post editing approach wherein spell checking of the output obtained from the ASR system

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. 4, Issue 5, May 2016

is done after the input speech is converted into text. The N-Gram dataset is being used since it is a huge repository of data with data being collected from world web pages and internet documents, which have data ranging from proper nouns, domain specific terms, special expressions, technical words, acronyms and terminologies; covering an ample number of words of the language.

With the help of N-Gram dataset, for the proposed system vocabulary dictionary and sentences are available which act as a huge repository for both training as well as testing phase of the proposed system. An algorithm called soundX is then created. This algorithm finds the best possible match on basis of co-articulation. Co-articulation is the situation wherein occurrence of a word depends on word before and after it. For eg: If the sentence being provided as input speech is: "Can I kiss the baby!", then the N-Gram dataset will generate possible available matches "kill" and "kiss". However, the soundX algorithm will take into picture the co-articulation effect and select the best of the possible matches i.e. "kiss" that will fit into the sentence. In other words, from all the possible matches returned by the N-Gram dataset, soundX selects the one that fits the best. This sentence thus selected with the help of N-Gram dataset and soundX algorithm is printed as text on screen.

This work has reduced human effort to a great extent as human can speak more quickly than typing via. a keyboard. Also while a speaker provide voice input he can utilize his hands for any other work. Moreover the system can also prove to be a blessing for visually challenged people or the ones which don't have hands.

V. EXPERIMENTAL SETUP

Fig. 4 below demonstrates the experimental setup for the proposed work.

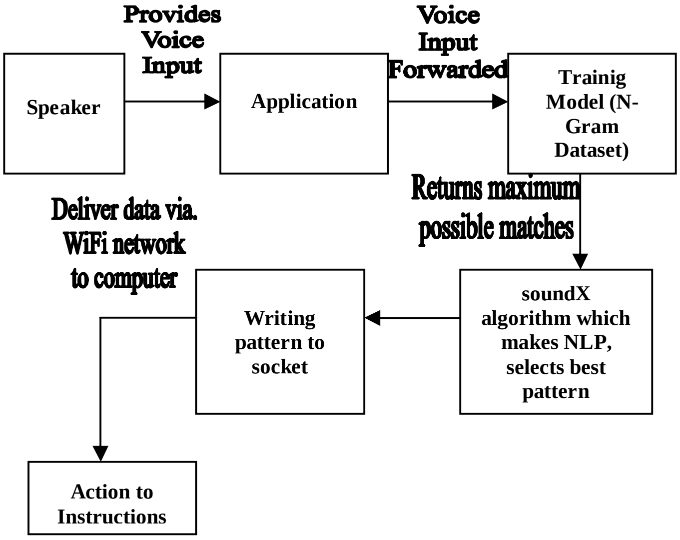

Fig. 4 Experimental setup for the proposed system

The objective of the proposed system is to print text on computer screen by providing voice input and not by using traditional means like keyboard or mouse but by providing input through android mobile phone. From the speech input, features are extracted using MFCC and the acoustic characteristics are recognized using HMM. MFCC and HMM are basically used to obtain the speech signal in it's maximum possible pure form i.e. devoid of noise or any other impurities that are fed along with speech input. This modified signal is fed to Google server which returns the maximum possible matches for the input speech signal by making use of N-Gram dataset. Then soundX algorithm is designed which makes use of NLP (Natural Language Processing).

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. 4, Issue 5, May 2016

VI. RESULTS

This voice input goes to the refinement process using MFCC and HMM and then goes to Application model comprising of grammar rule which makes use of N-Gram dataset and returns the best possible matches for the input provided, From these matches returned, the soundX algorithm designed during this proposed system, selects the best possible matches. Now if user again provides some input say “santa banta”, then after going through all the processing as before, this new input gets appended after punctuation list to another punctuation list.

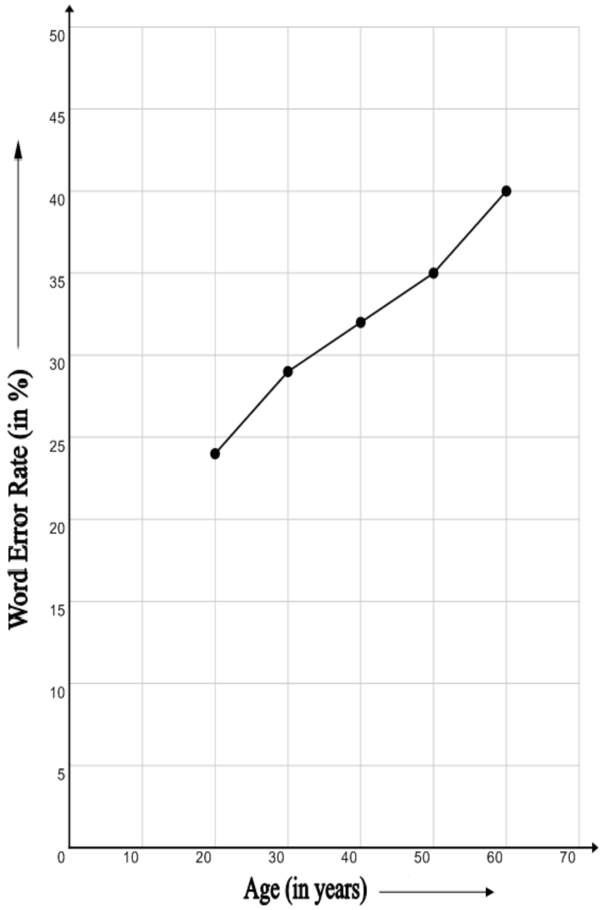

Fig. 5 Graph plotted between WER and Age

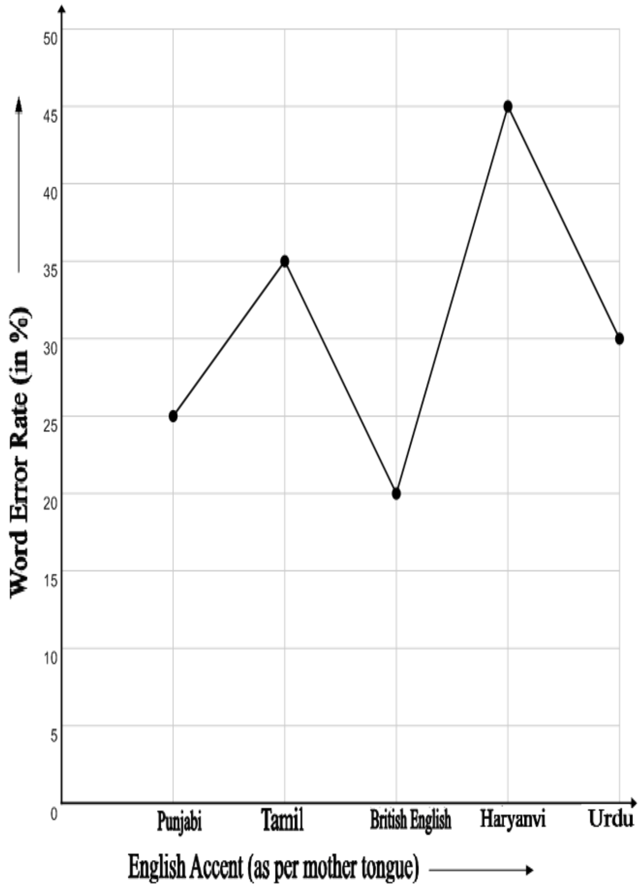

Fig. 7 Graph plotted between WER and English Accent (as per mother tongue)

 DOI: 10.15680/IJIRCCE.2016. 0405295

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. 4, Issue 5, May 2016

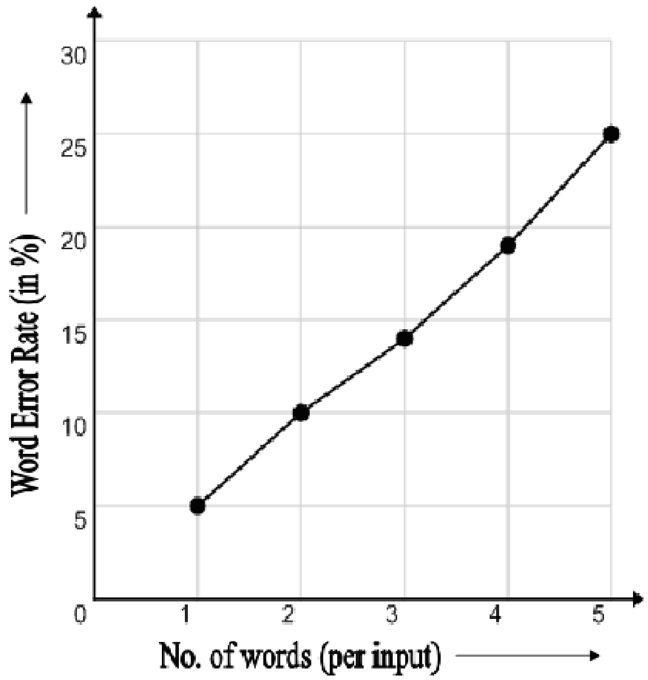

Fig. 6 Graph plotted between WER and No. of words per input

VII. CONCLUSION AND FUTURE WORK

More than fifty years is that the gift of speech recognition system. The ASR systems influence be helpful not just for blind individuals however additionally let ready individuals to try and do alternative or another work as their hands and eyes square measure liberated to bask in other activities. Moreover, these ASR systems is developed in and square measure existent in native languages like English, Tamil, Punjabi, Hindi, Chinese, etc; that breaks the obstacle of the person being educationally unfortunate to work the pc. Speaker freelance continuous speech recognition systems with giant vocabulary, square measure in-demand which might be consummated by victimization the feature extraction technique MFCC with the popularity technique HMM that facilitate in making very powerful systems that provide sensible speech recognition results.

For future work, a brand new system is designed on similar grounds wherever not solely text is being written on screen however most of the pc functioning is handled by providing voice input through itinerant. this could vary from writing URL in browser's address bar to computer's management functions like: Ctrl+C, Ctrl+X; to call some to attaching files to e-mails. Moreover, presently this projected system need each the mobile and laptop computer to be in same

REFERENCES

[1] Pradeep Kumar Jaisal, Pankaj Kumar Mishra, "A Review of Speech Pattern Recognition Survey", International Journal of Computer Science and Technology, 2012.
[2] Santosh K.Gaikward, Bharti W.Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications, 2010.
[3] "Speech Recognition Technology Choices", A Vocollect White Paper, 2010.
[4] Lisa Myers, "An Exploration of Voice Biometrics", SANS Institute Infosec, 2004.
[5] Rudan Bettelheim, David Steele, "Speech and Command Recognition", FreeScale White Paper, 2010.
[6] Bhupinder Singh, Neha Kapur, Puneet Kaur, "Speech Recognition with Hidden Markov Model : A Review", International Journal of Advanced Research in Computer Science and Software Engineering, 2012.
[7] Vibha Tiwari, "MFCC and it's Applications in Speaker Recognition", International Journal on Emerging Technique, 2010.

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. 4, Issue 5, May 2016

[8] Vimala C., Dr. V.Radha, "A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal, 2012. A.P.
[9] Henry Charles G. Devaraj, "Alaigal- A Tamil Speech Recognition", Tamil Internet Singapore, 2004.
[10] Er. Jaspreet Kaur, Er. Nidhi, Ms. Rupinder Kaur, "Issues Involved in speech To Text Conversion", International Journal of Computional Engineering Research, 2012.
[11] Kuldeep Kumar, R.K.Aggarwal, "Hindi Speech Recognition System using HTK", International Journal of Computing and Business Research, 2011.
[12] Han Zhiyan, Lun Shxian, Wang Jian, "Speech Emotion Recognition System based on Integrating Feature and Improved HMM", The 2nd International Conference on Computer Application and System Modeling, 2012.
[13] Ore A. Soluade, "A Comparative Analysis of Speech Recognition Platforms", Communications of the IIMA, 2009.
[14] M. Chandrashekhar, M.Ponnavaikko, "Tamil Speech Recognition: A Complete Model", Elctronic Journal Technical Acoustics, 2008.
[15] Radha. V, Vimala. C, Krishnaveni. M, "Continuous Speech Recognition System for Tamil Language Using Monophone-based Hidden Markov Model", CCSEIT-12, ACM, 2012.
[16] Zhao Lishuang, Han Zhiyan, "Speech Recognition System Based on Integrating Feature and HMM", IEEE DOI 10.1109/ICTMA.2010.298, 2010.
[17] Ghulam Muhammad, Yousef A. Alotaibi, Mohammad Nurul Huda, "Automatic Speech Recognition for Bangla Digits", IEEE 978-1-4244-6284-1109/ICCIT.2009, 2009.
[18] Marek Bohac, "Performance Comparison of Several Techniques to Detect Words in Audio Streams and Audio Scene", IEEE 54th International Symposium ELMAR-2012, 2012.
Bassam A.Q. Al-Qatab, Raja N. Ainon, "Arabic Speech Narayanana, "The Vera Am Mittag German Audio
Visual Emotional Speech Database", ICME 2008, (HTK)", IEEE 978-1-4244-6716-711 0, 2010.
[19] Sandipan Mandal, Biswajit Das, Pabitra Mitra, "Shruti-II: A Vernacular Speech Recognition System in Bengali and an Application for Visually Impaired Community",
IEEE 978-1-4244-5974-2/10, 2010.
[20] Xiaofeng Wu, Ryohei Nakatsu, "Vision-aided Speech Recognition System for a Small Four-Legged Robot", IEEE 978-1-4244-5858-5/10, 2010.
[21] Sanjana Primorac, Mladen Russo, "Android Application for Sending SMS messages with Speech Recognition Interface", MIPRO, 2012.
[22] Yanan Jian, Jianshe Jin, "An Interactive Interface between Human & Computer based on Pattern & Speech Recognition", International Conference on Systems & Informatics, IEEE 978-1-4673-0199-2122, 2012.
[23] Huda Sarfaraz, Sarmad Hussain, Riffat Bokhari, Agha Ali Raza, Inam Ullah, Zahid Sarfaraz, Sophia Pervez, Asad Mustafa, Iqra Javed, Rahila Praveen, "Large Vocabulary Continuous Speech Recognition for Urdu", ACM 978 1-4503-034202/10/12, 2010.
[24] Matus Pleva, Stanislav Ondas, Jozef Juhar, Anton Cizmar, Jan Papaj, Lubomir Dobos, "Speech & Mobile Technologies for Cognitive Communication & Information Systems", IEEE Proceedings, 2nd International Conference on Cognitive InfoCommunications (CogInfoCom), 2011.
[25] Oscal T.C. Chen, Jhen Jhan Gu, Ping-Tsung Lu and Jia-You Ke, "Emotion Inspired Age and Gender Recognition Systems", IEEE 978-1-4673-2527-1/12, 2012.
[26] Ahmad A.M. Abushariah, Teddy S. Gunawan, Mohammad A.M. Abushariah, Othman O. Khalifa, "English Digits Speech Recognition System Based on Hidden Markov Models", International Conference on Computer & Communication Engineering, 978-1-4244-6235-3/IEEE, 2010.
[27] A. Mukhopadhayay, S. Chakraborty, M. Choudhary, A. Lahiri, S. Dey, A.Basu, "Shruti: An Embedded text-to speech system for Indian Languages", IEEE Proceeding-Software Vol.153, No.2, April 2006.
[28] M.U. Akram, M. Arif, "Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling approach", 0-7803-8680-9/04, IEEE, 2004.
[29] Mohammad A.M. Abu Shariah, Raja N. Ainon, Roziati Zainuddin, Othmn O. Khalifa, "Human Computer Interaction Using Isolated Words Speech Recognition Technology", International Conference on Intelligent & Advanced Systems, 1-4244-1355-9/07, IEEE, 2007.
[30] Mohd Zaizu Ilyas, Salina Abdul Samad, Aini Hussain, Khairul Anuar Ishak, "Speaker Verification using Vector Quantization and Hidden Markov Model", The 5th Student Conference on Research and Development, 1-4244-1470-9/07, IEEE, 2007.
[31] Che Yong Yeo, S.A.R. Al-Haddad, Chee Kyun Ng, "Animal Voice Recognition for Identification (ID) Detection System", IEEE 7th International Colloquium on Signal Processing & it's Applications, 2011.
[32] Yousef Basil, Paul Semaan, "ASR Context-Sensitive Error Correction Based on Microsoft N-Gram Dataset", Journal ofComputing, 2012.
[33] Charl van Heerden, Etienne Barnard, Michael Feld, Christian Miller, "Combining Regression & Classification Methods for Improving Automatic Speaker Age Recognition", ICASSP 2010, IEEE, 2010.
Michael Grimm, Kristian Kroschel, Shrikanth
[34] Charles Corfield, "Demystifying Speech Recognition", nVoq White Paper.
[35] Z.Hachkar, B.Mounir, A.Farchi, J.El Abbadi, "Comparison of MFCC and PLP Parameterization", Canadian Journal on Artificial Intelligence, Machine Learning & Pattern Recognition, 2011.
[36] Lidia Mangu, Mukund Padmanabhan, "Error Corrective Mechanisms for Speech Recognition", IEEE 0-7803-7041-4/01, 2001.
[37] S.Basu, C.Neti, N.Rajput, A.Senior, L.Subramaniam, A.Verma, "Audio-Visual Large Vocabulary Continuous Speech Recognition in Broadcast Domain", IEEE 0-7803-5610-1/99, 1999.
[38] Michael Feld, Etienne Barnard, Charl van Heerden, Christian Muller, "Multilingual Speaker Age Recognition: Recognition Analyses on the Lwazi Corpus", IEEE 978-1-4244-5480-8/09, ASRU 2009.
[39] F. Reena Sharma, S. Geetanjali Wason, "Speech Recognition & Synthesis Tool: Assistive Technology for Physically Disabled Persons", International Journal of Computer Science & Telecommunications [Volume 3, Issue 4, April 2012].
[40] M. Jackson, "Automatic Speech Recognition: Human Computer Interface for Kinyarwanda Language", Master Thesis, Faculty of Computing & Information Technology, Makerere University, 2005
[41] Nelson Morgan, "Deep and Wide: Multiple Layers in Automatic Speech Recognition", IEEE Transactions on Audio, Speech & Language Processing, Vol. 20, No. 1, January 2012.