IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Speech Corpora, Feature Extraction Techniques and Classifiers with Special Reference to Automatic Speech Recognition

D. Dutta[1*], R.D.Choudhury[2], S.Gogoi[3]

[1]Department of IT, Gauhati University, Assam, India
[2,3]CS & IT, IDOL, Gauhati University, Assam, India

[*]Corresponding Author: dipankar.jrt1982@gmail.com, MN.: +09-98642-09351

DOI: https://doi.org/10.26438/ijcse/v7i2.372378 | Available online at: www.ijcseonline.org

Accepted: 14/Feb/2019, Published: 28/Feb/2019

Abstract— in the current years, speech recognition has emerged as an important research area. To carry out further research on automatic speech recognition, a comprehensive review of existing work in this domain stands useful and constructive for the researchers. This paper has presented a recent literature review on speech recognition considering various existing speech corpora, speech features and different models or classifiers used in speech recognition. Different speech databases have been compared in terms of the number of speakers, type of speakers such as native or acted, age and gender of speakers and speech recording environment. Various techniques for speech signal acquisition and pre-processing of the speech signals are also addressed in this work.

Keywords— Automatic speech recognition, boundary detection, Feature-extraction, classifier, Mel-Frequency coefficient, phonemes, Speech Filter

NOMENCLATURE

ASR −Automatic Speech Recognition, ASRS − Automatic Speech recognition System, DBN− Deep Belief Neural Network, HMM− Hidden Markov Model, LDA − Linear Discriminant analysis, GMM − Gaussian Mixture Model, DFT−Discrete Fourier Transformation, SVM−Support Vector Machine, DTW−Dynamic Time Warping, ANN − Artificial Neural Network, DNN − Deep Neural Network, RNN − Recurrent Neural Network, KNN − k− nearest− neighbour, PLP− Perceptually Based Linear Predictive, MFCCs −Mel-Frequency Cepstral Coefficients, VTLN −Vocal Tract Length Normalization, FIR −Finite Impulse Response Filter, LVSR− Large Vocabulary Speech Recognition, IIR −Infinite Impulse Response Filter, CNN − Convolutional Neural Network, ZCR − Zero-crossing rate.

I. INTRODUCTION

The primary and most efficient media for communication among human is natural languages in the form of the speech signal. Communication between a system and human being require the technological development on natural speech recognition. The main function of an ASRS is to translate an audio signal or voice, produced by a human into its corresponding textual version. The purpose of developing an ASRS is to build a robust system to accept any speech signal uttered by the human as an input and recognize it with complete accuracy. Researchers have been kept trying to increase the ability of a machine to understand speech signals, produced by the human. The process was started in the year 1952 by introducing a technique for digit recognition. Nowadays ASRS has been used in various fields such as virtual reality, multimedia searches, natural language understandings, fighter aircraft, helicopters, automatic translation, home automation, telephony, hands-free computing, etc.

1. Types of ASR system: Based on the various parameters, ASRS can be classified into the following categories –

1.1 Speech: Based on the type of utterances an ASR system has been classified into four different categories:

- isolated word: This type of ASRS is designed to recognize human voice with single utterances.
- Connected word recognition system: Similar to the isolated word recognizer except that it can process multiple utterances with least amount of pause among the words.
- Continuous word recognition system: This type of recognizer accepts naturally spoken sentences uttered by a human.
- Spontaneous speech recognition system: This type of recognition system can accept and recognize spontaneous speech such as incomplete sentences, utterances with a laugh,

coughing etc. Spontaneous speech is very difficult to acquire as well as recognize.

1.2 Vocabulary size: Depending on the volume of the speech database, the ASR system is classified into three different categories, such as — small vocabulary containing less than or equal to 100 numbers of utterances, medium vocabulary with more than 100 and less than or equal to 500 numbers of utterances and large vocabulary with of more than 500 numbers of utterances.

1.3 Speaker: On the basis of the known and unknown speakers, ASRS can be categorized as speaker dependent and speaker independent.

Speaker dependent: This type of ASR system can recognize only the pre-processed speech data, uttered by selected speakers.

Speaker independent: This type of speech recognizer can recognize speeches uttered by any speaker of a specific language. Thus, this type of recognition system provides flexibility to recognize the utterances made by unknown speakers of a particular language. Training is not required for the new unknown speakers.

Rest of the paper is organized as follows - Section II contains the introduction of the components of an ASR system with a block diagram, Section III presents a comparative study between ASR system and Human being. The Section IV concludes the research work with future directions.

II. COMPONENTS OF AN ASRS

ASR systems can be divided into two main parts – Pre-processing and Post processing. Pre-processing includes the speech signal acquisition, framing, windowing, features extractions, and training. Post-processing phase includes the classification of unknown speech signals using a suitable algorithm to generate the respective output. ASR systems can be depicted as shown in Fig .1.

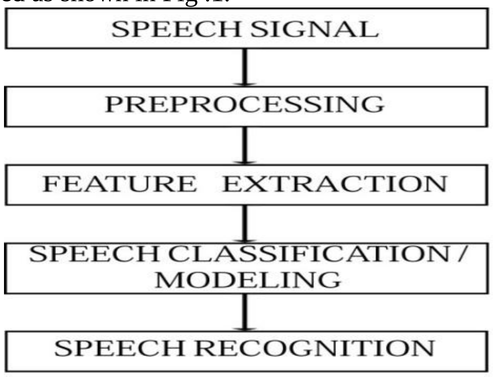

Fig 1 : Speech Recognition Process

It consists of five different modules such as

- Speech signal acquisition
- Pre-processing
- Feature extraction
- Acoustic Modelling
- Recognition

A. Speech signal acquisition

Speech signals are recorded during this phase of the ASR system. A transducer such as a microphone or a telephone is usually used to record the speech uttered by speakers who are told to sit in front of the microphone at a distance of 10-12 cm in a normal environment without noise or in a noisy environment. In general, a sampling rate from 8 KHz to 16 KHz with the mono channel is used during recording the speech. Different speech factors have to be considered during recording of speech signals, such as recording environment, whether it is noisy or quite enough, robustness of speech, speakers - age, gender, emotion, and health condition and type i.e. native/actor, size of the vocabulary, repetition of the utterances by different speakers, number of languages. TIMIT is a speech corpus used in many experiments [1], [2], [3] of the ASR system. Recent years, videos have also been recorded along with the speech to develop speech corpus. In [4], an audio visual corpus has been proposed with 34,000 utterances of six-word sentences in the English language. Speech, as well as video, has been collected from 34 different speakers including 18 numbers of males and 16 numbers of females of age between 18-49 years and of 1,000 utterances each. Some existing robust-noise databases are - a corpus BEN_ASR001 in Bengali language has been developed through recording telephonic conversations of 1,000 different speakers in a low background noise, a large vocabulary speech corpus EN1_ASR001 with 117,900 utterances recorded in a mixed environment from 50 numbers of speakers, MAC_ASR001 has been developed in Mandarin language with 200,000 utterances from 2,000 different speakers of age between 16-60 years. The recordings of those speech corpora have been done from mobile telephonic conversations. A real-world large vocabulary with 21,535 utterances has been used in [5] for continuous and spontaneous speech recognition. In this work, mobile users are allowed to do their required web search or their daily business through voice. This database consists of speech with background noise, music, hesitation, accent, sloppy pronunciation, repetition, and different audio channels. In [6], T1-NIST database has been proposed with 28,383 digits for digit recognition. Another database has been designed with 75,000 utterances, considering 5 different isolated words and being collected from different telephone customers [7]. In 2013, a speech dataset has been designed by U. Bhattacharjee [8], collecting 30,000 utterances from 27 males and 23 females for 30 different phonemes available in Assamese language. In the research work [9] around 75,000 isolated utterances had been collected from speakers to build the dataset, where 7,980 random isolated utterances are digitized and used for testing. In [7], the authors have been proposed a dataset of thirty-nine (39) isolated words, including English alphabets, three specific English words, and ten numbers of single digits. A medium size dataset was used in AT&T Bell Lab [10], with one hundred twenty-nine (129) words. In [11], authors have been proposed a dataset of spontaneous utterances recorded from mobile phone conversations across the US. Authors allowed the speakers to carried out their daily dealings and web browsing from their mobile phones using voice in a normal environment i.e. with different types of noise such as side speech, background music etc. in the surroundings.

B. Pre-processing

During speech signal acquisition lots of errors may occur, unwanted voice i.e. noise, voiceless signals may be recorded. Presence of unwanted signals in speech may decrease the recognition performance of the ASR system. To remove all unwanted signals from the speech dataset, researchers have to go through a set of processes known as the pre-processing phase which involves analog-to-digital conversion, end-point or boundary detection, filtering and windowing. Unvoiced signals/Silence parts can be removed manually through some software such as “PRAAT", "COOL EDIT PRO", etc. or by programmatically. The voice activity detection process used to detect the unvoiced, voiced or silent voice from a speech signal is known as ZCR. The ZCR may perform a significant role for end-point detection in the speech signal. To remove unwanted noise from the recorded signals, different types of filters are used. Some of the mostly used speech signals filters are FIR, IIR, Wiener filtering etc. To reduce the noise from the recorded speech signal, some of the past research works [12], [13], [14], [15], and [16] were applied spectral subtraction, Kalman filtering, and Wiener filtering. A filtering technique VTLN has been used to minimize the effect of variations in vocal-tract lengths among different speakers to improve the robustness of the system [16], [17], [18], [19], [20], [21], [22]. Noise adaptive training strategies and model adaptation techniques have also been investigated in [23], [24], [25], [26], [27]. At the end of this stage, the windowed version of each speech signal is found for extracting the required speech features.

C. Feature extraction

Features extraction is a method of extorting the features representing the acoustical information of the speech signal. Recognition performance of ASR system depends on speech features. Two types of features are widely addressed in literature namely temporal and spectral features which are computed in the time domain and frequency domain respectively. Different types of spectral analysis techniques are –

 Cepstral Analysis – It gives a very efficient technique to detach the excitation from the vocal tract shape. In ASRS, Cepstral analysis has been used to detect pitch value and formant frequency from the speech signal.

 Mel-frequency Cepstral Coefficients - MFCC has been proposed as a speech feature by Davis and Mermelstein in the year of 1980. It has been widely used in ASRS [1] ,[28], [29]. The number, location, and shape of the filters and the way of warping the power spectrum may affect the effectiveness of MFCC [30]. The computational steps of MFCC include - discrete Fourier transformation of the windowed speech segment to obtain the short-term power spectrum. This short-term power spectrum is warped into Mel-frequency such that it can percept the human ear functions, employing Mel filter bank algorithm and the inverse of DFT.

 LPC – LPC is a popular practice for investigating speech features. Comparatively slow and unreliable linear filters outputs are considered as more appropriate input samples for LPC, particularly if the filter is excited by irregular and concise pulses. The linear combinations of input and previous output samples can be used to get a better evaluation of the output sequence. The shape of the vocal-tract and speech formant can be estimated to analyse speech signal through LPC and the estimated vocal-tract parameter can be used in synthesizing of a speech signal.

 Linear Discriminant analysis - LDA and principal component analysis are applied [31] for optimizing the transformations and decrease the number of feature dimensions. The linear discriminant analysis technique can be used in a better way when the internal class frequencies are unequal and the performance of the inner class frequency depends on some randomly generated data.

D. Modeling

The primary element of an ASRS is the Acoustic model. The efficiency of the system solely depends on this model. Some of the modelling approaches used in ASR systems are –

i. The Acoustic Phonetic Approach – In speech recognition, the acoustic-phonetic approaches are used by the researchers for more than 40 years. In 1967, Hemdal and Hughes proposed the first acoustic-phonetic approach which addresses the presence of phonemes in natural languages. Acoustic properties of phonetic units are highly variable.

ii. Pattern Recognition approach – The two main parts of this approaches are - training with the pre-processed speech samples and matching the pattern of the testing dataset with the training data. Extracted features from each of the speech signal can be represented in terms of some pattern or templates. Each individual pattern represents a voice i.e. may be a sound or a word. During the recognition of speech signals, the pattern of the testing dataset is compared with the pattern of the training dataset.

iii. Knowledge Base Approaches – This type of approaches are used expertise knowledge. The variation observed in different utterances or speech signals are hand coded and feed into the knowledge base of an expert system.

iv. Statistical Based Approaches – Here, on the basis of speech features, speech signals are modeled using some standard statistical model like HMM and GMM.

In ASR system, different modeling techniques are presented in literature as classifier such as – HMM [7], [10], [16], ANN [32], [33], SVM [38], DNN [35] and some other Hybrid models such as ANN-HMM [36], [37], [38], DNN-HMM [1], [11], [39], [40], [41],[42], [69], [70], GMM-ANN [43], CD- KNN-RNN [44], GMM-HMM [45], [46], [47] etc.

❖ HMM – It is a statistical based approach. The HMM model can be characterized by five parameters - N, M, A, B and µ, where N represents the total number of states, M denotes the number of individual observations for each state, A denotes the state transition, B denotes the probability distribution for the observation variable and µ represents the initial state of the distribution. The performance of the HMM model for a noisy environment is well enough to some extent because the HMM model treats each of the speech signals individually. Based on the probability of sound transition from one state to another, the HMM model may guess the sound entity, even if it lost due to noise. The main disadvantage of this model is that the assumptions for the loss entity are considered on the basis of priori modeling and which may lead the model to be inaccurate and handicap the system performance. A sequence of time series data can be statistically modeled using a popular statistical tool HMM. In [16], HMM has been used to recognize speaker independent isolated word with the accuracy rate 98.5%, [7] to recognize English alphabets and digits from zero to nine with accuracy rate 100%, [10] to recognize 129 airlines vocabularies, etc. It has been observed that the efficiency of the HMM model can be improved through the addition of energy [48].

❖ Support vector machine – SVM is a powerful classification tool. SVM is applied for both speeches as well as speaker recognition. SVM has been used to classify robust speech signals and to verify unknown speakers. It has been observed that the combination of SVM with the statistical model GMM can increase the accuracy rate of classification or verification. It can be used to classify unknown data with a better performance in comparison to the other models. So SVM has been widely used in recognition of anonymous speaker and robust speech.

❖ DNN-HMM hybrid model – In recent years, DNN-HMM, a hybrid classification model has been proposed for phone recognition [39] [40] [49]. In [11], a context-dependent model has been proposed to recognize speech from a large vocabulary speech dataset using pre-trained DNN-HMM hybrid architecture with an accuracy rate of 68.2%. In the research articles [1] [41] [42] has been proposed a DNN-HMM based hybrid architecture. Here the analyzed speech features of unknown utterances has been used as the input for the DNN architecture and the output of this unit used as the input for the HMM architecture.

❖ ANN-HMM hybrid model – In this hybrid architecture, the output unit of the ANN is considered as the input for the HMM model [11]. From 1980 to the mid of 1990, the hybrid model ANN-HMM technique has been used as a classifier to increase the recognition rate in a large vocabulary speech recognition. It has been observed that the research work [36], [37], [38], [50] used the hybrid structure as a classifier to recognize speech. It has been observed that in [51], [52], [53], [54], [55], [56], [57] used ANN-HMM architecture to recognize speech signal.

❖ GMM-HMM hybrid model – GMM-HMM architecture has been used to recognize speech signals [45] [46] [47]. The output of a neural network is used as the input of the hybrid architecture to improve the recognition rate. A neural network has been trained on large vocabulary speech corpora consists of approximately more than 1000 hours of data.

❖ Dynamic Time Warping – It is a technique to measure similarities between two time-varying sequences.

To recognize isolated letters uttered by 100 speakers, in 1988, a time delay neural network has been proposed as a classifier and reported with 7.8% character error rate [58]. KNN-RNN based classifier has been used in [44] to compare the detection efficiency of vowels in Assamese language using the acoustic-phonetic feature. Recognition rate has been reported as 84.3% and 87% for RNN and KNN based algorithms respectively. To recognize vowel phonemes of Assamese language, ANN based algorithm [32] and RNN based algorithm [59] has been used to carry out the experiments. In [60], the recognition rate for fricative sound in Assamese speech has been reported 96% using RNN. An ANN based classifier was used in [33] to recognize the digits of Assamese language extracted through an Adaptive pre-emphasis filter.

III. COMPARIOSN OF THE PERFORMANCE BETWEEN ASR SYSTEM AND HUMAN

In recent years, the recognition accuracy of ASR has been increased significantly. However, the question arises here is that what will be the ultimate goal in case of the accuracy of ASR systems with efficient time complexity. Therefore, several studies have been done to compare the machine recognition performance with human recognition performance to derive the value of the performance difference. The performance of an ASR system is usually measured in terms of only recognition of utterances of different types of words, whereas speech uttered by a human can be measured by another human in terms of various factors such as facial expressions, body language, mode etc. The accuracy rate of an ASRS in terms of human listener and machine has been compared at several levels such as - words, phoneme, articulatory features, noisy speech, degraded speech etc. [61-68].

IV. CONCLUSION AND FUTURE SCOPE

Speech recognition is a very challenging research domain in the field of machine learning, due to the presence of different natural languages with different regional accent among people across the globe. If we consider only the North-eastern region of India then it has been observed that this part has 39 local languages among different tribes. Moreover, the same community people have the different accent depending on the region within the state. The accuracy of the ASR system is the most challenging issue for researchers. In this paper, we have presented the basic overview of the ASR system along with the speech corpus, speech features and classification techniques behind the systems addressed in various languages and the performance evaluation of the system. In the future, we will try to carry out research work on the Assamese language, an Indo-Aryan language which is spoken especially in Assam and Arunachal Pradesh of North-east India.

REFERENCES

[1] Mohamed, Abdel-Rahman, George Dahl, and Geoffrey Hinton. "Deep belief networks for phone recognition." Nips workshop on deep learning for speech recognition and related applications. Vol. 1. No. 9. 2009.

[2] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, pp. 6645-6649. IEEE, 2013.

[3] Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal processing magazine 29, no. 6 (2012): 82-97.

[4] Cooke, Martin, et al. "An audio-visual corpus for speech perception and automatic speech recognition." The Journal of the Acoustical Society of America 120.5 (2006): 2421-2424.

[5] Yu, Dong, Li Deng, and George Dahl. "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition." In Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2010.

[6] Mari, J-F., J-P. Haton, and Abdelaziz Kriouile. "Automatic word recognition based on second-order hidden Markov models." IEEE Transactions on speech and Audio Processing 5, no. 1 (1997): 22-25.

[7] Rabiner, L. R., and J. G. Wilpon. "Some performance benchmarks for isolated word speech recognition systems." Computer Speech & Language 2.3-4 (1987): 343-357.

[8] Utpal Bhattacharjee, "A comparative study of LPCC and MFCC features for the recognition of Assamese phonemes." International Journal of Engineering Research and Technology2,

[9] Wilpon and Jay G., "Automatic recognition of keywords in unconstrained speech using hidden Markov models." In IEEE Transactions on Acoustics, Speech, and Signal Processing38.11 (1990): 1870-1878.

[10] Rabiner, L. R., S. E. Levinson, and M. M. Sondhi. "On the use of hidden Markov models for speaker-independent recognition of isolated words from a medium-size vocabulary." AT&T Bell Laboratories Technical Journal 63.4 (1984): 627-642.

[11] Dahl and George E., "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." IEEE Transactions on audio, speech, and language processing 20.1 (2012): 30-42.

[12] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[13] M. Karam, H. F. Khazaal, H. Aglan and C. Cole, "Noise removal in speech processing using spectral subtraction", Journal of Signal and Information Processing, 2014.

[14] A. Agarwal and Y. M. Cheng, "Two-stage Mel-warped Wiener filter for robust speech recognition", In Proc. ASRU, vol. 99, pp. 67-70, 1999.

[15] G. R. Babu and R. Rao, "Modified Kalman Filter-based Approach in Comparison with Traditional Speech Enhancement Algorithms from Adverse Noisy Environments", International Journal on Computer Science and Engineering, vol. 3, no. 2, pp. 744-759, 2011.

[16] S.Gogoi and U. Bhattacharjee, "Vocal tract length normalization and sub-band spectral subtraction based robust Assamese vowel recognition system," In IEEE International Conference on Computing Methodologies and Communication (ICCMC), IEEE, pp. 32-35, 2017.

[17] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," Computer Speech & Language, vol. 20, no. 1, pp. 107–123, 2006.

[18] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, vol. 1, 1996.

[19] J. Lung et al., "Implementation of Vocal Tract Length Normalization for Phoneme Recognition on TIMIT Speech Corpus," in International Conference on Information Communication and Management, Singapore: IPCSIT, pp. 136–140, 2011.

[20] B . Widmer, "Implementation of Vocal Tract Length Normalization A Study of Methods". [Online]. Available: http : // ssli.ee.washington.edu / people / bwidmer /VTL_Talk/VTL_Talk.PDF. Accessed: 2013.

[21] S. Gogoi and U. Bhattacharjee, "Impact of Vocal Tract Length Normalization on the Speech Recognition Performance of an English Vowel Phoneme Recognizer for the Recognition of Children Voices," International Journal of Computer Trends and Technology (IJCTT), vol. 39, no. 2, pp. 105–109, 2016. [Online]. Available: http://www.ijcttjournal.org/2016/Volume39/number-2/IJCTT-V39P118.pdf. Accessed: Oct. 1, 2016.

[22] G. Garau, S. Renals, and T. Hain, "Applying Vocal Tract Length Normalization to Meeting Recordings," 2005. [Online]. Available: http://www.cstr.ed.ac.uk/downloads/ publications/2005/giuliagarau_eurospeech05.pdf.

[23] H. Jiang, K. Hirose and Q. Hue, "A minimax search algorithm for robust continuous speech recognition," IEEE Transactions on Speech and Audio Processing, 8(6): 688–694, 2000.

[24] B. Nasersharif and A. Akbari, "Improved HMM entropy for robust sub-band speech recognition," in13th European In Signal Processing Conference, IEEE, pp. 1–4, 2005.

[25] H. Xu, et al. , "Noise Condition-Dependent Training Based on Noise Classification and SNR Estimation," IEEE Transactions on Audio, Speech, and Language Processing, 15(8): 2431–2443, 2007.

[26] O. Kalinli, et al., "Noise adaptive training for robust automatic speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, 18(8): 1889–1901, 2010.

[27] J. Ganitkevitch, "Speaker adaptation using maximum likelihood linear regression,": Rheinish-Westflesche Technische Hochschule Aachen, the course of Automatic Speech Recognition, www-i6. informatik. rwthaachen. : [Online]. Available: http://www.cs.jhu.edu/~juri/pdf/mllr-rwth-2005.pdf. , 2015.

[28] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).

[29] Ghai, Wiqas, and Navdeep Singh. "Analysis of automatic speech recognition systems for indo-aryan languages: Punjabi a case study." Int J Soft Comput Eng 2.1 (2012): 379-385.

[30] Mousmita Sarma, and Kandarpa Kumar Sarma. "Segmentation and classification of vowel phonemes of assamese speech using a hybrid neural framework." Applied Computational Intelligence and Soft Computing 2012, 2012.sd

[31] Kwon, Oh-Wook, Kwokleung Chan, and Te-Won Lee. Speech feature analysis using variational Bayesian PCA. IEEE Signal Processing Letters 10, no. 5. 2003, pp.137-140.

[32] Mousmita Sarma, Krishna Dutta, and Kandarpa Kumar Sarma. "Speech corpus of assamese numerals extracted using an adaptive pre-emphasis filter for speech recognition." In Computer and Communication Technology (ICCCT), 2010 International Conference on, IEEE, 2010, pp. 461-466.

[33] Mousmita Sarma, Krishna Dutta, and Kandarpa Kumar Sarma. "Speech corpus of assamese numerals extracted using an adaptive pre-emphasis filter for speech recognition." In Computer and Communication Technology (ICCCT), 2010 International Conference on, IEEE, 2010, pp. 461-466.

[34] Mohammadi, Seyed Hamidreza, and Alexander Kain. "Voice conversion using deep neural networks with speaker-independent pre-training." In Spoken Language Technology Workshop (SLT),IEEE, 2014, pp.19-23.

[35] Hasegawa-Johnson, Mark, Jon Gunderson, Adrienne Perlman, and Thomas Huang. "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria." In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. IEEE International Conference on, vol. 3, pp. III-III. IEEE, 2006.

[36] Morgan, Nelson, and Herve Bourlard. "Continuous speech recognition using multilayer perceptrons with hidden Markov models." Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on. IEEE, 1990.

[37] Renals and Steve, "Connectionist probability estimators in HMM speech recognition." IEEE Transactions on Speech and Audio Processing 2.1 (1994): 161-174.

[38] Robinson, A. J., Cook, G. D., Ellis, D. P., Fosler-Lussier, E., Renals, S. J., & Williams, D. A. G. Connectionist speech recognition of broadcast news. Speech Communication, 2002. Pp. 37(1-2), 27-45.

[39] Mohamed, Abdel-Rahman, George Dahl, and Geoffrey Hinton. "Deep belief networks for phone recognition." Nips workshop on deep learning for speech recognition and related applications. Vol. 1. No. 9. 2009.

[40] Mohamed, Abdel-Rahman, Dong Yu, and Li Deng. "Investigation of full-sequence training of deep belief networks for speech recognition." Eleventh Annual Conference of the International Speech Communication Association. 2010.

[41] Dahl, George, Abdel-Rahman Mohamed, and Geoffrey E. Hinton. "Phone recognition with the mean-covariance restricted Boltzmann machine." Advances in neural information processing systems. 2010.

[42] Mohamed, Abdel-Rahman, George E. Dahl, and Geoffrey Hinton. "Acoustic modeling using deep belief networks." IEEE Transactions on Audio, Speech, and Language Processing20.1 (2012): 14-22.

[43] Shaoqin, Yao, and Zhang Linghua. "Voice Conversion Based on Mixed GMM-ANN Model." Journal of Data Acquisition and Processing 2, 2014.

[44] Mridusmita Sharma, and Kandarpa Kumar Sarma. "Dialectal Assamese vowel speech detection using acoustic phonetic features, KNN and RNN." In Signal Processing and Integrated Networks (SPIN), 2015 2nd International Conference on, pp. 674-678. IEEE, 2015.

[45] Acero and Alex, "Live search for mobile: Web services by voice on the cell phone." Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, 2008.

[46] Morgan and Nelson, "Pushing the envelope-aside [speech recognition]." IEEE Signal Processing Magazine 22.5 (2005): 81-88.

[47] Hwang, Mei-Yuh, and Xuedong Huang. "Shared-distribution hidden Markov models for speech recognition." IEEE Transactions on Speech and Audio Processing 1.4 (1993): 414-420.

[48] Rabiner and L. R., "Recognition of isolated digits using hidden Markov models with continuous mixture densities." Bell Labs Technical Journal 64.6 (1985): 1211-1234.

[49] Dahl, George, Abdel-Rahman Mohamed, and Geoffrey E. Hinton. "Phone recognition with the mean-covariance restricted Boltzmann machine." Advances in neural information processing systems. 2010.

[50] Hennebert and Jean, "Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems." (1997).

[51] Franzini M, Lee KF, Waibel A. Connectionist Viterbi training: a new hybrid method for continuous speech recognition. Acoustics, Speech, and Signal Processing, ICASSP-90., International Conference, IEEE, 1990, pp. 425-428.sd

[52] Levin E. Word recognition using hidden control neural architecture. InAcoustics, Speech, and Signal Processing, ICASSP-90., International Conference, IEEE, 1990, pp.433-436.

[53] Morgan, Nelson, and Herve Bourlard. Continuous speech recognition using multilayer perceptrons with hidden Markov models. In Acoustics, Speech, and Signal Processing, ICASSP-90.,International Conference, IEEE, 1990, pp.413-416.

[54] Niles, Les T., and Harvey F. Silverman. Combining hidden Markov model and neural network classifiers. In Acoustics, Speech, and Signal Processing, ICASSP-90., International Conference, IEEE, pp. 417-420.s

[55] Trentin, Edmondo, and Marco Gori. A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing 37, no. 1-4, 2001, pp. 91-126.

[56] Haffner P, Franzini M, Waibel A. Integrating time alignment and neural networks for high performance continuous speech recognition. InAcoustics, Speech, and Signal Processing, ICASSP-91., International Conference, IEEE, 1991, pp.105-108.

[57] Bengio, Yoshua, Renato De Mori, Giovanni Flammia, and Ralf Kompe. Global optimization of a neural network-hidden Markov model hybrid. IEEE transactions on Neural Networks3, no. 2, 1992, pp.252-259.

[58] Lang, K. J. The development of the time-delay neural network architecture for speech recognition. Technical Report CMU-CS, 1988, pp. 88-152.

[59] Mridusmita Sharma, Mousmita Sarma, and Kandarpa Kumar Sarma. "Recurrent Neural Network based approach to recognize assamese vowels using experimentally derived acoustic-phonetic features." In Emerging Trends and Applications in Computer Science (ICETACS), 2013 1st International Conference on,. IEEE, 2013, pp. 140-143.

[60] Patgiri, Chayashree, Mousmita Sarma, and Kandarpa Kumar Sarma. "Recurrent neural network based approach to recognize assamese fricatives using experimentally derived acoustic-phonetic features." In Emerging Trends and Applications in Computer Science (ICETACS), IEEE, 2013, pp. 33-37.

[61] Lippmann, R., Speech recognition by machines and humans. Speech Communication 22 (1), 1997. PP. 1–15.

[62] Van Leeuwen, D.A., van den Berg, L.G., Steeneken, H.J.M. Human benchmarks for speaker independent large vocabulary recognition

performance. In: Proceedings of Euro speech, Madrid, Spain, 1995 pp. 1461– 1464

[63] Meyer, B., Wesker, T., Brand, T., Mertins, A., Kollmeier, B. A human–machine comparison in speech recognition based on a logatome corpus. In: Proceedings of the Workshop on Speech Recognition and Intrinsic Variation, Toulouse, France, 2006.

[64] Sroka, J.J., Braida, L.D. Human and machine consonant recognition. Speech Communication 45, 2005. pp. 401–423.

[65] Moore, R.K. There's no data like more data: but when will enough be enough? In: Proceedings of the Acoustics Workshop on Innovations in Speech Processing, vol. 23 (3), Stratford-upon-Avon, UK, 2001. pp. 19– 26.

[66] Moore, R.K. A comparison of the data requirements of automatic speech recognition systems and human listeners. In: Proceedings of Euro speech, Geneva, Switzerland, 2003, pp. 2581–2584.

[67] Cooke, M. A glimpsing model of speech recognition in noise. Journal of the Acoustical Society of America 119 (3), 2006. pp. 1562–1573.

[68] Cutler, A., Robinson, T. Response time as a metric for comparison of speech recognition by humans and machines. In J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, Conference of the International Speech Communication Association. 2010, pp.68–73.

[69] Sujatha N. and Prakash K. An Efficient and Scalable Auto Recommender System Based on Users Behavior in Isroset-Journal (IJSRCSE) Vol.6 , Issue.6 , pp.35-40, Dec-2018.

[70] Mutkule Prasad R, Interactive Clothing based on IoT using QR code and Mobile Application Journal (IJSRNSC) , Vol.6 , Issue.6 , pp.1-4, Dec-2018.

Authors Profile

Dipankar Dutta pursed Master of Computer Application (MCA) from Dibrugarh University, in the year of 2007, Assam, India. He is currently pursuing Ph.D. and working as Assistant Professor in Department of Computer Science, North Eastern Regional Institute of Management (NERIM) since 2017. His main research work focuses on Speech Processing and AI. He has 10 years of teaching experience and 2 years of Research Experience.

Ridip Dev Choudhury received his Master of Science (M.Sc.) in Computer Science and Ph.D. in Computer Science from Gauhati University, Assam, India in the year of 2004 and 2014 respectively. Presently he is working as an Assistant Professor in Gauhati University Institute of Distance and Open Learning, Assam, India. His research interest is in the field of Speech Processing, Digital Image Processing and Expert System.

Swapnanil Gogoi received his Master of Science (M.Sc.) in Computer Science from Gauhati University, Assam, India and Ph.D. in Computer Science and Engineering from Rajiv Gandhi University, Arunachal Pradesh, India in the year of 2004 and 2017 respectively. Presently he is working as an Assistant Professor in Gauhati University Institute of Distance and Open Learning, Assam, India. His research interest is in the field of Speech Processing and Robust Speaker Recognition.