Smart Innovation, Systems and Technologies 435

Ramesh Chandra Poonia
Sugam Sharma
Ibrahim A. Hameed
Kamal Upreti *Editors*

Smart Cyber Physical Systems

Proceedings of ICSCPS 2024

KES International

Springer

Smart Innovation, Systems and Technologies

Volume 435

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago, DBLP.

All books published in the series are submitted for consideration in Web of Science.

Chapter 51
Study and Analysis of Automatic Speech Recognition Systems for Indian and Foreign Languages

Sagar P. Chitte, Madhukar N. Shelar, Sunil S. Nimbhore, and Aashish R. Lahase

Abstract Human–Machine Interface (HMI) through Automatic Speech Recognition (ASR) is a system that enables the interaction between humans and machines through software applications. Speech Recognition (SR) is an interdisciplinary branch of computer science and computational linguistics that creates methods and tools to allow computers to recognize and translate spoken language into text with the primary advantage of being able to search. Most of today's market is captured by human–machine interface devices, which use the concept of automatic speech recognition. Several researchers have developed ASR systems for different Indian regional and foreign languages, as well as for various dialects, using different techniques and tools. This study presents the different methods used by different researchers in the field of automatic speech recognition. It gives comparisons between different feature extraction methods and classification methods used in human–machine interfaces. It also explores the different challenges in this field. This study gives knowledge of speech recognition for isolated and connected words and continuous and spontaneous speech. This study will help other researchers develop technology-oriented solutions for human–machine interfaces. This study gives detailed information on how different researchers used different feature extraction methods and classification methods for different languages and it is observed that Mel Frequency Cepstrum Coefficients (MFCC) of feature extraction and Convolutional Neural Network (CNN) of classification methods give higher accuracy. The improvements and accuracy rates in ASR for both isolated and continuous speech in a variety of languages are highlighted in this paper.

Keywords Human–machine interface · Automatic speech recognition · Feature extraction · Classification

S. P. Chitte (✉) · M. N. Shelar
Commerce, Management and Computer Science (CMCS) College, Nashik, Maharashtra, India
e-mail: cmcsspc@gmail.com

S. S. Nimbhore · A. R. Lahase
UDCS&IT, Dr. BAM University, Chh. Sambhajinagar, Maharashtra, India

R. C. Poonia et al. (eds.), *Smart Cyber Physical Systems*, Smart Innovation, Systems and Technologies 435, https://doi.org/10.1007/978-981-96-2182-8_51

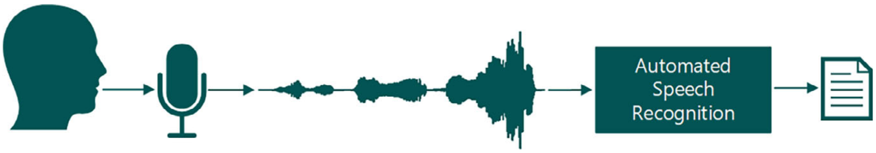

Fig. 51.1 Process of automatic speech recognition [1]

51.1 Introduction

ASR recognizes and processes human speech using computer hardware and software. It's employed to identify the words that someone has spoken. It needs the primary user(s)' predetermined or saved voices. By keeping voice samples and lexicons in the system, it must train the ASR system. The speech signal is converted into words by it. The identified terms may be used as Natural Language Processing's input or as output, as described in Fig. 51.1.

Some popular ASR systems available today are Google Assistant, Apple Siri, and Amazon Alexa. ASR is also used in many sectors such as automotive, healthcare, customer support, call center, security, media, etc. to control various functions using voice command thereby improving convenience and safety. This study reviews the many approaches that are currently available in ASR in Indian as well as foreign languages, containing different sections. Section 51.2 gives information on the phases of ASR, and information about work reviewed in the past on ASR is presented in Sect. 51.3. Section 51.4 gives insight into the work published in ASR until today based on the types of utterances and also summarizes different feature extraction and classification methods. Section 51.5 concludes with recommendations for further research.

51.2 Automatic Speech Recognition

ASR has two main phases as training and testing [2] as presented in Fig. 51.2.

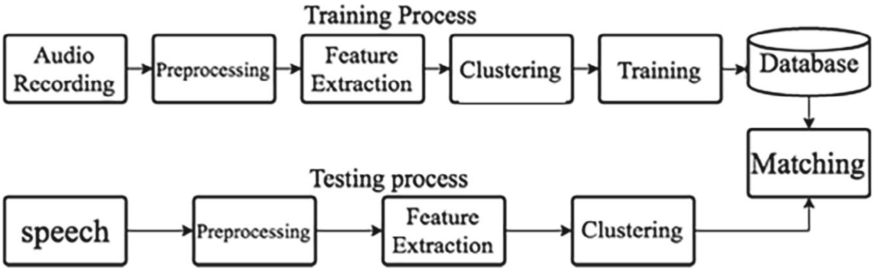

Fig. 51.2 Process of automatic speech recognition [2]

51.2.1 Training Phase

The training phase consists of the following steps:

- Audio Recording: Collection of speech samples from different people of different age groups.
- Preprocessing: Removing unwanted and background noise from the audio stream and identifying speech activity.
- Feature Extraction: Extracting relevant information and reducing the data size of the speech signal.
- Clustering: Arranging a collection of items so that those in the same group referred to as a cluster, resemble one another more than they do the objects in other groups (also known as clusters).
- Training: Creation of model or algorithm.
- Database: Speech Corpus.

51.2.2 Testing Phase

The testing phase consists of following steps:

- Speech: Human vocal communication.
- Preprocessing: Removing unwanted and background noise from the audio stream and identifying speech activity.
- Feature Extraction: Extracting relevant information and reducing the data size of the speech signal.
- Clustering: Arranging a collection of items so that those in the same group referred to as a cluster, resemble one another more than they do the objects in other groups (also known as clusters).
- Matching: Matching of spoken word or sentence with speech corpus and output.

Various techniques and tools are suitable for creating a speech recognition system, as given in Table 51.1. Several researchers have developed ASR systems for different Indian regional and foreign languages, as well as for various dialects, using these techniques and tools.

51.2.3 Feature Extraction Methods

Mel Frequency Cepstrum Coefficients:

The Mel frequency Cepstrum in sound curing is a depiction of a sound's short-term power band. It depends on the log power band's linear cos transform [3]. MFCCs are frequently employed as voice recognition characteristics [4] systems. The Mel frequency bands are uniformly distributed and closely resemble the human vocal

Table 51.1 Techniques and tools used in ASR

Feature extraction techniques	Classification techniques	Tools
Perceptual linear predictive (PLP)	Hidden Markov Model (HMM)	PRAAT
Linear predicted coding (LPC)	Dynamic Time Warping (DTW)	AUDACITY
Linear prediction cepstral coefficients (LPCC)	Artificial Neural Networks (ANN)	CSL
Mel frequency cepstrum coefficients (MFCC)	Time Delay Neural Network (TDNN)	SPHINX
Line spectral frequencies (LSF)	Deep Neural Network (DNN)	SCARF
Discrete wavelet transform (DWT)	Vector Quantization (VQ)	MICROPHONES
	K-Nearest Neighbor (K-NN)	
	Support Vector Machine (SVM)	
	Convolutional Neural Network(CNN)	
	Deep Convolutional Neural Network (DCNN)	
	Long Short-Term Memory (LSTM)	
	Feed-Forward Neural Network (FFNN)	

system. It can be used to recognize the sound [5]. MFCC coefficients are calculated using following formula [6, 7].

$$f_{mel} = 2595 \log_{10}\left[1 + \frac{f_{linear}}{700}\right] \tag{51.1}$$

One disadvantage of MFCC is that if there is background sound then it does not generate exact result [8].

Linear Predicted Coding:

With this feature extraction technique, substantial speech features are produced by modeling vocal tract of humans [9]. By estimating the formants, it assesses the audio clips. The speech is divided into two very independent parts using this analysis: the glottal excitation-LP residual and the vocal tract characteristics-LP coefficients. Any audio clip can be predicted using this analysis as a linear weighted collection of preceding samples. Following is the formula using LP analysis for audio processing [8]:

$$\hat{s} = \sum_{k=1}^{p} a_p s(n-k) \tag{51.2}$$

Here predicted sample is s and predictor coefficient is p.

It calculates exact estimates in faster manner of any audio clip. It is thought to be the most efficient method of a sound recognition system since it effectively chooses the vocal tract data from a given utterance. Conventional linear methods suffer from aliased autocorrelation parameters. LPC calculations are infamous for having vast sensitivity to quantization noise.

Linear Prediction Cepstral Coefficients:

LPCCs are Cepstral coefficients. These are calculated using the spectral envelope determined by LPC [9]. These coefficients are derived by applying fast Fourier transform [FTT]. It is then followed by logarithmic magnitude spectrum. The steps of computing LPCC coefficients are the same as LPC except the LPC parameter conversion to be done in the last step. When comparing these features to LPC features, the error rate is reduced. LPCC calculations are infamous for having vast sensitivity to quantization noise [8].

Perceptual Linear Predictive:

Loudness compression, equal loudness pre-emphasis, and the intensity of the sensitive band are combined in the PLP technique to extract relevant information for speech. It combines both spectral analysis and linear prediction analysis [9]. The bark scale frequency can be calculated using following formula:

$$bark(f) = \frac{26.81f}{1960 + f} - 0.53 \tag{51.3}$$

Here linear frequency is f. One of its advantages is that it can withstand noise, channel fluctuations, and microphones. Since speaker-dependent information is successfully suppressed, PLP offers more accurate recognition than LPC. Also it can also improve the functionality of a speaker-independent recognition system. The front end which is based on PLP is susceptible to variations of the formant frequency.

51.2.4 Classification Methods

Hidden Markov Model:

HMM makes the assumption that the system under study is a Markov process with hidden states. HMM requires an observable procedure whose results are “influenced” in a known manner by the results. A task-dependent lexicon word can be generated using HMM, where each hiding component state belongs to a sound [8].

Artificial Neural Networks:

ANN has several processing units known nodes or neurons, which are similar to nerve cells in the individual brain. It is able to both receive and send electrochemical signals.

Similar to the connections between neurons via synapses in the brain, every neuron is connected to numerous others by communication links of different strengths. ANNs are machine learning models that use an enormous amount of neurons to imitate how the human brain works. Three layers make up its organization: one for data, one (or more) hidden layer, and one for result [10]. The designs of neural networks have become increasingly popular because of their durability, flexibility, and nonlinearity [8].

Dynamic Time Warping:

Using DTW we can determine resemblance among sequences. With some limitations, the DTW approach provides the optimal match between the two sequences that are provided. Here the sequences are distorted non-linearly. Because of its adaptability to varying speech rates, it was determined to be the ideal method for recognition of speech.

Time Delay Neural Network:

It is another type of ANN, and it has a time lag in both its input and hidden neurons. TDNN has been proven a powerful network in handling the context information of speech signals [11–13].

Deep Neural Network:

It is a machine learning algorithm implemented by organizing a stack of layers of computing. It follows feed-forward neural network topology by including multiple hidden layers. The neurons are connected by communication links. Each neuron processes and then propagates the input signal it receives to the layer above it.

Convolutional Neural Network:

It does not use hidden layers. It introduces a unique structure that comprises alternating convolution and pooling layers [14, 15] (Table 51.2).

It is difficult to develop a Speech Corpus of the language if not available. Another challenge is to achieve good recognition accuracy using the ASR model. Research efforts to overcome these challenges carried out in the past may need further investigation for better performance of the ASR system.

51.3 Related Work

More Siddharth et al. [16] presented a review of Indian regional languages like Hindi, Sanskrit, Ahirani, Assamese, Punjabi, Tamil, Gujarati, Telugu, Manipuri, Kannada, etc. This paper gives insight into speech, speech recognition systems, speech text, types of speech, and speaker models. It explains the ASR framework with feature extraction methods. It gives a review of speech recognition based on Indian languages. This paper concludes that most of the researchers used the HTK toolkit and found it as a best method among all available. However, deep learning

Table 51.2 Work done on languages and numerals

Indian languages	Foreign languages
Marathi	Indonesian
Varhadi	Malay
Ahirani	Moroccan
Pali	Arabic
Hindi	
Punjabi	
Manipuri	
Kannada	
Assamese	
Sanskrit	
Bengali	
Gujarati	

instead of traditional techniques gives better results [17]. The paper concludes that the majority of researchers still use MFCC where they can try LPC. A combination of models is advised since studies have shown that utilizing a DNN model's HMM or GMM produces superior outcomes [18]. Research can be conducted using deep RNN models like LSTM as they are incredibly effective at recognizing sounds [19].

Deshmukh A. M. et al. [20] provides a literature survey on two key models, DNN-HMM and RNN-CTC. The study states that RNN-CTC is rapid in interpreting, simple to train and found best than DNN-HMM. The Maharashtra's native and non-native speakers' dialects were examined in [21]. This paper states that to measure the level of accuracy in ASR, a Word Error Rate (WER) metric can be used which is based the overall number of words, modifications, and removals. Olena Iosifovaa et al. [1] provides a comparison on two important approaches in ASR, namely, hybrid model and end-to-end model.

51.4 Literature Review

ASR is important nowadays in human–machine interaction as it reduces the effort of giving commands manually by typing. There are various methods of ASR, viz., classification and feature extraction methods. With ASR, voice commands may be seamlessly communicated with machines, greatly improving human–machine interfaces.

Several research papers and articles have been published on the aspects of ASR systems and maximizing speech recognition accuracy. These literatures are reviewed in detail and categorized based on the type of speech utterances used by researchers such as phonemes, numerals, words, and sentences.

51.4.1 Review Based on Phonemes

The tiniest distinguishable audio unit is called a phoneme. It distinguishes between words in a specific language. Every dialect has its distinctive phoneme system. You can think of all possible words as ordered phoneme sequences [22].

Rahul Laishram et al. [23] developed phoneme-based system by using MFCC, HMM for Manipuri language with overall system performance around 65.24% on 29 phonemes. Bansod Nagsen et al. [24] has also used MFCC as a feature extraction technique with Linear Perceptual Coding (LPC) to develop a speaker recognition system for the Varhadi language for a total of 1625 phonemes and found that MFCC is better feature extraction technique as compared to LPC. Abdel-Hamid et al. [14] used log energy calculated from MFSC characteristics, and came to know that limited weight sharing system is good as compared to MFCC. Further, he also found that an error rate reduction of 6–10% compared with DNNs can be achieved by using CNNs as a classification technique.

51.4.2 Review Based on Numerals

Saxena Babita et al. [25] developed Digit Speech Recognition (DSR) for the Hindi language by using the MFCC, HMM with an overall system performance of around 89.52%. Siddharth More et al. [26] has also used MFCC as a feature extraction method but instead of HTK, the author used K-NN classification on isolated Pali language digits (0–10) and found accuracy as 80.36% using 2200 speech samples by 12 male and 8 female (20*11*10). But [27] suggested MFCC, LSTM as good methods as compared to LPC and HMM. He developed ASR for decimal numbers in Indonesian language using LSTM, LPC, and MFCC. According to [28], using more MFCC coefficients or a bigger amount of training data could lead to more effective results. He proposed an Arabic-digit speech recognition model using RNN, LSTM, and MFCC. Sharmin Riffat et al. [29] developed ASR for the Bengali language spoken digits determining that CNN performs well than other comparable classification techniques and while collecting samples, it also came out that individual pronunciations vary depending on their gender, dialect, and generation. The same conclusion on CNN is given by [30] and discovered that the rate of accuracy for spoken digit recognition increases with dataset growth. Quality is also somewhat improved by the amount of epochs in CNN. A paper on a denoising method for speech synthesis of the Marathi numbers based on spectral subtraction was published in [31]. He discovered that the spectral subtraction approach has enhanced speech signal quality and decreased noise. This research focuses on using high-quality voice signals for additional processing. Because written languages do not always indicate how they should be pronounced, phonetic transcription is crucial, when [32] created the Speech Corpus for Assamese and transcription using MFCC and hidden Markov model toolkit and discovered some intriguing trends in the transcription of speech for Assamese language. Specifically, they noticed that while accuracy increases as trained data increases, for a few trained databases, accuracy eventually becomes static.

51.4.3 Review Based on Isolated Word

Word recognition is done by this system one at a time. There needs to be a pause in between each syllable. It is simple to compute end points in this system. So it is easy to implement [33].

Dua Mohit et al. [34] developed ASR for Punjabi language words, Patil [35] developed ASR for Ahirani language, [36] developed ASR for Malay language, [37] developed SR for Moroccan Dialect, and [38] developed ASR for Murmured by using HMM and Acoustic Analysis done by using Mel Frequency Cepstral Coefficient (MFCC). G. Hemakumar et al. [39] developed a Speaker Independent Isolated Kannada Word Recognizer by using Linear-Predictive coding (LPC) and found that MFCC gives better performance as compared to LPC. Pokhariya Jitendra et al. [40] used the LINUX platform to develop ASR for the Sanskrit language by using the MFCC, HMM. Siddarth More, Saksamudre Suman et al. [26, 41] used K-Nearest Neighbor (K-NN) as a classification technique but achieved less accuracy as compared to HMM and discovered that the accuracy of classification declines with vocabulary quantity. Dhanashri Dhavale et al. [33] employed a DBN pre-training approach to initialize DNNs and presented the isolated word ASR using DNN with HMM. Somnath Hase et al. [42] used a DCNN for isolated Marathi speech recognition system. Chakraborty Gautam et al. [8] used different classification techniques such as HMM, FFNN, RNN, TDNN, and CNN and found an increase in performance when trained and tested by the CNN algorithm.

51.4.4 Review Based on Continuous Speech

This system recognizes multiple words at a time. It does not require a pause between each word. In this system, it is difficult to calculate end points. So it is difficult to implement [16].

Supriya Handore et al. [43] developed ASR for the Marathi language by using MFCC and HMM with an overall system performance of around 80 to 90% on 910 Marathi sentences. Muhammad Hariz et al. [2] also used MFCC and HMM but used K-means clustering and developed ASR for the Indonesian language with overall system performance around 75.40% on uttering sentences with two or three words (Table 51.3).

Table 51.3 Comparison of various feature extraction and classification methods in ASR

References	Feature extraction technique/s	Classification technique/s	Language	Type of speech	Accuracy
[1]	–	HMM-DNN, CTC	–	–	HMM-DNN (Hybrid) better than CTC
[2]	MFCC	HMM	Indonesian	Sentence	75.40%
[3]	–	HMM, SVM	–	–	HMM better than SVM
[14]	MFSC	CNN, DNN		Phoneme	
[23]	MFCC	HMM	Manipuri	Phoneme	65.24%
[24]	MFCC, LPC	–	Varhadi	Phoneme	88%
[25]	MFCC	HMM	Hindi	Digits	89.52%
[26]	MFCC	K-Nearest Neighbor (K-NN)	Pali	Isolated Word, Digits	80.36%
[27]	MFCC, LPC	Long Short-Term Memory (LSTM)	Indonesian	Digits	96.58%
[28]	MFCC	RNN, LSTM	Arabic	Digits	69%
[29]	MFCC	CNN	Bengali	Digits	98.37%
[30]	MFCC	CNN	Gujarati	Digits	98.70%
[31]	VAD	–	Marathi	Digits	94.34
[32]	MFCC	HTK	Assamese	Phone	65.26%
[33]	MFCC	HMM, DNN, DBN	English	Isolated digits	86.06%
[34]	MFCC	HMM	Punjabi	Isolated word	94.08%
[35]	MFCC	HMM	Ahirani	Isolated word	94%
[36]	MFCC, LPC	HMM	Malay	Isolated words	94.86%
[37]	MFCC	HMM	Moroccan	Isolated words	90%
[38]	–	HMM	Murmur	Isolated words	81.20%
[39]	LPC	HMM	Kannada	Isolated words	91.66%
[40]	MFCC	HMM	Sanskrit	Isolated words	95.2% to 97.2%
[41]	MFCC	K-Nearest Neighbor (K-NN)	Hindi	Isolated word	89%

(continued)

Table 51.3 (continued)

References	Feature extraction technique/s	Classification technique/s	Language	Type of speech	Accuracy
[42]	MFCC	DCNN	Marathi	Isolated word	88%
[43]	MFCC	HMM	Marathi	Sentence	80% to 90%

51.5 Conclusion

Feature extraction methods and classification models in ASR are extensively researched by many researchers. RNN and CNNs are found to be the most excellent methods among all methods as per the literature review. There is no literature found that has implemented the ASR model for the Dangi language. Dangi is low-resource language. So there is a need for research in ASR for the Dangi language that will achieve better speech recognition accuracy to improve teaching learning process of tribal peoples.

This study illustrates, it is big challenge to develop ASR for low-resource language and noisy environment. Also it illustrates that the HMM is the most widely used classification technique till now. Many researchers have proposed HMM [16] as a classification technique. Furthermore, according to this analysis, MFCC is the most used feature extraction method. But this study also shows that by using other techniques than HMM and MFCC better speech recognition accuracy can be achieved.

References

1. Olena Iosifovaa,b, Ievgen Iosifova, Volodymyr Sokolovb, Oleh Romanovskyia, Igor Sukaylob.: (2021)
2. Muhammad, Hariz Zakka, Muhammad Nasrun, Casi Setianingsih, Muhammad Ary Murti.: Speech recognition for English to Indonesian translator using hidden Markov model. In: 2018 International Conference on Signals and Systems (ICSigSys), pp. 255–260. IEEE, (2018)
3. Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, Qi Tian.: HMM-based audio keyword generation (2004)
4. Ganchev, T., Fakotakis, N., Kokkinakis, G.: Comparative evaluation of various MFCC implementations on the speaker verification task Archived 2011–07–17 at the Wayback Machine. In: 10th International Conference on Speech and Computer (SPECOM 2005), **1**, pp. 191–194 (2005)
5. Wikipedia-https://en.wikipedia.org/wiki/Mel-frequency_cepstrum
6. Zhang Y.: Exploring neural network architectures for acoustic modeling, A Doctotal Thesis, MIT, USA, (2017). http://hdl.handle.net/1721.1/113981
7. Nacereddine, H.: Contribution to the automatic speech recognition of arabic language and its applications. A Doctotal Thesis, Badji Mokhtar-Annaba University, Algeria, 2013/2014
8. http://hdl.handle.net/10603/326001
9. Alim, S.A., N.K.A.: Rashid Some Commonly used Speech Feature Extraction Algorithms, A book chapter published on Dec., (2018). https://doi.org/10.5772/intechopen.80419

10. Rabiner, L.R.: A tutorial on hidden markov model and selected applications in speech recognition, Proc. IEEE, **77**, (1989)
11. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J.: Phoneme recognition using time delay neural network. IEEE Trans. Acoust. Speech Signal Process.Acoust. Speech Signal Process. **37**(3), 328–339 (1989)
12. Peddinti, V., Povey, D., Khudanpur, S.: A Time Delay Neural Network architecture for efficient modeling of Long Temporal Contexts, Interspeech, pp. 3214–3218 (2015)
13. Shaharin, R., Prodhan, U.K., Dr. Md. Rahman, M.: Time delay neural network for bengali speech recognition. Int J Adv Res Comput Sci Technol, **3**(4), (2015)
14. Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, Dong Yu.: Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**(10) (2014): 1533–1545
15. Martina Slıvova, Pavol Partila, Jaromır Tovarek, Miroslav Voznak.: Isol. Word Autom. Speech Recognit. Syst. (2020)
16. More, Siddharth, S., Prashantkumar, L., Borde, Sunil S., Nimbhore.: A review on automatic speech recognition system in Indian Regional languages. Int. J. Comput. Appl. 975: 8887
17. Nassif, Ali Bou, Ismail Shahin, ImtinanAttili, Mohammad Azzeh, Khaled Shaalan.: Speech recognition using deep neural networks: A systematic review. IEEE Access 7 (2019): 19143–19165
18. Shahin I, Nassif A.B, Hamsa S.: Novel cascaded Gaussian mixture model-deep neural network classier for speaker identication in emotional talking environments. Neural Comput. Appl., to be published
19. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
20. Deshmukh, A.M.: Comparison of hidden markov model and recurrent neural network in automatic speech recognition. Eur. J. Eng. Technol. Res. **5**(8), 958–965 (2020)
21. Pooja B. Abhang, Somanath Hase, Dr. Sunil S. Nimbhore.: Dialect analysis of native and non-native marathi speakers for continuous speech using ASR: A Review. High Technol. Lett. **27**(12), (2021) ISSN NO: 1006–674
22. Mousumi Malakar, Ravindra B. Keskar.: Progress of machine learning based automatic phoneme recognition and its prospect, Speech Communication, **135**, Pages 37–53, (2021) ISSN 0167–6393, https://doi.org/10.1016/j.specom.2021.09.006
23. Rahul, Laishram, Salam Nandakishor, L. Joyprakash Singh, Dutta, S.K.: Design of manipuri keywords spotting system using HMM. In: 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), pp. 1–3. IEEE, (2013)
24. Bansod, Nagsen, S., Siddharth, B., Dadhade, Seema, S., Kawathekar, Kale, K.V.: Speaker recognition using marathi (Varhadi) language. In: 2014 International Conference on Intelligent Computing Applications, pp. 421–425. IEEE, (2014)
25. Saxena, Babita, and Charu Wahi.: Hindi digits recognition system on speech data collected in different natural noise environments. In: International Conference on Computer Science, Engineering and Information Technology (CSITY 2015) February, pp. 14–15. (2015)
26. Siddharth, S., More, Prashant kumar, L., Borde, Sunil, S. Nimbhore.: Isolated Pali Word (IPW) feature extraction using MFCC & KNN Based on ASR. IOSR J. Comput. Eng. (IOSR-JCE), p-ISSN: 2278–8727, **20**(6), Ver. II, pp. 69–74 (2018)
27. Swedia, Ericks Rachmat, Achmad Benny Mutiara, Muhammad Subali.: Deep learning long-short term memory (LSTM) for Indonesian speech digit recognition using LPC and MFCC Feature. In: 2018 Third International Conference on Informatics and Computing (ICIC), pp. 1–5. IEEE, (2018)
28. Wazir, Abdulaziz Saleh Mahfoudh Ba, Joon Huang Chuah.: Spoken Arabic digits recognition using deep learning. In: 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), pp. 339–344. IEEE, (2019)
29. Sharmin, Riffat, Shantanu Kumar Rahut, Mohammad Rezwanul Huq.: Bengali spoken digit classification: A deep learning approach using convolutional neural network. Procedia Comput. Sci. **171**, 1381–1388 (2020)

30. Tailor, Jinal, H., Rajnish Rakholia, Jatinderkumar R. Saini, Ketan Kotecha.: Deep learning approach for spoken digit recognition in Gujarati language. Int. J. Adv. Comput. Sci. Appl. **13**(4) (2022)
31. Ramteke, G.D., Nimbhore, S.S., Ramteke, R.J.: Denoising speech of marathi numerals using spectral subtraction. (2012)
32. Sarma, Himangshu, Navanath Saharia, Utpal Sharma.: Development of Assamese speech corpus and automatic transcription using HTK. In: Advances in Signal Processing and Intelligent Recognition Systems, pp. 119–132. Springer International Publishing, (2014)
33. Dhanashri, Dhavale, Dhonde, S.B.: Isolated word speech recognition system using deep neural networks. In: Proceedings of the International Conference on Data Engineering and Communication Technology: ICDECT 2016, **1**, pp. 9–17. Springer Singapore, (2017)
34. Dua, Mohit, Aggarwal, R.K.: Virender Kadyan, and Shelza Dua. Punjabi automatic speech recognition using HTK. Int. J. Comput. Sci. Issues (IJCSI) **9**(4): 359 (2012)
35. Patil, Ajay, S.: Automatic speech recognition for ahirani language using hidden markov model toolkit (HTK). Int. J. Comput. Sci. Trends Technol. (IJCST) **2**
36. Maseri, Marlyn, Mazlina Mamat.: Malay language speech recognition for preschool children using Hidden Markov Model (HMM) system training. In: Computational Science and Technology: 5th ICCST 2018, Kota Kinabalu, Malaysia, 29–30 August 2018, pp. 205–214. Springer Singapore, (2019)
37. Mouaz, Bezoui, Beni Hssane Abderrahim, Elmoutaouakkil Abdelmajid.: Speech recognition of moroccan dialect using hidden Markov models. Procedia Comput. Sci. **151**: 985 991 (2019)
38. Kumar, Rajesh, Lakshmi Sarvani Videla, Soubraylu Siva Kumar, Asalg Gopala Gupta, Haritha, D.: Murmured speech recognition using hidden markov model. In: 2020 7th International Conference on Smart Structures and Systems (ICSSS), pp. 1–5. IEEE, (2020)
39. Hemakumar, G., Punitha, P.: Speaker Independent Isolated Kannada Word Recognizer. In: Multimedia Processing, Communication and Computing Applications: Proceedings of the First International Conference, ICMCCA, 13–15 December 2012, pp. 333–345. Springer India, (2013)
40. Pokhariya, Jitendra Singh, Sanjay Mathur.: Sanskrit speech recognition using hidden markov model toolkit. Int. J. Eng. Res. & Technol. (IJERT) **3**(10), 93–98, (2014)
41. Saksamudre, Suman, K., Deshmukh, R.R.: Comparative study of isolated word recognition system for Hindi language. Int. J. Eng. Res. Technol. **4**(07) (2015)
42. Somnath Hase, Sidhharth More, Dr. Sunil Nimbhore.: Isolated marathi word recognition using deep convolutional neural network (DCNN). Sambodhi (UGC Care Journal) ISSN: 2249–6661 **44**(03), (2021)
43. Supriya, S., Handore, S.M.: Speech recognition using HTK toolkit for Marathi language. In: 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 1591–1597. IEEE, (2017)