Chapter 14
Speech Recognition in Indian Languages—A Survey

Mousmita Sarma and Kandarpa Kumar Sarma

Abstract In this paper, a brief overview derived out of detailed survey of speech recognition works reported in Indian languages is described. Robustness of speech recognition systems toward language variation is the recent trend of research in speech recognition technology. To develop a system which can communicate with human in any language like any other human is the foremost requirement in order to design appropriate speech recognition technology for one to all. India is a country which has vast linguistic variations among its billion plus population. Therefore, it provides a sound area of research toward language-specific speech recognition technology. From the beginning of the commercial availability of the speech recognition system, the technology has been dominated by the hidden Markov model (HMM) methodology due to its capability of modeling temporal structures of speech and encoding them as a sequence of spectral vectors. Most of the work done in Indian languages also uses HMM technology. However, from the last 10–15 years after the acceptance of neurocomputing as an alternative to HMM, artificial neural network (ANN)-based methodologies have started to receive attention for application in speech recognition. This is a trend worldwide as part of which few works have also been reported by a few researchers.

Keywords Speech recognition system · Indian languages · Hidden Markov model (HMM) · Gaussian mixture model (GMM) · Artificial neural network (ANN)

M. Sarma (✉)
Department of Electronics and Communication Engineering,
Gauhati University, Guwahati, Assam, India
e-mail: go4mou@gmail.com

K.K. Sarma
Department of Electronics and Communication Technology, Gauhati University,
Guwahati 781014, Assam, India
e-mail: kandarpaks@gmail.com

K.K. Sarma et al. (eds.), *Recent Trends in Intelligent and Emerging Systems*,
Signals and Communication Technology, DOI 10.1007/978-81-322-2407-5_14

14.1 Introduction

The problem of automatic speech recognition (ASR) was at the forefront of research till 1930 when the first electronic voice synthesizer was designed by Homer Dudley of Bell Laboratories. After that, ASR lost its fascination among the speech processing community. Probably that was the starting of research in the direction of designing a machine that can mimic the human capability of speaking naturally and responding to spoken languages. Initial developments cover a simple machine that responds to isolated sounds. Recently, speech recognition technology has risen to such a height that a large community of people now talk to their smartphones, asking them to send e-mail and text messages, search for directions, or find information on the web. However, speech recognition technology is still far from having a machine that converses with humans on any topic like another human. In the present time, research in speech recognition concentrates on developing systems that can show robustness for variability in environment, speaker, and language. India is a linguistically rich country having 22 official languages and hundreds of other sublanguages and dialects spoken by various communities covering the billion plus population. Communication among human beings is dominated by spoken language. Therefore, it is natural for people to expect speech interfaces with computers which can speak and recognize speech in native language. But speech recognition technology in the Indian scenario is restricted to small amount of people who are both computer literate and proficient in written and spoken English. In this domain, extensive researches are going on all over India among various groups to make appropriate ASR systems in Indian languages.

Initial speech recognition systems were on isolated word recognition designed to perform special task. But in the last 25 years, certain dramatic progress in statistical methods for recognizing speech signals has been noticed. The statistical approach makes use of the four basic principles which are Bayes decision rule for minimum error rate, probabilistic models, e.g., hidden Markov models (HMMs), or conditional random fields (CRF) for handling strings of observations like acoustic vectors for ASR and written words for language translation, training criteria, and algorithms for estimating the free-model parameters from large amounts of data and the generation or search process that generates the recognition or translation result [1, 2]. The speech recognition research is dominated by the statistical approaches specifically by the HMM technology till the last one decade. It is the improvement provided by HMM technology for speech recognition in the late 1970s and the simultaneous improvement in speed of computer technology, due to which the ASR systems have become commercially viable in 1990s [3–5]. But recently in the last decade, all over the world, the ANN-based technologies are gaining attention. This is due to the fact that ANN models are composed of many nonlinear computational elements operating in parallel and arranged in the pattern of biological neural network. It is expected that human neural network like models may ultimately be able to solve the complexities of speech recognition system and provide human-like performance.

In this article, we have highlighted some works related to the research and development of ASR in Indian languages during the last decade so as to provide a picture of the fundamental progress that has been made in the large variety of Indian languages. The survey is divided into two groups based on statistical- and ANN-based technology. Initially, a glance of early speech recognition technology in world languages is also included.

14.2 Early Speech Recognition Technology

Early speech recognition systems used the acoustic-phonetic theories of speech to determine the feature [1]. Due to the complexity of human language, the inventors and engineers first focused on number or digit recognition. The first speech recognition system was built in Bell Laboratories by Davis, Biddulph, and Balashek in 1952 which could understand only isolated digits for a single speaker [6]. They used the formant frequency measured during vowel regions of each digit as a feature. During 1950–1970, laboratories in the United States, Japan, England, and the Soviet Union developed other hardware dedicated to recognizing spoken sounds, expanding speech recognition technology to support four vowels and nine consonants [7–13]. In the 1960s, several Japanese laboratories demonstrated their capability of building special-purpose hardware to perform a speech recognition task. Among them, the vowel recognizer of Suzuki and Nakata at the Radio Research Lab in Tokyo [7], the phoneme recognizer of Sakai and Doshita at Kyoto University [8], and the digit recognizer of NEC Laboratories [9] were most notable. The work of Sakai and Doshita involved the first use of a speech segmenter for analysis and recognition of speech in different portions of the input utterance. An alternative to the use of a speech segmenter was the concept of adopting a nonuniform time scale for aligning speech patterns [11, 12], dynamic programming for time alignment between two utterances known as dynamic time warping, in speech pattern matching [12] etc. Another milestone of 1960s is the formulation of fundamental concepts of linear predictive coding (LPC) [14, 15] by Atal and Itakura, which greatly simplified the estimation of the vocal tract response from speech waveforms. Development during 1970s includes the first speech recognition commercial company called Threshold Technology founded by Martin [1] and speech understanding research (SUR) program founded by advanced research projects agency (ARPA) of the U.S. Department of Defense [1]. Threshold Technology later developed the first real ASR product called the VIP-100 System [1] for some limited application and Carnegie Mellon University under ARPA developed Harpy system which was able to recognize speech using a vocabulary of 1011 words with reasonable accuracy. The Harpy system was the first to take advantage of a finite-state network to reduce computation and efficiently determine the closest matching string [16]. DRAGON system by Jim Baker was also developed during 1970s [17]. In 1980s, speech recognition turned toward prediction. Speech recognition vocabulary improved from about a few hundred words to several thousand words and had the potential to recognize an unlimited number of words.

The major reason for this up gradation is the new statistical method HMM. Rather than simply using templates for words and looking for sound patterns, HMM considered the probability of unknown sounds being words. The foundations of modern HMM-based continuous speech recognition technology were laid down in the 1970s by groups at Carnegie Mellon and IBM who introduced the use of discrete density HMMs [16–18], and then later at Bell Laboratories [19–21] where continuous density HMMs were introduced. Another reason of this drastic improvement of the speech recognition technology is the application of fundamental pattern recognition technology to speech recognition based on LPC methods in the mid 1970s by Itakura [22], Rabiner and Levinson [23], and others. Due to the expanded vocabulary provided by HMM methodology and the computer with faster processor, in 1990s speech recognition software become commercially available.

During 1980s, ANN technology was also introduced in the domain of speech recognition. The brains impressive superiority at a wide range of cognitive skills like speech recognition, has motivated the researchers to explore the possibilities of ANN models in the field of speech recognition in 1980s [24], with a hope that human neural network like models may ultimately lead to human-like performance. Early attempts at using neural networks for speech recognition centered on simple tasks like recognizing a few phonemes or a few words or isolated digits, with good success [25–27], using pattern mapping by multilayer perceptron (MLP). But at the later half of 1990, suddenly ANN-based speech research got terminated [24] after the statistical framework HMM come into focus, which supports both acoustic and temporal modeling of speech. However, it should be mentioned that the current best systems are far from equaling human-like performance and many important research issues are still to be explored. Therefore, the value of ANN-based research is still large and nowadays it is considered as the hot field in the domain of speech recognition.

14.3 Research of Speech Recognition in Indian Languages

The current speech researchers are focused on using technology to overcome the challenges in natural language processing, so that next-generation speech recognition system can provide easy and natural modes of interaction for its customer. Specifically, it has become the primary concern for the scientist and engineers to build systems that can be consumed by the common public and to enable natural language transactions between human and machine. The language-specific speech recognition is difficult mainly because the system requires knowledge of word meaning, communication context, and the commonsense. This variability includes the effect of the phonetic, phonology, syntax, semantic, and communication modes of the speech signal. While having the different meaning and usage patterns, words can have the same phonetic realization. If the words were always produced in the same way, speech recognition would be relatively easy. However, for various reasons words are almost always pronounced differently due to which it is still a challenge

to build a recognizer that is robust enough in case of any speaker, any language, and any speaking environment.

India is the country where vast cultural and linguistic variations are observed. Therefore, in such a multilingual environment there is a huge possibility of implementing speech technology. The constitution of India, has recognized 17 regional languages (Assamese, Bengali, Bodo, Dogri, Gujarati, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santhali, Sindhi, Tamil, Telugu, Urdu) along with Hindi which is the national language of India. However, till date the amount of work done in speech recognition in Indian languages has not reached the domain of rural and computer illiterate people of India [28]. Few attempts have been made by HP Labs India, IBM research lab, and some other research groups. Yet there is lots of scopes and possibilities to be explored to develop speech recognition system using Indian languages.

After the commercial availability of speech recognition system, the HMM technology has dominated the speech research. HMMs lie at the heart of virtually all modern speech recognition systems. The basic HMM framework has not changed significantly in the last decades, but various modeling techniques have been developed within this framework that has made the HMM technology considerably sophisticated [14, 15, 22]. At the same time, from the last one or two decades, ANN technology has also been used by various researchers. The current state has considered that HMM has given the best it could, but in order to improve the accuracy of speech recognition technology under language, speaker, and environmental variations, other technology is required. In the Indian language scenario, the speech recognition work can be reviewed in two parts: work done using statistical framework like HMM, gaussian mixture models (GMM) and a very few work done using ANN technology. Further a few hybrid technology-based work is also found in the literature. The following sections describe the speech recognition work developed in Indian languages over the last decade.

14.3.1 Statistical Approach

The basic statistical method used in speech recognition purpose is the HMM methodology. HMMs are a parametric model which can be used to model any time series but particularly suitable to model speech event. In HMM-based speech recognition, it is assumed that the sequence of observed speech vectors corresponding to each word is generated by a Markov model [29]. The forward backward reestimation theory called the Baum–Welch reestimation used in HMM-based speech recognition modifies the parameter in every iteration and the probability of training data increases until a local maxima reach. The success of HMM technology lies on its capability to estimate a extended set of unknown utterance from a known set of utterance given as training set [2, 30, 31]. The availability of well-structured software like hidden Markov model tool kit (HTK) [30] and CMUs Sphinx [31] which can successfully

implement HMM technology makes it easier for further research and development to incorporate new concepts and algorithms in speech recognition.

A few relevant work done in Indian languages using statistical framework like HMM are discussed below.

1. In the journal Sadhana in 1998, a work [32] has been reported by Samudravijaya et al., where they have presented a description of a speech recognition system for Hindi. The system follows a hierarchic approach to speech recognition and integrates multiple knowledge sources within statistical pattern recognition paradigms at various stages of signal decoding. Rather than making hard decisions at the level of each processing unit, relative confidence scores of individual units are propagated to higher levels. A semi-Markov model processes the frame level outputs of a broad acoustic maximum likelihood classifier to yield a sequence of segments with broad acoustic labels. The phonemic identities of selected classes of segments are decoded by class-dependent ANNs which are trained with class-specific feature vectors as input. Lexical access is achieved by string matching using a dynamic programming technique. A novel language processor disambiguates between multiple choices given by the acoustic recognizer to recognize the spoken sentence. The database used for this work consisted of sentences having 200 words, which are most commonly used in railway reservation enquiry task.
2. Another work [33] by Rajput et al. from IBM India Research Lab has been reported in 2000, where they have attempted to build decision trees for modeling phonetic context dependency in Hindi by modifying a decision tree built to model context dependency in American English. In a continuous speech recognition system, it is important to model the context-dependent variations in the pronunciations of phones. Linguistic-Phonetic knowledge of Hindi is used to modify the English phone set. Since the Hindi phone set being used is derived from the English phone set, the adaptation of the English tree to Hindi follows naturally. The method may be applicable for adapting between any two languages.
3. In 2008, Kumar et al. of IBM India Research Lab developed another HMM-based large vocabulary continuous speech recognition system for Hindi language. In this work [34], they have presented two new techniques that have been used to build the system. Initially, a technique for fast bootstrapping of initial phone models of a new language is given. The training data for the new language is aligned using an existing speech recognition engine for another language. This aligned data is used to obtain the initial acoustic models for the phones of the new language. Following this approach requires less training data. They have also presented a technique for generating baseforms, i.e., phonetic spellings for phonetic languages such as Hindi. As is inherent in phonetic languages, rules generally capture the mapping of spelling to phonemes very well. However, deep linguistic knowledge is required to write all possible rules, and there are some ambiguities in the language that are difficult to capture with rules. On the other hand, pure statistical techniques for baseform generation require large

amounts of training data, which are not readily available. But here they have proposed a hybrid approach that combines rule-based and statistical approaches in a two-step fashion.

4. For Hindi language, Gaurav et al. has reported another work recently in 2012 [35]. A continuous speech recognition system in Hindi is tailored to aid teaching geometry in primary schools. They have used the mel frequency cepstral coefficients (MFCC) as speech feature parameters and HMM to model these acoustic features. The Julius recognizer which is language independent was used for decoding.
5. Kumar et al. in 2012 has designed a feature extraction modules ensemble of MFCC, LPCC, perceptual linear predictive analysis (PLP) etc. to improve Hindi speech recognition system [36]. The outputs of the ensemble feature extraction modules have been combined using voting technique ROVER.
6. Bhuvanagirir and Kopparapu have reported another work on mixed language speech recognition [37] Hindi and English combination in 2012.
7. In 2008 Thangarajan et al. has reported a work in continuous speech recognition for Tamil Language [38]. They have built a HMM-based continuous speech recognizer based on word and triphone acoustic models. In this experiment, a word-based context-independent (CI) acoustic model for 371 unique words and a triphone-based context-dependent (CD) acoustic model for 1700 unique words have been built for Tamil language. In addition to the acoustic models, a pronunciation dictionary with 44 base phones and trigram-based statistical language model have also been built as integral components of the linguist. These recognizers give satisfactory word accuracy for trained and test sentences read by trained and new speakers.
8. In 2009, Kalyani and Sunithato worked toward the development of a dictation system like Dragon for Indian languages. In their paper [39], they have focused on the importance of creating speech database at syllable units and identifying minimum text to be considered while training any speech recognition system. They have also provided the statistical details of syllables in Telugu and its use in minimizing the search space during recognition of speech. The minimum words that cover maximum syllables are identified which can be used for preparing a small text for collecting speech sample while training the dictation system.
9. Another work in Telugu language is reported by Usha Rani and Girija in 2012. To improve the speech recognition accuracy on Telugu language, they have explored means to reduce the number of the confusion pairs by modifying the dictionary, which is used in the decoder of the speech recognition system. In their paper [40], they have described the different types of errors obtained from the decoder of the speech recognition system.
10. Das et al. reported a work [41] in Bengali language, where they have described the design of a speech corpus for continuous speech recognition. They have developed speech corpora in phone and triphone labeled between two age groups—20 to 40 and 60 to 80. HMM is used to align the speech data statistically and observed good performance in phoneme recognition and continuous word recognition done using HTK and SPHINX.

11. In Punjabi language, Dua, Aggarwal, Kadyan and Dua have reported a work in 2012, where they have attempted to develop a isolated word recognition system using HTK [42].
12. In 2013, another work has been reported by Mehta et al. where a comparative study of MFCC and LPC for Marathi isolated word recognition system is described [43].
13. Udhyakumar et al. reported a work [44] in 2004 for multilingual speech recognition to be used for information retrieval in Indian context. This paper analyzes various issues in building a HMM-based multilingual speech recognizer for Indian languages. The system is designed for Hindi and Tamil languages and adapted to incorporate Indian accented English. Language-specific characteristics in speech recognition framework are highlighted. The recognizer is embedded in information retrieval applications and hence several issues like handling spontaneous telephony speech in real-time, integrated language identification for interactive response and automatic grapheme to phoneme conversion to handle out of vocabulary words are addressed in this paper.
14. Some issues about the development of speech databases of Tamil, Telugu and Marathi for large vocabulary speech recognition system is reported in a work [45] by Anumanchipalli et al. in 2005. They have collected speech data from about 560 speakers in these three languages. They have also presented the preliminary speech recognition results using the acoustic models created on these databases using Sphinx 2 speech tool kit, which shows satisfactory improvement in accuracy.
15. During 2009, Hema A Murthy et al. described in [46, 47], a novel approach to build syllable-based continuous speech recognizers for Indian languages, where a syllable-based lexicon and language model are used to extract the word output from the HMM-based recognizer. The importance of syllables as the basic sub-word unit for recognition has been a topic for research. They have shown that a syllabified lexicon helps hypothesize the possible set of words, where the sentence is constructed with the help of N-gram based statistical language models. The database used for these works is the Doordarshan database, which is made of news bulletins of approximately 20 min duration of both male and female speakers.
16. Bhaskar et al. reported another work [48] on multilingual speech recognition in 2012. They have used a different approach to multilingual speech recognition, in which the phone sets are entirely distinct but the model has parameters not tied to specific states that are shared across languages. They have tried to build a speech recognition system using HTK for Telugu language. The system is trained for continuous Telugu speech recorded from native speakers.
17. In 2013, a work [49] has been reported by Mohan et al. where a spoken dialogue system is designed to use in agricultural commodities task domain using real-world speech data collected from two linguistically similar languages of India Hindi and Marathi. They have trained a subspace gaussian mixture model (SGMM) under a multilingual scenario [50, 51]. To remove acoustic, channel, and environmental mismatch between datasets from multiple languages, they

have used a cross-corpus acoustic normalization procedure which is a simple variant of speaker adaptive training (SAT) described by Mohan et al. in 2012 [52]. The resulting multilingual system provides the best speech recognition performance of 77.77 % for both languages .

14.3.2 Artificial Neural Network Based Approach

All the above-mentioned works available in open literature are based on HMM technology. All earlier theories of spoken word recognition [53–57] agree to the fact that the spoken word recognition is a complex, multileveled pattern recognition work performed by neural networks of human brain and the related speech perception process can be modeled as a pattern recognition network. Different levels of the language like lexical, semantic, phonemes can be used as the unit of distribution in the model. All the theories proposed that bottom up and top down processes between feature, phoneme, and word level combines to recognize a presented word. In such a situation, ANN models has the greatest potential, where hypothesis can be performed in a parallel and higher computational approach. ANN models are composed of many nonlinear computational elements operating in parallel and arranged in the pattern of biological neural network. The problem of speech recognition inevitably requires handling of temporal variation and ANN architecture like recurrent neural network (RNN), time delay neural network (TDNN) may proven to be handy in such situations. However, ANN-based speech recognition research is still at the preliminary state. A few works based on ANN technology are listed below.

1. Sarkar and Yegnanarayana have used fuzzy rough neural networks for Vowel Classification in a work reported in 1998. This paper [58] has proposed a fuzzy-rough set-based network which exploits fuzzy-rough membership functions while designing radial basis function neural networks for classification.
2. In 2001, Gangashetty and Yegnanarayana have described ANN models for recognition of consonant–vowel (CV) utterances in [59]. In this paper, an approach based on ANN models for recognition of utterances of syllable like units in Indian languages is described. The distribution capturing ability of an autoassociative neural network (AANN) model is exploited to perform nonlinear principal component analysis (PCA) for compressing the size of the feature vector. A constraint satisfaction model is proposed in this paper to incorporate the acoustic-phonetic knowledge and to combine the outputs of subnets to arrive at the overall decision on the class of an input utterance.
3. Khan et al. describe an ANN-based preprocessor for recognition of syllables in 2004. In this work [60], syllables in a language are grouped into equivalent classes based on their consonant and vowel structure. ANN models are used to preclassify the syllables into the equivalent class to which they belong. Recognition of the syllables among the smaller number of cohorts within a class is done by means of hidden Markov models. The preprocessing stage limits the

confusable set to the cohorts within a class and reduces the search space. This hybrid approach helps to improve the recognition rate over that of a plain HMM-based recogniser.

4. In 2005, Gangashetty, Chandra Sekhar, and Yegnanarayana have described an approach for multilingual speech recognition by spotting consonant–vowel (CV) units. The distribution capturing capability of AANN is used for detection of vowel onset points in continuous speech. Support vector machine (SVM) classifier is used as the classifier and broadcast news corpus of three Indian languages Tamil, Telugu, and Marathi is used [61].
5. Paul et al. in 2009 reported a work on Bangla speech recognition using LPC cepstrum features [62]. The self-organizing map (SOM) structure of ANN makes each variable length LPC trajectory of an isolated word into a fixed length LPC trajectory and thereby making the fixed length feature vector to be fed into the recognizer. The structures of the ANN are designed with MLP and tested with 3, 4, 5 hidden layers using the tan sigmoid transfer functions.
6. In 2012, Sunil and Lajish in a work [63] reported a model for vowel phoneme recognition based on average energy information in the zero-crossing intervals and its distribution using multilayer feedforward ANN. They have observed that the distribution patterns of average energy in the zero-crossing intervals are similar for repeated utterances of the same vowels and varies from vowel to vowel and this parameter is used as a feature to classify five Malayalam vowels in the recognition system. From this study, they have shown that the average energy information in the zero-crossing intervals and its distributions can be effectively utilized for vowel phone classification and recognition.
7. Pravin and Jethva recently in 2013 reported a work [64] on Gujrati speech recognition. MFCC of a few selected spoken words is used as feature to train a MLP-based word recognition system.
8. Chitturi et al. reported a work [65] in 2005, where they have proposed an ANN-based approach to model the lexicon of the foreign language with a limited amount training data. The training data for this work consisted of the foreign language with the phone set of three native languages, 49 phones in Telugu, 35 in Tamil, 48 in Marathi, and 40 in US English. The MLP with backpropagation learning algorithm learns how the phones of the foreign language vary with different instances of context. The trained network is capable of deciphering the pronunciation of a foreign word given its native phonetic composition. The performance of the technique has been tested by recognizing Indian accented English.
9. A work [66] by Thasleema and Narayanan in 2012 explores the possibility of multiresolution analysis for consonant classification in noisy environments. They have used wavelet transform (WT) to model and recognize the utterances of consonant–vowel (CV) speech units in noisy environments. A hybrid feature extraction module namely normalized wavelet hybrid feature (NWHF) using the combination of classical wavelet decomposition (CWD) and wavelet packet decomposition (WPD) along with z-score normalization technique is designed

in this work. CV speech unit recognition tasks performed for both noisy and clean speech units using ANN and k Nearest Neighborhood.

10. In 2010, Sukumar et al. has reported a work on recognition of isolated question words of Malayalam language from speech queries using ANN- and discrete wavelet transform (DWT)-based speech feature [67].
11. Sarma et al. has reported works on recognition of numerals of Assamese language in [68, 69] using ANN in 2009. In [68], the ANN models are designed using a combination of SOM and MLP constituting a learning vector quantization (LVQ) block trained in a cooperative environment to handle male and female speech samples of numerals. This work provides a comparative evaluation of several such combinations while subjected to handle speech samples with gender-based differences captured by a microphone in four different conditions viz. noiseless, noise mixed, stressed, and stress-free. In [69], the effectiveness of using an adaptive LMS filter and LPC cepstrum to recognize isolated numerals using ANN-based cooperative architectures. The entire system has two distinct parts for dealing with two classes of input classified into male and female clusters. The first block is formed by a MLP which acts like a class mapper network. It categorizes the inputs into two gender-based clusters.
12. In 2010, Bhattacharjee presented a technique for the recognition of isolated keywords from spoken search queries [70]. A database of 300 spoken search queries from Assamese language has been created. In this work, MFCC has been used as the feature vector and MLP to identify the phoneme boundaries as well as for recognition of the phonemes. Viterbi search technique has been used to identify the keywords from the sequence of phonemes generated by the phoneme recognizer.
13. A work by Dutta and Sarma [71] in 2012 describes a speech recognition model using RNN where linear predictive coding (LPC) and mel frequency cepstral coefficient (MFCC) are used for feature extraction in two separate decision blocks and the decision is taken from the combined architecture. The multiple feature extraction block-based model provides 10 % gain in the recognition rate in comparison to the case when individual feature extractor is used.
14. In 2013, Bhattacharjee in [72] provided a comparative study of linear predictive cepstral coefficients (LPCC) and MFCC features for the recognition of phones of Assamese language. Two popular feature extraction techniques, LPCC and MFCC, have been investigated and their performances have been evaluated for the recognition using a MLP-based baseline phoneme recognizer in quiet environmental condition as well as at different level of noise. It has been reported that the performance of LPCC-based system degrades more rapidly compared to the MFCC-based system under environmental noise condition whereas under speaker variability conditions, LPCC shows relative robustness when compared to MFCC though the performance of both the systems degrade considerably.
15. Sarma and Sarma in 2013, reported a work [73] where a hybrid ANN model is used to recognize initial phones from CVC-type Assamese words. An SOM-based algorithm is developed to segment the initial phonemes from its word

counterpart. Using a combination of three types of ANN structures, namely recurrent neural network (RNN), SOM, and probabilistic neural network (PNN), the proposed algorithm proves its superiority over the discrete wavelet transform (DWT)-based phoneme segmentation.

14.4 Conclusion

It can be concluded from the above literature that the speech recognition technology for Indian languages has not yet covered all the official languages. A few works are done in Hindi language by IBM research lab and a few other research groups have appreciable quality. A few other works have covered Marathi, Tamil, Telugu, Punjabi, Assamese, and Bengali languages which are widely spoken throughout the country. A few works are reported on multilingual speech recognition. However, ASR technologies are yet to be reported in some other mainstream languages like Urdu, Sanskrit, Kashmiri, Sindhi, Konkani, Manipuri, Kannada, Nepali etc. The HMM-based works have already supported the use of continuous speech. In contrast, ANN-based works are still centered around isolated words. But the scenario looks bright and many new success stories shall be reported in near future which shall take ASR technology in Indian languages to new heights.

References

1. Juang BH, Rabiner LR (2004) Automatic speech recognition—a brief history of the technology development. http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf
2. Gales M, Young S (2007) The application of hidden Markov models in speech recognition. Found Trends Sig Process 1(3):195–304
3. Levinson SE, Rabiner LR, Sondhi MM (1983) An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell Syst Tech J 62(4):1035–1074
4. Ferguson JD (1980) Hidden Markov analysis: an introduction, in hidden Markov models for speech. Institute for Defense Analyses, Princeton
5. Rabiner LR, Juang BH (2004) Statistical methods for the recognition and understanding of speech. In: Encyclopedia of language and linguistics
6. Davis KH, Biddulph R, Balashek S (1952) Automatic recognition of spoken digits. J Acoust Soc Am 24(6):637–642
7. Suzuki J, Nakata K (1961) Recognition of Japanese vowels preliminary to the recognition of speech. J Radio Res Lab 37(8):193–212
8. Sakai J, Doshita S (1962) The phonetic typewriter, information processing. In: Proceedings IFIP Congress, Munich
9. Nagata K, Kato Y, Chiba S (1963) Spoken digit recognizer for Japanese language. NEC Res Dev (6)
10. Fry DB, Denes P (1959) The design and operation of the mechanical speech recognizer at University College London. J British Inst Radio Eng 19(4):211–229

11. Martin TB, Nelson AL, Zadell HJ (1964) Speech recognition by feature abstraction techniques. Technical Report AL-TDR-64-176, Air Force Avionics Laboratory
12. Vintsyuk TK (1968) Speech discrimination by dynamic programming. Kibernetika 4(2):81–88
13. Viterbi AJ (1967) Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Trans Inf Theory 13:260–269
14. Atal BS, Hanauer SL (1971) Speech analysis and synthesis by linear prediction of the speech wave. J Acoust Soc Am 50(2):637–655
15. Itakura F, Saito S (1970) A statistical method for estimation of speech spectral density and formant frequencies. Electr Commun Japan 53(A):36–43
16. Lowerre BT (1976) The HARPY speech recognition system. Doctoral thesis, Carnegie-Mellon University, Department of Computer Science
17. Baker JK (1975) The Dragon system an overview. IEEE Trans Acoust Speech Sig Process 23(1):24–29
18. Jelinek F (1976) Continuous speech recognition by statistical methods. Proc IEEE 64(4):532–556
19. Juang BH (1984) On the hidden Markov model and dynamic time warping for speech recognition a unified view. AT T Tech J 63(7):1213–1243
20. Juang BH (1985) Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains. AT T Tech J 64(6):1235–1249
21. Levinson SE, Rabiner LR, Sondhi MM (1983) An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell Syst Tech J 62(4):1035–1074
22. Itakura F (1975) Minimum prediction residual principle applied to speech recognition. IEEE Trans Acoust Speech Sig Process 23:57–72
23. Rabiner LR, Levinson SE, Rosenberg AE, Wilpon JG (1979) Speaker independent recognition of isolated words using clustering techniques. IEEE Trans Acoust Speech Sig Process 27:336–349
24. Hu YH, Hwang JN (2002) Handbook of neural network signal processing. In: The electrical engineering and applied signal processing. CRC Press, Boca Raton
25. Lippmann RP (1990) Review of neural networks for speech recognition. In: Readings in speech recognition, pp 374–392. Morgan Kaufmann Publishers, San Mateo
26. Evermann G, Chan HY, Gales MJF, Hain T, Liu X, Mrva D, Wang L, Woodland P (2004) Development of the 2003 CU-HTK conversational telephone speech transcription system. In: Proceedings of ICASSP, Montreal, Canada
27. Matsoukas S, Gauvain JL, Adda A, Colthurst T, Kao CI, Kimball O, Lamel L, Lefevre F, Ma JZ, Makhoul J, Nguyen L, Prasad R, Schwartz R, Schwenk H, Xiang B (2006) Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system. IEEE Trans Audio Speech Lang Process 14(5):1541–1556
28. Languages. The Constitution of India, Eight Schedule, Articles 344(1) and 351:330. Available via http://lawmin.nic.in/coi/coiason29july08.pdf
29. Rabiner L, Juang BH (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3(1):4–16
30. Young S, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P (2000) The HTK book. http://htk.eng.cam.ac.uk/
31. Lee KF, Hon HW, Reddy R (1990) An overview of the SPHINX speech recognition system. IEEE Trans Acoust Speech Sig Process 38(1):35–45
32. Samudravijaya K, Ahuja R, Bondale N, Jose T, Krishnan S, Poddar P, Rao PVS, Raveendran R (1998) A feature-based hierarchical speech recognition system for Hindi. Sadhana 23(4):313–340
33. Rajput N, Subramaniam LV, Verma A (2000) Adapting phonetic decision trees between languages for continuous speech recognition. In: Proceedings of IEEE international conference on spoken language processing. Beijing, China
34. Kumar M, Rajput N, Verma A (2004) A large-vocabulary continuous speech recognition system for Hindi. IBM J Res Dev 48(5/6):703–715

35. Gaurav DS, Deiv G, Sharma K, Bhattacharya M (2012) Development of application specific continuous speech recognition system in Hindi. J Sig Inf Process 3:394–401
36. Kumar M, Aggarwal RK, Leekha G, Kumar Y (2012) Ensemble feature extraction modules for improved Hindi speech recognition system. Proc Int J Comput Sci Issues 9(3):359–364
37. Bhuvanagirir K, Kopparapu SK (2012) Mixed language speech recognition without explicit identification of language. Am J Sig Process 2(5):92–97
38. Thangarajan R, Natarajan AM, Selvam M (2008) Word and triphone based approaches in continuous speech recognition for Tamil Language. Wseas Trans Sig Process 4(3):76–85
39. Kalyani N, Sunitha KVN (2009) Syllable analysis to build a dictation system in Telugu language. Int J Comput Sci Inf Secur 6(3):171–176
40. Usha Rani N, Girija PN (2012) Error analysis to improve the speech recognition accuracy on Telugu language. Sadhana 37(6):747–761
41. Das B, Mandal V, Mitra P (2011) Bengali speech corpus for continuous automatic speech recognition system. In: Proceedings of international conference on speech database and assessments, pp 51–55
42. Dua M, Aggarwal RK, Kadyan V, Dua S (2012) Punjabi automatic speech recognition using HTK. Int J Comput Sci Issues 9(4):359–364
43. Mehta LR, Mahajan SP, Dabhade AS (2013) Comparative study of MFCC And LPC for Marathi isolated word recognition system. Int J Adv Res Electr Electr Instrum Eng 2(6):2133–2139
44. Udhyakumar N, Swaminathan R, Ramakrishnan SK (2004) Multilingual speech recognition for information retrieval in Indian context. In: Proceedings of the student research workshop at HLT-NAACL, pp 1–6
45. Anumanchipalli G, Chitturi R, Joshi S, Kumar R, Singh SP, Sitaram RNV, Kishore SP (2005) Development of Indian language speech databases for large vocabulary speech recognition systems. In: Proceedings of international conference on speech and computer
46. Lakshmi A, Murthy HA (2008) A new approach to continuous speech recognition in Indian languages. In: Proceedings national conference communication
47. Lakshmi SG, Lakshmi A, Murthy HA, Nagarajan T (2009) Automatic transcription of continuous speech into syllable-like units for Indian languages. Sadhana 34(2):221–233
48. Bhaskar PV, Rao SRM, Gopi A (2012) HTK based Telugu speech recognition. Int J Adv Res Comput Sci Softw Eng 2(12):307–314
49. Mohan A, Rose R, Ghalehjegh SH, Umesh S (2013) Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain. Speech Commun. http://dx.doi.org/10.1016/j.specom.2013.07.005
50. Povey D, Burget L, Agarwal M, Akyazi P, Kai F, Ghoshal A, Glembek O, Goel N, Karafiat M, Rastrow A (2011) The subspace Gaussian mixture model A structured model for speech recognition. Comput Speech Lang 25(2):404–439
51. Rose RC, Yin SC, Tang Y (2011) An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing
52. Mohan A, Ghalehjegh SH, Rose RC (2012) Dealing with acoustic mismatch for training multilingual subspace Gaussian mixture models for speech recognition. In: Proceedings of IEEE international conference on acoustics, speech and signal processing
53. Diehl RL, Lotto AJ, Holt LL (2004) Speech perception. Annu Rev Psychol 55:149–179
54. Eysenck MW (2004) Psychology-an international persepective. Psychology Press. http://books.google.co.in/books/about/Psychology.html?id=l8j_z5-qZfACredir_esc=y
55. Jusczyk PW, Luce PA (2002) Speech perception and spoken word recognition: past and present. Ear Hear 23(1):2–40
56. Bergen B (2006) Linguistics 431/631: connectionist language modeling. Meeting 10: speech perception. http://www2.hawaii.edu/bergen/ling631/lecs/lec10.htm
57. McClelland JL, Mirman D, Holt LL (2006) Are there interactive processes in speech perception? Trends Cogn Sci 10(8):363–369
58. Sarkar M, Yegnanarayana B (1998) Fuzzy-rough neural networks for vowel classification. IEEE Int Conf Syst Man Cybern 5:4160–4165

59. Gangashetty SV, Yegnanarayana B (2001) Neural network models for recognition of consonant-vowel (CV) utterances. In: Proceedings of international joint conference on neural networks
60. Khan AN, Gangashetty SV, Yegnanarayana B (2004) Neural network preprocessor for recognition of syllables. In: Proceedings of international conference on intelligent sensing and information processing
61. Gangashetty SV, Sekhar CC, Yegnanarayana B (2005) Spotting multilingual consonant-vowel units of speech using neural network models. In: Proceeding of international conference on non-linear speech processing
62. Paul AK, Das D, Kamal M (2009) Bangla speech recognition system using LPC and ANN. In: Proceedings of the seventh international conference on advances in pattern recognition, pp 171–174
63. Sunil KRK, Lajish VL (2012) Vowel phoneme recognition based on average energy information in the zerocrossing intervals and its distribution using ANN. Int J Inf Sci Tech 2(6)
64. Pravin P, Jethva H (2013) Neural network based Gujarati language speech recognition. Int J Comput Sci Manage Res 2(5)
65. Chitturi R, Keri V, Anumanchipalli G, Joshi S (2005) Lexical modeling for non native speech recognition using neural networks. In: Proceedings of international conference of natural language processing
66. Thasleema TM, Narayanan NK (2012) Multi resolution analysis for consonant classification in noisy environments. Int J Image Graph Sig Process 8:15–23
67. Sukumar AR, Shah AF, Anto PB (2010) Isolated question words recognition from speech queries by using artificial neural networks. In: Proceedings of international conference on computing communication and networking technologies, pp 1–4
68. Sarma MP, Sarma KK (2009) Assamese numeral speech recognition using multiple features and cooperative LVQ-architectures. Int J Electr Electr Eng 5(1):27–37
69. Sarma M, Dutta K, Sarma KK (2009) Assamese numeral corpus for speech recognition using cooperative ANN architecture. World Acad Sci Eng Technol 28:581–590
70. Bhattacharjee U (2010) Search key identification in a spoken query using isolated keyword recognition. Int J Comput Appl 5(8):14–21
71. Dutta K, Sarma KK (2012) Multiple feature extraction for RNN-based assamese speech recognition for speech to text conversion application. In: Proceedings of international conference on communications, devices and intelligent systems (CODIS), pp 600–603
72. Bhattacharjee U (2013) A comparative study Of LPCC and MFCC features for the recognition of assamese phonemes. Int J Eng Res Technol 2(1):1–6
73. Sarma M, Sarma KK (2013) An ANN based approach to recognize initial phonemes of spoken words of assamese language. Appl Soft Comput 13(5):2281–2291