Chapter 2
Literature Review

2.1 Introduction

Speech recognition is one of the most active areas of research from last six decades. Many important contributions in speech recognition research are reported in past 25 years. Several researchers have attempted to use articulatory and excitation source features to improve performance of speech recognition systems. There are a very limited number of works using excitation source features for speech recognition, while there are a good number of works exploring the AFs for speech recognition. Few prior works related to the use of articulatory and excitation source features for developing speech recognition systems are briefly discussed in this chapter. The organization of this chapter is as follows : The prior works related to the speech recognition are discussed in Sect. 2.2. Section 2.3 describes the prior works on the development of speech recognition systems using articulatory features (AFs). In Sect. 2.4, the prior works related to the development of speech recognition systems using excitation source features are explained. Section 2.5 summarizes this chapter.

2.2 Prior Works on Speech Recognition

From the existing literature, it is observed that there are lot of works available in the area of speech recognition. This section lists only few important contributions in speech recognition research. In 1959, D.B. Fry [1] presented about the future directions for speech recognition research. He has summarized the working of the human recognition system and described the importance of modeling the speech recognition by machine in a similar way to that of human recognition system. Since the human recognition mechanism depends on both acoustic cues and language characteristics, the language-specific information and spectral features must be combined to improve the performance of speech recognition systems.

K.S. Rao and Manjunath K.E., *Speech Recognition Using Articulatory and Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-49220-9_2

In 1973, Raj Reddy et al. [2] have developed a *HEARSAY* system for voice-chess application. The task of *HEARSAY* system is to recognize a spoken move in a given board position. The model used hypothesis and test paradigm with a set of cooperating independent parallel processes. The information from all the processes is collectively used to recognize the spoken utterance.

In 1976, G.M. White et al. [3] carried out isolated-word recognition using city names and alpha digits. The linear predictive analysis for preprocessing and dynamic programming for classification are used. It is observed that the use of data reduction techniques leads to the reduction in the performance of speech recognition systems.

In 1988, R.P. Lippmann [4] used neural networks for isolated-word recognition. The performance of neural networks is compared with conventional classifiers such as Gaussian and k-nearest neighbor classifiers. The vowel and digit classification experiments are performed. It is observed that neural networks perform better than conventional classifiers for both vowel and digit classification experiments.

In 1989, A. Waibel et al. [5] used time delay neural network (TDNN) for isolated phoneme recognition. The isolated phoneme recognizer was developed using 3 phonemes, namely /b/, /d/, /g/. Three-layered TDNN with error backpropagation is used. The phone recognition accuracy of 98.5% is reported.

In 1989, L.R. Rabiner [6] proposed hidden Markov models (HMMs) for continuous-speech recognition. Three basic problems of HMMs are addressed. Implementation issues related to use HMMs for developing speech recognition systems are explained. The connected-digit and isolated-word recognizers are developed. This is one of the very important contributions to speech recognition research.

In 1989, K.-F. Lee et al. [7] used HMMs for developing a continuous-speech recognizer. TIMIT speech corpus with 39 phones is used. Linear prediction cepstral coefficients (LPCCs) are used as spectral features, and Viterbi decoding was used for decoding the test utterances.

In 1990, F. Fallside et al. [8] have developed continuous-speech recognizer using neural networks. TIMIT corpus with 61 phones is used. The development of phoneme-to-word recognizer is described.

In 1994, H.A. Bourlard et al. [9] proposed hybrid HMM/multilayer perceptron (MLPs) approach for speech recognition. In hybrid HMM/MLP approach, the state emission probabilities of HMMs are estimated using MLPs. Speech recognition systems are developed using HMMs, MLPs, and combination of HMMs/MLPs. TIMIT speech corpus with 61 phones is used. Viterbi decoding is used for decoding test utterances. The performance of hybrid system developed using the combination of HMMs/MLPs is higher compared to other two systems.

In 2000, H. Hermansky et al. [10] proposed the development for tandem speech recognition systems. In tandem speech recognition systems, the output of the first stage is used as feature to develop the second stage. The posterior probabilities obtained from MLPs in the first stage are used as acoustic observations to develop the speech recognition system in the second stage using HMMs. This leads to the combination of discriminative feature processing ability of MLP in the first stage with distribution modeling ability of HMM in the second stage. A reduction of 35%

in the relative error rate compared to conventional Gaussian mixture model (GMM)-HMM-based system is observed.

In 2008, H. Ketabdar et al. [11] proposed a method for more accurate estimation of phone posteriors by the first stage of tandem speech recognition systems. The phone posteriors are better estimated by integrating phonetic and lexical knowledge along with discriminative knowledge. The phonetic and lexical knowledge is captured by using long temporal context. More accurately estimated phone posteriors resulted in the improvement of performance of tandem systems.

Much of the work is not reported in the context of Indian languages. Since the basic units in Indian languages are syllables, the syllable-based speech recognition systems are more appropriate for Indian languages. The syllable is more stable unit than phone as it captures the coarticulation effect well. Few works exploring the syllable-based speech recognition systems for Indian languages are listed below.

In 2004, S.V. Gangashetty et al. [12] have developed syllable-based speech recognition systems for three Indian languages, namely Telugu, Hindi, and Tamil. The syllables are generalized to consonant-vowel (CV) units. The CV units in the continuous speech are spotted using vowel onset points (VOPs) as the anchor points. Support vector machines (SVMs) and autoassociative neural networks (ANNs) are used for developing classification models.

In 2005, S.V. Gangashetty et al. [13] have proposed hybrid HMM/SVM systems by combining the evidences from HMMs and SVMs. The maximum-likelihood estimates of HMMs are combined with the discriminative knowledge captured by SVMs to recognize CV units more accurately. Hybrid HMM/SVM systems have outperformed both HMM-based and SVM-based systems.

In 2012, A.K. Vuppala et al. [14, 15] have proposed two-stage CV recognition system for improving the performance of syllable-based speech recognition system. Two-stage CV recognition system consists of HMMs in the first stage and SVMs in the second stage. HMMs are used for detecting vowel category, while the SVMs are used for detecting consonant category of the CV unit. Telugu broadcast news corpus is used to evaluate the performance of two-stage CV recognition system. It is found that two-stage CV recognition system outperformed the HMM-based and SVM-based single-stage systems. VOP detection methods are discussed in [16, 17]. Syllable-based speech recognition systems are reported in [18].

Few works related to isolated-word recognition systems in the context of Indian languages are listed below. In 2011, K. Kumar et al. [19] have developed isolated-word recognizer for Hindi using HMMs. In 2012, M. Dua et al. [20] have developed an isolated-word recognizer for Punjabi using HMMs.

In recent years, dramatic improvement in the performance of speech recognition systems is achieved by using deep neural networks (DNNs). In 2012, Abdel-rahman Mohamed et al. [21, 22] have used DNNs for speech recognition. DNNs have many layers of hidden units and very large number of parameters. DNNs take coefficients of several frames as input and produce posterior probabilities as output. It is shown that the HMMs with each state modeled using posterior probabilities of DNNs outperform the HMMs with each state modeled using the mixture of Gaussians.

In 2013, A. Graves et al. [23] have explored deep recurrent neural networks for speech recognition. Deep recurrent neural networks involve stacking of multiple recurrent hidden layers on top of each other. The obtained results are comparable with that of DNNs.

In 2013, Tara N. Sainath et al. [24] explored convolutional neural networks (CNNs) for large vocabulary speech recognition (LVCSR). The behavior of features obtained from CNNs is studied for different LVCSR tasks. The behavior of CNNs is compared with DNNs and GMMs. It is found that the CNNs have higher performance compared to DNNs and GMMs. The experiments are conducted using broadcast news corpus and switchboard corpus.

In 2014, Laszlo Toth [25] proposed the use of maxout activation function for CNNs to improve the performance of CNN-based speech recognition systems. It is found that the use of maxout active function resulted in the reduction of phone error rate up to 6%.

In general, speech can be broadly classified into read, extempore, and conversation modes of speech. Read speech involves reading out from the notes such as news reading. Extempore mode of speech is delivered without the aid of notes such as public speaking or delivering a lecture in a class. Conversation mode of speech is an interactive, spontaneous communication between two or more people. More details on read, extempore, and conversation modes of speech are given in Sect. 5.2. All the works described above have used read speech corpus. Few works related to extempore and conversation modes of speech are listed as below.

In 2003, J.L. Gauvain et al. [26] have developed conversational telephone speech recognition system using telephone conversational speech corpus. Speaker normalization and speaker adaptation techniques are employed to improve the performance of conversation speech recognition system.

In 2005, Florian Metze [27] has performed conversational speech recognition. The articulatory features are used to improve the performance of conversational speech recognition systems.

In 2013, Shridhara M V et al. [28] have developed a phone recognition system (PRS) for Kannada language using HMMs. Separate PRSs are developed for read,

Table 2.1 Summary of prior works on speech recognition

• Speech recognition research till 1989, mainly concentrated on the development of isolated-word recognizers
• Development of speech recognition systems using HMMs proposed by Lawrence R. Rabiner in 1989 is one of the major breakthroughs in speech recognition research
• Development of continuous-speech recognition systems started mostly after 1989
• Speech recognition systems are generally developed using HMMs, neural networks, and SVMs
• Tandem and hybrid approaches are most commonly used to improve the performance of speech recognition systems
• State-of-the-art LVCSR systems are developed using CNNs and large amount of training data using DNNs

extempore, and conversation modes of speech, and the results are compared. The phone recognition systems for Bengali and Odia are reported in [29, 30]. A summary of the prior works on speech recognition is provided in Table 2.1.

2.3 Prior Works on Speech Recognition Using Articulatory Features

There are some works exploring the AFs to improve the performance of speech recognition systems. Some of the recent ones are listed as follows: In 2002, Katrin Kirchhoff et al. [31] have used AFs to develop the robust speech recognition systems. The continuous-digit recognition using telephone speech and conversational speech recognition are carried out. It is shown that AF-based systems are capable of achieving superior performance at high noise levels. The combination of acoustic and AFs consistently leads to a significant reduction of word error rate across all acoustic conditions.

In 2005, Florian Metze [27] has used the AFs to improve the performance of conversational speech recognition systems. In 2007, O. Cetin et al. [32] have used AFs to develop the tandem PRSs. The AFs are derived by training MLPs using spectral features. Fisher and switchboard speech corpora are used. The derived AF evidences along with *perceptual linear prediction* features are used to improve the word error rate.

In 2007, Joe Frankel et al. [33] used AFs to develop tandem PRSs. MLP-based AF classifiers are trained using 2000 hours of telephone speech. The recognition accuracies of AF-tandem PRSs are higher than those of phone posterior-based tandem PRSs.

In 2009, Sabato Marco Siniscalchi et al. [34] have used the acoustic-phonetic information to develop speech recognition systems. The acoustic-phonetic information contained the place and manner of articulation. A bank of speech event detectors are used to score place and manner of articulation events using lattice rescoring approach, to derive acoustic-phonetic information. Three tasks, namely continuous-speech recognition, connected-digit recognition, and LVCSR, are carried out. It is found that in all the three cases, systems developed using acoustic-phonetic information have shown higher performance.

In 2013, Vikramjit Mitra et al. [35] have estimated articulatory trajectories from speech signals using neural networks. The articulatory trajectories indicate the place of constriction. The estimated articulatory trajectories are combined with MFCCs to develop LVCSR systems. Results show that the use of articulatory information improves the performance in both clean and noisy environments.

In all of the existing works, the AFs are mostly used as tandem features to improve the recognition accuracy of speech recognition systems. Hence, we have proposed weighted combination approach to combine the evidences derived from five different AF groups. The hybrid PRSs are developed using weighted combination of

Table 2.2 Summary of prior works on speech recognition using articulatory features

• AFs are used for developing robust speech recognition systems
• AFs are mostly used as tandem features to improve the performance of speech recognition systems
• There are no works exploring the AFs to improve the performance of speech recognition systems in the context of Indian languages
• In this book, a weighted combination approach is proposed to combine the evidences derived from different AF groups and AFs are explored in the context of Indian languages using Bengali

various AFs. The systematic analysis of the enhancement of phone-level accuracies contributed by each AF group is carried out. The analysis is carried out by developing separate hybrid PRSs based on the consonant AFs and vowel AFs. From the literature, it is observed that there are no works exploring the AFs to improve the performance of PRSs in the context of Indian languages. Hence, in this book, we have explored AFs in the context of Indian languages using Bengali. Since the AFs provide supplementary information for phone recognition, the combination of spectral and articulatory features may lead to significant improvement in the performance of PRSs. The objective of our study is to use AFs to improve the phone recognition accuracy of PRSs. A summary of the prior works on speech recognition using articulatory features is provided in Table 2.2.

2.4 Prior Works on Speech Recognition Using Excitation Source Features

There are very limited works exploring the excitation source features for speech recognition. Some of the recent works exploring the excitation source features for speech recognition are listed as follows. In 1996, Jialong He et al. [36] have used linear prediction (LP) residual features, containing excitation source information, to improve the performance of isolated-word recognizer. HMM-based speaker-independent isolated-word recognizer is developed using OGI-ISOLET speech corpus. An improvement of 13% was observed in the recognition accuracy. They have concluded that LP residual features contain useful information for speech recognition and act as complementary information to improve the recognition accuracy.

In 1998, Rathinavelu Chengalvarayan [37] has used LP residual features, containing excitation source information, to improve the performance of city name recognizer. A combination of LPCCs and LP residual features leads to the reduction of 8% in the string error rate [37]. In 2008, M. Chetouani et al. [38] claim that LP residual feature contains both linguistic and speaker information.

Table 2.3 Summary of prior works on speech recognition using excitation source features

• Excitation source features are mostly used for improving the performance of isolated-word recognition systems
• There are no works exploring the excitation source features for continuous-speech recognition
• Excitation source features for developing continuous-speech recognition systems are proposed

In 2011, N. Dhananjaya et al. [39] have hypothesized the manner of articulation (MOA) using excitation source information. HMM-based MOA recognizer is developed for five broad MOA categories using TIMIT speech corpus. The acoustic-phonetic information extracted from excitation source features is used to detect and correct the errors at the output of HMM-based MOA recognizer.

In all of the existing works, the excitation source features are mostly used for improving the performance of isolated-word recognition systems. In [36, 37], the excitation source features are used for improving the recognition accuracies of the isolated spoken letter recognizer and the city name recognizer, respectively. There are no works exploring the excitation source features for continuous-speech recognition. Hence, in this book, we have explored excitation source features for developing continuous-speech recognition systems. From the literature, it is observed that there are no works exploring the excitation source features to improve the performance of PRSs in the context of Indian languages. Hence, in this book, we have explored excitation source features in the context of Indian languages using Bengali. The objective of our study is to improve the performance of PRSs using the combination of vocal tract and excitation source features. A summary of the prior works on speech recognition using excitation source features is provided in Table 2.3.

2.5 Summary

In this chapter, overview of prior works on speech recognition and the existing works related to articulatory and excitation source features for developing speech recognition systems are briefly described. There are no works exploring the excitation source features for continuous-speech recognition. Articulatory features are mostly used as tandem features to improve the performance of speech recognition systems. There are no works exploring the articulatory and excitation source features to improve the performance of speech recognition systems in the context of Indian languages. Hence, in this book, articulatory and excitation source features are explored for an Indian language Bengali.

References

1. D.B. Fry, Theoretical aspects of mechanical speech recognition. J. B. Inst. Radio Eng. **19**, 211–218 (1959)
2. D. Raj Reddy, L.D. Erman, R.B. Neely, A model and a system for machine recognition of speech. IEEE Trans. Audio and Electroacoust. **AU-21**, 229–238 (1973)
3. G.M. White, R.B. Neely, Speech Recognition experiments with linear predication, bandpass filtering, and dynamic programming. IEEE Trans. Acoust. Speech Signal Process. **ASSP-24**, 183–188 (1976)
4. R.P. Lippmann, Neural network classifiers for speech recognition. Linc. Lab. J. **1**, 107–124 (1988)
5. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K.J. Lang, Phoneme recognition using time-delay neural networks. IEEE Trans. Acoust. Speech Signal Process. **37**, 328–339 (1989)
6. L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**, 257–286 (1989)
7. K.-F. Lee, H.-W. Hon, Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. **37**, 1641–1648 (1989)
8. F. Fallside, H. Lucke, T.P. Marsland, P.J. O Shea, M.S.J. Owen, R.W. Prager, A.J. Robinson, N.H. Russell, Continuous speech recognition for the TIMIT database using neural networks, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1990), pp. 445–448
9. H.A. Bourlard, N. Morgan, *Connnectionist Speech Recognition: A Hybrid Approach* (Kluwer Academic Publishers Norwell, USA, 1994)
10. H. Hermansky, D.P.W. Ellis, S. Sharma, Tandem connectionist feature extraction for conventional HMM systems, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2000), pp. 1635–1638
11. H. Ketabdar, H. Bourlard, Hierarchical integration of phonetic and lexical knowledge in phone posterior estimation, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2008), pp. 4065–4068
12. S.V. Gangashetty, C.C. Sekhar, B. Yegnanarayana, Spotting consonant-vowel units in continuous speech using autoassociative neural networks and support vector machines, in *IEEE Workshop on Machine Learning for Signal Processing* (2004), pp. 401–410
13. S.V. Gangashetty, C.C. Sekhar, B. Yegnanarayana, Combining evidence from multiple classifiers for recognition of consonant-vowel units of speech in multiple languages, in *IEEE International Conference on Intelligent Sensing and Information Processing* (2005), pp. 387–391
14. A.K. Vuppala, K. Sreenivasa Rao, S. Chakrabarti, Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points. Circuits Syst. Signal Process. **31**, 1459–1474 (2012)
15. A.K. Vuppala, K. Sreenivasa Rao, S. Chakrabarti, Improved consonant-vowel recognition for low bit-rate coded speech. Int. J. Adapt. Control Signal Process. **26**, 333–349 (2012)
16. A.K. Vuppala, J. Yadav, K. Sreenivasa Rao, S. Chakrabarti, Vowel onset point detection for low bit rate coded speech. IEEE Trans. Audio Speech Lang. Process. **20**, 1894–1903 (2012)
17. A.K. Vuppala, K. Sreenivasa Rao, S. Chakrabarti, Improved vowel onset point detection using epoch intervals. AEU - Int. J. Electron. Commun. **66**, 697–700 (2012)
18. Manjunath K.E., SBS Kumar, D. Pati, B. Satapathy, K. Sreenivasa Rao, Development of consonant-vowel recognition systems for Indian languages: Bengali and Oriya, in *IEEE INDICON* (2013)
19. K. Kumar, R.K. Aggarwal, Hindi Speech Recognition system using HTK. Int. J. Comput. Bus. Res. **2**, 1–12 (2011)
20. M. Dua, R.K. Aggarwal, V. Kadyan, S. Dua, Punjabi automatic speech recognition using HTK. Int. J. Comput. Sci. Issues **9**, 359–363 (2012)
21. A. Mohamed, G.E. Dahl, G. Hinton, Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**, 14–22 (2012)

22. G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. **29**, 82–97 (2012)
23. A. Graves, A.-R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013), pp. 6645–6649
24. T.N. Sainath, A. Mohamed, B. Kingsbury, B. Ramabhadran, Deep convolutional neural networks for LVCSR, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013), pp. 8614–8618
25. L. Toth, Convolutional deep maxout networks for phone recognition, in *International Speech Communication Association (INTERSPEECH)* (2014), pp. 1078–1082
26. J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, F. Lefevre, Conversational telephone speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2003), pp. 212–215
27. F. Metze, Articulatory features for conversational speech recognition. Ph.D. dissertation, Carnegie Mellon University, 2005
28. M.V. Shridhara, B.K Banahatti, L. Narthan, V. Karjigi, R. Kumaraswamy, Development of Kannada speech corpus for prosodically guided phonetic search engine, in *IEEE International Oriental COCOSDA (OCOCOSDA)* (2013), pp. 1–6
29. Manjunath K.E., K. Sreenivasa Rao, D. Pati, Development of phonetic engine for Indian languages: Bengali and Oriya, in *IEEE International Oriental COCOSDA* (2013)
30. Manjunath K.E., K. Sreenivasa Rao, Automatic phonetic transcription for read, extempore and conversation speech for an Indian language: Bengali, in *IEEE National Conference on Communications* (2014)
31. K. Kirchhoff, G.A. Fink, G. Sagerer, Combining acoustic and articulatory feature information for robust speech recognition. Speech Commun. **37**, 303–319 (2002)
32. O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, K. Livescu, An articulatory feature-based tandem approach and factored observation modeling, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2007), pp. 645–648
33. J. Frankel, M. Magimai-Doss, S. King, K. Livescu, O. Cetin, Articulatory feature classifiers trained on 2000 hours of telephone speech, in *International Speech Communication Association (INTERSPEECH)* (2007), pp. 36–41
34. S.M. Siniscalchi, C.-H. Lee, A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. Speech Commun. **51**, 1139–1153 (2009)
35. V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Richey, J. Yuan, M. Liberman, Articulatory trajectories for large-vocabulary speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013), pp. 7145–7149
36. J. He, L. Liu, G. Palm, On the use of residual cepstrum in speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1996), pp. 5–8
37. R. Chengalvarayan, On the use of normalized LPC error towards better large vocabulary speech recognition systems, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1998), pp. 17–20
38. M. Chetouani, M. Faundez-Zanuy, B. Gas, J.L. Zarader, Investigation on LP-residual representations for speaker identification. Pattern Recognit. **42**, 487–494 (2009)
39. N. Dhananjaya, B. Yegnanarayana, S.V. Gangashetty, Acoustic-phonetic information from excitation source for refining manner hypotheses of a phone recognizer, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2011), pp. 5252–5255

http://www.springer.com/978-3-319-49219-3

Speech Recognition Using Articulatory and Excitation Source Features

Rao, K.S.; K.E., M.

2017, XI, 92 p. 23 illus., 4 illus. in color., Softcover
ISBN: 978-3-319-49219-3