Continuous Speech Recognition Technologies—A Review

Shobha Bhatt, Anurag Jain, and Amita Dev

Abstract Speech recognition is the most emerging field of research, as speech is the natural way of communication. This paper presents the different technologies used for continuous speech recognition. The structure of speech recognition system with different stages is described. Different feature extraction techniques for developing speech recognition system have been studied with merits and demerits. Due to the vital role of language modeling in speech recognition, various aspects of language modeling in speech recognition were presented. Widely used classification techniques for developing speech recognition system were discussed. Importance of speech corpus during the speech recognition process was described. Speech recognition tools for analysis and development purpose were explored. Parameters of speech recognition system testing were discussed. Finally, a comparative study was listed for different technological aspects of speech recognition.

Keywords Speech recognition · Feature extraction · Continuous speech · Classification · Language model · HMM

1 Introduction

Speech recognition is the method of transforming spoken utterances into text. It has been researched for decades. Speech recognition can be applied for voice interfaces, voice-driven machines, speech-enabled browsers, and Internet-based services. For speech recognition, the speech signal is first transformed into parametric form, and acoustic models are generated from the extracted features. At the recognition

S. Bhatt (✉) · A. Jain
U.S.I.C.T, Guru Gobind Singh Indraprastha University, Sector-16, Dwarka, Delhi, India
e-mail: bhattsho@gmail.com

A. Jain
e-mail: anurag@ipu.ac.in

A. Dev
Indira Gandhi Delhi Technical University for Women, New Delhi, India
e-mail: amita_dev@hotmail.com

M. Singh and Y. Rafat (eds.), *Recent Developments in Acoustics*, Lecture Notes in Mechanical Engineering, https://doi.org/10.1007/978-981-15-5776-7_8

step, parametric observations are attained, which have higher probabilities matching corresponding to input speech. Automatic Speech Recognition (ASR) system recognizes the utterances from spoken words [1]. Speech recognition systems can be characterized into isolated word, connected word, continuous speech, spontaneous speech, and speaker adapted speech. The words are spoken in isolation for isolated word recognition, and words are spoken with some pauses in connected word recognition. Continuous speech recognition deals with the continuously spoken word. Spontaneous speech is a natural way; the humans speak.

Speaker-adaptive systems, the speech recognition system is developed independently, and then this model is used for distinctive characteristics of target speaker [2, 3]. Researchers have experimented with improving recognition with advancements in signal processing, acoustic modeling, language modeling, and search algorithms. State of the art has been shifted from isolated word to speaker adaptation, context-dependent modeling, and discriminative training [4] during the years. Progress in speech recognition was made possible due to better and accurate speech modeling techniques, different features, extraction methods, and efficient search algorithms [5, 6]. Speech recognition systems using Hidden Markov Model (HMMs), Artificial Neural Networks (ANNs), and hybrid of HMMs and ANNs have been experimented [7, 8].

Extracting useful information from speech signal is very significant phase toward speech recognition. It is required to transform the signal into low-dimensional space, which is called feature extraction [9]. Further, extracted features should be resilient to environmental changes. For taking out these features from the speech signal, perception- and production-based methods are widely used. Perception-based methods work on the principle of the human hearing system while production-based methods work on how speech is produced. Different perception-based scales such as Mel and Bark are used for processing due to the nonlinear perception of the human hearing system. Further, the selection of the windowing method and pre-emphasis are also important criteria for pre-processing of the speech [10–12]. Various methods were experimented using cepstral mean normalization, energy thresholding, Linear Discriminant Analysis (LDA), and application of filter banks to improve the basic features for speaker independence and speech recognition in different noisy conditions. Other improvements were carried out using the transformation of basic features by performing differentiation and concatenation on these static features. For this, first difference and second difference of static features are widely used [13, 14]. It is also a crucial decision to decide the sub-word unit for speech recognition. Different sub-word models, such as syllable based and phoneme based, are used [15]. Scholars in speech recognition face problems due to variability in speech [16, 17]. Classifiers are used in the speech recognition process during recognition. The classification methods such as HMMs, ANNs, and discriminative training are presented in literature [18–22]. The complexity of the ASR system is increased when it advances from isolated to spontaneous speech and speaker-dependent to the speaker-independent mode. Speaker independent system are more difficult to develop than speaker dependent system [25]. Another challenge is coarticulation effect in phone segments. The phone segments are affected by its neighboring phones [26]. Further, problems that

need addressing are lack of speech corpus, pronunciation dictionaries, and transcription files for under-resourced languages. Design of speech corpus is also crucial for the development of a speech recognition system [36, 37].

The purpose of writing this paper is to present a review of different technologies and trends in speech recognition. Readers will be able to know how the speech recognition system works, its challenges, and trends in speech recognition after going through this paper. It is an effort to highlight different aspects of speech recognition. The paper describes speech recognition issues such as feature extraction, classification methods, language modeling, speech corpus, widely used tools for the development of the spoken systems, and speech recognition parameters.

The paper is structured as follows. Section 2 describes continuous speech cognition and the structure of a speech recognition system. Section 3 explains various feature extraction methods. Section 4 describes speech corpus and widely used classification techniques. Section 5 deals with the role of language modeling in speech recognition. Section 6 describes different widely used speech recognition tools. Section 7 illustrates parameters for speech recognition testing. Finally, comparison of different classification and feature extraction techniques is presented.

2 Speech Recognition Structure

The continuously spoken words are converted into text during continuous speech recognition. Speech recognition system can be represented in Fig. 1 at a basic level. Feature extraction block converts speech signal into a suitable parametric feature. These features are used for generating acoustic models of speech utterances. The acoustic model is prepared from speech parameters. Language model block contains all the issues related to language modeling in speech recognition. The output from both the block is fed into a speech recognition engine. Speech recognition engine outputs the recognized word based on inputs from both the blocks.

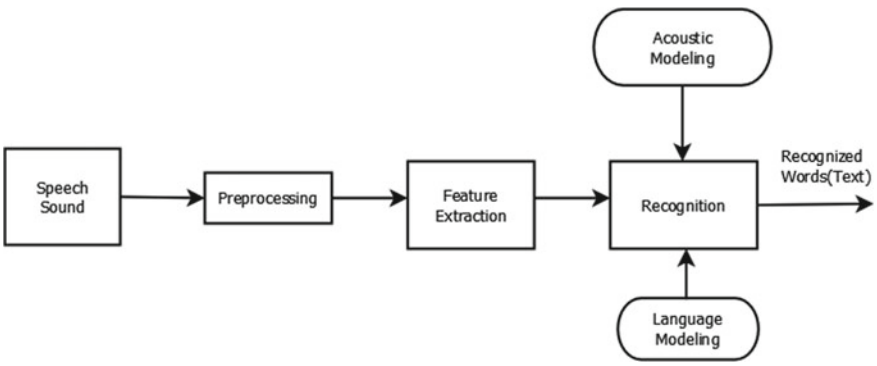

Fig. 1 Speech recognition process

3 Feature Extraction Techniques

Feature extraction is a crucial block in the speech recognition structure. Feature extraction is the process of calculating a set of meaningful properties from speech wave, which are significant for further processing. These properties are termed as features of the speech signal. Feature extraction approaches are divided into production based and perception based. Production-based features are calculated on the principle of how speech is produced, and perception-based features are based on how we perceive the sound. Linear Predictive Coding (LPC) is implemented using the concept of voice production mechanism while Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) perceptually motivated coefficients are used for feature extraction. Extracted features are further modified using different dynamic coefficients. Feature extraction methods are also categorized based on temporal analysis and spectral analysis [4, 38, 39].

For feature extraction, working of the vocal tract model and auditory model is captured. The main steps are pre-emphasis and windowing. The short-term spectrum is calculated using windowing methods such as Hamming. The spectrum can be represented as the signal's Fourier coefficients or as the set of power values at the outputs from a bank of filters. Further, the cepstrum is also calculated, which is the inverse Fourier transform of the logarithm of the spectrum. Different nonlinear logarithmic scales are used for the features based on the behavior of the auditory model [14, 36].

4 Speech Corpus and Classification Methods of Speech Recognition

Development of speech is essential corpus significant and crucial steps toward speech recognition are essential. For developing a speech recognition system, enough samples of speech are required. Different efforts have been made by researchers to develop standard speech corpuses such as from Linguistic Data Consortium (LDC) [27]. For low-resourced languages, classification is important to process in speech recognition. Here we have explained widely known models based on HMMS and artificial neural network-based model. The classification methods are used for the recognition process. Experiments have also been conducted to explore speech recognition using hybrid models with a combination of HMMs and ANNs.

4.1 Hidden Markov Model

HMMs represent the temporal nature of a speech signal. In HMMs, only observation can be seen, and the states behind observations cannot be directly observed.

Any HMM is represented by three features ($\prod$, A, B),

where $\prod$ is the initial state distribution vector,

A = [aij] is the state transition matrix which shows the probability of transition from state ai to state aj at a particular time, and

B = [bjk] is the matrix of observing symbols vk at current state qj.

The HMMs are used for observation sequence probability estimation, classification of utterances, and model parameter estimation by using forward–backward from Baum and Viterbi algorithm [5, 12, 28]. The most important phases using HMM-based speech recognition are training and testing. The advantages of using HMMs are that time and spectral variability can be modeled parallelly. Other advantages are HMMs can be initialized by using an initial estimate of the HMM parameters without labeling the data [4, 27–29, 35].

4.2 Artificial Neural Networks (ANNs)

In speech recognition, the artificial neural network has mostly used the method after HMM. These can be used independently or with the combination of HMMs. ANNs are very effective classifier. These classifiers learn and organize the data based on input data during the training phase. ANNs are capable of adapting when the data is unknown. Most popular ANN architectures are multilayer perceptron, self-organizing maps, and radial basis function [8, 31, 32].

5 Language Model

Knowledge of spoken language is production based for speech recognition. This includes knowledge of linguistic sounds, pronunciation variation, interpretation of the meaning of words, grammatical structure, and the possible meaning of sentences. By using language models, the search space for the recognition engine is reduced. Natural language processing is used for the implementation of these features in automatic speech recognition. Example of some of the steps is designing pronunciation dictionary and defining a grammar for the utterances [5, 7]. Nowadays, statistical language models are used that give probabilities of a sequence of words based on the previous set of words. Most widely used language models for speech recognition are N-gram language models, which predict the probabilities of words based on occurrences of its previous N-1 words [27].

6 Speech Recognition Tool Kit

There are different types of open-source speech recognition tools available. These systems have been developed using different programming languages like Java, C++, and C. These tool kits are based on HMMs, ANNs, and combination of both. The following section describes each in detail.

6.1 Htk

The HTK is an HMM-based tool for developing HMMs, especially for building the speech recognition systems. It was initially developed at Cambridge University by Speech Vision and Robotics Group. Main modules for developing the speech recognition systems are data preparation, training, and testing. The module HCopy is used for feature extraction with other tools. The module HCompv is used for flat start initialization. The module HeRest is used for re-estimation, and HVite is used for recognition. Speech recognition results can also be analyzed with tool HResult. The well-documented material is available as HTK book for developing different types of speech recognition system [3, 30].

6.2 Sphinx

Large vocabulary ASR systems were developed using Sphinx in the late 80s. Nowadays, versions of Sphinx which are available are 2, 3, 3.5, and 4 decoder versions and one common training tool which is named as Sphinx train. Sphinx can also be used to train context-dependent HMMs. It supports MFCC and PLP speech features with dynamic coefficients [5, 31, 33, 34].

6.3 Kaldi

Kaldi is a speech recognition tool kit available under Apache License. It is written in C ++. It was developed by Daniel Povey and others. Stable release was made in October 2013. Kaldi requires basic concepts of speech recognition and basic processing of speech signal. For better and efficient use of Kaldi, basic knowledge of scripting language like bash, Perl, and Python are also required. Kaldi has support for the deep neural network also [2].

7 Speech Recognition Evaluation

ASR systems are evaluated based on different parameters such as recognition accuracy, out of the vocabulary word rejection, response time, and error recovery [3]. For calculating speech recognition, accuracy reference string and recognized strings are used. There are three types of errors, namely, substitution, insertion, and deletion error [3]. [18, 36] is for speech recognition. Parameters are defined below:

$$\%\ \text{Correct} = (\text{N} - \text{D} - \text{S})/\ \text{N} \times 100 \tag{1}$$

$$\%\text{Accuracy} = (\text{N} - \text{D} - \text{S} - \text{I})/\ \text{N} \times 100 \tag{2}$$

where D denotes deletion error, I shows error due to insertion, S denotes the error caused by substitution, and N shows the entire number of labels in reference transcription. Another metric Word Error Rate (WER) is also used for describing the performance of the speech recognition system. It is specified as given below [35, 36]:

$$\%\text{WER} = (\text{D} + \text{S} + \text{I})/\ \text{N} \times 100 \tag{3}$$

8 Comparative Analysis of Classification and Feature Extraction Techniques

Every classification and feature extraction technique has its own merits and demerits. Table 1 [7, 33] shows a comparative analysis of HMM- and ANN-based classification techniques. HMM-based model best describes the temporal nature of the speech signal and ANN-based system is better in learning and needs fewer steps in comparison to other speech recognition techniques.

Table 2 [7, 33] shows a comparative analysis of different feature extraction techniques. LPC-based coefficients are easy to compute. The perception-based coefficients such as MFCCs and PLPs provide better recognition due to the use of different nonlinearity functions to mimic the auditory model.

9 Conclusion

This paper highlights technical details which are required during speech recognition. An attempt was made to review the speech recognition area during the past years. The focus was to review the different trends and techniques used in speech recognition. In the past year, several technical advancements occurred. Research findings show that

Table 1 Classification techniques [7, 33]

Classification technique	Advantages	Disadvantages
Hidden Markov model	It best describes the temporal nature of speech recognition	It is based on the probabilistic model. Present state probability depends on the previous state
	It is easy to model	
	It can work in both discrete and continuous modes	
Artificial neural network	It is self-learning and adaptive model	A large amount of training data is needed
	Many steps, like traditional speech recognition, are not required in ANNs	
	ANNs are adaptive to a new environment	
	Low-level features can also be used for speech recognition	

Table 2 Feature extraction techniques [7, 33]

Linear Predictive Coefficients (LPC)	Mel Frequency Cepstral Coefficients (MFCCs)	Perceptual Linear Prediction (PLP)
The speech signal model based on the human voice production system is well represented by LPCs Further, it provides linear characteristics and reasonable source vocal tract separation. LPC approaches are easy to understand and implement	MFCCs are perceptually motivated coefficients and offer good discrimination and a small correlation Further features of the gradually varying parts are focussed in the low cepstral coefficients MFCCs can be manipulated to generate different variants Individual features of MFCC appear just inadequately correlated, which serve for the development of a statistical-model-based system	PLP coefficients provide a better approximation of the speaker-independent system. PLP coefficients are generated in the short-term spectrum of the speech signal using perceptual scales PLP-dominant frequencies are quite insensitive to vocal tract length

significant work has been done toward different feature extraction, classification, and development of resources for research in speech recognition. Several standard speech corpora were designed to meet the requirements for the developments of the speech recognition system. It was also observed that different open-source tools based on HMM and ANN are also available for research work. This work also presented a comparative analysis of their specific features and classification techniques with their merits and demerits. Language models limit the search space in speech recognition, so the importance of language models has been explained. Finally, evaluation parameters for speech recognition have been discussed. However, despite a lot of research

work, there is a need for the development of robust speech recognition system, especially for under-resourced language to bridge the gap of the digital divide. Surely this research paper will help the research community to go deeper in speech recognition research.

Acknowledgments The authors would like to acknowledge the Ministry of Electronics and Information Technology (MeitY), Government of India, for providing financial assistance for this research work through "Visvesvaraya Ph.D. Scheme for Electronics and IT".

References

1. Sarma BD, Mahadeva Prasanna SR (2017) Acoustic–phonetic analysis for speech recognition: a review. IETE Tech Rev 1–23
2. kaldi-asr.org/doc/kaldi_for_dummies.html
3. Furui S (2007) Speech and speaker recognition evaluation. In: Dybkjær L, Hemsen H, Minker W (eds) Evaluation of text and speech systems. Text, speech and language technology, vol 37. Springer, Dordrecht
4. Saon George, Chien Jen-Tzung (2012) Large-vocabulary continuous speech recognition systems: a look at some recent advances. IEEE Signal Process Mag 29(6):18–33
5. Kacur J, Rozinaj G (2008) Practical issues of building robust HMM models using HTK and SPHINX systems, speech recognition, France Mihelic and Janez Zibert (ed), InTech. https://doi.org/10.5772/6376
6. Bahl LR et al (1999) Context dependent modeling of phones in continuous speech using decision trees. HLT
7. Cutajar M et al (2013) Comparative study of automatic speech recognition techniques. IET Signal Process 7(1):25–46
8. Lippmann Richard P (1989) Review of neural networks for speech recognition. Neural Comput 1(1):1–38
9. Vimala C, Radha V (2015) Isolated speech recognition system for Tamil language using statistical pattern matching and machine learning techniques. J Eng Sci Technol (JESTEC) 10(5):617–632
10. Picone Joseph W (1993) Signal modeling techniques in speech recognition. Proc IEEE 81(9):1215–1247
11. Fook CY et al (2013) Comparison of speech parameterization techniques for the classification of speech disfluencies. Turkish J Electric Eng Comput Sci 21(1):983–1994
12. Scharenborg OE, Bouwman AGG, Boves LWJ (2000) Connected digit recognition with class specific word models
13. Nieuwoudt C, Botha EC (1999) Connected digit recognition in Afrikaans using hidden Markov models
14. Bhiksha R, Singh R (2011) Design and implementation of speech recognition systems. Carniege Mellon School of Computer Science
15. Davel M, Martirosian O (2009) Pronunciation dictionary development in resource-scarce environments
16. Wu T (2009) Feature selection in speech and speaker recognition. Katholieke Universiteit Leuven
17. Kumar K, Kim C, Stern RM (2011) Delta-spectral cepstral coefficients for robust speech recognition. In: 2011 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE
18. Aggarwal RK, Dave M (2012) Integration of multiple acoustic and language models for improved Hindi speech recognition system. Int J Speech Technol 15(2):165–180

19. Bush M, Kopec G (1987) Network-based connected digit recognition. IEEE Trans Acoust Speech Signal Process 35(10):1401–1413
20. Singhal S, Dubey RK (2015) Automatic speech recognition for connected words using DTW/HMM for English/Hindi languages. In: 2015 Communication, control and intelligent systems (CCIS). IEEE
21. He ZG, Liu ZM (2012) Chinese connected word speech recognition based on derivative dynamic time warping. In: Advanced materials research, vol 542. Trans Tech Publications
22. Bernardis G, Bourlard H (1998) Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems. In: Fifth international conference on spoken language processing
23. Bourlard H, Morgan N (1998) Hybrid HMM/ANN systems for speech recognition: overview and new research directions. In: Adaptive processing of sequences and data structures. Springer, Berlin, pp 389–417
24. Livescu Karen, Fosler-Lussier Eric, Metze Florian (2012) Subword modeling for automatic speech recognition: past, present, and emerging approaches. IEEE Signal Process Mag 29(6):44–57
25. Renals S, McKelvie D, McInnes F (1991) A comparative study of continuous speech recognition using neural networks and hidden Markov models. In: 1991 International Conference on Acoustics, Speech, and Signal Processing. ICASSP-91. IEEE
26. Saini P, Kaur P, Dua M (2013) Hindi automatic speech recognition using htk. Int J Eng Trends Technol (IJETT), 4(6), 2223–2229 versité de Aix-en-Provence, 1998
27. Makhoul John, Schwartz Richard (1995) State of the art in continuous speech recognition. Proc Natl Acad Sci 92(22):9956–9963
28. Klatt Dennis H (1977) Review of the ARPA speech understanding project. J Acoust Soc Am 62(6):1345–1366
29. Jelinek Frederick (1976) Continuous speech recognition by statistical methods. Proc IEEE 64(4):532–556
30. 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
31. htk.eng.cam.ac.uk
32. Dev Amita S, Agrawal S, Roy Choudhury D (2003) Categorization of Hindi phonemes by neural networks. AI & SOCIETY 17(3–4):375–382
33. Anusuya MA, Katti SK (2011) Front end analysis of speech recognition: a review. Int J Speech Technol 14(2):99–145
34. https://cmusphinx.github.io/
35. Bhatt S, Dev A, Jain A Hindi speech vowel recognition using hidden Markov model. In: Proceedings of The 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages, pp 196–199
36. Bhatt, Shobha, Dev, Amita Jain, Anurag. Hidden Markov Model Based Speech Recognition-A Review. In: 12 th INDIACom 2018, 5th International conference on "computing for sustainable global development, 1st to 3rd March, 2018. http://bvicam.ac.in/news/INDIACom%202018%20Proceedings/Main/papers/712.pdf
37. Bhatt S, Jain A, Dev A (2017) Hindi Speech recognition: issues and challenges. In: 11th INDIACom 4rd International conference on computing for sustainable global Development. 1st to 3rd March, 2017. http://bvicam.ac.in/news/INDIACom%202017%20Proceedings/Main/papers/936.pdf
38. Agrawal SS, Prakash N, Jain A (2010) Transformation of emotion based on acoustic features of intonation patterns for Hindi speech. Afr J Math Comput Sci Res 3(10): 255–266
39. Madan A, Gupta D (2014) Speech feature extraction and classification: a comparative review. Int J Comput Appl 90(9)