Complex & Intelligent Systems (2022) 8:2623–2661
https://doi.org/10.1007/s40747-022-00665-1

SURVEY AND STATE OF THE ART

Computational intelligence in processing of speech acoustics: a survey

Amitoj Singh[1] · Navkiran Kaur[2] · Vinay Kukreja[3] · Virender Kadyan[4] · Munish Kumar[2]

Received: 14 April 2021 / Accepted: 22 January 2022 / Published online: 17 February 2022

Abstract
Speech recognition of a language is a key area in the field of pattern recognition. This paper presents a comprehensive survey on the speech recognition techniques for non-Indian and Indian languages, and compiled some of the computational models used for processing speech acoustics. An immense number of frameworks are available for speech processing and recognition for languages persisting around the globe. However, a limited number of automatic speech recognition systems are available for commercial use. The gap between the languages being spoken around the globe and the technical support available to these languages are very few. This paper examined major challenges for speech recognition for different languages. Analysis of the literature shows that lack of standard databases availability of minority languages hinder the research recognition research across the globe. When compared with non-Indian languages, the research on speech recognition of Indian languages (except Hindi) has not achieved the expected milestone yet. Combination of MFCC and DNN–HMM classifier is most commonly used system for developing ASR minority languages, whereas in some of the majority languages, researchers are using much advance algorithms of DNN. It has also been observed that the research in this field is quite thin and still more research needs to be carried out, particularly in the case of minority languages.

Keywords Speech recognition · Discriminative training · Deep neural network · Machine learning · Acoustic modeling

Introduction

The most continually evolving and explored the area in speech processing is Automatic Speech Recognition (ASR). Researchers have invested a lot to improve the performance of the real-time speech processing applications like recognizing digits, transcriptions broadcasting large vocabulary news, speech dictation systems, dialogue systems, etc. Apart from the advancements being made in the field throughout the world, ASR still poses an enormous number of challenges to the area and languages that are yet to be explored efficiently. A limited number of automatic speech recognition systems are available for commercial use, e.g., Apple Siri support 22 speech verities. This is exactly the number of languages officially recognized in India. The gap between the languages being spoken around the globe and the technical support available to these languages are very few. Some of the challenges include complication in designing systems for large vocabulary databases especially for minority languages where there is a scarcity of standard speech corpus, the articulator and phonetic features of the speaker in speaker-independent ASR, spontaneous patterns like ah, um, silence starts, out-of-vocabulary words, in the speech signal, and effect of external environmental factors like noise or distortion through the channel. These problems are prevalent in the real-time applications of ASR. Apart from all the interferences, the speech processing community has been successful to develop highly effective applications for dic-

✉ Amitoj Singh
amitoj.pb@gmail.com

Vinay Kukreja
vinay.kukreja@chitkara.edu.in

Virender Kadyan
vkadyan@ddn.upes.ac.in

Munish Kumar
munishcse@gmail.com

[1] School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala, Punjab, India

[2] Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

[3] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab Rajpura, India

[4] Speech and Language Research Centre, School of Computer Science, University of Petroleum and Energy Studies, Dehra Dun, Uttarakhand, India

tation which automatically generates written text from the input signal, database accesses, interfaces for human communication, machine control, and accessing automatic remote services over a dial-up connection for some majority languages.

Framework of ASR

The following block diagram depicts an automatic speech recognizer. The front-end processing includes pre-emphasis and extracting features. These feature extraction methods are explained in later sections of the paper. The speech sample is decoded at the backend with the help of knowledge gained from the acoustic model, language model, and pronunciation model. This transforms the input speech signal into a text string in a readable format. The grammar rules are fed into the language model. The front end corresponds to the training phase and the back end corresponds to the testing phase.

To represent real-world data, ASR is attributed to different dominant and impressive probabilistic modeling techniques including speech signals and documents of spoken language which are assembled from the applications in the real world. ASR problem-solving approach can be devised from the fundamental statistical classification approach, based on pattern recognition [73, 83]. From a given sequence of allowed words in the dictionary, classes are defined based on which the speech signal is represented parametrically. To find the sequence of words W that maximizes the factor Pr($W|X$) can be termed as the classification problem in ASR. Bayes' Theorem is used to factorize the latter factor [73] given as

$$\Pr(W|X) = \frac{\Pr(X|W)\Pr(W)}{\Pr(X)}. \quad (1)$$

For an acoustic input sequence X, the output factor Pr($W|X$) can be maximized by searching for a class W, through which the numerator on the right-hand side of Eq. (1), i.e., Pr($X|W$)Pr(W), can be maximized. The Language Model (LM) [257] which is represented by the factor Pr(W) is based on high-level coercions and the language-based information about the dataset of the words taken for a task. The acoustic model is given by the factor Pr($X|W$) which illustrates the sequence statistics of the parameterized acoustic studies in the given feature space provided that the uttered word phonemes are given. In the 90s, Hidden Markov Model (HMM) is proved to be the best possible acoustic model for the efficient modeling of sub-word including phonemes, syllables, etc. as well as complete speech sentences [257, 121]. The Markov Chain model also coined as the n-gram model is popularly used in language models for sentences and word classification in text documents. For the HMMs as well as n-gram models, learning different model parameters from a wide training data with the help of appropriate training criteria becomes imperative. Results have proved that the efficiency of the data-driven models is mostly based on the performance of the estimated models and the different modeling techniques adopted to train the data. Although HMMs have been proven to perform effectively for ASR acoustic modeling applications, including efficient pattern recognition, they suffer from major drawbacks. Although HMMs have contributed a lot to different fields of research, especially, in speech recognition, it suffers from major intrinsic limitations. Owing to this, researchers have decided to follow different approaches in ASR application building. A hybrid of Artificial Neural Networks (ANNs) with HMMs is jotted down [37] to conquer these drawbacks. Either existing Continuous Density Hidden Markov Models (CDHMMs) were trained with the help of forward–backward or Viterbi pseudocodes which presented reduced discriminative efficiency among several other techniques as they are trained on Machine Learning (ML) criteria which are discriminative. Also, the number of parameters in HMMs hinders their implementation in hardware. Consequently, to overcome such limitations, ANNs when trained discriminatively can perform non-parametric assessments among a sequence of patterns. ANNs use a specific number of constants, which makes it feasible for developing neural microchips. However, the hybrid of ANN–HMM systems has outperformed the earlier systems. Advancements in computing hardware over the past 2 decades have led to the acoustic learning of ANN systems much simpler. ANNs with numerous numbers of hidden units and training data require subtle and efficient hardware like General Programming Units (GPU) for massive computations. Typically, related feature input vectors based on wide temporal contexts of the acoustic frames are handles easily by ANNs as compared to the standard GMM-HMMs. Log mel-frequency spectral coefficients are used directly by the hybrid ANN–HMMs excluding a de-correlating discrete cosine transform (DCT) [217, 183] which were used earlier by Gaussian mixture model (GMMs). These elements proved the efficiency of the systems. Combining HMMs and ANNs have been started in the 90s and termed as *hybrid* HMM/ANN [81, 99, 182]. With advancement in computational power, advance form of ANN named DNN gained popularity. There are number of architectures that are useful in implementing the deep learning concept: Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), Auto Encoder (AE), and Convolutional Neural Network (CNN). RBM is extensively used in deep learning. They are used for generating stochastic models of ANN. RBM have variant of Boltzmann Machines (BMs). BMs are NN having stochastic processing units which are connected bidirectional, whereas DBN has many layers of RBM which used greedy layer algorithm for training. DBN is probabilistic generative model having stochastic, hidden units. Undirected and symmetric connections are present between the two top layers. The layer above

provides top–down connection to the lower layer. RBM is formed from, every two adjacent layers. Each RBM's visible layer is connected with previous RBM's hidden layer and above two layers are non-directional. The upper layer and the lower layer connection is directed and is in top–bottom manner. Various RBM layers present in DBN are sequentially trained CNN which is subpart of discriminative deep learning architecture. CNN has two types of various layers. One is convolution layers also called c-layers and another sub-sampling layers also known as s-layers. Alternately c-layers and s-layers are connected and help in the framework of middle part of the network. The inputs are convolved with trainable filter to produce feature maps in first c-layer. In every filter, a layer of connection weights is present. Additional feature maps in the first s-layer are produced, and this procedure continues and feature maps in the following c-layers and s-layers are obtained.

Approaches to automatic speech recognition

In the field of machine learning, different statistical learning approaches have been studied thoroughly. To build an effective classifier for pattern recognition applications, there exist two distinct learning approaches, namely, generative learning and discriminative learning. The generative learning technique uses the density estimation models to compute the probability of each class for the distribution of the data. However, inherit dependencies among the data can be easily exploited with generative learning methods using structural constraints. Apart, the major limitation for these models lies in the fact that such models require a true distribution of the data to build an optimal classifier. On the other side, discriminative learning schemes have recently gained immense popularity in artificial intelligence or machine learning applications. Since it does not employ modeling the underlying distribution of data and directly involves optimizing a mapping function from the input class to the output class labels, whereas discriminative models have been adopted widely to train different systems. Optimization of the mapping function can be achieved by implementing different training criteria. However, latent variables cannot be handled directly to reduce the underlying structure of the data in discriminative training. Furthermore, computations from simultaneous classes are carried out in parallel which enhances the computational complexity of the technique. Pure discriminative techniques are only used as a substitute component featuring neural networks for extraction of features in the front-end phase. Researchers have associated generative and discriminative learning techniques that were termed as discriminative learning of generative models. Ample research has been conducted in the past years to propose efficient algorithms, which can learn generative models in a discriminative manner for machine learning and ASR applications. These hybrid techniques include research which is being carried out since early 1980 where HMMs were trained discriminatively using various training methods like Maximum-Mutual Information Estimation (MMIE), CMLE/MMIE, Minimum Classification Error (MCE), Maximum Likelihood Estimation (MLE), Minimum Phone Error (MPE), etc. In this survey article, the author reviews the most relevant work in the literature including training techniques for recognizing speech models and especially, concerning discriminative training of various models used for building speech recognition systems. Apart from it also focused on the literature of feature vectors, which is one of the crucial elementals in any discriminative learning system. The author avoids any technical or experimental errors in detailing out the results. The ensuing article is organized into the following sections: In section "Challenges and issues", a study on the prevalent challenges for ASR systems has been presented. The consecutive unit, Section "Motivation", includes motivation areas for developing better ASR systems with improved recognition accuracy. Section "Quality assessment" deals with the quality assessment measures that have been taken by the authors. Section "Recognition of non-Indian ASR systems" presents the research work carried out by researchers around the globe for foreign languages. In section "Recognition of Indian ASR Systems", the author presented different ASR systems and techniques used in research for Indian languages. Section "Recognition results of non-Indian and Indian languages" compares the recognition accuracies and Word Error Rate (WERs) achieved by different research work by exploring the tables cited by the author. The tables are segregated into non-Indian, Indian languages and the speech research work is conducted on publicly available datasets. Section "Synthesis analysis" provides some suggestions for the researchers to carry forward the work in this field. The author concludes in sections "Suggestions on future directions" and "Conclusion" by analyzing different techniques and results.

Objectives of speech recognition applications

With the development and advancement in the digital signal processing hardware as well as software, the speech recognition system has achieved great strides. Still, the machine cannot match the performance of humans in terms of accuracy as well as in terms of speech. The suitability of each method is shown to the application. The following are the various objectives of the Speech Recognition application:

- To know Speech Recognition and the way it works.
- To view Speech Recognition applications in various areas.
- To know Speech Recognition implementation as a single application.

- To check how Speech Recognition is faster than writing by hand and checking hand-free capability of Speech Recognition.
- To make Speech Recognition useful for the mental and physical disabled person
- To used Speech Recognition for various applications like the voice dialing industry, dictation, and navigation.

Challenges and issues

An extensive study of the existing literature has been carried out to identify the various techniques and methodologies along with the existing challenges available for speech recognition that could trigger further research in the field. Speech has been classified into isolated or connected words, continuous or spontaneous speech. The basic modes of speech include speaker-dependent and speaker-independent. Each speech recognition system is attributed to several challenges. First, there is a wide variation of the speakers in uttering a word leading to the pronunciation difference. This variation is further attributed to the age, gender, and dialect of the physical appearance of the speaker. Secondly, background noise can add to the problems of recognizing speech accurately in different environments and real-time applications. The next challenge includes the physiological aspect of pronouncing the words by how stress is given on different syllables, phones, and vowels. This particularly affects the speech recognition of tonal languages. A continuous speech system is rather difficult to implement owing to the uninterrupted speech we use in real life. This further poses problems in speech recognition systems. Other factors contributing to the lower recognition rates include poor microphone quality, position, and direction of the microphone relative to the speaker. Despite these challenges, environment variation, channel variation, style of speaking, age, and gender contributes to the challenging task of speech recognition, e.g., Kirchhoff and Vergyri [158] mentioned that the Arabic language script falls short of the vowels as well as other information related to the phones. The major difficulties, which posed issues for speech recognition of the Russian language, are the variations in the informal speech of the language and the non-existence of standard dictionaries [123]. There is some speech corpus that has been collected by some researchers/research agencies in India, but this corpus has not been made available to the researchers for carrying out their research in the field. Thus, a lack of a standard database for Indian languages is a dearth. Shanthi Therese and Lingam [299] reported that for designing an efficient speech recognition system, selecting and extracting the most relevant parametric information are very crucial. Segmentation of words into corresponding phonemes in Indian languages is a tedious task to pursue. Owing to the linguistic variations in different languages of India, a single language may have many scripts and multiple languages may have one script. Furthermore, different people of different regions speak one common language in different ascents or tones. For instance, Punjabi, being a tonal language, is spoken differently in different parts of Punjab. Although a lot of work has been done in other languages, Punjabi, being one of the most popularly used languages across the globe, needs some attention from the researchers in terms of speech processing. Accuracy, noise removal, information retrieval, and varying bit rate are some of the most considerable parts of speech recognition challenges.

Motivation

Discriminative learning schemes have gained immense popularity in artificial intelligence or machine learning applications. A lot of research has already been carried out on foreign languages using different discriminative criteria for training purposes. However, Indian languages still need to employ discriminative training techniques for improved performance of ASR systems. Future study should be focused on determining whether dialect-ID systems are robust against speaker's variability, or whether systems incorporating prosodic information are required to provide further improvements. The development of a successful ASR engine depends upon a few factors. The selection of an appropriate feature may affect the training and testing phase of the system. These feature vectors depend upon various conditions. These discriminate feature needs were classified to an appropriate classifier. The selection of a classifier also plays a crucial role. Each classifier has its limitations and functionalities. Performance recognition of the testing phase is affected due to training through feature and their observation stores in a different classifier. The training of the system is focused mainly upon its corpus. The size and methods of speech corpus collection may drastically affect the system output. Also, the language models should be developed appropriately depending on the size and the nature of the data. Such factors significantly affect the accuracy of the system. The prevalent ASR systems for the majority of languages spoken across the globe have still not achieved the required efficiency [168]. Speech recognition techniques including Discriminative training of the language models, for Large Vocabulary Continuous Speech Recognition systems (LVSCR), need to be implemented more efficiently to improve the recognition accuracy.

Quality assessment

After the inclusion/exclusion criterion was employed to find out the relevant paper, the quality assessment was per-

formed on the remaining papers. The quality assessment form ("Appendix 1") includes all the papers indulge in the review, which contains high-quality speech recognition research. The questions included in Sect. 1 of "Appendix 1" served as a basis for screening the study. After the research paper was included, the study was done for the classification based on section "Inclusion/exclusion criteria". Then, we preceded to sections Recognition of non-Indian ASR systems and "Recognition of Indian ASR Systems".

Data extraction

During the starting of the review, we faced numerous problems. Extraction of all the relevant data ("Appendix 2") from many studies was very difficult. Due to this problem, it is needed to contact various researchers to find the needed details where we could not be inferred from the search paper:

- The steps carried in the data extraction are as follows.
- All the papers were surveyed first, and then, from the primary studies, data were extracted.
- Another researcher checked data extraction consistency by performing the data extraction on primary data. The samples were randomly selected, and the crossing checking of results was done.
- During cross-checking if any disagreement was noticed, then authors were used to resolving them in the consensus meeting.

Inclusion/exclusion criteria

A systematic review of literature is a method to identify, evaluate, and interpret the available literature in the form of research papers, articles, journals, etc., so that the studied literature can be summarized, research gaps can be identified and a base for carrying out future research can be formulated [162]. When the review procedure was being carried out, the authors focused on the following research questions:

- Question: What types of standard databases have been used to carry out the experiments?
- Question: What terminologies have been applied to formulate new databases? For example number of speakers included, environment variations, bias on gender utterances, etc.
- Question: What challenges have been identified for a technique for Indian as well as non-Indian languages?
- Question: What feature extraction techniques have been employed?
- Question: What results have been reported in the findings of the experiment?

The following datasets were mentioned in the studies:

Different other types of research questions could have been identified and involved in carrying out the review procedure, but the above-mentioned questions clearly defined the insights and the focus of the research which needs to be carried out in the field of ASR. From the immense amount of existing literature regarding speech recognition and its terminologies, primary research articles focusing on the central idea are filtered out with the help of database search using significant keywords. The authors then refine the selected papers manually. Irrelevant papers for the research were discarded manually based on the information in the title of the paper. The following facets are identified depending on the focus of the research and the research questions, which triggered the inclusion–exclusion principle for the research: speech recognition, feature extraction, large vocabulary continuous speech recognition, and ASR systems. The study followed a systematic approach to include quantitative and qualitative research articles, which have been published. This made the database search more comprehensive. Finally, the primary studies were included in the review that depends on the abstract and the full text. Our systematic approach and strict inclusion criteria likely reduced heterogeneity, but did not eliminate biases in the original studies, diversity in study design and population, and publication bias.

Recognition of non-Indian ASR systems

Jiang [131] summarized the discriminative learning training techniques of HMMs for automatic speech recognition. The author presented discriminative training criteria available in the literature as well as the optimization methodologies employed for discriminative learning. Gales et al. [87] review the different forms of discriminative learning models including maximum entropy Markov models, hidden conditional random fields, and conditional augmented models. The application of these models with respect to large vocabulary continuous speech recognition systems has been discussed. Hinton et al. [111] presented a survey to compare and summarize the existing methodologies used in different stages of speech recognition. The paper focuses on different feature extraction techniques including LPC, MFCC, etc., and several approaches to speech recognition. Apart, the evaluation of different ASR techniques has been demonstrated. Hemakumar and Punitha [110] provided a technological overview of the fundamental research carried out in the field of ASR over the past many years. The paper discussed different problems persisting in ASR and different methodologies developed so far for feature extraction, classification, models employed for a different database, and strategies used for performance evaluation. Johnson et al. [201] performed a systematic review of literature that referred to speech recognition (SR) in health care settings from 2000 to 2014. Six medical

databases are searched by a qualified health librarian. They conclude that SR, although not as accurate as human transcription, does deliver reduced turnaround times for reporting and cost-effective reporting, although equivocal evidence of improved workflow processes. Debatin et al. [64] reviewed offline voice recognition on Android mobile devices, they revealed that research priorities of offline SR are on reducing the error rate, developing neural networks for language models, and research advance statistical model. However, very few solutions have been offered for offline voice recognition on Android mobile devices. Clark et al. [56] analyzed that speech human–computer interaction work focuses on nine key topics: system speech production, modality comparison, user speech production, assistive technology and accessibility, design insight, and experiences with interactive voice response (IVR) systems. Nassif et al. [230] conducted a review of literature on speech recognition from 2006 to 2019. They concluded that most of the studies reported still used MFCCs as feature extraction for speech signals. MFCCs were profoundly used in traditional classifiers (HMM and GMM). 75% of DNN models were standalone models where only 25% of the models used hybrid models. This paper tried to explore state-of-art speech recognition with respect to different feature and modeling approaches in Indian and non-Indian languages across the world. Singh et al. [306] conducted a review of the spoken languages of India. The survey was conducted based on the relevant research articles published from 2000 to 2018. The purpose of this systematic survey is, to sum up, the best available research on automatic speech recognition of Indian languages. Kaur et al. [147] reviewed the status of speech recognition research conducted on tonal languages spoken around the globe. Authors observed that a lot of work has been done for the Asian continent tonal languages, i.e., Chinese, Thai, Vietnamese, Mandarin, but little work been reported for the Mizo, Bodo, Indo-European tonal languages like Punjabi, Latvian, Lithuanian as well for the African continental tonal languages, i.e., Hausa and Yoruba (Tables 1, 2).

Table 1 Speech corpus

Languages	Agencies/speech corpus
English	TIMIT, CTIMIT, NTIMIT, SWITCHBOARD, CHiME, OGI AlphaDigit, VOA, KEELE, IEMOCAP, DIRHA, ABI-1 corpus, AMI meeting corpus, Fisher, LibriSpeech
Japanese	CSJ
Multiple	NIST LRS 2005, 2006, 2007,DARPA, IARPA Babel
Mandarin	Hub4, DidiCallcenter, DidiReading
Russian	SPIIRAS
Indian	IIIT-H, Shruti, Cdac-Pune, Noida, BENG_YO, BENG_OLCRBL, CRBLP, IWAMSR, FIRE
Czech	WEB IT 5-g, SYN2006PUB 5-g, CRaTd
Dutch	Polyphone,Spoken Dutch, Folktale, JASMIN-CGN, Gesproken,
Finnish	News Wire, SPEECON, FinDialogue, FinnTreeBank-3, SAPU
French	Phone Book, MediaParl, French–English BTEC, ESTER
German	OGI Numbers95, Strange Corpus 10-SC10, Common Voice German
Arabic	Arabic Speech Corpus, King Saud University Arabic Speech Database, Tunisian_MSA, MediaSpeech

Table 2 Facets for inclusion/exclusion criteria

Facet	Relevant topics
Speech recognition	The system is developed for speech recognition; uses different approaches to speech recognition
Feature extraction	Studies including different types of feature extraction techniques or using hybrid feature extraction methods
Large vocabulary continuous speech recognition	Emphasizes speech recognition for LVCSR systems
ASR systems	Included results about the ASR systems so developed

English

English, an Indo-European language, has been accepted globally and spoken by people around the world for communication. The English language has many accents British, American, Indian, etc. A few standard databases have been formulated by the researchers to research speech analysis and recognition. Some of them have been made public also. Results of research work conducted on these databases have been discussed in Table 3. Richardson and Campbell [269] developed an SVM classifier for NIST 2007 language corpus. Benzeghiba et al. [29] presented a comprehensive analysis of different terminologies and methodologies used for automatic speech recognition. The review included factors like speaking accent, physiology of the speaker, age, emotions, and the style of speech. Also, different modeling techniques were presented. A comparative study of various model architecture-using hybrid HMM) was examined on TIMIT phone recognition. It was employed by global discriminative training methods that outrun slightly better than its baseline HMM approach. Variants of neural networks have been discussed that are used for automatic recognition of speech. Saon and Chien [279] introduce Bayesian sensing hidden Markov models (BS-HMMs) to perform Bayesian sensing and model regularization for heterogeneous training data. Results of BS-HMM on an LVCSR exhibited improvements over conventional HMMs based on Gaussian mixture models. Cai et al. [41] applied acoustic maxout neural net-

Table 3 Recognition results for non-Indian languages

Authors	Language/Corpus/data set	Feature extraction technique	Acoustic modelling	Accuracy/WER (%)
Ordelman et al. [237]	Dutch, 65 K words	–	RNN-HMM	WER 39
Nouza et al. [233]	Czech, 200 k	MFCC	HMM	OOV<3
Pui-Fung and Man-Hung [255]	Chinese, HKU93 corpus	MFCC	HMM	Tonal-syllable error rate 13.5
Byrne et al. [40]	Czech and English, 10,000 h	MFCC	HMM	WER 40
Deemagaran and Kawtrakul [65, 66]	Connected Digit Thai, 2000 utterances	MFCC, delta–delta MFCC	CDHMM	Acc. 70.33
Pylkkönen and Kurimo (2004)	Finnish, 12 h	–	HMM	WER 19.8
Eng and Ahmad [79]	Malay, 15 syllables	LPC	SOM and MLP	WER 21
Suebvisai et al. [317]	NECTEC, 6 h	Tone and pitch features	HMM	WER 16-19
Ververidis and Kotropoulos [347]	64 emotion speech data	Pitch, MFCC and Teager-Energy operator	Multichannel HMM	53%
Chien and Huang [50]	Mandarin, TCC300 (300 speakers)	MLLR, MAPLR, MCELR, CMLLR, and AAPLR	HMM	SER30.7
Afify et al. [5]	Iraqi Arabic, 90 K	MFCC + LDA + MLLT	HMM	WER improvement 13
Kirchhoff et al. [159]	LDC CallHome (CH),	FLM, GA	HMM, N-best list	WER 39.6
Hoffmeister et al. [114]	Mandarin,Hu b4 and TDT4, 440 h and 872 h	MFCC	CMLLR, MLLR	CER 17.7
Rajnoha and Pollak [262]	Czech, 63 k phonetically rich sentence, SPEECON set and TEMIC set,	MFCC, PLP, MFCC,RPLP	HMM	WER 19.4
Zeng et al. (2008)	LDC CSLU corpus, 50,191 utterances	MFCC	GMM-LM model	Acc. 78
Seman and Jusoff [294, 295]	Buletin Utama TV3 Broadcast News, 550 utterances	PVD	HMM	WER 14.57
Patil and Basu [241]	Marathi, Hindi, Urdu and Oriya 200 Marathi, 180 Hindi, 150 Oriya, 70 Urdu	MFCC	HMM	Acc. 49.75
Martincic-Ipsic et al. [203]	Croatia, VEPRAD 13 h, 9431 utterances	MFCC	HMM	WER 10.61
Wand et al. [350]	Chinese, 1642 h of BN and BC speech data of LDC	MPL, MPE MFCC, PLP	RWTH + SRIGD-GR-A	CER 7.7
Despres et al. [67]	Dutch, Conversational Telephone Speech (CTS)	PLP, MLP	CD-HMM, GMM	WER35.1
Kitaoka et al. [161]	Japanese, CENSREC-1-C	MFCC	HMM	Acc. 65.74
Mitankin et al. [214]	Bulgarian, 147 speakers	–	HMM	Acc. 92.63
Seman et al. [296]	Malaysian Parliament Session Data, 34,466 utterances	STE, ZCR and EEF	DHMM	Acc. 80.76
Kolar and Liu [164]	Czech: (CZ) broadcast conversation English: ICSI meeting corpus	–	HMM, MaxEnt, BoosTexter	Acc. 69.12

Table 3 continued

Authors	Language/Corpus/data set	Feature extraction technique	Acoustic modelling	Accuracy/WER (%)
Liu et al. [186]	Arabic, Read and spontaneous, 100 speakers,, 20 h	MFCC shifted delta cepstra	GMM, MLE	WER 15.26
Valente et al. [339]	Mandarin, 1600 h, GALE 2007	MFCC, PLP	–	CER 6-13
Nouza et al. [234]	Czech and Slovak, 350 K Czech, 170 K Slovak words	MFCC	HMM	WER SK 23.95 CZ 25
Huet et al. [122]	French, ESTER, 200,000-word	–	HMM	WER 22
Qian and Liu [256]	Chinese, DB863 Wall Street Journal (WSJ1) 171 h	MFCC	STA using MPE, fMPE	Chinese 97.13 English 95.37
Boril and Hansen [35]	Czech, CAR2E database Czech Lombard Speech Database (CLSD)	MFCC. PLP	HMM Codebook	WER 29.8
Burget et al. [38]	German, Spanish, English Callhome corpora, Total 50 h of corpus	MFCC	SGMM HMM	WER 59.8
Theera-Umpon et al. [325]	Thai, 2700 single syllables, 30 speakers	MFPLP and MFCC	HMM	Tone recognition 66.4
Vazhenina and Markov [341]	IPA phonetic alphabets, 16,350 utterances	MFCC	3-state HMM	Higher phoneme recognition
Procházka et al. [253]	Czech, Web1T5-gramcorpus(WEB1T),SYN2006PUB5-gramCorpus	MFCC	HMM	Acc. 71.32
Kuo et al. [176]	Phase 5 DARPA GALE Evaluation, 1500 h	DLM features	DLM t	WER 8.5
Gulic et al. [98]	CRO-AN corpus, 41 words	–	HMM	WER 7.74
Karpov et al. [142]	Russian, SPIIRAS	MFCC	HMM-GMM	5% relative improvement
Decker et al. [3]	English, German, French, 64 k words	–	CDHMM Gaussian mixtur	WER 56.6 (Pooled)
Shi et al. [301]	Dutch, 5050 speakers	MFCC	CDHMM	16% accuracy improvement
Saon and Chien [279]	English broadcast new, 50 h	PLP, LDA	VTLN, FMLLR, GMM, BS-HMM	BS-HMMs WER 17.6
Kawahara [148]	Japanese, 200 h in speech or 2.4 M words	–	LSV	CER10
Pan et al. [239]	LVSCR Chinese, 700 h, Hub01, Hub98	Logarithmic spectrumVTLN	CD-DNN	WER 22.8
Shi et al. [302]	Dutch, Spoken Corpus, 44,368 unique words	–	RNNLM + POS	WPA 23.11
Yang et al. [359]	Chinese, King-ASR-018 (corpus), 150 h, 850 speakers	Conditional random fields	HMM	Acc. 79.2
Kipyatkova et al. [156]	Russian, 1068 words	MFCC	CD-HMM	WER 33
Karpov et al. (2013)	GlobalPhone and SPIIRAS, 200 K words	MFCC	Bigram LM	WER 26.9
Liu and Sim [187]	Malay, 35 k + utterances (74.5 h)	MFCC, TVWR	HMM	WER13.1
Heigold et al. [109]	Romanian, 220 h	MFCC	DNN	WER11.7

Table 3 continued

Authors	Language/Corpus/data set	Feature extraction technique	Acoustic modelling	Accuracy/WER (%)
Enarvi and Kurimo [77]	FinDialogue and SPEECON, 44 min	–	VTLN and MLLR	WER 39.8
Seltzer et al. [293]	English, Aurora 4 of 7137 utterances	MFCC and FBANK	DNN–HMM	7.5% improvement over standard DNN
Tuske et al. [337]	German, French, 150 h for each language	MFCC,BN	HMM-GMM	WER GER 35.3 FN 45.5
Karpov et al. [143]	Russian, GlobalPhone, and SPIIRAS	MFCC	n-gram LM	WER 26.9
Ali et al. [11]	GALE data, 200 h	MFCC + LDA + MLLT	GMM-HMM	WER 26.95
Nouza et al. [233]	Czech and Slovak, 83,000 h	MFCC	GMM	20% improvement in WER
Kertkeidkachorn et al. [151]	Thai, CU-MFEC Corpus	PLP features	HMM	WER 23.85
Swietojanskiet al. [318]	English, AMI meeting corpus	40,dimensional Mel filter bank	GMM	9.7improvement in WER
Kapralovaet al. [141]	Russian, Google, 25 h	–	DNN	WER 25.1
Abdel-Hamid et al. [1]	English, TIMIT	MFCC	CNN	WER 6–10
Sainath et al. [273]	English, EARS	40-dimensional features	CNN using HMM/GMM	12–14% improvement in WER
Mateju et al. [205]	Czech, 550kspeech recordings	MFCC Filter Bank	DNN HMM	WER 14.66%
Radeck-Arneth et al. [258, 259]	German, 175 sentences (German Wikipedia), 567 utterances (German European Parliament)	MFCC	DNN	WER 20.5%
Botros et al. [36]	French Quaero Corpus	Multilingual features	LSTM RNN-LM	61.7% WER
Gonzalez-Dominguez et al. [95]	German, French, Finnish Japansese, Dutch, Romanian, Mandrian, Russian French Czech, Google 5 M LID, 150 k utterances by 34	–	DNN-LID	WER FR12.88 JP 20.13 DE15.58 RU 25.72 CN 18.14
Ljubesic et al. [191]	Croatian, hrWaC	MRR	SVM, RF	SVM 70.4 RF 59.8
Chaloupka et al. [43]	Czech, CRaT	MFCC	HMM and GMM	Acc. 68.30
Lopez-Moreno et al. [194]	English, NIST LRE 2009 and VOA	MFCC	Hybrid DNN/ I-Vector	C_{avg}45% improvement
Sailor and Patil [272]	English, TIMIT and AURORA 4	ConvRBM- CCConvRBM-BANK	HMM-DNN	WER 4.8–13.65 WER 1.25–3.85
Caranica et al. [42]	Romania, 100 audio clips, 112 spoken swords	MFCC PLP	HMM/GMM	CER 9.6
Medennikov and Prudnikov [207]	Russian, 390 h Russian spontaneous speech dataset	MFCC	SDBN-DNN BLSTM score fusion	WER 19.5 19.8 17.8
Razavi et al. [267]	French, PhoneBook, and MediaParl	PLP Cepstral features	KL-HMM	Acc. 74.2
Bérard et al. [30]	French, 9218 French Tokens	MFCC	SMT	WER 23–26
Chunwijitra et al. [54]	Thai, BEST LOTUS-BN and HIT-BTEC	MFCC	GMM, MPE RNNLM	WER Relative 1.54

Table 3 continued

Authors	Language/Corpus/data set	Feature extraction technique	Acoustic modelling	Accuracy/WER (%)
Bahdanau et al. [23]	English, Wall Street Journal	MFCC	RNN	WER 9.3
Huang et al. [121]	Mandarin, 863 project 110 h	LDA-MLLT	Bidirectional RNN	Acc. 88.2
Kipyatkova and KArpov [157]	Russian, 327 phonetically balanced phrases, 50 speakers 21 h	MFCC	RNN	WER 22.87
Wang et al. [351]	Thai, 800 h, 800 speakers	MFCC	HMM-DNN	WER 37.6
Watanabe et al. [355]	Dutch, Voxforge, 6,739 utterance	filterbank	CNN BLSTM	CER23.2
Hori et al. [115]	English, CHiME-3	MFCC,DOCC, MMeDuSA	DNN/RNN	WER 5.05
Smit et al. [310]	Finnish, Multiple, 150 h, 425 speaker	–	RNNLM TDNN	WER 22.79
Enarvi et al. [78]	Finnish, 85 h	–	n-gram models	WER 27.1
Renjith and Manju [105]	Tamil, Telgu, corpus: Amritaemo	LPCC, Hurst	K-NN, ANN	Acc. 75.7
Khokhlov et al. [154]	Russian, 100 h	FBANK, PLP,MFCC	Deep max-out networks (DMN)	WER 39.4
Spille et al. [311]	German, 18 h	LDA, MLLT, fMLLR	DNN	Recognition Threshold 1.9 db
Zou et al. [366]	HKUST corpus, 150-h speech	Log-Mel filter-banks	RNN-LM	4.8% relative improvement
Kaewprateep and Prom-on [136]	Thai, 100 h, 200 speakers LOTUSCELL 2.0 corpora	MFCC	CNN, LSTM	CNN 67.5
Chen et al. [49]	English, 300 h Switchboard corpus	36-dimensional filterbank	5 layer LSTM	WER 19.4%
Fukuda et al. [82]	Japanese Telephone conversations, 500 samples	MFCC, DCT	GMM-SVM	CER 13
Markovnikov et al. [200]	SPIIRAS, 30 h	MFCC, filter bank	DNN	WER 25.53
Seki et al. [291]	Dutch, 201 tokens	Mel filter bank and pitch features	CTC architecture	CER 32.2
Milde and Köhn [211]	German, Spoken Wikipedia Corpora	MFCC	TDNN-HMM	WER 20.04
Li et al. [184]	English, Supra CHLOE	Syllable based prosodic features	MD-DNN	Acc. 90.2
Russo et al. [271]	Croatia, 12 male speakers and contains 673 sentences in Croatian	MFCC Gammatone filterbank Gammatone HairCell	HMM	Acc. 93
Tong et al. [330, 331]	German, French, BREF and GlobalPhone corpora, German Broadcast News	–	ML-DNN-LHUC	WER FR7.3 GE 8.6
Iakushkin et al. [125]	Russian, 'yt-vad-1 k corpus, 1000 h	MFCC	LSTM DeepSpeech	WER 18
Ciobanu et al. [55]	Swiss-German, 14,647 training samples	Character and word features	SVM	62.03F1 score
Milde and Köhn [211]	German, 3,50,029 words	–	GMM-HMM-TDNN	WER 16.49
Tong et al. (2018a, b	French, English and German, 21 h	Log-mel filterbank features	CTC based	12% relative improvement
Scharenborg et al. [288]	Dutch, 64 h	FilterBank f	DNN	6.62% improvement

Table 3 continued

Authors	Language/Corpus/data set	Feature extraction technique	Acoustic modelling	Accuracy/WER (%)
Burileanu et al. [39]	Romanian, 3000–10,000 isolated words	MFCC	HMM	Acc. 52-70
Cucu et al. [58]	Romanian, 600 thousand words	–	ASR with LM	WER 30
Vergyri et al. [346]	LDC CallHome corpus of ECA, 120 speakers	52 MFCC	Morphology based LM	WER improvement 1.8

works to the Switchboard phone call data. Experiments were carried to minimize the effect of underfitting. The results reported that maxout networks converged faster than linear networks. Lopez-Moreno et al. [194] used DNN as an end-to-end LID classifier and extracted bottleneck features. Experiments were carried out in two separate outlines: the complete NIST Language Recognition Evaluation dataset 2009 (LRE'09) and Voice of America (VOA) data from LRE'09. DNN-based systems significantly outperform the *i*-vector system when dealing with short-duration utterance. Khademian and Homayounpour [152] developed a joint-token passing algorithm, and used deep neural networks for joint-speaker identification and their gain estimation. It achieved 5.3% absolute task performance improvement. Badino et al. [22] experiment with DNN–HMM phone recognition systems that use measured articulatory information. Evaluations on both the MOCHA-TIMIT mask and the mngu0 datasets show that the recovered AFs reduce phone error rate (PER) in both clean and noisy speech conditions. Moore et al. [223] evaluate the performance of the CHiME3 baseline ASR system in a diverse range of acoustic conditions using the ACE Challenge database of AIRs and noise. The evaluation exploits the recently released ACE. The benefit of speech enhancement processing has been demonstrated, with a reduction of WER up to 82%. Hanani et al. [101] worked on language identification of 14 regional accents of British English in the ABI-1 corpus. This system achieves a recognition accuracy of 89.6%, compared with 95.18% for the ACCDIST-based system. Sailor and Patil [272] projected an unsupervised learning model based on convolutional restricted Boltzmann machine (RBM) with rectified linear units. Experiments on the TIMIT and AURORA 4 databases show that ConvRBM can more general representations of the speech signals. Hori et al. [115] proposed a system with end-to-end recurrent neural networks (RNNs). The beamformed signal is processed by a single-channel long short-term memory (LSTM) enhancement network, which is used to extract stacked mel-frequency cepstral coefficients (MFCC) features. recurrent neural network-based were extending by applying beamforming, noise-robust feature extraction techniques, and large-scale LSTM RNN language models and achieved 5.05% WER for the real-test data. Maas et al. [195] found that increasing model size and depth are simple but effective ways to improve WER performance. Experiments suggest that the DNN architecture is quite competitive with specialized architectures such as DCNNs and DLUNNs. The DNN architecture outperformed other architecture variants in both frame classification and final system WER. Sainath et al. [273] introduced a joint CNN/DNN architecture to allow speaker-adapted features to be used, and authors investigated a strategy to make dropout effective after HF sequence training. Experiments on 3 LVCSR tasks, namely a 50 and 400 h BN task and a 300 h SWB task, indicate that a CNN with the proposed speaker-adapted and ReLU + dropout ideas allow for a 12%–14% relative improvement in WER over a strong DNN system. Seide et al. [290] present CN-DNN–HMM for speech recognition tasks. Discriminatively trained Gaussian-mixture HMMs on the Switchboard corpus reduced the word-error rate. Swietojanski et al. [318] investigate CNNs for large vocabulary distant speech recognition using the AMI meeting corpus and found that CNNs improve the WER by 6.5% relative compared to conventional DNN models. Similar results were also found by Li et al. [184], Abdel-Hamid et al. [1]. Weng et al. [356] introduced DNN to experiment on CHiME. DNN was trained using the MFCC features. The authors achieved significant improvement in WER as compared to the basic DNN system. Sainath et al. [273] applied CNN to LVCSR systems. Experiments were conducted on 50 h of data of English Broadcast News Speech Corpora. Advanced features were extracted to train the system. Zhang et al. [364] proposed a hybrid approach of CNN and CTC (Connectionist Temporal Classification) to overcome the limitations of CNN. 40-dimensional log Mel filter bank coefficients were extracted, and results were shown on the TIMIT database. Beck et al. [27] introduced Hidden Conditional Random Fields (CRF) for large vocabulary systems to overcome poor modeling features of HMM. Experiments were conducted on Switchboard 300 h speech corpus and the results were reported on a neural network that was trained on Gammatone features. Bahdanau et al. [23] developed LVSR systems with the help of RNN that performs sequence prediction directly at the character level. Two methods were

proposed to speed up searching operation, in the first scan to a subset of most promising frames was restricted, and in the second pooling, the information contained in neighboring frames helps in reducing source sequence length. Chen et al. [49] proposed a modular training framework of E2E ASR, while end-to-end decoding is retained. The results show that the proposed system performance gap between CI-phone CTC and the A2W model is reduced. Ravanelli et al. [266] performed experiments on TIMIT, DIRHA, CHiME, and LibriSpeech databases. Hybrid feature extraction techniques were employed using MFCC, fBANKS, and FMLLR to train the RNN–HMM system, and a significant improvement over standard RNN systems has been reported. [Seltzer et al. [293], Jing et al. [132]].

Mandrian

Liu et al. (2010 applied different existing discriminative training approaches like MPE and fMPE on bilingual speech corpus. The results showed that the STA phone clustering technique is better than the existing phone clustering criteria. The bilingual corpus combined Mandarin and English. Chien and Huang [50] gave a discriminative linear regression adaptation algorithm for HMM for speech recognition. Aggregate aposteriori linear regression (AAPLR) was proposed for discriminative adaptation when the classification errors of adaptation data have to be minimized. Hwang et al. [124] pointed out that since Mandarin is a tonal language, adding pitch information might help in speech recognition. Therefore, they added pitch information into the input of the Tandem neural nets. The system builds with confusion network combination yield 9.1% CER on the DARPA GALE 2007 dataset.

Hoffmeister et al. [114] developed the RWTHLVCSER system for Mandarin. The proposed system integrated additional feature streams such as tone and NN-based posteriors and the combination of multiple systems. Plahl et al. [246] again developed RWTH bases LVCSR system for Mandarin. A new reduced toneme set is developed. This helps in a reduction in character error rate by about 3% relative. Wang et al. [352] explore the use of multifactor clustering for training data and the use of MPE–MAP and fMPE–MAP acoustic model adaptations. A 6% relative reduction in recognition error rate compared to a Mandarin recognition system that does not use genre-specific acoustic models was achieved. Valente et al. [339] investigate all prevalent frontends for scalability. Results reveal that the MLP features produce relative improvements at the different steps of a multi-pass system. Yang et al. [359] proposed a hybrid technique for automatically generating Chinese abbreviations and perform vocabulary expansion using the output of the abbreviation model for voice search. An improvement from 16.9 to 79.2% was achieved by incorporating the top-10 abbreviation candidates into the vocabulary. Li et al. [185] conducted experiments on the Hub4 Chinese broadcast news database. A 3-g language model was trained for the experiments. Chen et al. [48] developed large-scale Mandarin speech corpora and studied its pronunciation patterns. The system was evaluated with the help of multi-speaker read-speech mode. Research has been carried out on the tonal aspect of the language using continuous speech and reported improvement in recognition rate (Lei et al. 2016). Huang et al. [121] conducted experiments on Mandarin speech recognition by extracting pitch-related features and training the DNN–HMM-based acoustic model. They proposed an Encoder-Classifier framework for modeling Mandarin tones using RNN. The resulted show that the proposed network improves tone classification accuracy. Zou et al. [366] presented the contrasting behavior of CTC and attention-based encoder–decoder models. The experiments were conducted on the DidiCallcenter dataset and DidiReading dataset.

Japanese

Large vocabulary corpus of Spontaneous Japanese Speech, telephone-based name recognition, and MIT JUPITER weather information continuous speech data was examined by McDermott et al. [206]. Minimum Classification Error (MCE) has been evaluated on the available corpus which outperformed the baseline methods in calculating the word error rate.

Shimizu et al. [303] performed experiments on Chinese—Japanese and Chinese–English datasets to achieve a recognition accuracy of 82–94%. Nakamura [229] used a neural network-based model on multiple languages that included English and Japanese. Kinoshita et al. [155] concentrate on dealing with the effect of late reverberations. The projected technique initially estimates late reverberations using long-term multi-step linear prediction, and afterward reduces the late reverberation effect by employing spectral subtraction. The proposed technique showed significant improvements in the performance of the ASR in real recordings under severe reverberant conditions. Ichikawa et al. [126] proposed dynamic features for different types of data that outperformed MFCC features. Hotta [117] worked on a database of 5240 words. The experiments were conducted on the HTK toolkit. FBank and MFCC features were extracted from the data. Kawahara [148] defines an automatic transcription system in the Japanese Parliament. The authors proposed a lightly supervised training scheme based on statistical language model transformation that fills the gap between faithful transcripts of spoken utterances and final texts for documentation and achieved character accuracy of nearly 90%. Moriya et al. [224] use covariance matrix adaptation evolution strategy (CMA-ES) with a multi-objective Pareto optimization to tune DNN–HMM-based large vocabulary speech recognition

systems. The experiments were performed on the Spontaneous Japanese corpus. The proposed technique optimizes systems for achieving high accuracy. Mufungulwa et al. [225] proposed an algorithm in speech modulation spectrum for Running Spectrum Analysis (RSA) and was applied to speech data. Accuracy in the noisy environment increases about 4% compared to current conventional methods. Fukuda et al. [82] focused on detecting breathing sounds in continuous speech of Japanese telephone conversations. The authors achieved high-rate accuracy with GMM- and SVM-based models.

Russian

Kipyatkova et al. [156] proposed a language model that integrated statistical and syntactical text analysis. HMMs were used for acoustic modeling and the phonemes were modeled using continuous HMM. A speech corpus consisted of 100 continuous utterances and 1068 words. Ronzhin et al. [270] presented a comprehensive survey of the Russian language, the applied methods, and models for speech recognition techniques in Russia and foreign countries. Karpov et al. [142] described an ASR for large vocabulary systems in the Russian language. A hybrid of knowledge-based and statistical approaches was being used to build an acoustic model. To develop the language model, a novel method combines the syntactical and statistical aspects of the text for training data. The results were computed on two distinct Russian databases. The proposed language model was evaluated on 204 thousand words vocabulary and proved to be efficient than the existing models. On similar grounds, a speech recognition system based on phonetic decoding technique was developed by Savchenko [287]. Vazhenina and Markov [341] focused on a technique to select the phonemes based on the hybrid of phonological and statistical analysis. The proposed technique when applied to the IPA Russian phonetic set with the reduced number of phonemes achieved better results.

Karpov et al. [143] build the acoustic model with the combination of knowledge-based and statistical approaches to create several different phoneme sets. The analysis was conducted with 204 thousand words vocabulary and the performance of standard statistical *n*-gram LMs and the language models created using our syntactico-statistical method were compared. The results confirmed that the proposed language modeling approach is reducing word recognition errors. Kapralova et al. [141] re-decode speech logged by production recognizer to improve the quality of ground truth transcripts used for training alignments. A fully unsupervised approach to the acoustic model was described that took advantage of a large amount of traffic of Google's speech recognition products. Yanzhou and Mianzhu [360] optimized the recognition algorithm by implementing different feature extraction techniques. Prudnikov et al. [254] and Smirnov et al. [309] proposed a system to detect keywords from LVCSR systems where the experiments were conducted on the CMU-Sphinx platform. Similarly, LVCSR for the Russian language has been developed [142], Tatarnikova et al. (2006). Medennikov and Prudnikov [207] developed a speech recognition system with DNN combined with deep Bidirectional Long Short-Term Memory. The proposed techniques achieve a WER of 16.4%. Potapova and Grigorieva [250] conducted the perceptual–auditory analysis at various levels of Russian and German speech utterances in a noisy environment.

Kipyatkova and Karpov [157] constructed neural network-based models with a different number of elements in the hidden layer and perform linear interpolation of neural network models with the baseline trigram language model. The authors reviled that the application of RNN-based LMs reduced the WER when compared to baseline systems. Khokhlov et al. [154] used different acoustic modeling techniques like i-vectors, multilingual speaker-dependent bottleneck features, and a combination of feedforward and recurrent neural networks for building ASR in Russian. The study reviled that fully connected DNNs with max-out activations outperformed TDNN and BLSTM mode. Markovnikov et al. [200] presented an end-to-end speech recognition system for recognizing the extra-large vocabulary of Russian speech. The researchers have applied CTC and attention-based encoder–decoder with DNN modeling. SPIIRAS dataset was used for training the data with a length of speech corpus of 30 h. KALDI and TensorFlow toolkits were used for conducting the experiments. Kaya and Karpov [149] proposed computationally efficient feature normalization strategies for the challenging task of cross-corpus acoustic emotion recognition. The use of suprasegmental features' normalization strategies shows enhancement in performance over benchmark normalization approaches. Iakushkin et al. [125] developed a Russian-language speech recognition system based on DeepSpeech. They analyzed the utility of TensorFlow technology for optimizing linear algebra computations in neural network training. The proposed system generates a WER of 18%.

Romanian

Burileanu et al. [39] discussed the speech recognition model of a dialogue system for the Romanian language. MFCC and PLP features were employed to train the language model. The medium-sized vocabulary system was tested for efficiency only on a limited number of Romanian words. The speech data used for training consisted of 54 h of speech uttered by 17 male and 12 female speakers for the test data. Chiopu and Oprea [51] employed the use of neural networks for a discriminative speech recognition system in the Romanian language. Mean Squared Error (MSE), Minimum

Classification Error (MCE), Maximum-Mutual Information (MMI), and Minimum Phone Error (MPE) discriminative training frameworks were used to improve the recognition accuracy of the system. The minimum error computed for MLP with MSE was 0.0889. Dumitru and Gavat [74] presented a comparative study on continuous speech recognition systems in the Romanian language. The database for training consisted of 3300 phrases. TheSpeeD (Speech and Dialogue) Research Association [91] came through a significant result in developing LVSCR in the Romanian language. Different configurations of acoustic and language models were used on the speech corpora gave by Speed in 2014. Employing DNN–HMM hybrid models further improved the WER. The system was accurately tested on live transcription in Romanian. Militaru et al. [212] developed a ProtoLOGOS ASR in Romanian speech recognition. The acoustic model was trained on the statistical features of HMM and the language model was a bi-gram model. Perceptual Linear Prediction (PLP) features were used for the first time in Romanian ASR. Militaru et al. [212] developed a Romanian language automatic speech recognition system 'ProtoLOGOS'-based Hidden Markov models and speech signals were modeled using Perceptual Linear Prediction (PLP). Heigold et al. [109] experimented on the cross- and multi-lingual network of 11 Romance languages using 10 k hours of speech corpus. The average relative gains over the monolingual baselines are 42%.

Cucu et al. [58] present the improvements authors have brought to the SpeeD automatic speech recognition system. They discussed a noise robustness approach for the ASR system and experimented that proposed acoustic features and feature-transforms improve the accuracy. Caranica et al. [42] created an automatic speech recognition system (ASR) for spoken Romanian connected digits. HMMs and a finite state grammar language model are used, to build and optimize a fully functional digit recognizer system in the Romanian language. Tufiș & Dan [334] in their paper talked about the latest development in the Romanian language. CoRoLa project has more than 152 h of pre-processed speech recordings. Georgescu et al. (2018) presented a GMM–UBM modeling technique on RoDigits corpus of Romanian language. A feature vector of MFCC and LPC features was used to model the system with GMM–UBM techniques (Fig. 1).

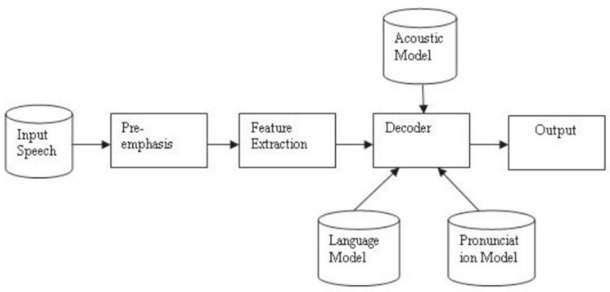

Fig. 1 Block diagram of ASR

Arabic

Vergyri et al. [346] evolved the use of models based on morphology for speech recognition at different stages in conversational Arabic. Language models, i.e., class models and single-stream factored models, were combined with the N-best list re-scoring framework. A large vocabulary recognition system was taken into consideration to evaluate the proposed techniques [276]. Satori et al. [286] presented an Arabic ASR, which worked on Open Source CMU-Sphinx-4. Hello_Arabic_Digit application employed the use of the proposed system. Hsiao et al. [118] suggested improvements on Generalized Discriminative Feature Extraction (GDFT) that was called regularized GDFT (rGDFT). The new system was evaluated on Iraqi and Arabic ASR tasks. MFCC features were extracted and combined with Linear Discriminant Analysis (LDA) frames. Ali et al. [11] prepared language resources to train and test the Arabic ASR. 200 h of GALE data was used in the system which is publicly available by LDC. The whole experiment was conducted on Kaldi and achieved good results. Baig et al. [24] applied discriminative training criteria, Maximum Likelihood (ML) on Holy Quran followed by MPE. Afify et al. [5] introduced an algorithm for simple word decomposition provided with a text corpus and an affix list. Lexicons were developed with the help of the proposed technique and a relative improvement in WER was observed. Kirchhoff and Vergyri [158] considered the cross-dialectal variations in Arabic, Modern Standard Arabic, and Egyptian Colloquial Arabic. The missing information was replaced with the help of morphological, contextual, and acoustic methods. Kuo et al. [176] investigated Discriminative Language Modeling (DLM) on large vocabulary Arabic broadcast ASR which were further employed for the use in Phase 5 DARPA GALE Evaluation. A typical detail of the minimum Bayes risk (MBR) method for DLS was given. Zarrouk et al. [363] worked on hybrid systems to identify isolated words from a large multi-dialect dataset of Arabic vocabulary. The authors achieved a higher accuracy rate using hybrid MLP/HMM. El-Amrani et al. [76] conducted experiments on simplified Arabic phonemes to develop a language model for the Holy Quran. WER of 1.5% was achieved by training a small dataset of audio files on the Sphinx tool. Speech corpus to carry out research in Arabic speech recognition has been developed and reported [12, 209, 279]. Telmem and Ghanou [323] presented a CMU-Sphinx system based on HMM for 11,220 audio files in the native language. Alsharhan and Ramsay [14] worked on the phonological aspects of the Arabic language. A set of language-dependent grapheme-to-allophone rules was formulated. In addition to

this, the stress features were extracted aiming to improve the acoustic modeling.

Malay

Seman and Jusoff [294] handled pronunciation variations in the Standard Malay (SM) system of speech recognition. The Standard Speech Malay corpus employed in the research included utterances from Buletin Utama TV3 Broadcast News which had around 550 words of 4 h length. PVD reduced the word error rate significantly. However, the proposed technique did not prove effective for handling phone changes. Fook et al. [80] cited a brief review on ASR technologies implemented on Malay Corpus. The authors have focused on the prevalent issues in speech recognition systems for noisy environments and the techniques to overcome these problems. Jamal and Shanta [130] developed an ASR system for curing patients of aphasic. Eng and Ahmad [79] proposed a hybrid technique of Self Organized Maps (SOM) and Multiplayer Perceptron (MLP) to recognize speech in Malay. LPC and a two-dimensional SOM feature map were used for feature extraction. Experiments were computed on 15 syllables of Malay that improved the recognition accuracy by 4%. Rosdi and Ainon (2008) developed a speech recognition system for isolated words in the Malay language based on HMM acoustic model. The research was carried out on five isolated phoneme word structures where 88% recognition accuracy was reported. Seman and Jusoff [295] proposed a system to segment and transcribe the spontaneous speech signal automatically instead of employing a manually annotated speech database. Evaluations on Standard Malay Television (TV3) news of local and non-local speakers were reported to be 42.53% and 30.8%, respectively. Seman et al. [296] proposed an endpoint detection technique to detect isolated words in Malay from Parliament Sessions in Malaysia. The algorithms combined Short-Term Energy (STE), Zero Crossing Rate (ZCR), frame-based Teager's Energy (FTE), and Energy Entropy feature (EEF). A Discrete-Hidden Markov Model (DHMM) classifier attained considerable results. HMM, models were also developed for evaluation of the speech of children suffering from language disorders such as stuttering (Tan et al. 2007). Sakti et al. [277] developed a system named ad A-STAR which is a network-based speech translation system of Asian languages.

Liu and Sim [187] investigate the proposed Temporally Varying Weight Regression (TVWR) method for cross-lingual speech recognition. Experiment using the Czech, Hungarian, and Russian posterior features conducted on Malay speech, TVWR was found to consistently outperform the tandem systems trained on the same features. Rahman et al. [260] worked on a small dataset of 390 sentences to recognize child speech in the Malay language. A speech recognition accuracy of 76% on the HTK toolkit was achieved. Apandi and Jamil [20] developed a speech corpus in the Malay language using different emotions such as happiness, anger, and sadness with the help of 30 speakers including children, young adults, and middle-aged adults. MSTAT (Malay Speech Therapy Assistance Tools) assisted the therapist to diagnose children with such disorders. Malay speech recognition systems for children's speech have also been developed by Ting et al. [329], Rahman et al. [260], Ting and Yunus [328] and Mustafa et al. [227]. Draman et al. [70] extracted voice samples from Telekom Malaysia's call center to develop an ASR for Malay speech. N-gram models were then implemented on the transcribed data. Maseri and Mamat [204] implemented a speech recognition system for Malay with the help of HMM training to recognize preschool children's speech. MFCC features were extracted for the available dataset.

Thai

Suebvisai et al. [317] constructed a Thai speech recognizer by applying the rapid bootstrap technique to build the acoustic model. Pronunciation variations for the words improved the accuracy instead of consonantal cluster phones. Schultz et al. [289] contributed to the research in the Speech-to-Speech translation system. The system was bootstrapped, and a translation component was built. The prototype was built in English for the doctor and in Thai for the patient which then translates the speech input. Deemagaran and Kawtrakul [65, 66] presented a speaker-independent Thai connected digit speech recognition system. To extract features, MFCC, delta MFCC, delta–delta MFCC, delta energy, and delta–delta energy techniques were implemented. Continuous density HMM was employed in the speech recognition procedure. Charoenpornsawat et al. [46] proposed a graphene-based speech recognition system for the Thai language. Tri-graphene with 500 acoustic models achieved better results. Theera-Umpon et al. [325] implemented a new technique for the classification of the tonal accent of the syllables. Speech samples from 30 speakers were collected. MFPLP and MFCC features were extracted from the samples. 66.4% accuracy on tonal accents was reported. Srisuwan et al. [313] focused on implementing surface electromyography (sEMG) for classifying Thai tonal speech. It was reported that the sEMGs were able to classify the tones into different categories better than other existing techniques. Thangthai et al. [324] worked on open-vocabulary Thai LVCSR by developing a hybrid language model to eliminate the out-of-vocabulary problem. BEST, LOTUS-BN, and HIT-BTEC were used to develop the language model. Hu et al. [119] incorporated tonal features such as F_0 and FFV to improve the recognition accuracy of Thai speech recognition. CNN was trained for acoustic modeling of the ASR system. Srijiranon and Eiamkanitchat [312] followed a Neuro-Fuzzy approach

to recognize Thai speech. PLP features were extracted for the speech samples and recognition accuracy of 20% for the test set was reported. Sertsi et al. [297] described the performance based on computational capability and recognition accuracy on the mobile device. The proposed offline system achieves a lower RTF by 24% compared online system on the mobile device. Wang et al. [351] introduced Automatic Speech Recognition (ASR) systems for Southeast Asian languages. The speech corpus collected for Thai and other languages was applied to build ASR systems. Deep learning techniques such as Bidirectional Long Short-Term Memory Networks and Time Delay Neural Networks may be used in acoustic modeling. Torres et al. [332] worked on the NIST 2016 SRE system which included Thai as one of the languages in the dataset. Tantibundhit et al. [321] developed a Thai speech recognition system based on different factors, namely, phonemic balance, familiarity, reliability, list equivalency, and homogeneity. The limitations of the system were also discussed. Speech recognition systems focused on the tonal aspects of the Thai language were also developed by Kertkeidkachorn et al. [151].

Chunwijitra et al. [54] proposed a syllable-based unit called pseudo-morpheme (PM) and a hybrid recurrent neural network language model (RNNLM) framework for Thai. The presented hybrid lexicon constituted open vocabulary for Thai LVCSR that reduced OOV rate around 1% just using 42% of the vocabulary size. The hybrid RNNLM obtained 1.54% relative WER reduction when compared with a conventional word-based RNNLM. Kaewprateep et al. [136] perform experiments on small-scale deep learning neural networks for Thai speech recognition. CNN and LSTM were built with a relatively small speech corpus. The result shows that CNN outperformed LSTM for small-scale deep learning. Tantisatirapong et al. [322] compare feature extraction techniques for accent-dependent Thai speech. Four frequency analysis methods were explored: Energy Spectral Density (ESD), Power Spectral Density (PSD), Mel-Frequency Cepstral Coefficients (MFCC), and Spectrogram (SPT). The corpus of isolated words from 60 speakers was recorded. Results reviled that the MFCC-based feature gives better accuracy than ESD, PSD, and SPT.

Croatian

Tadić and Fulgosi [319] developed a Croatian Language Lexicon (CML) from two different universities and included two sub-lexicons. Martincic-Ipsic et al. [203] developed context-dependent acoustic modeling using context-dependent triphone hidden Markov models and Croatian phonetic rules. The proposed system for Croatian acoustic modeling was developed as parts of speech interfaces in a spoken dialogue system for the weather forecast domain. Gulić et al. [98] collected letters and digits from two sources of data and developed a Sphinx model for the Croatian language. The results were listed on WER and SER. Nouza et al. [235] demonstrated the cost-effective approach for developing LVCSR systems for the Croatian language. The authors used 39 MFCC features on audio data of 320 h. The training data were collected from three different data sets. WER on the respective data sets was reported. Dunder (2014) worked on English and the Croatian language in a combined approach for the business domain. Dunder [191] worked on the Croatian corpus, hrWaC, including 2 billion words. SVM and Random Forest (RF) classifiers were used to experiment to achieve an accuracy of 70.4% and approximately 59.28%, respectively. Kacur and Rozinaj (2011) used HMM models for building large vocabulary recognition systems using the MASTER training scheme. MFCC and PLP techniques were used for feature extraction. The voicing feature was tested and surprisingly average improvement for PLP was 19.96% and 24.51% for MFCC. Nouza et al. [234] developed two corpora for Croatian and Slovene form the available corpus hrWaC and slWaC consisting of 59,212 tokens and 37,032 tokens, respectively, for Croatian and Slovene languages. An overall 10% improvement was observed on datasets of both languages. Agić and Ljubešić and Klubička [189] developed the first linguistically annotated data of Croatian language. The data were extracted from SETTIMES parallel corpus. [188] developed a Croatian training corpus, hr500k including 5,00,000 tokens. Russo et al. [271] proposed a speech recognition system based on cochlear behavior emulated by the filtering operations of the gammatone filterbank and subsequently by the Inner Hair cell (IHC) processing stage. Results reviled that proposed Gammatone Hair Cell (GHC) coefficients are lower for clean speech conditions but show a substantial increase in performance in noisy conditions.

Czech

Byrne et al. [40] worked on 10,000 h of annotated spontaneous speech to achieve a WER of 40% for English as well as Czech. MFCC features were extracted on the HTK toolkit. Nouza et al. [233] developed the first speech recognition system in Czech which can translate spoken broadcast programs into the language. A vocabulary of 200 k words was built and a language model of 300 M word text was extracted. Results on different types of broadcast programs were listed. Ircing et al. [128] worked using rich morphological tags on a class-based n-gram language model with many-to-many word-to-class mapping. This model improved recognition accuracy over the word-based baseline system. Boriland and Hassen [35] presented an unsupervised frequency domain and cepstral domain equalizations that increase ASR resistance to the Lombard effect. The proposed system provides an absolute word error rate (WER) reduction of 8.7%. Kolar and Liu [164] combine three statistical models—HMM, max-

imum entropy, and a boosting-based model BoosTexter. The result revealed that superior outcomes are achieved when all the three models are combined through posterior probability interpolation. Rajnoha and Pollak [262] build a speech recognition system with HMM that recognizes digits in a noisy environment. The results depicted that bark-frequency scaling, equal loudness pre-emphasis, and intensity-loudness power law in the MFCC brought improvement in noise robustness in the system. Nouza et al. [236] worked on Slavic languages: Czech and Slovak. Experiments were performed on 350 K words in Czech and 170 K words in Slovak. Distinctive features of these languages were also discussed. Procházka et al. [253] examined publicly available n-gram corpora for the creation of language models (LM) applicable. Results showed that the Web1T-based LMs, even after rigorous cleaning and normalization procedures, cannot compete with those made of smaller but more consistent corpora. Kombrink et al. [165] developed speech recognition systems for under-resourced languages; they applied machine translation to translate English transcripts of telephone speech into the Czech language to improve a Czech CTS speech recognition system. Přibil and Přibilová [252] investigate emotional types of spectral and prosodic features for Czech and Slovak emotional states of speech classification based on Gaussian mixture models (GMM). Experiments were conducted with four emotional states (joy, sadness, anger, and a neutral state). Testsexhibited the principal importance of correct classification of the speaker gender in the first level, which had a heavy influence on the resulting recognition score of the emotion classification. Nouza et al. [233] worked on transcribing 83,000 h of data of Czech and Czechoslovak Radio archives. The dictionaries involved 41 Czech and 48 Slovak phones. Baselines MFCC features were extracted and 32 mixture GMMs were equipped for modeling triphone state output pdfs. Chaloupka et al. [43] transcribed 80.000 h of Czech Radio audio archive of CRaT database. HMM- and GMM-based frameworks were trained for a different set of experiments. Mateju et al. [205] trained various DNNs with different training strategies inner structure and kinds of features. The resulting strategies for training of DNNs use the ReLU activation function, filter-bank-based features. Šturm and Volín [315] worked on testing the behavior of phonotactic structures of the Czech language. Šturm [316] worked on Czech syllables. 174 disyllabic Czech words were used for the experiment from 30 speakers.

Dutch

Ordelman et al. [237] developed a language model of 65 K words collected from Dutch newspapers. A hybrid of RNN and HMM systems was used for automatic speech recognition. Hämäläinen et al. [100] studied whether longer length acoustic units were better suited for modeling pronunciation variation and long-term temporal dependencies in speech than traditional phoneme-length units. A hierarchical method that was a mixture of word-, syllable-, and phoneme-length units was used and revealed that the presented approach did increase the word accuracy. Despres et al. [67] developed a system for Northern (NL) and Southern (VL) varieties of Dutch in the joint 'LIMSIVecsys' speech-to-text transcription systems for broadcast news (BN) and conversational telephone speech (CTS). Word error rates under 10% were obtained on BN development data. Pelemans et al. [245] explored three new application areas in ASR for Dutch conducted on toolkit SPRAAK. Scharenborget al. [288] trained an ASR system in Dutch and Mboshi languages. 40-dimensional Filterbank features were extracted on the Kaldi toolkit and trained using the multi-layer DNN. A relative improvement of 6.62% was reported. Pelemans et al. [244] experimented layered architecture approach to check whether it works for a large lexicon (400 k words) and language models (5-g), as well. The outcome shows that the architecture is already competitive and can be applied to acoustic models, language models, and lexicons. Shi et al. [301] used The Corpus Spoken Dutch and included 44,368 different words. 2% accuracy in predicting the words was reported. Cucchiarini and Van hamme [57] started the JASMIN-CGN corpus project to extend Corpus Gesproken Nederlands in three dimensions: age, mother tongue, and interaction mode. In total, 111 h and 40 min of speech were collected. Shi et al. [302] undertook rescoring experiments on a challenging corpus of spoken Dutch and investigate sentence and word length), to measure the performance of conventional language models. The results of experiments on CGN 32 data and WSJ data show that integrating sentence length and word length can achieve improvement. Watanabe et al. [355] introduce a model that recognizes 10 different languages, by directly performing grapheme (character/chunked-character). The proposed model is based on hybrid attention/connectionist temporal classification (CTC) architecture. Seki et al. [291] developed an ASR for multiple languages. 80-Dimensional Mel filter bank features were extracted in concatenation with the pitch features on the KALDI Toolkit.

Finnish

Pylkkönen and Kurimo (2004) worked on HMM-based phone duration modeling. To carry out the experiments, the speech recognition models used were speaker-dependent triphone models. 12 h of training data were collected from a Finnish book spoken by a female. 19.8% WER was cited. Siivola et al. [305] collected 300 million words from newspapers, books, and magazines from which 12 h of data were used for training the acoustic model. 56.4% WER on words was reported. Turunen and Kurimo [335] worked to improve

the performance of morph-based spoken document retrieval. Audio data of 288 h of spoken news in Finnish were collected. 26 h of speech data were used to train the acoustic model and 40.89% WER was reported. Kurimo and Turunen [179] worked on recovering speech recognition errors from spoken documents. The experiment was conducted on 270 news spoken stories uttered by a female speaker. Hirsimäki et al. [113] worked on large vocabulary language models of 40 million words collected from two different speech corpora. Significant reductions in error rate were listed. Enarvi and Kurio [77] worked on the available corpus SPEECON and FinDialogue. Finnish conversations from 13 distinct speakers, podcasts by 5 speakers, and recordings for 67 students were collected for transcription. 1.0% WER reduction was observed on the small data. An overall 55.6% WER was reported on the web data of the existing data sets.

Ginter et al. [93] developed a speech corpus named FinnTreeBank-3, from the existing resources. Parsing was done based on morphological tagging and dependency parsing. Enarvi et al. [78] worked on LVCSR systems for Finnish and Estonian speech recognition. The training data for Finnish included 85 h of speech data from three different sources. For Estonian, training data of 164 h from broadcasts, news, and lectures were used. Mansikkaniemiet al. [199] developed an ASR system for Finnish as well as Arabic. Three distinctive data sets for acoustic model training, i.e., Acoustic model training, were carried out on KALDI and language models on VariKN toolkit. RNNLMs were trained with the help of the TheanoLM toolkit. Behravan et al. [28] handle leveling in Finnish regional dialects using SAPU (Satakunta in Speech) corpus. Authors use attributes features like manner and place of articulation for leveling dialects. Experiments conducted with an i-vector system revealed that attribute features achieve higher dialect recognition accuracy and were less sensitive against age-related leveling. Varjokallio et al. [340] introduced a novel language model look-ahead technique using the class bi-gram model. This technique gave better results over the unigram look-ahead model.

French

Adda-Decker et al. [4] studied the radio interviews in French, and analyzed the syllabic structures and the corresponding variations. The authors also put a light on using ASR systems in French as a linguistic tool. The corpus used for experimentation includes 30 h of data with 254 k words. New terminologies to analyze W-syllables and S-syllables were introduced. Razavi et al. [267] related the graphemes and phonemes by training HMM. Experimental studies were confined to two databases: Phone Book and MediaParl Corpus. G2P approach was used for the pronunciation of the words. Dimulescu and Mareüil [69] worked to determine the origin of a speaker from an uttered speech sample by analyzing the phones. It was observed that the system worked well with Arabic speakers and poorly for speakers of Portuguese. Accents that were commonly mistaken were Spanish–Italian and English–German. Bérard et al. [30] used the French–English BTEC corpus to perform experiments on Machine Translation. The proposed system performed closely to the baseline systems. Kocabiyikoglu et al. [163] worked on the segmentation and augmentation of the existing speech corpus, LibriSpeech, and presented a large-scale corpus of 236 h. Venail et al. [344] developed software using word lists gathered from the speaker's lexicons and were then assessed based on manual as well as automatic scoring. Decker et al. [3] derived the text data for experimentation from the CHAMBER discussions and sampled. Three acoustic models for English, French, and German were built separately. 65-word lists were used for training. An OOV rate of 2.4% was reported on the CHAMBER text. Botros et al. [36] studied different clustering methods and compared their performance on the available data sets to improve the keyword search and WER further. Huet et al. [122] re-ordered the N-best lists to extract morpho-syntactic details in the post-processing of ASR speech signals. 2,00,000-word data set was extracted from the ESTER training corpus, while evaluation was carried out on 11,300 words. A WER of 22% and overall accuracy of 95% were reported. Imseng et al. [127] presented a bilingual database 'MediaParl' containing recordings in both French and German. Experiments were conducted using HMM/GMM systems. Multi-Layer Perceptron (MLP) and Rasta-based multi-lingual bottleneck features were examined for acoustic modeling in German and French languages [336, 337]. Results from experiments showed that multi-lingual BN features offered better cross-lingual portability [2], and Christodoulides et al. [53] presented a multi-level annotator 'DisMo' for spoken language corpora. It integrates part-of-speech tagging with basic disfluency detection and annotation and multi-word unit recognition. The proposed system was trained and tested on the 57 k-token corpus. Experiments revealed that 'DisMo' achieves a precision of 95–96.8%. Tong et al. [330, 331] examine multi-lingual (French, German) CTC training in context to adaptation and regularization techniques that were proved to be beneficial in more conventional contexts. Learning Hidden Unit Contribution (LHUC) was inspected to make language adaptive training. The performance of the universal phoneme-based CTC system was improved by applying dropout and LHUC.

German

Larson and Eickeler [180] demonstrated the use of syllable-based indexing features and how they outperform the word-based indexing features on large-vocabulary German-

language radio documentaries. Burget et al. [38] used a different approach to develop a multi-lingual speech recognition system, and they used entirely different phone sets, but the model had parameters not tied to specific states and are shared across languages. They use Subspace Gaussian Mixture Model for the experiments and obtained significant WER improvements with this approach. Siniscalchi et al. [307] developed a technique to design 'language-universal' acoustic models for phone recognition systems under the 'automatic speech attribute transcription' framework. Specifically, a phone recognizer that can decode languages with minimal available target-specific training data was built.

Weninger et al. [357] present a manually segmented and annotated speech corpus of over 160 h of German broadcast news and proposed an evaluation framework of LVCSR systems. The proposed framework achieved a word error rate of 9.2%. Radeck-Arneth et al. [258, 259] collected corpus in a controlled and clean environment for German distant speech. A total of 36 h of the corpus were recorded from 180 different speakers. Kaldi tool kit was used for the development of the ASR system and a WER of 20.5% was recorded for German distant speech recognition possible.

Gonzalez-Dominguez et al. [95] present end-to-end multilanguage ASR architecture deployed at Google. This helps in the selection of arbitrary combinations of spoken languages. Acoustic information was exploited by the DNN-based LID classifier. Ali [13] in their study aims to check conventional speech features to detect voice pathology, and if it can relate to voice quality. An automatic detection system based on MFCC was developed and tested on three different voice disorders. The accuracy of the MFCC-based system differs from database to database. The detection rate for the intra-database ranges from 72 to 95%. Ciobanu et al. [55] used SVM-based ensembles for Swiss and German data sets including distinct speakers. A 62.03% F1 score was reported. Milde et al. [211] performed experiments on three distinct open-source databases available for the German language. The experiment was conducted on the KALDI toolkit. Acoustic models like GMM–HMM and TDNN were used. 16.49% WER was cited by the authors. Spille et al. [311] tried to predict the speech recognition threshold (SRT) for normal-hearing people. For this, deep neural network (DNN)-based system was employed to convert the acoustic input into phoneme predictions, and ASR was trained on matched and multi-condition training. The best predictions are obtained for multi-condition training that employed amplitude modulation features. Milde and Kohn [211] train ASR system for German on Kaldi toolkit with two datasets. A total of 412 h of German read-speech data from Wikipedia corpus were taken and the system achieved a relative word error reduction of 26%.

Recognition of Indian ASR systems

Gaikwad et al. [86] have presented a state-of-the-art survey about the techniques available in speech recognition in Indian languages. He presented a survey of available speech recognition and feature extraction techniques. A speech recognition system was modeled in four working stages: Analysis, feature, extraction, modeling, and testing. The authors concluded to propose an interface for the Marathi language. Shanthi and Lingam [299] attempted to feature the advances made so far to extract features for speech recognition. Feature extraction techniques including Cepstral Analysis, Mel Cestrum Analysis, MFFCC, LDA, Fusion MFCC, LPC, perceptually based Linear Predictive Analysis (PLP) were discussed. For large vocabulary speech recognition systems, an Indian language database is established [172, 173]. Speech data from 560 speakers were collected for building ASR systems in Tamil, Telugu, and Marathi. Grapheme to phoneme conversion is performed which is followed by text selection. The acoustic models built on the language models were tested on Sphinx 2 tool kit. Hemakumar and Punitha [110] demonstrated contributions made by the researchers in developing ASRs for Indian languages. A broad view of the existing technologies and toolkits which are used for different process of recognizing speech is identified and presented chronologically. It was recognized that only a few Indian languages constituting Hindi, Marathi, Malayalam, Tamil, Telugu, and Bengali have well-developed ASRs, while other languages are yet under research. As compared to non-Indian languages, the research on speech recognition of Indian languages has not achieved that perfection yet. Therefore, research in the field of speech recognition of local Indian languages is still ongoing. In Indian languages, there is mainly an HMM-based model. Limited attempts have been made for the recognition of tones and discriminative analysis using deep learning models of speech.

Hindi

A good number of researchers have worked for the recognition of speech in the Hindi language. Wani et al. [354] developed a Hindi ASR. For feature extraction, MFCC was used, and the isolated words in Hindi were recognized using K-Nearest Neighbor (K-NN) and Gaussian Mixture Model (GMM). The speech corpus for training and testing was prepared by distinct male and female speakers in Hindi. The research proved to be satisfactory and useful for disabled and illiterate people for recognizing Hindi words. Sharma et al. [300] presented hybrid features that combine linear prediction and multi-resolution wavelet features. The classifier was based on linear discriminant function and HMM for both speaker-dependent and independent isolated Hindi speech data. The authors concluded that higher recogni-

tion accuracy was obtained using 3-level WBLPC features, while 4-level WBLPC features gave higher accuracy in the case of speaker-dependent. Kumar et al. [169] built a Hindi speech recognition system for connected words. A database of 102 words from 12 speakers was selected to develop the system on the Hidden Markov model toolkit (HTK). A comparative approach was used by the authors to manifest the improved results. An innovative technique to extract features using ensemble modules for speech recognition in Hindi was proposed [169]. The outputs of the ensemble classifier were collected with the help of the ROVER method of voting. The suggested system outperformed the baseline ASRs in Hindi. Aggarwal and Dave [7] combined the most efficient attributes of conventional, hybrid, and segmental HMMs using the ROVER technique, i.e., three distinct recognizers were developed and then merged with their unique set of features and classifiers. The WER was significantly reduced as compared to the traditional ASR systems. Aggarwal and Dave [8] proposed a method to integrate the existing feature extraction methodologies including MFCC, PLP, and gravity centroids to improve the performance of Hindi ASR systems. The results reported improved accuracy for medium-sized lexicons containing 600 words. Aggarwal and Dave [6] surveyed the existing reduction techniques and the areas of application. The techniques studied along with their advantages and disadvantages were PCA, LDA, HLDA, etc. All the techniques were implemented on Hindi speech data. Experiment results were cited down. A discriminative approach to train the HMM for continuous speech systems in Hindi was proposed by [72]. The feature extraction technique ensembles MFCC and PLP feature. For acoustic training of the model, MMIE and MME discriminative techniques were adopted. The proposed ensemble features with MPE gave better results than other feature extraction and discriminative techniques. A database of 100 speakers and 1000 sentences was used. Kumar et al. [170] developed an LVCSR in Hindi trained on 40 h of speech data of 120 speakers and 26,000 sentences. The language model which was trigram in nature was trained on 3 million words. Dua et al. [71] presented Differential Evolution (DE) technique for optimizing the filters in MFCC, GFCC, and BFCC. The performance of the proposed technique was evaluated both in a noise-free and noisy environment. A total of 100 speakers were used for the process where 80 speakers were used for training and 20 for testing. Upadhyaya et al. [338] had proposed a Context-Dependent Deep Neural network HMMs (CD-DNN–HMM) for large vocabulary Hindi speech using Kaldi automatic speech recognition toolkit. The experiments were performed on the AMUAV database. It demonstrates that CD-DNN–HMMs outperform the conventional CD-GMM–HMMs model and provide improvement in the WER of 3.1% over the conventional triphone model.

Bangla

Paul et al. [243] developed a Bangla speech recognition system where LPC was used for feature extraction and ANN for pattern recognition. It was cited that MLP with 5 layers was more robust than MLP with 3 layers. Muhammad et al. [226] proposed a novel ASR system to recognize digits in Bangla. The speech data were recorded with the help of the local people of Bangladesh and were medium in size. Feature extraction employed MFCC features and an HMM-based classifier to recognize the digits. It was also reported that due to dialect variations, the recognition accuracy of the system was affected. Mandal et al. [196] developed a speech corpus for Bengali. The proposed algorithm for text selection can be used for optimum text selection, whereas triphone or diphone can be used as the selection parameter. Hasnat et al. [104] reported the use of pattern classification by HMM model incorporated with the stochastic language model. Adaptive noise cancellation and endpoint detection were used to preprocess the signals. For every speech input signal, spectral feature vectors were extracted. The research was conducted on isolated as well as continuous speech sentences. Das et al. [61] developed a corpus for Bengali speech to recognize speech in speaker-independent continuous speech systems. The speech data were categorized into different classes depending on age and language. Phone and triphone-based speech data were prepared and incorporated with statistical modeling techniques. The collected data were implemented with 39 features on HTK. Banerjee et al. [25] compared the acoustic models based on triphone and monophonic. Methods were described to develop clusters of triphones with the help of decision-based tree methodology. 4000 recorded sentences were used for training and 600 distinct sentences from the same speakers were used for testing. The authors were concerned about developing speech systems front end which could be used for segmentation and clustering the continuous speech sentences in Bangla into the desired number of clusters [261]. Six different speakers recorded the speech data and the system was tested on 758 words from 120 sentences. Das et al. [62] performed experiments on Bengali corpus BENG_YO and BENG_OL using speaker adaptation techniques. Hossain et al. [116] implemented BPN (Back Propagation Network) for a speech recognition system in Bangla. Ten speakers were used to record ten digits in Bangla. MFCC features for 5 speakers' data were used for training and the other 5 for testing. The ASR worked well for speaker-dependent and speaker-independent systems. Bhowmik et al. [34] used CRBLP and C-DAC speech corpora for training DNN. MFCC features were extracted for the speech samples. Bhowmik and Mandal [33] applied DNN to Bengali continuous speech samples from C-DAC corpus. The recognition accuracy of 89.40% was reported on the TIMIT database. Reza et al. [268] devel-

oped a Bengali ASR based on isolated word datasets using HMM and Gaussian emissions with DNN. 96.67% accuracy was achieved using the HMM–GMM classifier. Pal et al. [238] developed an ASR in the Bengali language for handling queries regarding agricultural commodities. The experiments were conducted on the KALDI toolkit using the speech corpus of local people. Nahid et al. [228] have shown that a deep LSTM network can model Bengali speeches effectively. The context of phones is taken into consideration while modeling phoneme-based speech recognition. The authors have solely emphasized detecting individual Bengali words. They have achieved a word detection error rate of 13.2% and phoneme detection error rate of 28.7% on the Bangla-RealNumber audio dataset. Al Aminet al. [10] used DNN–HMM and GMM–HMM-based models, which have been implemented in the Kaldi toolkit, for continuous Bengali speech recognition benchmarking on a standard and publicly published corpus called SHRUTI. The study has been shown using Kaldi-based feature extraction recipes with DNN–HMM and GMM–HMM acoustic models have achieved performances WER 0.92% and WER 2.02%. Popli and Kumar [249] observed that training not only improves Query-by-Example Spoken Term Detection (QbE-STD) in the language of the same language family like Hindi but also other Indian languages like Tamil and Telugu.

Marathi

Kayte and Gawali [150] discussed and reviewed the existing terminologies for speech recognition in Marathi. Different approaches, methods, and tools, techniques, and applications of speech synthesis were demonstrated. The database used for the research was named IWAMSR which contained five speakers from diverse age groups, gender, and race out of which three were male and two females [153]. Three different databases of IWAMSR were constructed. Performance evaluation of the three databases resulted that database 3 of IWAMSR has improved recognition accuracy. Gaikwad et al. [84] studied the feature extraction and classification techniques in Marathi continuous speech recognition systems. The authors researched different feature extraction hybrid techniques. When evaluated on the accuracy, MFLDWT proved to the best hybrid model. Gaikwad et al. [85] proposed a hybrid feature extraction methodology that combined MFCC and LDA. The results of the proposed feature extraction method were cited and compared with traditional feature extraction techniques. A comparative analysis of the traditional feature extraction methods including MFCC, vector quantization (VQ), and LPC for Marathi was reported [208]. The authors detailed the Marathi database creation which contained 120 samples of Marathi vowels and 360 samples of Marathi conson ants recorded by one male and one female speaker. Gawali et al. [89] proposed the creation of a Marathi database containing 175 samples. The speakers were local Marathi people and aged 22–35. A speech recognition system based on the combination of MFCC and DTW feature extraction techniques was developed. Waghmare et al. [349] worked on emotional speech recognition exploring MFCC feature extraction techniques and classifying with the help of LDA. A Marathi database was constructed with samples of data obtained from 5 Marathi movies. The results were presented in 5 classes of Happy, Anger, Sad, Afraid, and Surprise. Darekar and Dhande [60] implemented a novel technique to recognize emotions using a hybrid PSO-FF algorithm in Marathi speech. Cepstral, NMF, and MFCC feature extraction techniques were used to extract features from Marathi and benchmark databases. The whole experiment was conducted on MATLAB. Patil et al. [242] has recorded a database of 5300 phonetically balanced Marathi sentences to train the context-dependent HMM. The subjective quality measures (MOS and PWP) show that the HMMs with seven hidden states can give an adequate quality of synthesized speech as compared to five states and with less time complexity than seven state HMMs. Yi et al. [362] propose a language-adversarial transfer learning technique to improve the performance of low-resource speech recognition tasks. Experiments were conducted on IARPA Babel datasets. The author proposed adversarial learning which was used to ensure that the shared layers of the SHL-Model would learn more language invariant features. Bhanja et al. [31] investigated the language discriminating ability of various acoustic features like pitch Chroma, mel-frequency Cepstral coefficients (MFCCs), and their combination. The system performance has been analyzed for features extracted using different analysis units, like, syllables and utterances.

Tamil

Tamil is a Dravidian language. Saraswathi and Geetha [280] proposed an enhanced morpheme-based language model for Tamil with limited vocabulary size. The text and speech corpora for Tamil were collected from newspapers and magazines, and articles for political issues from newspapers. The enhanced morpheme-based trigram language model with back-off smoothing technique performed better for the two corpora in Tamil. Plauche et al. [248] proposed an affordable approach for collecting linguistic resources from literate and illiterate workers of agriculture in three districts of Tamil Nadu and developed an ASR for the farmers. Chandrasekar and Ponnavaiko [44] presented a continuous speech recognition system in Tamil. An approach based on the segmentation of the speech signal followed by BPN for classification was deployed. 247 Tamil characters were tested in the system. Saraswathi and Geetha [281] implemented language models at various steps in speech recognition systems for Tamil, i.e., for segmentation, recognition, and error correc-

tion. The error rate was significantly improved at various phases of the phoneme, syllable, and word recognition for Tamil. Charles et al. [45] proposed an enhanced continuous speech recognition system in Tamil which was independent of the speaker and the device, namely *Alaigal*. The acoustic model is comprised of three basic steps: Feature Extraction, HMM modeling, and Estimation of parameters using Gaussian Estimation. Premkumar et al. [251] developed an LVCSR based on distinct factors, namely, pronunciation dictionary, language modeling, and front-end. 21.34% syllable error rate was reported. Chen et al. [47] presented approaches for keyword search (KWS) system using conversational Tamil provided by the IARPA Babel program. Strategies like optimization data selection through Gaussian component indexed N-grams, keyword aware language modeling, and Subword modeling of morphemes and homophones can help in tackling low-resource challenges. Manohar et al. [198] used DNN acoustic models to obtain a WER of 0.5%. The Fisher English Corpus was used to carry out the experiments. Sivaranjani and Bharathi [308] developed a speech recognition system in Tamil by extracting MFCC features and building the model using HMM. 95% accuracy has been reported. A system based on the triphone decision tree clustering method was developed by Lokesh et al. [192]. The authors conducted experiments on the FIRE dataset 2011 and implemented BRNN-SOM on the dataset to achieve an accuracy of 93.6%. Sarma et al. [284] explored the usage of the monolingual Deep Neural Network model to address the problem of speech recognition (LR) in the I-vector framework. Time Delay Deep Neural Network (TDDNN) architecture was used. Experiments showed that the proposed system gave low average cost performance as compared to the GMM–UBM-based system.

Telugu

Hegde et al. [106, 108] used a hybrid of Modified Group Delay Feature (MODGDF) and MFCC for computing the joint features of continuous speech recognition of Tamil and Telugu. The proposed features gave better results when compared to MFCC alone. Hegde et al. [107] modified the group delay function to extract cepstral features and were named modified group delay features (MODGDF). Results were evaluated on DBIL Tamil and Telugu, TIMIT, OGI_MLTS, and NTIMIT databases. The modified features outperformed the baseline features. Ramamohan and Dandapat [263] presented an emotional speech recognition system in Telugu and English. 30 native male speakers recorded the data for Telugu. It was reported that the sinusoidal features outperformed linear prediction and cepstral features for speech emotion recognition. Venkateswarlu et al. [345] used both MLP and TLRN models to train and test the speech recognition model. Results of Multilayer Perceptron (MLP) and Time Lagged Recurrent Neural Network (TLRN with the features of LPCC and MFCC are compared for better results. Renjith and Manju [105] build a system to recognize emotions in Tamil and Telugu languages. Linear Predictive Cepstral Coefficients (LPCC) and Hurst Parameter were extracted from the emotions. K- Nearest Neighbor (K-NN) and Artificial Neural Network (ANN) classifiers were used to identify the emotions. Hurst parameter gave more accuracy as compared to LPCC. In 2018, Vegesna et al. [342] again worked on emotion recognition. Telugu speech corpus of 64,464 utterances to extract MFCC and prosody features. A hybrid of GMM–HMM classifier was employed. Results showed that adapted emotive speech models have yielded better performance over the existing neutral speech models. Mannepalli et al. [197] prepared a speech corpus with the help of varying accents of local speakers of Telugu. MFCC feature extraction technique was followed by GMM classification. The accents considered were coastal Andhra, Telangana, and Rayalaseema.

Kannada

Kannada is one of the most widely spoken languages of Southern India with over 50 million people in India. Punitha and Hemakumar [110] designed Kannada continuous speech recognition for speaker-dependent mode. LPC coefficients were extracted as features. K-means clustering was used to categorize the classes. Anusuya and Katti [17] presented a review of distinct feature extraction techniques with or without Wavelet transform for Kannada speech recognition. For noise-free data, Discrete Wavelet Transforms (DWT) and for noisy data, Wavelet Packet Decomposition (WPD) was adopted for preprocessing the signal. For testing, 500 different samples were recorded by 10 female speakers. Harisha et al. [102] cited the existing ASR techniques for Indian languages with research efforts to develop isolated digit recognition. MFCC features were extracted followed by the ANN classifier with a back propagation algorithm for speech recognition. Five distinct speakers of age group between 20 and 35 recorded the speech database. Anusuya and Katti [19] used vector quantization to eliminate silence from the input speech sample. For training, 100 speech signals were used. Removing noise and silence from the signals reduces the WER. Cutajar et al. [59] reviewed different techniques and technologies of ASR systems. Antony et al. [16] developed an isolated word Kannada speech recognition system. A hybrid of DWT and PCA has been used. Sajjan and Vijaya [274] developed a speech recognition system using phoneme modeling, wherein each phoneme was characterized by tristate HMM where each state was represented by GMM. The performance was tested for monophone and word-internal triphone modeling. Pardeep and Rao [240] compared some baseline speech recognition sys-

tems like HMM–GMM, HMM–ANN, and HMM–DNN. Experiments showed that the HMM–DNN baseline system gave an improvement of about 7–8% than other baseline systems. Yadava and Jayanna [358] demonstrated a spoken query system that could be used to access the latest agricultural commodity prices and weather information in Kannada. MFCC was used to extract features from the voice samples. The 80 and 20% of validated speech data were used for system training and testing, respectively. Sajjan and Vijaya [275] presented a speech recognition system by employing decision tree-based clustering to build context-dependent triphone HMMs. It was observed that clustering of triphones using a universal list of articulatory questions performs well compared with manually created phonetic question lists. Geethashree and Ravi [90] developed an emotional Speech corpus. Different classifiers were used for building the system like Mean Opinion Score (MOS), K-NN (K-Nearest Neighbour), and LVQ (Learning Vector Quantization) classifiers. Kannadaguli and Bhat [140] evaluated the performance of Bayesian and HMM-based techniques to recognize emotions in Kannada speech. Kumar et al. [171] developed a speech recognition system in a noisy environment and speech corpus was collected from 2400 speakers. The acoustic models were built using the monophone, triphone1, triphone2, triphone3, subspace Gaussian mixture models (SGMM), DNN–HMM, and a combination of DNN and SGMM. DNN–HMM system gave the least WER.

Malayalam

Malayalam is the eighth most widely spoken language in India. Mohamed and Nair [215] developed a small vocabulary speaker-independent continuous speech recognition system in Malayalam. A CDHMM was used for modeling the phonemes and acoustic features were extracted with the use of MFCC features. Baum-Welch and Viterbi algorithms were used for training. Antony et al. [16] used 1080 words in the speech corpus for the experiment. The SVM classifier was used for speech recognition. Uncertainties in the lexical tones were identified. Kurian and Balakrishnan [178] proposed a speaker-independent system to recognize the digits. Acoustic features were extracted with the help of MFCC followed by HMM classifier for speech recognition. Krishnan et al. [168] employed the use of four different types of wavelets for extracting features. The classifier for pattern recognition was ANN. 160 speech samples were used to conduct the research. Mohamed and Nair [215] developed a speech recognition system of small corpus and the system was trained using continuous density HMM, used to model phonemes, where the observation probability density functions (pdfs) were continuous. The presented system produced a word accuracy of 94.67%. In the next year, Mohamed and Nair again proposed a context-dependent, small vocabulary, continuous speech recognition in Malayalam. HMMs were combined with ANNs. 108 sentences with 540 words and a total of 3060 phonemes were used for training. Kurian and Balakriahnan [177] used. HMM and GMM tied states to evaluate the ASR system. Bigram and trigram models were used for performance evaluation. Anand et al. [15] developed an LVCSR for visually impaired people. 30 h of speech data from 80 native speakers were used in the system. HMM was used for acoustic modeling. The pronunciation variations in the dictionary were handled by combining rule-based and statistical methods. Mohamed and Ramachandran Nair [216] explored the use of the pairwise neural network as an alternative to multi-class neural network systems for estimating emission probabilities of the states of HMM. Results showed that the pairwise system outperforms the multi-class recognition system. Thennattil and Mary [326] developed a system implementing phonetic engine (PE) for continuous speech. A speech corpus of more than 1 h was collected and transcribed using International Phonetic Alphabets. Phonemes were mapped to 40 frequently occurring phonemes and then modeled using continuous HMM. Mohamed and Lajish [218] proposed methods for recognizing vowel phonemes using nonlinear speech parameters like maximal Lyapunov exponent and Phase Space Anti-diagonal Point Distribution (PSAPD). Phase Space Ant-diagonal Point Distribution when combined with MFCCC gave better accuracy of 80.44%.

Oriya

Very few researchers have contributed to developing ASR systems in Oriya. Mohanty and Swain [219] proposed an Oriya isolated speech recognition system for visually impaired students to appear in examinations. The system could translate isolated answers into isolated text. For training, a set of 1800 isolated Oriya words spoken by 30 speakers were used followed by HMM-based recognition. Mohanty and Swain [221] proposed a method for emotion recognition classified as anger, sadness, surprise, astonished, fear, happiness, and neutral. Fuzzy k-means clustering was adopted on the collected speech data from 35 native speakers of age between 22 and 58 years. Mohanty and Bhattacharya [220] proposed wavelet neural networks for speech recognition in Oriya. Two speakers recorded the speech data followed by noise cancellation. ANN was used for recognition. Patil and Basu [241] collected corpora for various tasks in the text-independent speaker identification. Experiments confirmed that MFCC performed better than LPC and LPCC. Londhe and Kshirsagar [193] developed a new speech corpus of 100 different isolated words including 67 sentences from 478 different speakers. The dataset was collected in the native region and included words from English as well as Chattisgarhi

language including scripts from literature and newspaper articles.

Kumar et al. [174] developed Prosody and phonetically Rich Transcribed speech corpus for Bengali and Oriya languages. Ten hours of reading speech, 5 h of conversation speech, and 5 h of extempore speech have been collected. International Phonetic Alphabet (IPA) was used to transcribe the speech corpus. Mohanty and Swain [222] presented aHMM-based Speech input–output System for the Indian farming community to retrieve the cultivation information like availability of seeds and fertilizers. The word accuracy was found to be 75.13%. Dash et al. [63] developed a system for four Indian languages: Hindi, Marathi, Bengali, and Oriya by integrating articulatory information into acoustic features. Both speaker-dependent and -independent recognition experiments were conducted using GMM–HMM, DNN–HMM, and LSTM–HMM. DNN–HMM system outperformed the other models.

Assamese

Sarma et al. [283] developed a speech corpus generation technique with the help of an LMS filter and LPC cepstrum. ANN was used for recognizing. Sarma and Sarma [285] adopted LPC and PCA features for handling the mood and gender diversity of the native Assamese speakers. ANN was modeled using a hybrid of SOM, LVQ, and MLP. Dutta and Sarma [75] proposed a hybrid of LPC and MFCC features for recognizing Assamese speech. Recurrent Neural Network (RNN) framework was deployed for speech recognition. Kalita [137] reported the study of six Bodo vowels and eight Assamese vowels. Ten male and female speakers of each language recorded speech samples. LPC and MFCC techniques were employed to extract features. Kandali et al. [138] worked on emotion speech recognition in Assamese using MFCC features and a GMM classifier. 14 male and 13 female speech data were used for training. Kandali et al. [139] extracted features based on WPCC2 (Wavelet Pocket Cepstral Coefficients computed by method 2), MFCC, tfWPCC2 (Teager Energy operated in Transform Domain), and tfMFCC. The classifier was based on GMM. MESDNEI (MultiLingual Emotional Speech Database of North East India) was developed for research. Shahnawazuddin et al. [298] developed a spoken query system in the Assamese language for the farmers in the agriculture commodity. Data were collected from the local farmers. MFCC features were extracted for the collected data and trained on HTK toolkit using HMM/GMM. Research work on the Assamese language has been combined with other languages as well which was reported by Tuske et al. [337] and HartMann et al. [103]. Misra et al. [213] classified vowels with the help of GMM and compared the result with ANN. ANN achieved a 4% higher accuracy rate than GMM. Dey et al. [68] developed an ASR for agricultural commodities using SGMM- and DNN-based acoustic models. A relative improvement of 32% in WER was observed as compared to the GMM–HMM ASR system. Ismail and Singh [129] worked on Kamrupi and Goalparia dialect identification of the language by extracting MFCC features from the speech samples. Sarma et al. [282] developed a speech recognition system using a small training data set of 20 h in three different modes. 78.05% accuracy was achieved using the HMM-HTK toolkit. Bharali and Kalita [32] researched isolated words spoken by 15 speakers. Variants of MFCC features were used to train HMM, VQ, and I-vector models. Yi et al. [361] showed an adversarial end-to-end acoustic model for low-resource languages. The attention-based adversarial end-to-end language identification carries enough language information. Experiments were conducted on IARPA Babel datasets. The end-to-end model was trained with a low CTC loss function. The experiment proposed has gained a 9.7% relative word error rate reduction.

Punjabi

Kumar and Singh [175] developed a speaker-independent autonomous speech recognition system in Punjabi. The corpus included 1433 sentences and the authors reported an accuracy of 90.8% using MFCC features. GUI in Java Programming was built for the language model. Kaur and Singh [145] presented an effective speech model based on PNCC features for continuous Punjabi speech recognition. 34 phones for training 158 words were used. HMM, the model was used for training the system. An accuracy of 71.92% in a noisy environment has been observed. Kumar and Singh [169] built an isolated speech recognition system in Punjabi using LPC features. VQ and DTW have been used for recognizing the speech. 94% accuracy has been reported. Kaur and Singh [146] compared three feature extraction techniques, such as PNCC, PLP, and MFCC. The system was trained using HMM. 34 phones for the Punjabi language were used to break each word into small sound frames. MFCC gave the best results in a noise-free environment with 86.05% accuracy over PLP and PNCC. Guglani and Mishra [97] extracted PLP as well as MFCC features for continuous speech samples of Punjabi. The experiment was conducted on the KALDI toolkit. The performance of triphone and monophone models was compared. Results showed that MFCC features improved the performance of the ASR system and the triphone model performed better than the monophone model. Kadyan et al. [134] tested three different combinations at speech feature vector generation and two-hybrid classifiers. MFCC, RASTA-PLP, and PLP were randomly combined with GA + HMM and DE + HMM techniques to produce refine model parameters. Results from experiments showed that MFCC and DE + HMM technique improved the accuracy when compared

with RASTA-PLP and PLP using hybrid HMM classifiers. Kadyan et al. [135] used DNN against GMM. Baseline MFCC and GFCC methods were integrated with cepstral mean and variance normalization for feature extraction. Hybrid classifiers: GMM–HMM, and DNN–HMM were used to obtain performance improvement. Kadan et al. [353] worked on the role of prosody-modification-based out-of-domain data augmentation on children speech speech corpus. In addition to these, they also studied the effect of varying the number of senones, the number of hidden nodes, and hidden layers as well as early stopping resulting in 32.1% of Relative Improvement (RI) in comparison to the baseline system with varied senones.

Zhang Goyal et al. [365] tackle with an issue of dialect classification on the basis of tonal aspects of laryngeal phoneme [h] on four major dialects of Indian Punjabi language with two key parameters, namely F0 variation, and acoustic space, which are calculated using two formant frequencies: F1, and F2. Further work was extended through processing of acoustic information at feature level by comparing the performance analysis using basic or hybrid Linear Predictive Cepstral Coefficients feature extraction methods. The result showed that the hybrid LPCC + F0 system achieved a Relative Improvement (R.I.) of 6.94% on Subspace Gaussian Mixture Model in comparison to that of basic LPCC approach, respectively.

Recognition results of non-Indian and Indian languages

This section presents a brief report on the WER and accuracies achieved by researchers so far for speech recognition in different languages. A review of the prevalent feature extraction techniques is represented in Table 3. Spoken language processing system such as automatic speech recognition depends mainly upon a speech corpus. Only a few languages in India currently enjoy the advantages of language technologies such as successful speech recognition engines like Hindi and English. ASR in Indian languages is still at its early stage of research and getting more attention nowadays. Slightly, a few numbers of languages to date can assemble speech corpus in their native resource.

Building speech corpora for an Indian language is a difficult task. Statistical approaches used for modeling an ASR engine depend upon a large amount of training corpus that helps in the recognition of uttered signals. Thus, it is mandatory to have a database that comprises all characteristics of the typical user who spoke data in a realistic environment. A speech dataset is of two types: a dataset collected in a particular task domain and a general-purpose dataset. Previously [52], a Hindi language corpus was built with 50,000 Devnagari scripts to develop an ASR system. The speech corpus of different language ideas was designed and evaluated for Marathi by TIFR and IIT Bombay [94], and Hindi language travel domain dataset by C-DAC Noida [21]. And Telugu language dataset for the Mandi information system by IIIT Hyderabad, English, Hindi, and Telugu dataset for travel and emergency services was collected by IIT Hyderabad. Another general-purpose corpus of Telugu, Hindi, Tamil, and Kannada was prepared by IIT Kharagpur [265]. Hindi, Indian English corpus was developed by KIIT, Bhubaneswar and supported by Nokia research center China [304]. These corpora have been studied to analyze the attempt used for the development of an ASR system in Indian languages. Three language optimal databases were constructed for Tamil, Telugu, and Marathi languages that catalyze research activities. Two methods were adopted to collect the corpus for a landline and mobile dataset [172–173] tested on the Sphinx-II toolkit. The corpus was collected in the coordination of IIT Hyderabad and HP Labs, India. The Indian English and Hindi language corpus was collected [9] to involve the versatility by mobile communication environment of 100 speakers. The constructed database was employed in mobile speech recognition services. A Punjabi speech corpus of Malwai dialect was collected from 50 speakers that help in the development of the Punjabi Speech synthesis system. The recorded dataset was labeled phonemically to obtain phonemics and its sub-phonemic information [26]. Issues in the construction of speech corpus were observed and studied [160] for Indian languages. MFCC and LPCC are the two most useful and promising techniques for extracting features of the speech samples. MFCC has been used by many numbers of researchers for the feature extraction stage (Dua et al. 2006). However, Cutajar et al. [59] reported that MFCC signals were not robust to the noisy environment. Furthermore, MFCC technique believes that a frame of a speech contains the information of a single phoneme at a time. However, this may not be true for continuous speech recognition systems where the frame may contain information of two or more phonemes.

Another significant drawback includes that the features in MFCC are extracted from the power spectra of the speech sample, thus, not including the phase spectra. Owing to this, speech enhancement needs to be performed to the speech signalbefore extracting the features [166]. The Discrete Wavelet Transform methodology includes only the temporal information along with the frequency information. DWTs have been explored [264]. Researchers have tried using the hybrid of MFCC and DWT to enhance the recognition, known as MFDWC (mel-frequency discrete wavelet coefficients) [333]. LPCC features have been used to overcome the demerits of LPC and MFCC [18]. The extensive use of LPCC was studied. Lee and Hung [181] mentioned the comparison analysis of MFCC and LPC. Perceptual Linear Prediction (PLP) feature considered the three aspects. PLP

features are well suited for noisy environments [88]. RelAtive SpecTrA PLP (RASTA) enhances the PLP robustness. RASTA-PLP features are best to use in noisy environments instead of noise-free environments. The Vector Quantization (VQ) method involves computation errors in quantization. VQ can be merged with MFCC [120] as well as DWT. Principal Component Analysis (PCA) improves the robustness of the system for noisy data [181, 320, 343]. Linear Discriminant Analysis (LDA) involves a supervised mechanism, unlike PCA. Variations to the conventional LDA were mentioned [210, 343]. A brief review of the different techniques used in identifying and recognizing speech for ASR systems in non-Indian languages is presented in the table. The table contains information about the data sets being used, feature extraction methodologies, and the recognition results.

A comprehensive view of the speech recognition terminologies on Indian languages is presented in Table 4.

A through study has been performed on the research papers that have been used in this study. Most of the major languages have been covered in this review paper. Research data have been put in tabular form for better understanding. In the next section, synthesis analyses of the studies have been presented. All the studies included have been critically reviews and analysis has been put forth for the readers.

Synthesis analysis

The authors list the following findings from the literature studied. The most relevant finding here is that to attain better accuracy results, a researcher should use hybrid feature extraction techniques. Such techniques provide efficient and useful information for the input to a classifier.

(a) The method to combine different classifiers is dependent on the information which is provided by a single classifier.
(b) Owing to the variations in the style of speaking of different individuals of different regions, speech recognition has been a demanding research area.
(c) The accuracy of the speech recognition and identification process relies on the features which have high discriminating power. Thus, there is a drastic requisite to study variant feature selection algorithms that can achieve good accuracy.
(d) Due to distinct characteristics and variations in the tone including the diverse speaking style of the speaker, it sometimes becomes a ponderous task for the researcher to research different speech corpus with good accuracy.
(e) Since the performance of a classifier relies on the extraction of the features at the feature extraction stage, it is thus important to carefully select the methods of feature extraction and the classifiers.
(f) Most of the work on Indian languages has been carried out on languages such as Hindi, Bangla, Telugu, and Tamil. Languages like Punjabi, Devanagari, Dogri, Kashmiri, Gujarat, and languages of the northeast part of India are yet to be explored.
(g) The lack of standard databases available for the researchers is significantly low in Indian languages. This presents a future direction to perform various experiments.
(h) It is observed that the most commonly used feature extraction methods for speech recognition and identification task are PLP, LFCC, MFCC, and RASTA.
(i) It is also noted that the most commonly used classifiers for speech recognition and identification task are HMM, DNN, and DNN–HMM.
(j) It is also observed from the literature studied that the HMM classifier is very commonly used for speech recognition with good accuracy. The toolkits commonly used by the researchers are Sphinx and HTK. But again, some findings proved that accuracy rates can be increased with DNN or DNN–HMM models.

Suggestions on future directions

In the field of speech recognition, an immense number of directions can be explored to carry out further research as the proposed techniques being used for extracting the features of the speech samples can be extended further by combining different techniques to improve the speech recognition rate and WER. The authors mention the following suggestions on future research directions for speech recognition [131]:

(a) The researchers should develop standard databases for Indian languages. The LVCSR database systems should be focused more. These databases should be made accessible to the researchers, so that future research can nurture.
(b) Indian languages have not been employed efficient feature extraction techniques. Researchers have only focused on baseline MFCC features as studied in the prevalent literature. Hybrid features like MFCC + LDA + MLLT, MFCC + BFCC + GFCC, LDA + MFCC, MF + PLP, RASTA-PLP, etc. may be implemented.
(c) More research can be carried out methodologies for different regional languages.
(d) Researchers in the studied literature have reported their results of recognition accuracy in clean, i.e., noise-free environments, especially for Indian languages. However, in real-time applications, background noise and

Table 4 Recognition results of Indian languages

Authors	Language/corpus data set	Feature extraction	Acoustic modeling	Accuracy
Hegde et al. [106, 108]	Tamil and Telugu, 20 news bulletins	MODGDF MFCC	HMM	Acc. 86–87
Ramamohan and Dandapat [263]	English and Tamil, 150 files	MFCC	Sinusoidal features	Tamil 92.3 Eng. 87.1
Hegde et al. [107]	TIMIT, OGI_MLTS and NTIMIT, LVSCR	MODGDF	HMM	Acc. 96
Plahl et al. [246]	Mandrain Hub4 TDT4, 1534 h	MFCC, PLP	VLTn, CMLLR	CER 9.8
Banerjee et al. [25]	Bengali, 4000 sentences	MFCC	Single Gaussian HMM	Acc. 81
Kandali et al. [138]	Assamese, 27 speakers	MFCC	GMM	Acc. 74.4
Sarma et al. [283]	Assamese, 100 samples	LPC	ANN	Acc. 95.1
Kandali et al. [139]	MESDNEI, 140 sentences	WPCC2, MFCC, tfWPCC2 tfMFCC	GMM	Acc. 83.33 88.10, 90.48 respectively
Hwang et al. [124]	Mandarin, LDC Chinese Gigaword corpus,	MLP, PLP	HMM	CER 9.1
Mohanty and Swain [221]	Oriya, 900 phrases	pitch, first two formants, jitter	Fuzzy k-means clustering	Acc. 65.16
Saraswathi and Geetha [281]	Tamil, 200 sentences and 100 speakers	LPC, PLP	HMM	WER 12.9
Mohamed and Nair [215]	Malayalam, 108 sentences with 540 words, 12 speakers, 25 h	MFCC + delta-delta	HMM	WER 5.3
Gaikwad et al. [85]	Marathi, 625 sentences	MFCC,LDA	HMM	Acc. 89.24
Mohamed and Ramachandran Nair [216]	Continuous speech in Malayalam, 108 sentences	MFCC	HMM	Acc. 94.67
Sarma and Sarma [285]	Assamese, 1000 samples	LPC + PCA	ANN	Acc. 96.2
Aggarwal and Dave [7]	Hindi, 19 h broadcast news	RASTA-PLP MF-PLP MFCC	Generic HMM Segmented HMM ANN-HMM	WER 13
Das et al. [62]	Bengali BENG_OL and BENG_YO	MFCC	MLLR, MLLT, LDA	Phone Recognition accuracy 75.3
Cucu et al. [58]	Romanian, 23.8 k utterances 38.1 h	MFCC, PNCC	HMM	WER 39.4
Shahnawazuddin et al. [298]	Assamese, 250 inqueries	MFCC	HMM/GMM	WER 14.10
Chen et al. [47]	Tamil, Babel IARPA	MFCC LDA + MLLT + fMLLR	DNN	Relative Gain 14.9
Pardeep and Rao [240]	Kannada, 8 h	MFCC, MLF	HMM–DNN	PER 43.45
Kaur and Singh [146]	Punjabi, 154 unique words, 3 h of speech corpus	MFCC, PLP, PNCC	HMM	86.5
Dua et al. [72]	Hindi, 1000 sentences	MF + PLP	HMM–GMM MMIE + MPE	WER 35.2
Kumar and Singh [175]	Punjabi, 1433 sentences	MFCC	HMM	Acc. 90.8
Kadyan et al. [134]	Punjabi, 5000 most frequent words, 45,000 utterances of 32 h	MFCC, PLP. Rasta	HMM, DNN	Acc. 67.38
Yadava and Jayanna [358]	Kannada, 2000 farmers, 232 words, 100 mandis and 104 commodities	MFCC	SGMM	WER 17.18

Table 4 continued

Authors	Language/corpus data set	Feature extraction	Acoustic modeling	Accuracy
Sarma et al. [282]	Assamese, 20 h	–	DNN	Acc. 78.05
Lokesh et al. [192]	FIRE dataset, 3000 sentences	MAR + PLP	BRNN–SOM	Acc. 93.6
Vegesna et al. [342]	IIIT-H Telugu speech corpus 64,464 utterances	MFCC and prosody features	GMM–HMM	Acc. 75
Kadyan et al. [135]	Dataset1:21,764 words, 13 speaker Dayaset 2: 422 unique phonetically rich sentences	GFCC, MFCC, MLLT, LDA	GMM–HMM DNN–HMM	DNN–HMM WER 5.2
Sarma et al. [284]	CALLFRIEND, Tamil Hindi	MFCC	DNN–UBM	WER 3.19
Yi et al. [361]	Assamese 10-h	BLSTM-CTC	DNN, BLSTM	97
Nahid et al. [228]	Bangla 2000 Samples	MFCC, LSTM	HMM, GMM	WER 13.2
Al Amin et al. [10]	Bangla 21.64 h	LDA + MLT + SAT	DNN, GMM–RNN	Acc. 92
Popli and Kumar [249]	Bangla 7000 Samples	LDAM + MPL	QbE-STD	WER 38.08
Yi et al. [362]	Marathi 10 h	BLSTM, LSTM	GMM–HMM	WER 10.1
Bhanja et al. [31]	Marathi 72 h	LSTM	GMM, MFCC, HMM, RNN	Acc. 90.1 and 88.1
Kumar et al. [171]	Kannad, 2400 speakers	MFCC	DNN–HMM	WER 4.10
Dash et al. [63]	132 phrases, 2 speakers	MFCC, delta-delta	GMM–HMM, DNN–HMM, LSTM–HMM	DNN–HMM PER27.7

other factors affect the accuracy of recognizing the speech.

(e) Tonal aspects of languages have not been put forward by most of the researchers. As per the literature, the authors could only find tonal research done on languages like Mandarin Chinese, Bangla, Arabic, and other tonal languages that lack the research initiative.

(f) Languages like Punjabi, Dogri, and other under-resourced languages need a standard database to work on. Also, the tonal features of the languages have not been considered yet.

(g) Research on extracting efficient and suitable tonal features should be conducted for developing ASR systems for tonal languages.

Conclusion

In this paper, the authors have presented an extensive review and analysis of different feature extraction techniques employed for speech recognition for non-Indian and Indian languages. Different evaluation parameters in which the researchers have reported their work have been mentioned with the help of different tables. These parameters were WER, PER, SER, accuracy, recognition rate, and comparative analysis of different techniques. The authors have cited a comprehensive study of the work done for different foreign languages across the world. These languages include Chinese, Japanese, Russian, Romanian, Malay, Thai, and Arabic. Also, a summary of the speech recognition research on Indian languages, i.e., Hindi, Bangla, Oriya, Tamil, Telugu, Malayalam, Punjabi, Kannada, Marathi, and Assamese has been mentioned. The authors suggest that efficient feature extraction techniques implemented on Non-Indian languages can be implemented for Indian languages, as well. This is because the research area for Indian languages is too wide. Also, not many accurate ASR systems have been developed for Indian languages. Furthermore, LVCSR systems should be explored more by the researchers to improve the accuracy of recognizing speech for such systems, so that applications for real-time use can be developed.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix 1: Quality assessment forms

Screening question
Section-1
Does the research paper refer to speech recognition? Yes
Consider:
This paper contains the study of speech recognition. All forms of studies like case study, experimental study, or the research paper are included
Section 1—is evaluated first. If the positive reply is received, then proceed to section 2
Screening question
Section-2
Key Sub–Area Categorization
Is the research paper focused on speech recognition of tonal language? Yes
Consider
Is the study's focus or main focus is on speech recognition or not?
Did the study fit in the above-mentioned subareas categorized?

If the study's main focus is on speech recognition, then proceed to Section 3, else proceed to Section 4.

Detailed questions
Section-3
Findings
Is there a clear statement of findings? Yes
Consider
Did the study mention the approach/speech recognition
Has the speech recognition technique reported?
What is the corresponding transformation technique, finding, i.e., source representation
Comparison
Were the data reported sufficient for comparative analysis? Yes
Consider
Are the necessary parameters for comparison discussed?
Is the study referring speech recognition explicitly?

Detailed questions
Section-4
Findings
Did the study mention the type of speech recognition and the data set used? Yes
Consider
How well the speech recognition is categorized?
Did the study explicitly mention the type/process of speech recognition, or is to be inferred from the study

Appendix 2: Data items extracted from all papers

Data item	Description
Study identifier	Unique ID for the study
Bibliographic data	Author, year, title, source
Type of article	Journal article, Conference article, workshop paper
Study aims/context/application domain	What are the aims of the study, i.e., search focus, i.e., the research areas the paper focus on
Study design	Classification of study—feature extraction, classification, speech recognition, comparative analysis, etc
What is the speech recognition technique of tonal languages?	Spoken words are transcribed by the computers into readable text
How was comparison carried out?	Values of important parameters for speech recognition, i.e., Fundamental frequency, tone, energy, gain, amplitude, power, frequency response, voiced and unvoiced feature
Subject system	How the data was collected: it refers to the subject system and its size
Data analysis	Data analysis, i.e., corresponding source representation and error rate, recognition accuracy
Developer of the tool and usage	It refers to the speech recognition tool, developer and usage of the tool
Study findings	Major findings or conclusions from the primary study like percentage of speech's recognition accuracy

References

1. Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(10):1533–1545
2. Adda J, Dustmann C, Stevens K (2017) The career costs of children. J Polit Econ 125(2):293–337
3. Adda-Decker M, Lamel L, Adda G, Lavergne T (2011) A first LVCSR system for Luxembourgish, a low-resourced European language. Lang Technol Conf 2011:479–490
4. Adda-Decker M, Boula de Mareuil P, Adda G, Lamel L (2005) Investigating syllabic structures and their variation in spontaneous french. Speech Commun 46:119–139
5. Afify M, Sarikaya R, Kuo HKJ, Besacier L, Gao Y (2006) On the use of morphological analysis for dialectal Arabic speech recognition. In: Ninth international conference on spoken language processing, pp 270–280
6. Aggarwal RK, Dave M (2011) Projected features for hindi speech recognition system. Int J Adv Res Comput Sci 2(3):129–134
7. 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
8. Aggarwal RK, Dave M (2013) Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system. Telecommun Syst 52(3):1457–1466
9. Agrawal SS, Sinha S, Singh P, Olsen JØ (2012) Development of text and speech database for hindi and indian english specific to mobile communication environment. In: LREC, pp 3415–3421
10. Al Amin MA, Islam MT, Kibria S, Rahman MS (2019) Continuous Bengali speech recognition based on deep neural network. In: International conference on electrical, computer and communication engineering, IEEE, pp 1–6
11. Ali A, Zhang Y, Cardinal P, Dahak N, Vogel S, Glass J (2014) A complete kaldi recipe for building arabic speech recognition systems. In: Spoken language technology workshop (SLT), 2014, IEEE, pp 525–529
12. Ali M, Elshafei M, Al-Ghamdi M, Al-Muhtaseb H, Al-Najjar A (2008) Generation of Arabic phonetic dictionaries for speech recognition. Innov Inf Technol 2008:59–63
13. Ali M (2018) Character level convolutional neural network for German dialect identification. In: Proceedings of the fifth workshop on NLP for similar languages, varieties and dialects, pp 172–177
14. Alsharhan E, Ramsay A (2019) Improved Arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions. Inf Process Manage 56(2):343–353
15. Anand AV, Devi PS, Stephen J, Bhadran VK (2012) Malayalam Speech Recognition system and its application for visually impaired people. In: India Conference (INDICON), 2012, Annual IEEE, pp 619–624
16. Antony PJ, Mohan SP, Soman KP (2010) SVM based part of speech tagger for Malayalam. In: Recent trends in information, telecommunication and computing (ITC), 2010, IEEE, pp 339–341
17. Anusuya MA, Katti SK (2011a) Comparison of different speech feature extraction techniques with and without wavelet transform to Kannada speech recognition. Int J Comput Appl 26(4):19–24
18. Anusuya MA, Katti SK (2011b) Front end analysis of speech recognition: a review. Int J Speech Technol 14(2):99–145
19. Anusuya MA, Katti SK (2012) Speaker independent kannada speech recognition using vector quantization. In: IJCA proceedings on national conference on advancement in electronics and telecommunication engineering NCAETE, pp 32–35
20. Apandi N, Jamil N (2016) An analysis of Malay language emotional speech corpus for emotion recognition system. In: Industrial electronics and applications conference (IEACon), 2016, IEEE, pp 225–231
21. Arora K, Arora S, Roy MK (2013) Speech to speech translation: a communication boon. CSI Trans ICT 1(3):207–213
22. Badino L, Canevari C, Fadiga L, Metta G (2016) Integrating articulatory data in deep neural network-based acoustic modeling. Comput Speech Lang 36:173–195
23. Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y (2016) End-to-end attention-based large vocabulary speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), Shanghai, pp 4945–4949
24. Baig MMA, Qazi SA, Kadri MB (2015) Discriminative training for phonetic recognition of the Holy Quran. Arab J Sci Eng 40(9):2629–2640
25. Banerjee P, Garg G, Mitra P, Basu A (2008)Application of triphone clustering in acoustic modeling for continuous speech recognition in Bengali. In: Pattern recognition, 2008, ICPR, pp 1–4
26. Bansal S, Sharan S, Agrawal SS (2015) Corpus design and development of an annotated speech database for Punjabi. In: IOriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation, IEEE, pp 32–37
27. Beck E, Hannemann M, Dötsch P, Schlüter R, Ney H (2018) Segmental encoder-decoder models for large vocabulary automatic speech recognition. In: Proc. Interspeech, pp 766–770
28. Behravan H, Hautamaki V, Siniscalchi SM, Khoury E, Kurki T, Kinnunen T, Lee CH (2014) Dialect levelling in Finnish: a universal speech attribute approach. In: iInterspeech, 2014, pp 2165–2169
29. Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C, Rose R (2007) Automatic speech recognition and speech variability: a review. Speech Commun 49(10–11):763–786
30. Bérard A, Pietquin O, Servan C, Besacier L (2016) Listen and translate: a proof of concept for end-to-end speech-to-text translation. arXiv:1612.01744
31. Bhanja CC, Bisharad D, Laskar RH (2019) Deep residual networks for pre-classification based Indian language identification. J Intell Fuzzy Syst 36(3):2207–2218
32. Bharali SS, Kalita SK (2018) Speech recognition with reference to Assamese language using novel fusion technique. Int J Speech Technol 2018:1–13
33. Bhowmik T, Mandal SKD (2016) Deep neural network based phonological feature extraction for Bengali continuous speech. In: Signal and information processing (IConSIP), pp 1–5
34. Bhowmik T, Mukherjee S, Mandal SKD (2015) Detection of attributes for bengali phoneme in continuous speech using deep neural network. In: Signal processing and integrated networks (SPIN), pp 103–108
35. Boril H, Hansen JHL (2010) Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments. IEEE Trans Audio Speech Lang Process 18(6):1379–1393
36. Botros R, Irie K, Sundermeyer M, Ney H (2015) On efficient training of word classes and their application to recurrent neural network language models. In: INTERSPEECH-2015, pp 1443–1447
37. Bourlard HA, Morgan N (2012) Connectionist speech recognition: a hybrid approach, 247. Springer Science & Business Media, Berlin
38. Burget L, Schwarz P, Agarwal M, Akyazi P, Feng K, Ghoshal A, Rastrow A (2010) Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models. In: 2010 IEEE international conference on acoustics, speech and signal processing, pp 4334–4337
39. Burileanu C, Popescu V, Buzo A, Petrea CS, Ghelmez-Haneş D (2010) Spontaneous speech recognition for Romanian in spoken dialogue systems. Proc Roman Acad 11(1):8391

40. Byrne W, David D, Martin F, Samuel G, Jan H, Douglas O, Michael P, Josef R, Bhuvana R, Dagobert S, Todd W, Jing R (2004) Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Trans Speech Audio Process 12:420–435
41. Cai M, Shi Y, Liu J (2013) Deep maxout neural networks for speech recognition. In: automatic speech recognition and understanding (ASRU), pp 291–296
42. Caranica A, Cucu H, Andi B, Corneliu B (2016) On the Design of an automatic speech recognition system for Romanian language. Control Eng Appl Inf 18:65–76
43. Chaloupka J, Nouza J, Malek J, Silovsky J (2015) Phone speech detection and recognition in the task of historical radio broadcast transcription. In: 38th international conference on telecommunications and signal processing (TSP), Prague, pp 1–4
44. Chandrasekar M, Ponnavaikko M (2008) Tamil speech recognition: a complete model. Electron J Tech Acoust 2008:20
45. Charles AH, Devaraj G (2004) Alaigal-A Tamil speech recognition. Tamil Internet 2004:5
46. Charoenpornsawat P, Hewavitharana S, Schultz T (2006) Thai grapheme-based speech recognition. In: Human language technology conference of the NAACL, Companion Volume: Short Papers, pp 17–20
47. Chen NF, Ni C, Chen IF, Sivadas S, Xu H, Xiao X, Wang L (2015) Low-resource keyword search strategies for Tamil. In: IEEE international conference on acoustics, speech and signal processing, pp 5366–5370, IEEE
48. Chen NF, Wee D, Tong R, Ma B, Li H (2016) Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: analysis on iCALL. Speech Commun 84:46–56
49. Chen Z, Qi L, Hao L, Kai Y (2018) On modular training of neural acoustics-to-word model for lvcsr. In: ICASSP, pp 1–5
50. Chien J, Huang C (2006) Aggregate a posteriori linear regression adaptation. IEEE Trans Audio Speech Lang Process 14(3):797–807
51. Chiopu D, Oprea M (2014) Using neural networks for a discriminant speech recognition system. Int Conf Dev Appl Syst 2014:165–169
52. Chourasia V, Samudravijaya K, Chandwani M (2005) Phonetically rich hindi sentence corpus for creation of speech database. In: Proc. O-Cocosda, pp 132–137
53. Christodoulides, G, Avanzi, M, Goldman JP (2018) DisMo: a morphosyntactic, disfluency and multi-word unit annotator. In: An evaluation on a corpus of French spontaneous and read speech. arXiv:180202926
54. Chunwijitra V, Chotimongkol A, Wutiwiwatchai C (2016) A hybrid input-type recurrent neural network for LVCSR language modeling EURASIP. J Audio Speech Music Process 2016(1):15
55. Ciobanu AM, Malmasi S, Dinu LP (2018) German dialect identification using classifier ensembles. arXiv:1807.08230
56. Clark E, Doyle P, Garaialde D, Gilmartin E, Edlund J, Aylett M, Cabral J, Munteanu C, Cowan B (2018) The State of Speech in HCI: trends, themes and challenges. arXiv:1810.06828
57. Cucchiarini C, Van hamme H (2013) The JASMIN speech corpus: recordings of children, non-natives and elderly people In: Spyns P, Odijk J (eds) Essential speech and language technology for dutch theory and applications of natural language processing. Springer, Berlin, Heidelberg
58. Cucu H, Buzo A, Petrică L, Burileanu D, Burileanu C (2014) Recent improvements of the SpeeD Romanian LVCSR system. In: 10th international conference on communications (COMM), Bucharest, pp 1–4
59. Cutajar M, Gatt E, Grech I, Casha O, Micallef J (2013) Comparative study of automatic speech recognition techniques. IET Signal Proc 7(1):25–46
60. Darekar RV, Dhande AP (2018) Emotion recognition from Marathi speech database using adaptive artificial neural network. Biol Inspired Cogn Archit 23:35–42
61. Das B, Mandal S, Mitra P (2011) Bengali speech corpus for continuous auutomatic speech recognition system. In: 2011 international conference on speech database and assessments (Oriental COCOSDA), pp 51–55, IEEE
62. Das B, Mandal S, Mitra P, Basu A (2013) Aging speech recognition with speaker adaptation techniques: study on medium vocabulary continuous Bengali speech. Pattern Recogn Lett 34(3):335–343
63. Dash D, Kim M, Teplansky K, Wang J (2018) Automatic speech recognition with articulatory information and a unified dictionary for Hindi, Marathi, Bengali and Oriya. Interspeech, pp 1046–1050
64. Debatin L, Haendchen Filho A, Dazzi L (2018) Offline speech recognition development—a systematic review of the literature. Int Conf Enterprise Inf Syst 2:551–558
65. Deemagarn A, Kawtrakul A (2004a) Thai connected digit speech recognition using hidden markov models. In: International conference on speech and computer, pp 731–735
66. Deemagarn A, Kawtrakul A (2004b) Thai connected digit speech recognition using hidden Markov models, SPECOM-2004, pp 731–735
67. Despres J, Fousek P, Gauvain JL, Gay S, Josse Y, Lamel L, Messaoudi A (2009) Modeling northern and southern varieties of Dutch for STT, tenth annual conference of the international speech communication association, pp 96–99
68. Dey A, Shahnawazuddin S, Deepak KT, Imani S, Prasanna SRM, Sinha R (2016) Enhancements in Assamese spoken query system: enabling background noise suppression and flexible queries. In: Twenty second national conference on communication (NCC), pp 1–6, IEEE
69. Dimulescu VB, Mareüil PB (2006) Perceptual identification and phonetic analysis of 6 foreign accents in french. In: INTERSPEECH'2006, pp 441–446
70. Draman M, Tee DC, Lambak Z, Yahya MR, Yusoff MM, Ibrahim SH, Saidon S, Haris NA, Tan TP (2017) Malay speech corpus of telecommunication call center preparation for ASR. In: 5th international conference on information and communication technology (ICoIC7), pp 1–6 IEEE
71. Dua M, Aggarwal RK, Biswas M (2018) Performance evaluation of Hindi speech recognition system using optimized filterbanks. Eng Sci Technol Int J 21(3):389–398
72. Dua M, Aggarwal RK, Biswas M (2017) Discriminative training using heterogeneous feature vector for hindi automatic speech recognition system. In: 2017 international conference on computer and applications, pp 158–162, IEEE.
73. Duada RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York
74. Dumitru CO, Gavat I (2006) A comparative study of feature extraction methods applied to continuous speech recognition in romanian language. In: 48th international symposiumon multimedia signal processing and communications, pp 115–118, IEEE
75. Dutta K, Sarma KK (2012) Multiple feature extraction for RNN-based assamese speech recognition for speech to text conversion application. In: 2012 International conference on communications, devices and intelligent systems (CODIS), (pp 600–603), IEEE
76. El-Amrani MY, Rahman MH, Wahiddin MR, Shah A (2016) Building CMU Sphinx language model for the Holy Quran using simplified Arabic phonemes. Egypt Inf J 17(3):305–314
77. Enarvi S, Kurimo M (2013) A novel discriminative method for pruning pronunciation dictionary entries. In: 7th conference on speech technology and human—computer dialogue (SpeD), Cluj-Napoca, pp 1–4

78. Enarvi S, Smit P, Virpioja S, Kurimo M (2017) Automatic speech recognition with very large conversational finnish and estonian vocabularies. IEEE/ACM Trans Audio Speech Lang Process 25(11):2085–2097
79. Eng GK, Ahmad AM (2005) Malay speech recognition using self-organizing map and multilayer perceptron. In: Postgraduate annual research seminar, pp 233–237
80. Fook CY, Hariharan M, Yaacob S, Adom AH (2012) A review: Malay speech recognition and audio visual speech recognition. In: International conference on biomedical engineering, pp 479–484, IEEE
81. Franzini M, Lee KF, Waibel A (1990) Connectionist Viterbi training: a new hybrid method for continuous speech recognition. In: Acoustics, speech, and signal processing, 1990. ICASSP-90, 1990 International Conference on (pp. 425–428), IEEE
82. Fukuda T, Ichikawa O, Nishimura M (2018) Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition. Speech Commun 98:95–103
83. Fukunaga K (1999) Statistical pattern recognition. In: Handbook of pattern recognition and computer vision, pp 33–60
84. Gaikwad S, Gawali B, Mehrotra SC (2012) Novel approach-based feature extraction for Marathi continuous speech recognition. In: Proceedings of the international conference on advances in computing, communications and informatics, pp 795–804 ACM
85. Gaikwad S, Gawali B, Yannawar P, Mehrotra S (2011) Feature extraction using fusion MFCC for continuous marathi speech recognition. In: India Conference (INDICON), pp 1–5, IEEE
86. Gaikwad SK, Gawali BW, Yannawar P (2010) A review on speech recognition technique. Int J Comput Appl 10(3):16–24
87. Gales MJ, Diehl F, Raut CK, Tomalin M, Woodland PC, Yu K (2007) Development of a phonetic system for large vocabulary Arabic speech recognition. In: IEEE workshop on automatic speech recognition & understanding, pp 24–29, IEEE
88. Ganapathy S, Thomas S, Hermansky H (2009) Modulation frequency features for phoneme recognition in noisy speech. J Acoust Soc Am 125(1):EL8–EL12
89. Gawali BW, Gaikwad S, Yannawar P, Mehrotra SC (2011) Marathi isolated word recognition system using MFCC and DTW features. ACEEE Int J Inf Technol 1(01):21–24
90. Geethashree A, Ravi DJ (2018) Kannada emotional speech database: design, development and evaluation international conference on cognition and recognition. Springer, Singapore, pp 135–143
91. Georgescu A, Cucu H, Burileanu C (2017) SpeeD's DNN approach to Romanian speech recognition. In: International conference on speech technology and human-computer dialogue (SpeD), Bucharest, pp 1–8
92. Georgescu AL, Cucu H (2018) GMM-UBM Modeling for speaker recognition on a Romanian large speech corpora. In: International conference on communications (COMM), pp 547–55, IEEE
93. Ginter F, Nyblom J, Laippala V, Kohonen S, Haverinen K, Vihjanen S, Salakoski T (2013) Building a large automatically parsed corpus of Finnish, 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), vol 85, pp 291–300
94. Godambe T, Samudravijaya K (2011) Speech data acquisition for voice based agricultural information retrieval. In: 39th All India DLA Conference, Punjabi University, Patiala, June
95. Gonzalez-Dominguez J, Eustis D, Lopez-Moreno I, Senior A, Beaufays F, Moreno PJ (2015) A real-time end-to-end multilingual speech recognition architecture. IEEE J Sel Top Signal Process 9(4):749–759
96. Graves A, Mohamed AR, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: International conference on acoustics, speech and signal processing, pp 6645–6649, IEEE
97. Guglani J, Mishra AN (2018) Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. Int J Speech Technol 21(2):211–216
98. Gulic M, Lucanin D, Simic A (2011) A digit and spelling speech recognition system for croatian language. In: Proceedings of the 34th international convention, MIPRO, Opatija, Croatia, pp 1673–1678
99. Haffner P, Franzini M, Waibel A (1991), April. Integrating time alignment and neural networks for high performance continuous speech recognition In: Acoustics, speech, and signal processing, 1991. ICASSP-91, 1991 International Conference on (pp 105–108), IEEE
100. Hämäläinen A, Teixeira A, Almeida N, Meinedo H, Fegyó T, Dias MS (2015) Multilingual speech recognition for the elderly: the AALFred personal life assistant. Procedia Comput Sci 67:283–292
101. Hanani A, Russell MJ, Carey MJ (2013) Human and computer recognition of regional accents and ethnic groups from British English speech. Comput Speech Lang 27(2013):59–74
102. Harisha SB, Amarappa S, Sathyanarayana DS (2015) Automatic speech recognition-a literature survey on indian languages and ground work for isolated kannada digit recognition using MFCC and ANN. Int J Electron Comput Sci Eng 4(1):91–105
103. Hartmann W, Le VB, Messaoudi A, Lamel L, Gauvain JL (2014) Comparing decoding strategies for subword-based keyword spotting in low-resourced languages. In: Fifteenth annual conference of the international speech communication association, pp 2764–2768
104. Hasnat M, Mowla J, Khan M (2007) Isolated and continuous bangla speech recognition: implementation, performance and application perspective. In: International Symposium On Natural Language Processing (SNLP), Hanoi, Vietnam
105. Hegde Renjith S, Manju KG (2017) Speech based emotion recognition in Tamil and Telugu using LPCC and hurst parameters-A comparative study using KNN and ANN classifiers. In: 2017 international conference on circuit, power and computing technologies (ICCPCT), Kollam, pp 1–6
106. Hegde RM, Murthy HA, Gadde VRR (2004a) Continuous speech recognition using joint features derived from the modified group delay function and MFCC. In: Eighth international conference on spoken language processing
107. Hegde RM, Murthy HA, Gadde VRR (2007) Significance of the modified group delay feature in speech recognition. IEEE Trans Audio Speech Lang Process 15(1):190–202
108. Hegde RM, Murthy HA, Rao GR (2004b) Application of the modified group delay function to speaker identification and discrimination. In: 2004 IEEE international conference on acoustics, speech, and signal processing, pp I-517 IEEE
109. Heigold G, Vanhoucke V, Senior A, Nguyen P, Ranzato M, Devin M, Dean J (2013) Multilingual acoustic models using distributed deep neural networks. In: IEEE international conference on acoustics, speech and signal processing, Vancouver, BC, pp 8619–8623
110. Hemakumar G, Punitha P (2013) Speech recognition technology: a survey on Indian languages. Int J Inf Sci Intell Syst 2(4):1–38
111. Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97
112. Hirsimäki T, Kurimo M (2004) Decoder issues in unlimited Finnish speech recognition. In: Proc. of the 6th nordic signal processing symposium, pp 320–323
113. Hirsimäki T, Creutz M, Siivola V, Kurimo M, Virpioja S, Pylkkönen J (2006) Unlimited vocabulary speech recognition with morph language models applied to Finnish. Comput Speech Lang 20:515–541

114. Hoffmeister B, Plahl C, Fritz P, Heigold G, Loof J, Schluter R, Ney H (2007) Development of the RWTH Mandarin LVCSR system. In: IEEE workshop on automatic speech recognition and understanding (ASRU), Kyoto, pp 455–460
115. Hori T, Chen Z, Erdogan H, Hershey JR, Le Roux J, Mitra V, Watanabe S (2017) Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend. Comput Speech Lang 46:401–418
116. Hossain M, Rahman M, Prodhan UK, Khan M (2013) Implementation of back-propagation neural network for isolated bangla speech recognition. arXiv:1308.3785
117. Hotta H (2011) Japanese speaker-independent homonyms speech recognition. Procedia-Soc Behav Sci 27:306–313
118. Hsiao R, Metze F, Schultz T (2010) Improvements to generalized discriminative feature transformation for speech recognition. In: Eleventh annual conference of the international speech communication association, pp 1361–1364
119. Hu X, Saiko M, Hori C (2014) December Incorporating tone features to convolutional neural network to improve Mandarin/Thai speech recognition, signal and information processing association annual summit, pp 1–5, IEEE
120. Hu X, Zhan L, Xue Y, Zhou W, Zhang L (2011) Spoken Arabic digits recognition based on wavelet neural networks. In: international conference on systems, man, and cybernetics (SMC), pp 1481–1485, IEEE
121. Huang H, Hu Y, Xu H (2017) Mandarin tone modeling using recurrent neural networks. arXiv:171101946
122. Huet S, Gravier G, Sébillot P (2010) Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition. Comput Speech Lang 24(4):663–684
123. Huijbregts M, Wooters C, Ordelman R (2007) Filtering the unknown: speech activity detection in heterogeneous video collections. In: Eighth annual conference of the international speech communication association, pp 2925–2928
124. Hwang M, Peng G, Ostendorf M, Wang W, Faria A, Heidel A (2009) Building a highly accurate mandarin speech recognizer with language-independent technologies and language-dependent modules. IEEE Trans Audio Speech Lang Process 17(7):1253–1262
125. Iakushkin O, Fedoseev G, Shaleva SA, Degtyarev A, Sedova SO (2018) Russian-language speech recognition system based on Deepspeech
126. Ichikawa O, Fukuda T, Nishimura M (2010) Dynamic features in the linear-logarithmic hybrid domain for automatic speech recognition in a reverberant environment. IEEE J Sel Top Signal Process 4(5):816–823
127. Imseng, D, Bourlard, H, Caesar, H, Garner PN, Lecorvé G, Nanchen A (2012) MediaParl: Bilingual mixed language accented speech database In: 2012 IEEE spoken language technology workshop (SLT) (pp 263–268) IEEE
128. Ircing P, Psutka JV, Psutka J (2009) Using morphological information for robust language modeling in Czech ASP system. IEEE Trans Audio Speech Lang Process 17(4):840–847
129. Ismail T, Singh LJ (2017) Dialect identification of assamese language using spectral features. Indian J Sci Technol 10:20
130. Jamal N, Shanta S, Mahmud F, Shaabani MNAH (2017) Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: a review. AIP Conf Proc 1883:020028
131. Jiang H (2010) Discriminative training of HMMs for automatic speech recognition: a survey. Comput Speech Lang 24(4):589–608
132. Jing S, Mao X, Chen L, Comes MC, Mencattini A, Raguso G, Ringeval F, Schuller B, Di Natale C, Martinelli E (2018) A closed-form solution to the graph total variation problem for continuous emotion profiling in noisy environment. Speech Commun 104:66–72
133. Kačur J, Rozinaj G (2011) Building accurate and robust HMM models for practical ASR systems. Telecommun Syst 52(3):1683–1696
134. Kadyan V, Mantri A, Aggarwal RK (2017) A heterogeneous speech feature vectors generation approach with hybrid hmm classifiers. Int J Speech Technol 20(4):761–769
135. Kadyan V, Mantri A, Aggarwal RK, Singh A (2019) A comparative study of deep neural network based Punjabi-ASR system. Int J Speech Technol 22(1):111–119
136. Kaewprateep J, Prom-on S (2018) Evaluation of small-scale deep learning architectures in Thai speech recognition. In: 2018 International ECTI Northern Section Conference on electrical, electronics, computer and telecommunications engineering
137. Kalita SK (2010) Nonlinearity and cepstral/mel cepstral measure of the spectral characteristics of assamese and bodo phonemes. Int J Open Problems Compt Math 3(5):151–165
138. Kandali AB, Routray A, Basu TK (2008) Emotion recognition from Assamese speeches using MFCC features and GMM classifier. In: TENCON 2008–2008 IEEE Region 10, pp 1–5, IEEE
139. Kandali AB, Routray A, Basu TK (2009) Vocal emotion recognition in five native languages of Assam using new wavelet features. Int J Speech Technol 12(1):1–13
140. Kannadaguli P, Bhat V (2018) A comparison of Bayesian and HMM based approaches in machine learning for emotion detection in native Kannada speaker. In: IEEMA Engineer Infinite Conference, pp 1–6, IEEE
141. Kapralova O, Alex J, Weinstein E, Moreno P, Siohan O (2014) A big data approach to acoustic model training corpus selection. In: Annual conference of the international speech communication association, INTERSPEECH, pp 2083–2087
142. Karpov A, Kipyatkova I, Ronzhin A (2011) Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In: Twelfth annual conference of the international speech communication association, pp 3161–3164
143. Karpov A, Markov K, Kipyatkova I, Vazhenina D, Ronzhin A (2014) Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Commun 56:213–228
144. Kat P, Hemakumar G (2014) Speaker dependent continuous Kannada speech recognition using HMM. In: International conference on intelligent computing applications, pp 402–405
145. Kaur A, Singh A (2016a) Power-normalized cepstral coefficients (PNCC) for Punjabi automatic speech recognition using phone based modelling in HTK. In: 2nd international conference on applied and theoretical computing and communication technology, pp 372–375, IEEE
146. Kaur A, Singh A (2016b) Optimizing feature extraction techniques constituting phone based modelling on connected words for Punjabi automatic speech recognition. In: International conference on advances in computing, communications and informatics (ICACCI), Jaipur, pp 2104–2108
147. Kaur J, Singh A, Kadyan V (2020) Automatic speech recognition system for tonal languages: state-of-the-art survey. Arch Comput Method Eng. https://doi.org/10.1007/s11831-020-09414-4
148. Kawahara T (2012) Transcription system using automatic speech recognition for the japanese parliament (Diet), Twenty-fourth innovative appications of artificial intelligence conference,pp 2224–2228
149. Kaya H, Karpov AA (2018) Efficient and effective strategies for cross-corpus acoustic emotion recognition. Neurocomputing 275:1028–1034
150. Kayte S, Gawali DB (2015) Marathi speech synthesis: a review. In: international journal on recent and innovation trends in computing and communication, pp 2321–8169
151. Kertkeidkachorn N, Punyabukkana P, Suchato A (2014) Using tone information in Thai spelling speech recognition. In: 28th

Pacific Asia conference on language, information and computing, pp 178–184
152. Khademian M, Homayounpour MM (2018) Monaural multi-talker speech recognition using factorial speech processing models. Speech Commun 98:1–16
153. Khetri GP, Padme SL, Jain DC, Fadewar DH, Sontakke DB, Pawar DVP (2012) Automatic speech recognition for marathi isolated words. Int J Appl Innov Eng Manag 1(3):69–74
154. Khokhlov Y, Medennikov I, Romanenko A, Mendelev V, Korenevsk M, Prudnikov A, Tomashenko N, Zatvornitskiy A (2017) The STC keyword search system for OpenKWS 2016 evaluation 3602–3606 1021437/Interspeech, pp 2017–1212
155. Kinoshita K, Delcroix M, Nakatani T, Miyoshi M (2009) Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. IEEE Trans Audio Speech Lang Process 17(4):534–545
156. Kipyatkova I, Karpov A, Verkhodanova V, Železný M (2012) Analysis of long-distance word dependencies and pronunciation variability at conversational Russian speech recognition. In: Federated conference on computer science and information systems, pp 719–725, IEEE
157. Kipyatkova IS, Karpov AA (2017) A study of neural network Russian language models for automatic continuous speech recognition systems. Autom Remote Control 78:858
158. Kirchhoff K, Vergyri D (2005) Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. Speech Commun 46(1):37–51
159. Kirchhoff K, Vergyri D, Bilmes J, Duh K, Stolcke A (2006) Morphology-based language modeling for conversational Arabic speech recognition. Comput Speech Lang 20(4):589–608
160. Kiruthiga S, Krishnamoorthy K (2012) Design issues in developing speech corpus for Indian languages—a survey. In: International conference on computer communication and informatics, pp 1–4, IEEE
161. Kitaoka N, Yamada T, Tsuge T, Miyajima C, Yamamoto K, Nishiura T, Nakayama M, Denda Y, Fujimoto M, Takiguchi T, Tamura S, Matsuda S, Ogawa T, Kuroiwa S, Takeda K, Nakamura S (2009) CENSREC-1-C: an evaluation framework for voice activity detection under noisy environments. Acoust Sci Technol 30(5):363–371
162. Kitchenham BA (2007) Guidelines for performing systematic literature reviews in software engineering technical report EBSE-2007–01 Keele University
163. Kocabiyikoglu AC, Besacier L, Kraif O (2018) Augmenting Librispeech with French translations: a multimodal corpus for direct speech translation evaluation. arXiv:180203142
164. Kolar J, Liu Y (2010) Automatic sentence boundary detection in conversational speech: a cross-lingual evaluation on English and Czech. In: 2010 IEEE international conference on acoustics, speech and signal processing, pp 5258–5261
165. Kombrink S, Mikolov T, Karafiát M, Burget L (2012) Improving language models for ASR using translated in-domain data. In: IEEE international conference on acoustics, speech and signal processing, pp 4405–4408
166. Korba MCA, Messadeg D, Djemili R, Bourouba H (2008) Robust speech recognition using perceptual wavelet denoising and mel-frequency product spectrum cepstral coefficient features. Informatica 32:3
167. Krishnan VV, Anto PB (2009) Feature parameter extraction from wavelet subband analysis for the recognition of isolated malayalam spoken words. Int J Comput Netw Secur 1(1):52–55
168. Krishnan VV, Jayakumar A, Babu AP (2008) Speech recognition of isolated malayalam words using wavelet features and artificial neural network. In: 4th IEEE international symposium on electronic design, test and applications, pp 240–243, IEEE
169. Kumar K, Aggarwal RK, Jain A (2012) A Hindi speech recognition system for connected words using HTK. Int J Comput Syst Eng 1(1):25–32
170. Kumar M, Rajput N, Verma A (2004) A large-vocabulary continuous speech recognition system for Hindi. IBM J Res Dev 48(56):703–715
171. KumarAdava PG, Jayanna HS (2019) Continuous Kannada speech recognition system under degraded condition. Circ Syst Signal Process. https://doi.org/10.1007/s00034-019-01189-9
172. Kumar R, Kishore S, Gopalakrishna A, Chitturi R, Joshi S, Singh S, Sitaram R (2005a) Development of Indian language speech databases for large vocabulary speech recognition systems, SPECOM, pp 1–4
173. Kumar R., Kishore S, Gopalakrishna A, Chitturi R, Joshi S, Singh S, Sitaram R (2005b) Development of Indian language speech databases for large vocabulary speech recognition systems, SPECOM
174. Kumar S, Rao SB, Pati D (2013) Phonetic and Prosodically Rich Transcribed speech corpus in Indian languages: Bengali and Odia. In: International conference oriental COCOSDA Held Jointly with 2013 conference on asian spoken language research and evaluation, pp 1–5
175. Kumar Y, Singh N (2017) An automatic speech recognition system for spontaneous Punjabi speech corpus. Int J Speech Technol 20(2):297–303
176. Kuo HKJ, Arisoy E, Mangu L, Saon G (2011) Minimum Bayes risk discriminative language models for Arabic speech recognition. In: IEEE workshop on automatic speech recognition and understanding, pp 208–213, IEEE
177. Kurian C, Balakriahnan K (2012) Continuous speech recognition system for Malayalam language using PLP cepstral coefficient. J Comput Business Res 3(1):1–23
178. Kurian C, Balakrishnan K (2009) Speech recognition of Malayalam numbers world congress on nature and biologically inspired computing, pp 1475–1479, IEEE
179. Kurimo M, Turunen V (2005) To recover from speech recognition errors in spoken document retrieval. In: 9th European conference on speech communication and technology, pp 605–608
180. Larson M, Eickeler S (2003) Using syllable-based indexing features and language models to improve German spoken document retrieval. Interspeech 2003:1217–1220
181. Lee JY, Hung JW (2011) Exploiting principal component analysis in modulation spectrum enhancement for robust speech recognition. In: Eighth international conference on fuzzy systems and knowledge discovery, vol 3, pp 1947–1951, IEEE
182. Levin E (1990) April Word recognition using hidden control neural architecture. In: Acoustics, speech, and signal processing, 1990 ICASSP-90, 1990 International Conference on (pp 433–436) IEEE
183. Li J, Yu D, Huang JT, Gong Y (2012) Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN–HMM. In:Spoken Language Technology Workshop, pp 131–136, IEEE
184. Li K, Mao S, Li X, Wu Z, Meng H (2018) Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. Speech Commun 96:28–36
185. Li X, Yang Y, Pang Z, Wu X (2015) A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary Chinese speech recognition. Neurocomputing 170:251–256
186. Liu G, Lei Y, Hansen JH (2010) Dialect identification: impact of differences between read versus spontaneous speech. In: 18th European signal processing conference, pp 2003–2006, IEEE
187. Liu S, Sim KC (2013) Multi-stream temporally varying weight regression for cross-lingual speech recognition. In: IEEE work-

shop on automatic speech recognition and understanding, Olomouc, pp 434–439
188. Ljubesic N, Agic Z, Klubicka F, Batanovic V, Erjavec T (2018) hr500 K—a reference training Corpus of Croatian. In: Language Technologies & Digital Humanities, Ljubljana, Solvenia, pp 154–160
189. Ljubešić N, Klubička F (2014) {bs,hr,sr}WaC - Web Corpora of Bosnian, Croatian and Serbian. In: 9th Web as Corpus Workshop (WaC-9), pp 29–35
190. Ljubešić N, Erjavec T (2011) hrWaC and slWaC: compiling web corpora for Croatian and Slovene. In: International conference on text, speech and dialogue. Springer, pp 395–402
191. Ljubesic, N, Dobrovoljc K, Fiser D (2015) MWELex - mwe lexica of croatian, slovene and serbian extracted from parsed corpora, vol 39, pp 293–300
192. Lokesh S, Kumar PM, Devi MR, Parthasarathy P, Gokulnath C (2018) An automatic Tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. Neural Comput Appl 2018:1–11
193. Londhe ND, Kshirsagar GB (2018) Chhattisgarhi speech corpus for research and development in automatic speech recognition. Int J Speech Technol 21(2):193–210
194. Lopez-Moreno I, Gonzalez-Dominguez J, Martinez D, Plchot O, Gonzalez-Rodriguez J, Moreno PJ (2016) On the use of deep feedforward neural networks for automatic language identification. Comput Speech Lang 40:46–59
195. Maas AL, Qi P, Xie Z, Hannun AY, Lengerich CT, Jurafsky D, Ng AY (2017) Building DNN acoustic models for large vocabulary speech recognition. Comput Speech Lang 41:195–213
196. Mandal S, Das B, Mitra P, Basu A (2011) Developing Bengali speech corpus for phone recognizer using optimum text selection technique. In: International conference on asian language processing, pp 268–271, IEEE
197. Mannepalli K, Sastry PN, Suman M (2016) MFCC-GMM based accent recognition system for Telugu speech signals. Int J Speech Technol 19(1):87–93
198. Manohar V, Povey D, Khudanpur S (2015) Semi-supervised maximum mutual information training of deep neural network acoustic models. In: Sixteenth annual conference of the international speech communication association, pp 1–5
199. Mansikkaniemi A, Smit P, Kurimo M (2017) Automatic construction of the Finnish parliament speech corpus. In: Proceedings of the annual conference of the international speech communication association, interspeec, pp 3762–3766
200. Markovnikov N, Kipyatkova I, Lyakso E (2018) End-to-end speech recognition in Russian. In: International conference on speech and computer. Springer, Cham, pp 377–386
201. Johnson M, Lapkin S, Long V (2014) A systematic review of speech recognition technology in health care. BMC Med Inform Decis Mak 14(1):94
202. Martinčić-Ipšić S, Ribarić S, Ipšić I (2008) Acoustic modelling for Croatian speech recognition and synthesis. Informatica 19:227–254
203. Martincic-Ipsic S et al (2009) Automatic evaluation of synthesized speech. In: 2009 international conference on information technology interfaces, pp 305–310
204. Maseri M, Mamat M (2018) Malay language speech recognition for preschool children using hidden markov model (HMM) system training, computational science and technology. Springer, Singapore, pp 205–214
205. Mateju L, Cerva P, Zdansky J (2015) Investigation into the use of deep neural networks for LVCSR of Czech. In: 2015 IEEE international workshop of electronics, control, measurement, signals and their application to mechatronics, pp 1–4
206. McDermott E, Hazen TJ, Le Roux J, Nakamura A, Katagiri S (2007) Discriminative training for large-vocabulary speech recognition using minimum classification error. IEEE Trans Audio Speech Lang Process 15(1):203–223
207. Medennikov I, Prudnikov A (2016) Advances in STC Russian spontaneous speech recognition system lecture notes in computer science, pp 116–123. 101007/978-3-319-43958-7_13
208. Mehta LR, Mahajan SP, Dabhade AS (2013) Comparative study of MFCC and LPC for Marathi isolated word recognition system. Int J Adv Res Electr Electron Instrum Eng 2(6):2133–2139
209. Menacer MA, Mella O, Fohr D, Jouvet D, Langlois D, Smaïli K (2017) Development of the Arabic Loria automatic speech recognition system (ALASR) and its evaluation for Algerian dialect. Procedia Comput Sci 117:81–88
210. Messaoud ZB, Hamida AB (2011) Combining formant frequency based on variable order LPC coding with acoustic features for TIMIT phone recognition. Int J Speech Technol 14(4):393
211. Milde B, Köhn A (2018) Open source automatic speech recognition for German In: Speech communication; 13th ITG-Symposium, pp 1–5
212. Militaru D, Gavat I, Dumitru O, Zaharia T, Segarceanu S (2009) ProtoLOGOS, system for Romanian language automatic speech recognition and understanding (ASRU). In: 5-th conference on speech technology and human-computer dialogue, Constant, pp 1–9
213. Misra DD, Dutta K, Bhattacharjee U, Sarma KK, Goswami PK (2015) Assamese vowel speech recognition using GMM and ANN approaches. In: Recent trends in intelligent and emerging systems. Springer, pp 163–170
214. Mitankin P, Mihov S, Tinchev T (2009) Large vocabulary continuous speech recognition for bulgarian. In: Proceedings of the RANLP, pp 246–250
215. Mohamed A, Nair KR (2012) HMM/ANN hybrid model for continuous Malayalam speech recognition. Procedia Eng 30:616–622
216. Mohamed A, Ramachandran Nair KN (2015) Connectionist approach for emission probability estimation in malayalam continuous speech recognition. In: Mandal J, Satapathy S, Kumar Sanyal M, Sarkar P, Mukhopadhyay A (eds) Information systems design and intelligent applications advances in intelligent systems and computing. Springer, New Delhi, p 339
217. Mohamed AR, Hinton G, Penn G (2012) Understanding how deep belief networks perform acoustic modelling. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4273–4276
218. Mohamed FK, Lajish VL (2016) Nonlinear speech analysis and modeling for Malayalam vowel recognition. Procedia Comput Sci 93:676–682
219. Mohanty R, Swain BK (2010a) Emotion recognition using fuzzy K-means from Oriya speech. In: 2010 for International Conference [ACCTA-2010], pp 3–5
220. Mohanty S, Bhattacharya S (2008) Recognition of voice signals for Oriya language using wavelet neural network. ACM Int J Expert Syst Appl 34(3):2130–2147
221. Mohanty S, Swain BK (2010b) Markov model based Oriya isolated speech recognizer-an emerging solution for visually impaired students in school and public examination. Spec Issue IJCCT 2(2):3
222. Mohanty S, Swain BK (2013) Double ended speech enabled system in Indian travel & tourism industry. In: IEEE international conference on computational intelligence and computing research, pp 1–7
223. Moore AH, Parada PP, Naylor PA (2017) Speech enhancement for robust automatic speech recognition: evaluation using a baseline system and instrumental measures. Comput Speech Lang 46:574–584
224. Moriya T, Tanaka T, Shinozaki T, Watanabe S, Duh K (2015) Automation of system building for state-of-the-art large vocabulary speech recognition using evolution strategy. In: IEEE

workshop on automatic speech recognition and understanding (ASRU), Scottsdale, pp 610–616
225. Mufungulwa G, Tsutsui H, Miyanaga Y, Abe S, Ochi M (2017) Robust speech recognition for similar Japanese pronunciation phrases under noisy conditions. In: International symposium on signals, circuits and systems (ISSCS), pp 1–4
226. Muhammad G, Alotaibi YA, Huda MN (2009) Automatic speech recognition for Bangla digits. In: 12th international conference on computers and information technology, pp 379–383
227. Mustafa MB, Salim SS, Rahman FD (2016) A two-stage adaptation towards automatic speech recognition system for Malay-speaking children. Int J Comput Electr Autom Control Inf Eng 10:3
228. Nahid MMH, Purkaystha B, Islam MS (2017) Bengali speech recognition: a double layered LSTM-RNN approach, 20th International Conference of Computer and Information Technology, pp 1–6, IEEE.
229. Nakamura S (2014) Towards real-time multilingual multimodal speech-to-speech translation, spoken language technologies for under-resourced languages, pp 13–15
230. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K (2019) Speech recognition using deep neural networks: a systematic review. IEEE Access 7:19143–19165
231. Neti C, Rajput N, Verma A (2002) A large vocabulary continuous speech recognition system for Hindi. In: Proceedings of the national conference on communications, Mumbai, pp 366–370
232. Nouza J, Červa P, Jan S (2013) Adding controlled amount of noise to improve recognition of compressed and spectrally distorted speech. Int Conf Acoust Speech Signal Process 1988:8046–8050
233. Nouza J, Cerva P, Zdansky J, Blavka K, Bohac M, Silovsky J, Rott M et al (2014) Speech-to-text technology to transcribe and disclose 100,000+ hours of bilingual documents from historical Czech and Czechoslovak radio archive. In: Fifteenth annual conference of the international speech communication association
234. Nouza J, Cerva P, Zdansky J, Kucharova M (2012) A Study on adapting czech automatic speech recognition system to croatian language. In: Proceedings of the 54th international symposium, Zadar, Croatia, pp 227–230
235. Nouza J, Safarik R, Cerva P (2016) ASR for South Slavic languages developed in almost automated way. In: Interspeech, pp 3868–3872
236. Nouza J, Zdansky J, Cerva P (2010) System for automatic collection, annotation and indexing of Czech broadcast speech with full-text search. In: 15th IEEE MELECON Conference, Malta, pp 202–205
237. Ordelman R, Hessen AV, Jong FD (2003) Compound decomposition in Dutch large vocabulary speech recognition. In: Eighth European conference on speech communication and technology, pp 225–228
238. Pal M, Roy R, Khan S, Bepari MS, Basu J (2018) PannoMulloKathan: voice enabled mobile app for agricultural commodity price dissemination in Bengali language. In: Proc interspeech 2018, pp 1491–1492
239. Pan J, Liu C, Wang Z, Hu Y, Jiang H (2012) Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: why DNN surpasses GMMs in acoustic modeling. In: International symposium on Chinese spoken language processing (ISCSLP), pp 301–305, IEEE
240. Pardeep R, Rao KS (2016) Deep neural networks for kannada phoneme recognition. In: 2016 Ninth international conference on contemporary computing (IC3). https://doi.org/10.1109/ic320167880202
241. Patil HA, Basu TK (2008) Development of speech corpora for speaker recognition research and evaluation in Indian languages. Int J Speech Technol 11(1):17–32
242. Patil SP, Lahudkar SL (2019) Hidden-Markov-model based statistical parametric speech synthesis for Marathi with optimal number of hidden states. Int J Speech Technol 22(1):93–98
243. Paul AK, Das D, Kamal MM (2009) Bangla speech recognition system using LPC and ANN. In: Seventh international conference on advances in pattern recognition, pp 171–174, IEEE
244. Pelemans J, Demuynck K, Wambacq P (2012) A layered approach for dutch large vocabulary continuous speech recognition. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4421–4424
245. Pelemans J, Demuynck K, Wambacq P (2014) Speech recognition web services for Dutch, Proceedings LREC, pp 3041–3044
246. Plahl C, Hoffmeister B, Hwang MY, Lu D, Heigold G, Lo J, Ney H (2008) Recent improvements of the RWTH GALE Mandarin LVCSR system. In: Int conf on speech communication and technology, Brisbane, Australia, pp 2426–2429
247. Plauché M, Prabaker M (2006) Tamil market: a spoken dialog system for rural India. In: Working papers in computer-human interfaces
248. Plauche, M, Nallasamy, U, Pal, J, Wooters, C, Ramachandran D (2006) Speech recognition for illiterate access to information and technology. In: Information and communication technologies and development, 2006 ICTD'06 International Conference on (pp 83–92) IEEE
249. Popli A, Kumar A (2015) Query-by-example spoken term detection using low dimensional posteriorgrams motivated by articulatory classes. In: 17th international workshop on multimedia signal processing, pp 1–6
250. Potapova R, Grigorieva M (2017) Crosslinguistic intelligibility of Russian and German speech in noisy environment. J Electr Comput Eng 2017:5
251. Premkumar MJJ, Vu NT, Schultz T (2013) Experiments towards a better LVCSR System for Tamil. Training 30(33):1012
252. Přibil J, Přibilová A (2013) 2013 Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech. EURASIP J Audio Speech Music Process 1:8
253. Procházka V, Pollak P, Žďánský J, Nouza J (2011) Performance of Czech speech recognition with language models created from public resources. Radioengineering 20(4):1002–1008
254. Prudnikov A, Medennikov I, Mendelev V, Korenevsky M, Khokhlov Y (2015) Improving acoustic models for Russian spontaneous speech recognition. In: International conference on speech and computer. Springer, pp 234–242
255. Pui-Fung W, Man-Hung S (2004) Decision tree based tone modeling for Chinese speech recognition. In: IEEE international conference on acoustics, speech, and signal processing, pp I-905. IEEE
256. Qian Y, Liu J (2010) Phone modeling and combining discriminative training for mandarin english bilingual speech recognition. In: IEEE international conference on acoustics speech and signal processing (ICASSP), pp 4918–4921
257. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
258. Radeck-Arneth S, Milde B, Lange A, Gouvêa E, Radomski S, Mühlhäuser M, Biemann C (2015) Open source german distant speech recognition: Corpus and acoustic model. In: International conference on text, speech, and dialogue, pp 480-488. Springer, Cham
259. Radeck-Arneth S, Milde B, Lange A, Gouvêa E, Radomski S, Mühlhäuse, M, Biemann C (2015b) Open source german distant speech recognition: Corpus and acoustic model. In: International conference on text, speech, and dialogue, Springer, pp 480–488
260. Rahman FD, Mohamed N, Mustafa MB, Salim SS (2014) Automatic speech recognition system for Malay speaking children. In:

Third ICT international student project conference (ICT-ISPC), pp 79–82, IEEE
261. Rahman MM, Khan MF, Moni MA (2010) Speech recognition front-end for segmenting and clustering continuous Bangla speech Daffodil International University. J Sci Technol 5(1):67–72
262. Rajnoha J, Pollak P (2007) Modified feature extraction methods in robust speech recognition. In: 17th international conference Radioelektronika, Brno, pp 1–4
263. Ramamohan S, Dandapat S (2006) Sinusoidal model-based analysis and classification of stressed speech. IEEE Trans Audio Speech Lang Process 14(3):737–746
264. Ranjan S (2010) A discrete wavelet transform based approach to Hindi speech recognition. In: International conference on signal acquisition and processing, pp 345–348
265. Rao KS (2011) Application of prosody models for developing speech systems in Indian languages. Int J Speech Technol 14(1):19–33
266. Ravanelli M, Serdyuk D, Bengio Y (2018) Twin Regularization for online speech recognition. arXiv:180405374
267. Razavi M, Rasipuram R, Magimai-Doss M (2016) Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework. Speech Commun 80:1–21
268. Reza M, Rashid W, Mostakim M (2017) Prodorshok I: a bengali isolated speech dataset for voice-based assistive technologies: a comparative analysis of the effects of data augmentation on HMM-GMM and DNN classifiers, humanitarian technology conference, pp 396–399, IEEE
269. Richardson FS, Campbell WM (2008) Language recognition with discriminative keyword selection. In: IEEE international conference on acoustics, speech and signal processing, pp 4145–4148 IEEE
270. Ronzhin AL, Yusupov RM, Li IV, Leontieva AB (2006) Survey of russian speech recognition systems. In: 11th international conference SPECOM, pp 54–60
271. Russo M, Stella M, Sikora M, Pekić V (2019) Robust Cochlear-model-based speech. Recogn Comput 8(1):5
272. Sailor HB, Patil HA (2016) Novel unsupervised auditory Filterbank learning using convolutional RBM for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 24(12):2341–2353
273. Sainath TN, Kingsbury B, Saon G, Soltau H, Mohamed AR, Dahl G, Ramabhadran B (2015) Deep convolutional neural networks for large-scale speech tasks. Neural Netw 64:39–48
274. Sajjan SC, Vijaya C (2016) Continuous Speech recognition of Kannada language using triphone modeling. Int Conf Wirel Commun Signal Process Netw. https://doi.org/10.1109/wispnet20167566174
275. Sajjan SC, Vijaya C (2018) Kannada speech recognition using decision tree based clustering. In: Shetty N, Patnaik L, Prasad N, Nalini N (eds) Emerging research in computing, information, communication and applications ERCICA 2016. Springer, Singapore
276. Sak H, Saraçlar M, Gungor T (2012) Morpholexical and discriminative language models for Turkish automatic speech recognition. IEEE Trans Audio Speech Lang Process 20(8):2341–2351
277. Sakti S, Paul M, Finch A, Sakai S, Vu TT, Kimura N, Hori C, Sumita E, Nakamura S, Park J, Wutiwiwatchai C (2013) A-STAR: toward translating Asian spoken languages. Comput Speech Lang 27(2):509–527
278. Samsudin NH, Kong TE (2004) A Simple Malay speech synthesizer using syllable concatenation approach, MMU international symposium on information and communications technologies, pp 1–4
279. Saon G, Chien J (2012) Bayesian sensing Hidden Markov models. IEEE Trans Audio Speech Lang Process 20(1):43–54
280. Saraswathi S, Geetha TV (2007) Comparison of performance of enhanced morpheme-based language model with different word-based language models for improving the performance of tamil speech recognition system. ACM Trans Asian Lang Inf Process 6(3):9
281. Saraswathi S, Geetha TV (2010) Design of language models at various phases of Tamil speech recognition system. Int J Eng Sci Technol 2(5):244–257
282. Sarma H, Saharia N, Sharma U (2017) Development and analysis of speech recognition systems for Assamese language Using HTK. ACM Trans Asian Low-Resourc Lang Inf Process 17(1):7
283. Sarma M, Dutta K, Sarma KK (2009) Assamese numeral corpus for speech recognition using cooperative ANN architecture. Int J Electr Electron Eng 3(8):456–465
284. Sarma M, Sarma KK, Goel NK (2018) Language recognition using time delay deep neural network. arXiv:180405000
285. Sarma MP, Sarma KK (2011) Assamese numeral speech recognition using multiple features and cooperative LVQ-architectures. Int J Electr Electron 5:1
286. Satori H, Harti M, Chenfour N (2007) Introduction to Arabic speech recognition using CMUSphinx system. arXiv:07042083
287. Savchenko AV (2013) Phonetic words decoding software in the problem of Russian speech recognition. Autom Rem Control 74(7):1225–1232
288. Scharenborg O, Besacier C, Black A, Hasegawa-Johnson M, Metze F, Neubig G, Stuker S, Godard MM, Lucas O et al (2018) Linguistic unit discovery from multi-modal inputs in unwritten languages: summary of the Speaking Rosetta JSALT 2017 workshop. arXiv:1802.05092
289. Schultz T, Alexander D, Black AW, Peterson K, Suebvisai S, Waibel A (2004) A Thai speech translation system for medical dialogs. In: Demonstration papers at HLT-NAACL association for computational linguistics, pp 34–35
290. Seide F, Li G, and Yu D, 2011 Conversational speech transcription using context-dependent deep neural networks, Twelfth annual conference of the international speech communication association, 430–440
291. Seki H, Watanabe S, Hori T, Roux JL, Hershey JR (2018) An end-to-end language-tracking speech recognizer for mixed-language speech. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), Calgary, AB, pp 4919–4923
292. Seljan S, Dunder I (2014) Combined automatic speech recognition and machine translation in business correspondence domain for English–Croatian. Int J Ind Syst Eng 8(11):1980–1986
293. Seltzer ML, Yu D, Wang Y (2013) An investigation of deep neural networks for noise robust speech recognition. In: IEEE international conference on acoustics, speech and signal processing, pp 7398–7402
294. Seman N, Jusoff K (2008a) Automatic segmentation and labeling for spontaneous standard Malay speech recognition. In: International conference on advanced computer theory and engineering, 59–63, IEEE
295. Seman N, Jusoff K (2008b) Acoustic pronunciation variations modeling for standard Malay speech recognition. Comput Inf Sci 1(4):112
296. Seman N, Bakar ZA, Bakar NA (2010) An evaluation of endpoint detection measures for malay speech recognition of an isolated words. In: International symposium in information technology (ITSim), vol 3, pp 1628–1635, IEEE
297. Sertsi P, Chunwijitra V, Chunwijitra S, Wutiwiwatchai C (2016) Offline Thai speech recognition framework on mobile device. In: International joint conference on computer science and software engineering (JCSSE), pp 1–5
298. Shahnawazuddin S, Deepak KT, Sarma BD, Deka A, Prasanna SM, Sinha R (2015) Mannepalli. J Signal Process Syst 81(1):83–97

299. Shanthi Therese S, Lingam C (2013) Review of feature extraction techniques in automatic speech recognition. Int J Sci Eng Technol 2(6):479–484
300. Sharma A, Shrotriya MC, Farooq O, Abbasi ZA (2008) Hybrid wavelet based LPC features for Hindi speech recognition. Int J Inf Commun Technol 1(3–4):373–381
301. Shi Y, Wiggers P, Jonker CM (2012) Towards recurrent neural networks language models with linguistic and contextual features. In: Thirteenth annual conference of the international speech communication association, pp 1664–1667
302. Shi Y, Larson M, Pelemans J, Jonker CM, Wambacq P, Wiggers P, Demuynck K (2015) Integrating meta-information into recurrent neural network language models. Speech Commun 73:64–80
303. Shimizu T, Ashikari Y, Sumita E, Zhang J, Nakamura S (2008) NICT/ATR Chinese-Japanese-English speech-to-speech translation system. Tsinghua Sci Technol 13(4):540–544
304. Shrishrimal PP, Deshmukh RR, Waghmare VB (2012) Indian language speech database: a review. Int J Comput Appl 47(5):17–21
305. Siivola V, Hirsimäki T, Virpioja S (2007) On growing and pruning Kneser-Ney smoothed n-gram models. IEEE Trans Audio Speech Lang Process 15(5):1617–1624
306. Singh A, Kadyan V, Kumar M, Bassan N (2019) ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. Springer, Berlin
307. Siniscalchi SM, Lyu D, Svendsen T, Lee C (2012) Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data. IEEE Trans Audio Speech Lang Process 20(3):875–887
308. Sivaranjani C, Bharathi B (2016) Syllable based continuous speech recognition for tamil language. Int J Adv Eng Tech 7(1):4
309. Smirnov V, Ignatov D, Gusev M, Farkhadov M, Rumyantseva N, Farkhadova M (2016) A Russian keyword spotting system based on large vocabulary continuous speech recognition and linguistic knowledge. J Electr Comput Eng. https://doi.org/10.1155/2016/4062786
310. Smit P, Virpioja S, Kurimo M (2017) Improved subword modeling for WFST-based speech recognition. In: INTERSPEECH 2017—18th annual conference of the international speech communication Association. Stockholm, Sweden.
311. Spille C, Ewert SD, Kollmeier B, Meyer BT (2018) Predicting speech intelligibility with deep neural networks. Comput Speech Lang 48:51–66
312. Srijiranon K, Eiamkanitchat N (2015) Thai speech recognition using Neuro-fuzzy system. In: 12th international conference on electrical engineering/electronics, computer, telecommunications and information technology, pp 1–6, IEEE
313. Srisuwan N, Phukpattaranont P, Limsakul C (2012) Feature selection for Thai tone classification based on surface EMG. Procedia Eng 32:253–259
314. Stüker S, Schultz T (2004) A grapheme based speech recognition system for Russian. In: 9th Conference Speech and Computer, pp 1–7
315. Šturm P, Volín J (2016) P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment. J Phon 55:38–52
316. Šturm P (2018) Experimental evidence on the syllabification of two-consonant clusters in Czech. J Phon 71:126–146
317. Suebvisai S, Charoenpornsawat P, Black A, Woszczyna M, Schultz T (2005) Thai automatic speech recognition. In: IEEE international conference on acoustics, speech, and signal processing, pp I-857, IEEE
318. Swietojanski P, Ghoshal A, Renals S (2014) Convolutional neural networks for distant speech recognition. IEEE Signal Process Lett 21(9):1120–1124
319. Tadić M, Fulgosi S (2003) Building the Croatian morphological lexicon. In: Workshop on morphological processing of Slavic Languages, association for computational linguistics, pp 41–46
320. Takiguchi T, Ariki Y (2007) PCA-based speech enhancement for distorted speech recognition. J Multimed 2(5):13–18
321. Tantibundhit C, Onsuwan C, Munthuli A, Sirimujalin P, Anansiripinyo T, Phuechpanpaisal S, Wright N, Kosawat K (2018) Development of a Thai phonetically balanced monosyllabic word recognition test: derivation of phoneme distribution, word list construction, and response evaluations. Speech Commun 103:1–10
322. Tantisatirapong S, Prasoproek C, Phothisonothai M (2018) Comparison of feature extraction for accent dependent Thai speech recognition system IEEE seventh international conference on communications and electronic, pp 322–325
323. Telmem M, Ghanou Y (2018) Estimation of the optimal HMM parameters for amazigh speech recognition system using CMU-Sphinx. Procedia Comput Sci 127:92–101
324. Thangthai K, Chotimongkol A, Wutiwiwatchai C (2013) A hybrid language model for open-vocabulary Thai LVCSR. Interspeech, pp 2207–2211
325. Theera-Umpon N, Chansareewittaya S, Auephanwiriyakul S (2011) Phoneme and tonal accent recognition for Thai speech. Expert Syst Appl 38(10):13254–13259
326. Thennattil JJ, Mary L (2016) Phonetic engine for continuous speech in malayalam. IETE J Res 62(5):679–685
327. Thomas S, Hermansky H (2009) Modulation frequency features for phoneme recognition in noisy speech. J Acoust Soc Am 125(1):8–12
328. Ting HN, Yunus J (2004) Speaker-independent Malay vowel recognition of children using multi-layer perceptron. In: IEEE Region 10 Conference, pp 68–71
329. Ting HN, Zourmand A, Chia SY, Yong BF, Hamid BA (2012) Formant frequencies of Malay vowels produced by Malay children aged between 7 and 12 years. J Voice 26(5):664-e1
330. Tong S, Garner PN, Bourlard H (2018a) Cross-lingual adaptation of a CTC-based multilingual acoustic model. Speech Commun 104:39–46
331. Tong S, Garner PN, Bourlard H (2018b) Cross-lingual adaptation of a CTC-based multilingual acoustic model. Speech Commun. https://doi.org/10.1016/jspecom201809001
332. Torres-Carrasquillo PA, Richardson F, Nercessian S, Sturim D, Campbell W, Gwon Y, Vattam S, Dehak N, Mallidi H, Nidadavolu PS, Li R (2017) The MIT-LL, JHU and LRDE NIST 2016 speaker recognition evaluation system, Interspeech, pp 1333–1337
333. Tufekci Z, Gowdy JN, Gurbuz S, Patterson E (2006) Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition. Speech Commun 48(10):1294–1307
334. Tufiș D, Dan C (2018) A Bird's-eye view of language processing projects at the Romanian academy. In: Eleventh international conference on language resources and evaluation, pp 2446–56
335. Turunen VT, Kurimo M (2007) Indexing confusion networks for morph-based spoken document retrieval. In: 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 631–638
336. Tuske Z, Nolden D, Schluter R, Ney H (2013) Multilingual hierarchical MRASTA features for ASR, INTERSPEECH-2013, pp 2222–2226
337. Tuske Z, Nolden D, Schluter R, Ney H (2014) Multilingual MRASTA features for low-resource keyword search and speech recognition systems. In: IEEE international conference on acoustics, speech and signal processing, pp 7854–58
338. Upadhyaya P, Farooq O, Abidi MR (2018) Block energy based visual features using histogram of oriented gradient for bimodal hindi speech recognition. Procedia Comput Sci 132:1385–1393

339. Valente F, Doss MM, Plahl C, Ravuri S, Wang W (2010) Comparative large scale study of MLP features for mandarin ASR,Interspeech'10, Brisbane, Australia, pp 2630–2633
340. Varjokallio M, Virpioja S, Kurimo M (2018) First-pass techniques for very large vocabulary speech recognition FF morphologically rich languages 2018 IEEE spoken language technology workshop, pp 227–234
341. Vazhenina D, Markov K (2011) Phoneme set selection for Russian speech recognition. In: 7th international conference on natural language processing and knowledge engineering, pp 475–478, IEEE
342. Vegesna VVR, Gurugubelli K, Vuppala AK (2018) Application of emotion recognition and modification for emotional telugu speech recognition mobile networks and applications, pp 1–9
343. Veisi H, Sameti H (2011) The integration of principal component analysis and cepstral mean subtraction in parallel model combination for robust speech recognition. Dig Signal Process 21(1):36–53
344. Venail F, Legris E, Vaerenberg B, Puel JL, Govaerts PJ, Ceccato JC (2016) Validation of the French-language version of the OTOSPEECH automated scoring software package for speech audiometry. Eur Ann Otorhinolaryngol Head Neck Dis 133(2):101–106
345. Venkateswarlu RLK, Teja RR, Kumari RV (2012) Developing efficient speech recognition system for Telugu letter recognition. In: 2012 international conference on computing, communication and applications, Dindigul, Tamilnadu, pp 1–6
346. Vergyri D, Kirchhoff K, Duh K, Stolcke A (2004) Morphology-based language modeling for Arabic speech recognition SRI International Menlo Park United States, pp 1–4
347. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: Resources, features, and methods. Speech Commun 48(9):1162–1181
348. Viszlay P, Juhár J, Pleva M (2012) Alternative phonetic class definition in linear discriminant analysis of speech. In: 19th international conference on systems, signals and image processing, pp 637–640, IEEE
349. Waghmare VB, Deshmukh RR, Shrishrimal PP, Janvale GB (2014) Emotion recognition system from artificial marathi speech using MFCC and LDA techniques. In: Fifth international conference on advances in communication, network, and computing–CNC, pp 1–9
350. Wand M, Toth A, Jou SC, Schultz T (2009) Interspeech, Brighton, United Kingdom. In: 2009 impact of different speaking modes on EMG-based speech recognition, pp 648–651
351. Wang L, Tong R, Leung C, Sivadas S, Ni C, Ma S (2017) Cloud-based automatic speech recognition systems for Southeast Asian Languages. In: International conference on orange technologies, pp 147–150
352. Wang W, Mandal A, Lei X, Stolcke A, Zheng J (2008) Multifactor adaptation for mandarin broadcast news and conversation speech recognition. Interspeech 2008:2103–2102
353. Kadyan V, Shanawazuddin S, Singh A (2021) Developing children's speech recognition system for low resource Punjabi language. Appl Acoust 2021:178
354. Wani P, Patil UG, Bormane DS, Shirbahadurkar SD (2016) Automatic speech recognition of isolated words in Hindi language. In: International conference on computing communication control and automation, pp 1–6, IEEE
355. Watanabe S, Hori T, Hershey JR (2017) Language independent end-to-end architecture for joint language identification and speech recognition. In: IEEE automatic speech recognition and understanding workshop (ASRU), Okinawa, pp 265–271
356. Weng C, Yu D, Watanabe S, Juang BHF (2014) Recurrent deep neural networks for robust speech recognition. In: IEEE international conference on acoustics, speech and signal processing, pp 5532–5536, IEEE
357. Weninger F, Schuller B, Eyben F, Wöllmer M, Rigoll G (2014) A broadcast news corpus for evaluation and tuning of german LVCSR systems. arXiv:14124616
358. Yadava TG, Jayanna HS (2017) A spoken query system for the agricultural commodity prices and weather information access in Kannada language. Int J Speech Technol 20(3):635–644
359. Yang D, Pan Y, Furui S (2012) Vocabulary expansion through automatic abbreviation generation for Chinese voice search. Comput Speech Lang 26(5):321–335
360. Yanzhou M, Mianzhu Y (2014) Russian speech recognition system design based on HMM. In: International conference on logistics, engineering, management and computer science, pp 377–380
361. Yi J, Tao J, Bai Y (2019) Language-invariant Bottleneck features from adversarial end-to-end acoustic models for low resource speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), UK, pp 6071–6075
362. Yi J, Tao J, Wen Z, Bai Y (2018) Language-adversarial Transfer Learning for Low-resource Speech Recognition. IEEE/ACM Trans Audio Speech Lang Process 2018:1–1
363. Zarrouk E, BenAyed Y, Gargouri F (2015) Graphical models for multi-dialect arabic isolated words recognition. Procedia Comput Sci 60:508–516
364. Zhang Y, Pezeshki M, Brakel P, Zhang S, Bengio CLY, Courville A (2017) Towards end-to-end speech recognition with deep convolutional neural networks. arXiv:170102720
365. Zhang Goyal K, Singh A, Kadyan V (2021) A comparison of Laryngeal effect in the dialects of Punjabi language. J Ambient Intell Hum Comput 2021:1–14
366. Zou W, Jiang D, Zhao S, Li X (2018) A comparable study of modeling units for end-to-end Mandarin speech recognition. arXiv:180503832