Soft Computing (2021) 25:1617–1630
https://doi.org/10.1007/s00500-020-05248-1

METHODOLOGIES AND APPLICATION

AutoSSR: an efficient approach for automatic spontaneous speech recognition model for the Punjabi Language

Yogesh Kumar[1] · Navdeep Singh[2] · Munish Kumar[3] · Amitoj Singh[3]

Published online: 10 August 2020

Abstract
In this article, the authors have presented the design and development of automatic spontaneous speech recognition of the Punjabi language. To dimensions up to the natural speech recognizer, the very large vocabulary Punjabi text corpus has been taken from a Punjabi interview's speech corpus, presentations, etc. Afterward, the Punjabi text corpus has been cleaned by using the proposed corpus optimization algorithm. The proposed automatic spontaneous speech model has been trained with 13,218 of Punjabi words and more than 200 min of recorded speech. The research work also confirmed that the 2,073,456 unique in-word Punjabi tri-phoneme combinations present in the dictionary comprise of 131 phonemes. The performance of the proposed model has grown increasingly to 87.10% sentence-level accuracy for 2381 Punjabi trained sentences and word-level accuracy of 94.19% for 13,218 Punjabi words. Simultaneously, the word error rate has been reduced to 5.8% for 13,218 Punjabi words. The performance of the proposed system has also been tested by using other parameters such as overall likelihood per frame and convergence ratio on various iterations for different Gaussian mixtures.

Keywords Gaussian mixtures · MFCC · Recognition accuracy · Spontaneous speech · Acoustic model

1 Introduction

From human prehistory to the new media of the future, speech communication has been and will be the leading mode of human social bonding and information exchange. In addition to human communication, the human inclination for spoken language communication finds a replication in human–machine interaction as well. The area of automatic speech recognition (ASR) exploded in the last decades since people incline to be more and more occupied and look after hands-free and eyes-free alternative medium to devices. ASR is the process of taking a sound of dialogue as an input, recorded by a microphone, a phone, etc., and renovates it into transcription as close as likely to the actual spoken data (Akyildiz et al. 2002). Different human beings have a different style of speaking and environmental disturbances that are the main difficulties for implementation of an ASR system. The transfer of the speech signal into a text message is the main task of the ASR system. It is performed by independent of the device, speaker or the environments in a precise and competent way. There are lots of changes needed to be done in spontaneous speech and have the need to record every spoken word by spontaneous speech recognition (Stouten et al. 2006). Since user's sound is typically natural and unplanned, they are demonstrated by replication, artificial initiate, incomplete words and terminate in the core, restarts or there is a further linguistic existence such as cough. A microphone is used by human beings to convert articulate them into vocalizations identification method that comes under Punjabi

Communicated by V. Loia.

✉ Munish Kumar
munishcse@gmail.com

Yogesh Kumar
yogesh.arora10744@gmail.com

Navdeep Singh
navdeep_jaggi@yahoo.com

Amitoj Singh
amitoj.ptu@gmail.com

[1] Department of Computer Science and Engineering, Chandigarh Group of Colleges, Landran (Mohali), Punjab, India

[2] Department of Computer Science, Mata Gujri College, Fatehgarh Sahib, Punjab, India

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

language spontaneous speech recognition scheme. Then a speech is obtained and understand using a computer and to get fine articulation or accurate output. The representation divergence makes use of speech recognition complicated for spontaneous speech. The sounds are usually impulsive and non-planned and are usually labeled by replications, preservation, false start, incomplete words and unplanned words, silence gap, etc. in case of spontaneous speech (Saini and Kaur 2013). In this article, the focus is on the implementation of the live speech model of spontaneous speech for the recognition of the Punjabi language. The graphical user interface for automatic spontaneous speech model for the Punjabi language has also been created and tested for the Punjabi text database (Ankita and Kawahara 2010).

1.1 Novelty in this article

The field of automatic speech recognition (ASR) has been explored in the last decades as people tend to be busier and has been looking for hands-free and eyes-free interfaces to devices. The primary concern of the ASR is to record an acoustic signal of dialogue and regulate the words that were articulated by pattern matching. To do this, a set of acoustic and linguistic models have to be stored in a computer database that signifies the definite outlines. These language models are then compared with captured signals. Our contribution in the field of ASR system is to the development of automatic spontaneous live speech recognition system for the Punjabi language. We have explored many ASR systems which are available in Indian and non-Indian languages for recognizing isolated, connected, continues and spontaneous speech. Based on the prior study, we come to conclude that there is a scope to the development automatic speech model which recognizes the spontaneous Punjabi live speech. The developed spontaneous Punjabi live speech model is independent of the speaker. The model trained with very large vocabulary Punjabi text corpus and tested in both noisy and noise-free environments. The developed spontaneous speech model for Punjabi language is independent and has also been tested using various performance parameters and compared with other languages speech recognition system.

2 Related work

The progress in speech recognition research permits to build a spoken dialogue system which is not constrained by the rules describing the expected behavior of the user (Ali et al. 2015). Various researchers have presented the work in the development of automatic speech recognition system for different languages as shown in Table 1.

3 Design and implementation of spontaneous speech recognition system for Punjabi language

In this section of the paper, the design and implementation of automatic spontaneous speech recognition system for Punjabi are discussed. The section elaborates about how to train the system to build up an acoustic model for the Punjabi language. The corpus optimization algorithm is also presented to optimize the Punjabi speech corpus for the adoption and utilization of the Punjabi corpus. The automatic spontaneous live speech recognition system for the Punjabi language has been implemented in the article. In the output of the language model, word error rate and sentence error rate for Punjabi speech corpus have also been computed (Hofmann et al. 2010; Izzad et al. 2013).

3.1 Automatic spontaneous Punjabi speech model development

So far, no work has been done for spontaneous speech recognition of Punjabi language. The primary purpose of the proposed system is to the development of an automatic spontaneous speech recognition system for Punjabi language. To implement automatic spontaneous Punjabi speech recognizer, the various modeling techniques for ASR which have been used for building other Indian languages and given in the literature have been studied (Abushariah et al. 2010; Braathen et al. 2002). The software simulation tools and different mapping techniques of data preparation for the Punjabi language have been explored (Hoesen et al. 2016). The followings are the steps to utilize the Punjabi speech corpus:

- To design a speech dataset for Punjabi spontaneous speech (read speech corpus).
- To employ the speech corpora to develop a speaker specific AutoSSR system for the Punjabi language.
- To analyze the proposed system for accuracy based on standard parameters for speech recognition. The proposed system design for the spontaneous Punjabi speech model development is depicted in Fig. 1.

3.1.1 Development of the Punjabi speech corpus

The Punjabi text dataset has been read out and note down in organized environmental circumstance.

- The proposed corpus optimization algorithm is applied to recorded speech to optimize and utilize the Punjabi speech corpus.
- Feature vectors are generated by applying the MFCC feature extraction mechanism (Kalaivani 2013).

Table 1 Reported work in the field of automatic speech recognition system

Authors	Target language	Reported work
Hernandez-Mena et al. (2017)	Mexican Spanish	The authors worked on CIEMPIESS corpus for spontaneous speech recognition in Mexican Spanish. The language model composed by 1,505,491 words extracted from the 2489 newsletters of university. The phonetic dictionary is having the 53,169 unique words
Vijayendra and Thakar (2016)	Gujarati speech model	The authors have used the Mel-frequency cepstral coefficients (MFCCs) and real cepstral coefficient (RC) for feature extraction process of the Gujarati words. Different algorithm has also used for removing unwanted silence from the recorded speech
Menacer et al. (2017)	Algerian dialect	They proposed the development of an ASR system for Algerian dialect for modern standard Arabic. The acoustic model of the target language is based on a DNN method and extracted the n-gram features. The model has been tested and performed the improvement for word error rate of 24% for recorded speech because no transcribed speech data are available for Algerian dialect
Vimala and Radha (2012)	Tamil language	They have presented the proposed work on isolated speaker-independent speech recognition system for Tamil language. HMM model has been used for the implementation of the presented work. Computed word accuracy is 88% for trained and tested speech by different speakers
Sarma et al. (2017)	Assamese language	The authors developed the speech system for Assamese language using HTK toolkit based on 20 h of speech. The deep neural network is used for classification purpose the reported accuracy is 78.05%
Singh et al. (2016)	Manipuri language	The authors proposed the language-independent rule-based formulation using phonetic segmentation based on entropy for Manipuri language. The reported output in terms of accuracy is 96%
Lokesh et al. (2019)	Tamil language	The authors presented the work in the field of automatic Tamil speech recognition system using deep learning-based recurrent neural network. The proposed method reported the accuracy is 93.6%
Ali et al. (2015)	Urdu digits	The authors investigated three classifications approaches for Urdu speech recognition system. The reported accuracy is 73% when classified by using SVM and trained by using MFCCs and 63% when trained by using random forest and tested by using linear discriminant analysis (LDA)
Sajjan and Vijaya (2016)	Kannada language	The authors presented the work in development of continuous speech recognition system for Kannada language. The speech system is having 46 phonemes in which 12 phonemes represented the vowels and trained by using MFCCs
Yu et al. (2019)	Tujia language and Chinese phonetic	The authors proposed work for low-resource Tujia language based on Chinese corpus. The CNN model has been used to extract the cross-linguistic features for automatic speech recognition model for Tujia language. The presented work reported the 46.19% recognition error rate
Patil et al. (2016)	Hindi language	The authors presented the isolated speech recognition system for Hindi language. The presented system accepted the voice from computer and transcribed text conversion of it. The MFCCs have been used for training and KNN machine learning model used for classification. The achieved accuracy for isolated Hindi speech model is 94.31%

- The training files are created and configured to build the Punjabi acoustic speech model.
- Linguistic modeling is used to decode the acoustic model for spontaneous Punjabi speech.

3.1.2 Punjabi text corpus

The first step in automatic spontaneous speech recognition system is the acquisition of the Punjabi text corpus from various resources such as telephonic conversation, interview corpus, presentations, newspaper and Punjabi university Web site. The main objective of the development of a phonetically rich Punjabi text corpus is to make sure that it signifies all sounds that arise in the Punjabi language. This allows in providing a baseline for training an automatic spontaneous speech recognition system. In spontaneous speech, the system must be trained for the speech in which the spoken words are not planned. Spontaneous speech is more realistic than planned speech. Different types of speakers have different ways of speaking style depending on collective attributes such as a different age-groups, genders and cultures. The phonetic lexicon consisted of 13,218 trained Punjabi words which are taken from various resources. This was done to guarantee that many common Punjabi words are present in the dictionary.

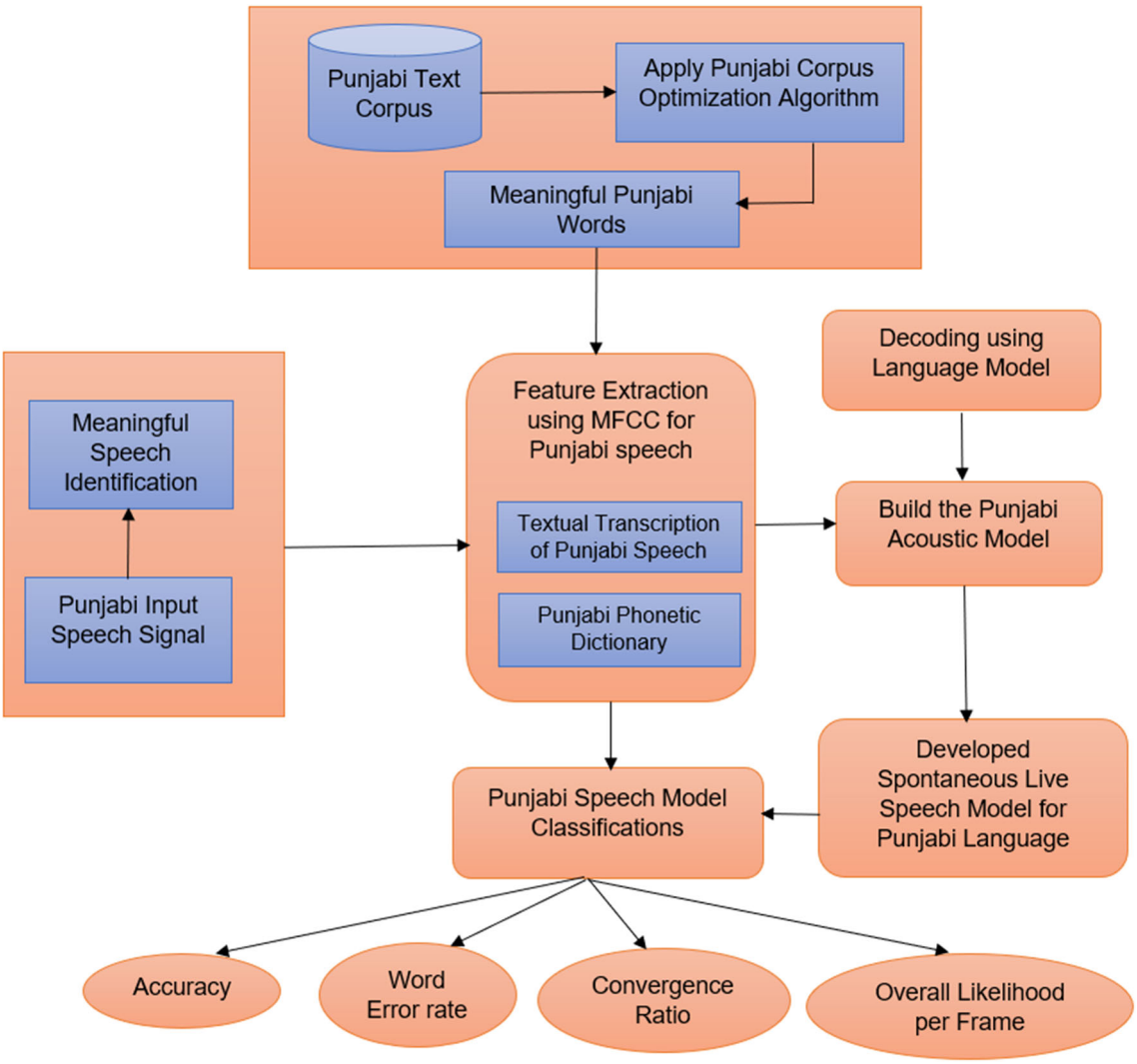

Fig. 1 Framework for the research work

The following are the training requirements for speech recognition process (Table 2).

3.1.3 Punjabi corpus optimization algorithm

The corpus optimization algorithm is applied to produce the list that comprises each Punjabi word tri-phonemes. So, we need to create tri-phone set using the formula:

$$\text{Total Tri} - \text{phone} = ph1 \times pht \times ph2$$

where ph1 is the first the Punjabi character in the Punjabi word and ph2 is the last character where pht (target phoneme) is present between all the occurrences of Punjabi words.

Table 2 Punjabi speech training requirements

Sr. no.	Process	Description
1	Speaker	Male voice
2	Recording tools	Microphone smart voice recorder, wave recording software
3	Sampling rate of the audio	16 kHz
4	Bit rate	16
5	Channel	Mono
6	Feature computational standard	39 double delta MFC coefficient

3.1.4 MFCC file for feature extraction of Punjabi acoustic spontaneous speech model

Mel-frequency cepstral coefficient (MFCC) features are extracted from the speech signals for training and recognition of the Punjabi spontaneous model (Fohr et al. 2017; Hoesen et al. 2016). The MFCCs store the features in binary format with the extension.mfc.

- if there are N characters in our Punjabi.phone file, then total possible words in Punjabi.dic file must be $N \times N \times N$.
- In the present work, we have 132 characters including 1 silent character in Punjabi.phone file.
- Based on the meaningful 131 phonemes, there can be an entire of 2,299,968 possible tri-phoneme combinations presented in the work (counting silence as a phoneme). In order to search the tri-phoneme combinations that actually are present in the Punjabi language, the phonetic dictionary is analyzed for tri-phonemes and their frequency of occurrence. The analysis is to search all the unique tri-phonemes that arise within Punjabi words. It is assumed that all words are pursued and preceded by silence which is dealt as a separate phoneme signified by the symbol #. The analysis showed that the dataset contains total 2,073,456 unique in-word Punjabi tri-phonemes.
- In order to figure out the Punjabi word set, a Punjabi corpus optimization approach is used. The primary objective of the algorithm is to allow the utility in the automatic spontaneous speech recognition system (Dahl et al. 2012) where the Punjabi word list utilized for training of the dialogue model contains a minimal Punjabi word set that can exploit the number of tri-phonemes Digalakis (2003a). Figure 2 presents the algorithm for corpus optimization.

3.1.5 Steps for training the acoustic model for Punjabi language

Acoustic model for Punjabi language has been created by using the Sphinx toolkit and Java programming language. The file structure hierarchy is given in Fig. 3 for the creation of acoustic model in the Punjabi language.

- *Dictionary File for Punjabi corpus* (Independent words are stored in it) Dictionary files map every word to a sequence of sound units, associated with each signal. And in filler dictionary rejected noise is stored in it.
- *Phone file for Punjabi corpus* Phone file is a record of individual sound units that need to make a word. It is actually representing the different sound units that we have in the Punjabi language.
- *Transcript and field file (path of wav files)* Transcription file (Punjabi_train.transcription and Punjabi_test.transcription) is a text file listing the transcription for each audio file.
- *Sphinx_Train.Test file* This is the configuration file where we need to configure the path for all required files (for field, transcript, etc.).

3.1.6 Language model for Punjabi speech corpus

Word-level construction and verbal communication are presented by linguist language model section. On behalf of

// Punjabi Corpus Optimization Algorithm

Step1: XPun is the input Punjabi word corpus in the trained Punjabi text corpus UPun.

Step2: WPun is the Punjabi word which is not present in the Punjabi trained words dictionary.

Step3: OPun is the output dictionary of the words.

Step 4: Select Punjabi word (WPun) from total trained Punjabi word corpus (UPun)

Step 5: while the Upun (trained Punjabi text corpus) is not empty

Step6: Apply the condition to check the given input Punjabi word is present in the output dictionary (OPun) else need to trained the word.

Step 7: if given input Punjabi word (WPun) is not present in the output dictionary (OPun)

Step 8: Add the input word in output dictionary.

Step 9: Finally, output dictionary (OPun) which have only meaningful words after checking from the trained Punjabi word corpus. Because the major problems are the Punjabi wordlist have repetition of words.

Fig. 2 Punjabi corpus optimization algorithm

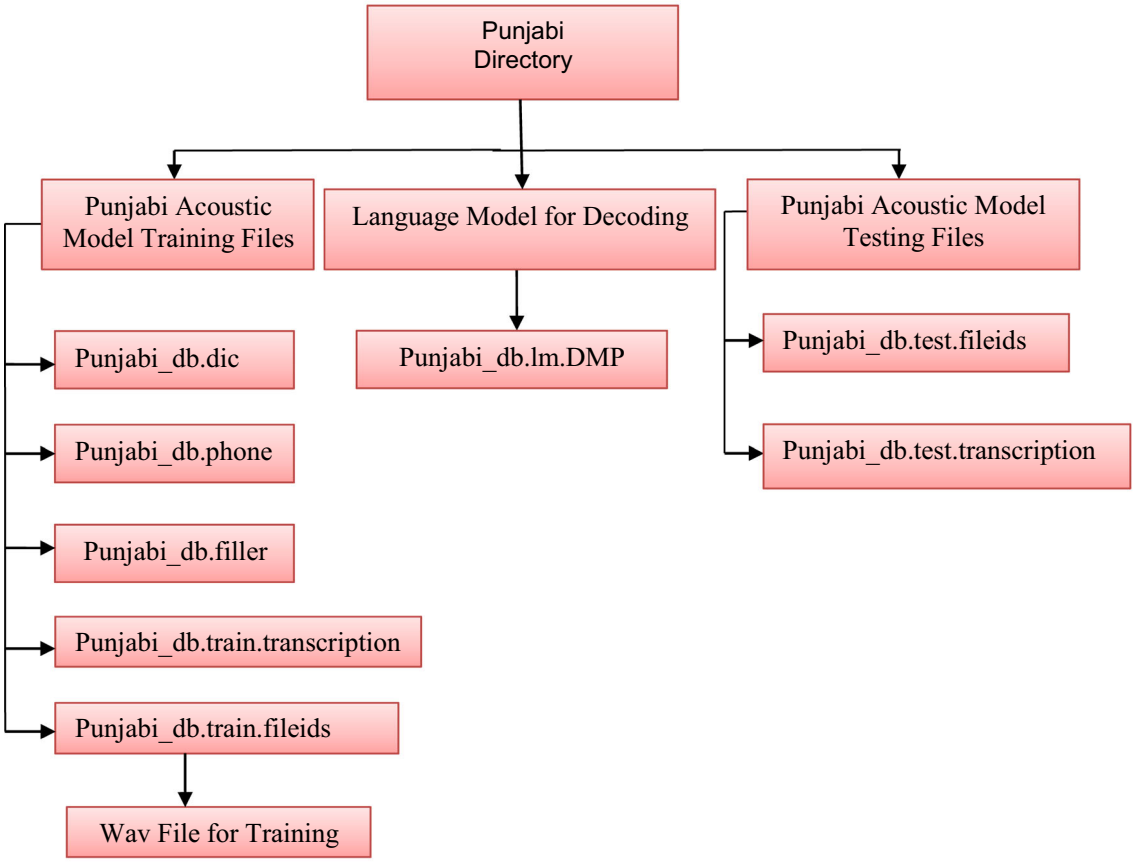

Fig. 3 Training and testing of the Punjabi acoustic model

articulation, a series of acoustic units are mapped using language model (Beke and Gosy 2012; Furui 2003). Decoding is the method to compute a series of Punjabi words that are expected to match the acoustic signal signified by the feature vectors. An acoustic model which is built using the Sphinx toolkit for each set of Punjabi word (phoneme or word), Punjabi words dictionary and language model with a set of Punjabi text corpus or word sequence likelihood has three information sources (Kumar and Singh 2016; Hendy and Farag 2013). Steps followed to create the language model for the Punjabi language are as follows:

Step I Input the Punjabi datasets to create the language model.
Step II Generate the Punjabi vocab file for Punjabi dataset.
Step III Lastly, the linguistic model is formed with extension lm.dmp, and further it applies for decoding purpose.

4 Testing the automatic spontaneous Punjabi recognition system

The first step is to use the Sphinx platform and Java language to develop the spontaneous Punjabi live speech model for testing purpose. Spontaneous Punjabi speech recognition model generates the transcriptions from the voice of the speaker. The voice which is more relevant at the time of recognition will be the outcome of the live Punjabi speech system. The proposed automatic spontaneous speech recognition model has also an option to construct the document while speaking and further save it into the document file. The performance of the proposed automatic spontaneous speech recognizer for the Punjabi language is evaluated by using various parameters such as recognition accuracy, word error rate, sentence error rate, convergence ratio and overall likelihood per frame.

4.1 Automatic spontaneous speech recognizer for Punjabi live speech

The language model and Punjabi trained set corpus have been used to build the final user interface for automatic

spontaneous live speech recognition system. The user interface of spontaneous live speech system has two distinct approaches for recognition of Punjabi speech. In the first approach, we have live speech test option, and secondly, we have a wave file test method in which we can also browse the wave file for Punjabi speech and then recognize it. We also have an option to utilize the recognized Punjabi speech corpus by saving into the file for future use.

Figure 4 shows the GUI of the spontaneous Punjabi speech model. In the Punjabi speech model, the output window is divided into three different sections. In the first section, categories of the recognized Punjabi words are displayed. When we click on the category, the Punjabi text corpus will be displayed in the middle section of the output window. In the third section, we have two different ways to recognize the speech. The first one is a live speech test option, and the second one is a wave file test.

4.1.1 Case 1: Live speech test

In this method, firstly we set the live speech test option and click on the start recognizer button and then use the mike to speak.

Figure 5 shows the output of the Punjabi live speech model that recognizes the Punjabi live speech for different categories of speech corpus.

4.1.2 Case 2: Wav file test

In the wav file test option, we browse the Punjabi wav file stored in the system and then click on the start recognizer button. It will recognize and transcribe the Punjabi speech corpus.

Figure 6 shows the output in the form of a transcription of the Punjabi spontaneous speech model by browsing the Punjabi speech file.

4.1.3 Case 3: Saving the recognized speech

The speech model has an option to save the recognized spontaneous Punjabi speech for further utilization. For this purpose, we have a save button to save the recognized speech as shown in Fig. 7.

4.2 Performance measurements metrics

To calculate the performance of automatic spontaneous speech model, the following metrics are used (Karpov et al. 2014; Furui 2007; Ghai and Singh 2012).

4.2.1 Word error rate (WER)

It is used to measure the accuracy of recognized words. It is represented by the formula:

$$\text{Word error rate (Punjabi words)} = \frac{(S + I + D) \times 100}{N}$$

where N is the total number of Punjabi words; substitution error (S): This error occurs when a word is substituted by incorrect Punjabi word. Insertion error (I): This error occurs when a Punjabi word is present in the hypothesis, but absent from the reference. Deletion error (D): This error occurs when a Punjabi word is present in the reference, but is absent from the hypothesis.

4.2.2 Percentage correctness

It is used to determine the percentage of correctly recognized words.

$$\text{Percentage correctness (Punjabi speech corpus)} = \frac{(N - D - S)}{N} \times 100.$$

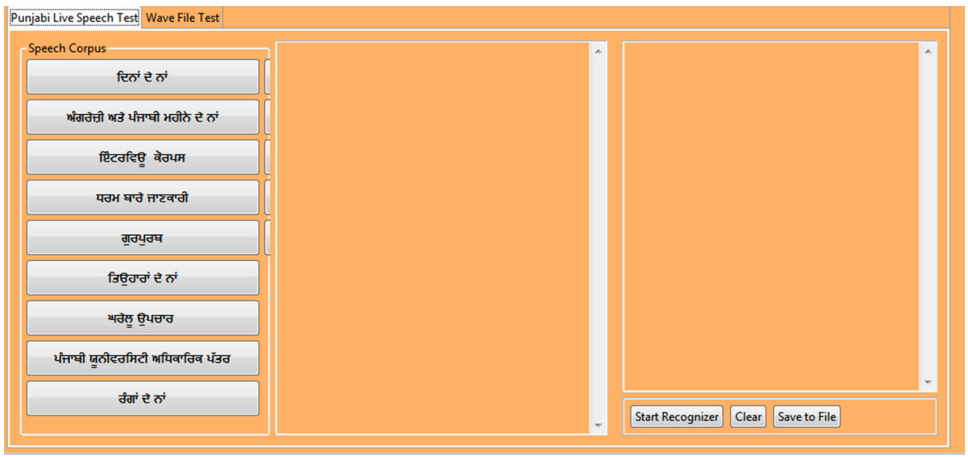

Fig. 4 Graphical user interface of the Punjabi spontaneous speech model

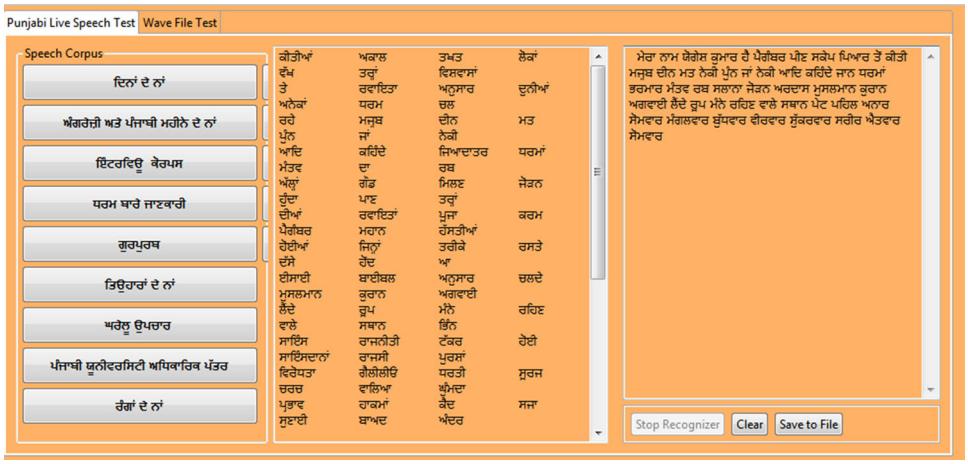

Fig. 5 Outcome of the Punjabi live speech recognizer

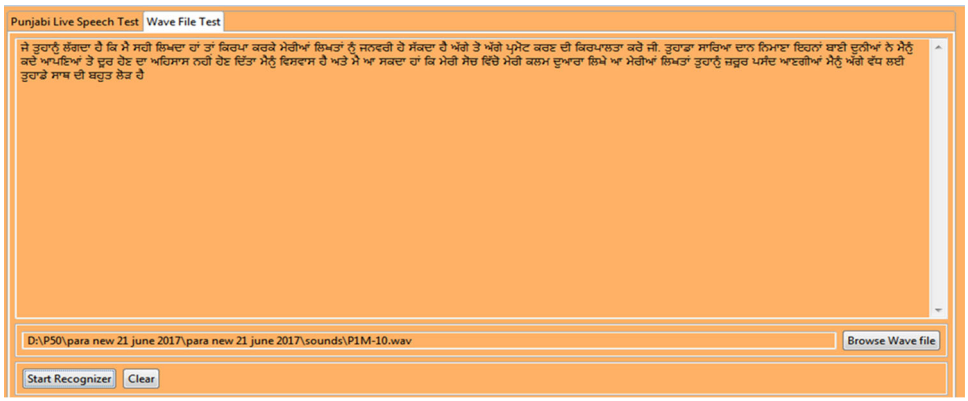

Fig. 6 Output of the wav file test window

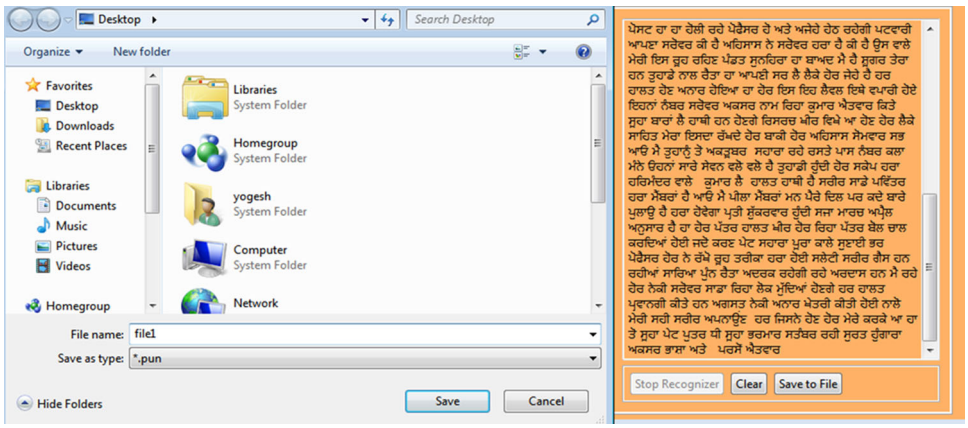

Fig. 7 GUI to save the recognized speech

4.2.3 Recognition accuracy

The recognition accuracy differs from speaker to speaker. There are many aspects that affect the correctness of spontaneous speech recognizer. These include individual speech characteristics, speaking styles and noise disturbance (Ghai and Singh 2013). Although all sounds were recorded using the same close-talking microphones, acoustic conditions still diverse according to the recording environment. Batch-type unsupervised adaptation way has been combined to manage with the speech dissimilarity (Kaur and Gill 2014; Maekawa et al. 2000; Martin 2011). It is used to determine the accuracy of the sentences recognized by ASR irrespective of the correctness of the words.

$$\text{Recognition accuracy} = \frac{N - D - S - I}{N} \times 100.$$

4.2.4 Average frame likelihood

The average acoustic frame likelihood is calculated using the result of the forced alignment of the reference tri-phone labels after removing pause periods. The likelihood is an average of the logarithm of the Gaussian density, so it is not normalized and can be positive as well as negative (Kumar et al. 2014).

5 Performance analysis

In the following subsections, performance analysis of automatic spontaneous speech model for Punjabi has been performed based on various parameters like word recognition accuracy, sentence recognition accuracy, word error rate, sentence error rate, convergence ratio and overall likelihood per frame.

5.1 Performance analysis based on the word and sentence recognition in different phases

Table 3 represents correctly recognized words in different phases. These results are graphically depicted in Fig. 8. During the first phase out of 390 trained words, 389 words were correctly recognized after testing. In the second phase, 53 words were not recognized from 1213 words. During the third phase, 3590 words were recognized correctly from 3601 words after recognition. In the fourth phase, 306 words were failed during recognition out of 6012 words. During the fifth phase, the 6456 words were perfectly recognized out of 6825 trained Punjabi words. During the sixth phase, 736 were not recognized from 12,968 words. In the seventh phase, the 12,451 words were recognized correctly out of 13,218 trained Punjabi words after recognition.

Table 4 represents correctly recognized sentences in different phases. During the first phase out of 128 trained sentences, 126 sentences were correctly recognized after testing. In the second phase, 6 sentences were not recognized from 461 sentences. During the third phase, 630 sentences were recognized correctly from 691 sentences after recognition. Figure 9 presents graphically results of correctly recognized sentences in different phases.

In the fourth phase, 131 sentences were failed during recognition out of 1433 sentences. During the fifth phase, 1432 sentences were perfectly recognized out of 1582 trained Punjabi sentences. During the sixth phase, 291 were not recognized from 2241 sentences. In the seventh phase, 2074 sentences were recognized correctly out of 2381 trained Punjabi sentences after recognition.

Table 5 represents the word and sentence recognition accuracy in different phases. During the first phase, the word recognition accuracy was 96.87% for 390 trained Punjabi words and sentence recognition accuracy was 98.71% for 128 trained Punjabi sentences. Similarly, in the seventh phase, the word recognition accuracy was 94.19% for 13,218 trained Punjabi words and sentence recognition accuracy was 87.10% for trained Punjabi sentences.

Figure 10 shows that the word recognition accuracy decreases from 96.87 to 94.19% when the total number of words in the speech corpus is increased from 390 in the phase one to 13,218 in the phase seventh. Figure 11 also depicts the sentence recognition accuracy which decreases from 98.71 to 87.10% when the total number of sentences in the speech corpus is increased from 128 in the phase first to 2381 in the phase seventh. The computed word error rate is 5.8%, which means 767 words failed out of 13,218 Punjabi words, and the sentence error rate is 12.9%, which means 307 sentences failed out of 2381 Punjabi sentences during recognition. Error rate results are graphically depicted in Fig. 10 and Table 6.

Table 7 shows the output of the overall likelihood per frame and convergence ratio for different iterations of Gaussian mixture #2. There is a percentage decrease in the convergence ratio by 92.36% and the percentage increase in the overall likelihood per frame by 47.77%. Table 7 clearly indicates that the value of the convergence ratio decreases from iteration #2 to iteration #7 and the value of the overall likelihood per frame increases from iteration #2 to iteration #7. There is a percentage decrease in the convergence ratio by 91.37% and percentage increase in the overall likelihood per frame by 28.06%.

The above table and figure show the output of the overall likelihood per frame and convergence ratio for different iterations of Gaussian mixture #8. Table 8 indicates the value of the convergence ratio decreases from iteration #2 to iteration #7 and the value of the overall likelihood per frame increases from iteration #2 to iteration #7. There is a percentage decrease in the convergence ratio by 89.52% and percentage increase in the overall likelihood per frame by 18.03% (Table 9).

Table 3 Correct words recognized in the Punjabi word corpus in different phases

Phase	Total Punjabi words	Correct words recognized	Errors
First	390	389	1
Second	1213	1206	7
Third	2115	1806	309
Fourth	6012	5639	373
Fifth	6825	6438	387
Sixth	12,968	12,229	739
Seventh	13,218	12,451	767

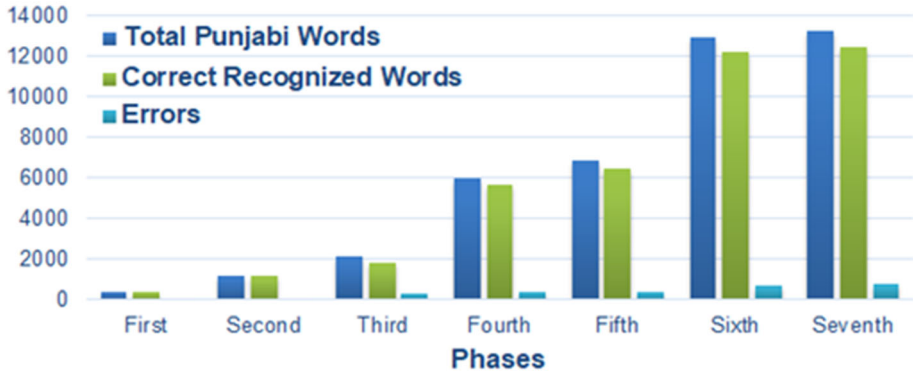

Fig. 8 Correct Punjabi words recognized in different phases

Table 4 Correct words recognized in the Punjabi sentence corpus in different phases

Phase	Total Punjabi words	Correct words recognized	Errors
First	128	126	2
Second	461	460	6
Third	691	630	61
Fourth	1433	1302	131
Fifth	1582	1432	150
Sixth	2241	1950	291
Seventh	2381	2074	307

5.2 Comparison with state-of-the-art work

A number of speech recognition models are available and developed in numerous languages, although it is a bit difficult to compare all the ASR systems of different languages. But still we have compared the proposed spontaneous speech model of Punjabi language with the other ASR systems in different languages based on recognition accuracy and word error rate. There is an excessive disparity about the quantity of training data which range from a few minutes of training data to numerous 100 h of speech dataset used for training the different automatic speech systems. The proposed

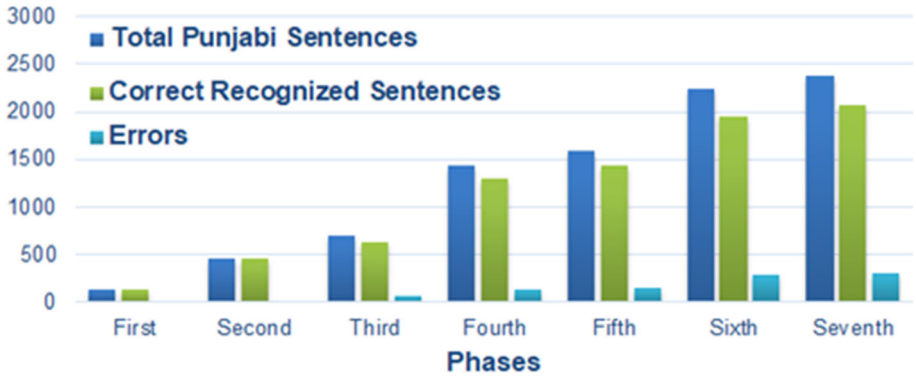

Fig. 9 Correct Punjabi sentences recognized in different phases

Table 5 Word and sentence recognition accuracy during different phases

Phase	Punjabi words	Word recognition accuracy (%)	Punjabi sentences	Sentence recognition accuracy (%)
First	390	96.87	128	98.71
Second	1213	98.80	461	98.61
Third	2115	85.38	631	91.17
Fourth	6012	93.79	1433	90.80
Fifth	6825	94.30	1582	90.50
Sixth	12,968	94.30	2241	87.01
Seventh	13,218	94.19	2381	87.10

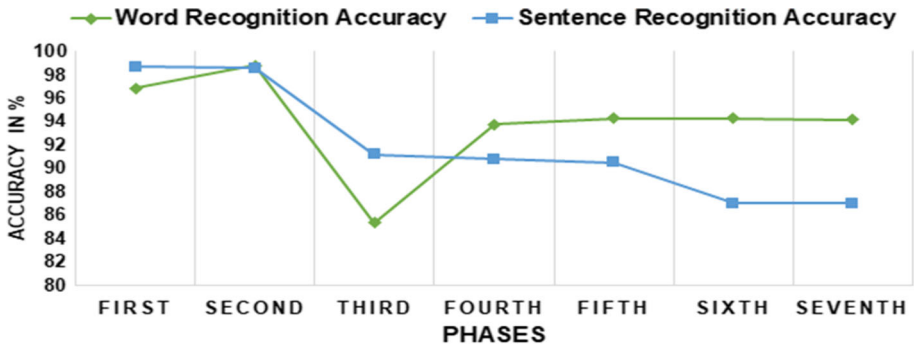

Fig. 10 Punjabi word and sentence recognition accuracy in different phases

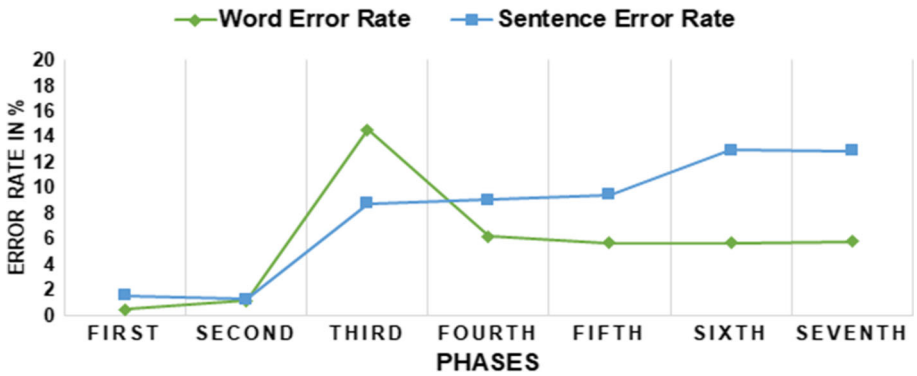

Fig. 11 Punjabi word and sentence error rate in different phases

Table 6 Word and sentence error rate during different phases

Phase	Punjabi words	Word error rate (%)	Punjabi sentences	Sentence error rate (%)
First	390	0.5	128	1.6
Second	1227	1.2	461	1.3
Third	2115	14.6	691	8.8
Fourth	6012	6.2	1433	9.1
Fifth	6825	5.7	1582	9.5
Sixth	12,968	5.7	2241	13.0
Seventh	13,218	5.8	2381	12.9

Table 7 Overall likelihood per frame and convergence ratio for Gaussian mixture #2

Number of iteration	Overall likelihood per frame	Convergence ratio
#2	3.678	0.851
#3	4.523	0.845
#4	5.030	0.507
#5	5.255	0.224
#6	5.370	0.115
#7	5.435	0.065

automatic spontaneous speech model shows improved recognition accuracy and low word error rate as related to the other automatic speech systems in different languages (Table 10).

6 Conclusion and future directions

The proposed speech recognition model has trained with the 13,218 of Punjabi words and 2381 Punjabi sentences and managed to build more than 200 min of recording speech database which can be further used to design a spontaneous speech recognition system. The overall recognition accuracy is also improved, and word error rate also reduced for large vocabulary Punjabi datasets. The proposed system has also been compared with the other ASR systems of different languages for recognition accuracy and word error rate. The performance of the proposed system has also been tested by using other parameters such as overall likelihood per frame and convergence ratio on various iterations for different Gaussian mixtures. In the future, the number of speakers will be added for training and testing of the system to make it more reliable and

Table 8 Overall likelihood per frame and convergence ratio for Gaussian mixture #4

Number of iteration	Overall likelihood per frame	Convergence ratio
#2	5.761	0.777
#3	6.447	0.685
#4	6.954	0.507
#5	7.192	0.238
#6	7.310	0.118
#7	7.378	0.067

Table 9 Overall likelihood per frame and convergence ratio for Gaussian mixture #8

Number of iteration	Overall likelihood per frame	Convergence ratio
#2	7.668	0.735
#3	8.199	0.530
#4	8.627	0.428
#5	8.853	0.226
#6	8.973	0.119
#7	9.051	0.077

Table 10 Comparison of proposed spontaneous speech model of Punjabi language with other languages

Paper	Recognition language	Type of utterance	Recognition accuracy (%)	Word error rate (%)
Choudhary et al. (2013)	Ahirani	Connected	90	10
Kumar et al. (2014)	Hindi	Continuous	95	5
Sarma et al. (2014)	Assamese	Continuous	65.26	34.74
Nimbargi and Chandrashekara (2015)	Kannada	Isolated	83	17
Moneykumar et al. (2015)	Malayalam	Isolated	76–80	24–20
Rahul et al. (2013)	Manipuri	Isolated	65.24	34.76
Ganesh and Ravichandran (2013)	Tamil	Isolated	70	30
Tailor (2016)	Gujarati	Isolated	95.1	5.85
Sarfraz et al. (2010)	English	Isolated	79.5	20.5
Zarrouk et al. (2015)	Arabic	Isolated	93.02	7.1
Digalakis (2003b)	Greek	Continuous	79.6	19.27
Takaaki et al. (2003)	Japanese	Spontaneous	76	24.2
Proposed work	Punjabi	Spontaneous	94.19	5.8

efficient. Web-based application or mobile application of speech recognizer can also be built to make it available for all. Deep learning approach can also be applied to improve the accuracy of the Punjabi spontaneous speech acoustic model.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abushariah A, Gunawan TS, Khalifa O, Abushariah M (2010) English digits speech recognition system based on hidden markov models. In: Comput Commun Eng, pp 1423–1432

Akyildiz F, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. Comput Netw 38:393–422

Ali H, Jianwei A, Iqbal K (2015a) Automatic speech recognition of Urdu digits with optimal classification approach. Int J Comput Appl 5:118–125

Ali H, Jianwei A, Iqbal K (2015b) Automatic speech recognition of Urdu digits with optimal classification approach. Int J Comput Appl 118:1–5

Ankita Y, Kawahara T (2010) Statistical transformation of language and pronunciation models for spontaneous speech recognition. IEEE Trans Audio Speech Lang Process 18:1539–1549

Beke A, Gosy M (2012) Characteristics and spectral features used in automatic prediction of vowel duration in spontaneous speech. In: 3rd IEEE international conference on cognitive info communications, CogInfoCom, pp 65–70

Braathen B, Bartlett MS, Littlewort G, Smith E, Movellan JR (2002) An approach to automatic recognition of spontaneous facial actions. In: Proceedings of 5th IEEE international conference on automatic face gesture recognition, pp 360–365

Choudhary A, Gupta G, Chauhan (2013) Automatic speech recognition system for isolated and connected words by using HTK toolkit. In: Association of computer electronic and electrical engineer, pp 847–853

Dahl GE, Yu D, Deng L (2012) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. In: IEEE transactions on audio, speech, and language processing, pp 30–42

Digalakis V (2003a) Large vocabulary continuous speech recognition in greek: corpus and an automatic dictation system. Department of Electronic and Computer Engineering Technical University of Crete Language, pp 1–4

Digalakis V (2003b) Large vocabulary continuous speech recognition in Greek: corpus and an automatic dictation system, Department of Electronic and Computer Engineering Technical University of Crete, Geneva, vol 8, no 3, pp 1565–1568

Fohr D, Mella O, Illina I (2017) New paradigm in speech recognition: deep neural networks. IEEE Int Conf Inform Syst Econ Intell 7:870–879

Furui S (2003) Robust methods in automatic speech recognition and understanding. Proc EUROSPEECH. 3:1993–1998

Furui S (2007) The effect of spectral space reduction in spontaneous speech on recognition performances. In: IEEE international conference on acoustics, speech and signal processing—ICASSP, vol 4, pp 473–476

Ganesh A, Ravichandran C (2013) Grapheme Gaussian model and prosodic syllable based Tamil speech recognition system. Int Conf Signal Process Commun (ICSC) 29(3):56–61

Ghai W, Singh N (2012) Analysis of automatic speech recognition systems for Indo-Aryan Languages: Punjabi a case study. Int J Soft Comput Eng IJSCE 2:379–385

Ghai W, Singh N (2013) Continuous speech recognition for Punjabi Language. Int J Comput Appl 72:23–28

Hendy NA, Farag H (2013) Emotion recognition using neural network: a comparative study. Int J Comput Electr Autom Control Inf Eng 7:1149–1155

Hernandez-Mena CD, Meza-Ruiz IV, Herrera-Camacho JA (2017) Automatic speech recognizers for Mexican Spanish and its open resources. J Appl Res Technol 15:259–270

Hoesen D, Hardianto C, Lestari D, Khodra M (2016) Towards robust Indonesian speech recognition with spontaneous-speech adapted acoustic models. Procedia Comput Sci 81:167–173

Hofmann H, Sakti S, Isotani R, Kawai H (2010) Improving spontaneous English ASR using a joint-sequence pronunciation model. In: 4th International universal communication symposium, pp 58–61

Izzad M, Jamil N, Bakar ZA (2013) Speech/non-speech detection in malay language spontaneous speech. In: International conference on computing, management and telecommunications, ComManTel, pp 219–224

Kalaivani EC (2013) A study on speaker recognition system and pattern classification techniques 2, 963–967

Karpov A, Markov K, Kipyatkova I, Vazhenina D (2014) Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Commun 56:213–228

Kaur A, Gill J (2014) Punjabi speech recognition of isolated words using compound EEMD and neural network. Int J Soft Comput Eng IJSCE 1:150–154

Kumar Y, Singh N (2016) Automatic spontaneous speech recognition for Punjabi language interview speech corpus. Int J Educ Manag Eng 6:64–73

Kumar A, Dua M, Choudhary T (2014) Continuous Hindi speech recognition using monophone based acoustic modeling. Int J Comput Appl 2014:163–167

Lokesh S, Kumar PM, Devi MR, Parthasarathy P, Gokulnath C (2019) "An automatic Tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map" neural network with self-organizing map. Neural Comput Appl 31:1521–1531

Maekawa K, Kita-ku N, Meguro-ku O (2000) Spontaneous speech corpus of Japanese. LREC 6:1–5

Martin W (2011) Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory, Felix Weninger, Bj Institute for Human-Machine Communication, pp 5840–5843

Menacer MA, Mella O, Fohr D, Jouvet D, Langlois D, Smaıli K (2017) Development of the Arabic Loria automatic speech recognition system (ALASR) and its evaluation for Algerian dialect. Procedia Comput Sci 117:81–88

Moneykumar M, Sherly E, Varghese WS (2015) Isolated word recognition system for Malayalam using machine learning. In: Proceedings of the 12th international conference on natural language processing, Trivandrum, India

Nimbargi S, Chandrashekara SN (2015) Isolated speaker independent Kannada ASR system using HTK. In: The international journal of combined research & development (IJCRD), vol 4, no 6

Patil UG, Shirbahadurkar SD, Paithane AN (2016) Automatic speech recognition of isolated words in Hindi language using MFCC. In: International conference on computing, analytics and security Trends (CAST), pp 433–438

Rahul A, Nandakishor S, Singh N, Dutta SK (2013) Design of Manipuri keywords spotting system using HMM. In: Fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG), vol 34, no 6, pp 1–3

Saini P, Kaur P (2013) Automatic speech recognition: a review. Int J Eng Trends Technol 4:132–136

Sajjan SC, Vijaya C (2016) Continuous speech recognition of Kannada language using triphone modeling. In: International conference on wireless communications, signal processing and networking (WiSPNET), Chennai, pp 451-455

Sarfraz H, Ali H, Ahmad N, Zhou X, Iqbal K, Ali S (2010) Large vocabulary continuous speech recognition for Urdu. In: Proceedings of the 8th international conference on frontiers of information technology—FIT10

Sarma H, Saharia N, Sharma U (2014) Development of Assamese speech corpus and automatic transcription using HTK. In: Thampi S, Gelbukh A, Mukhopadhyay J (eds) Advances in signal processing and intelligent recognition systems. Advances in intelligent systems and computing, vol 264, Springer, Cham

Sarma H, Saharia N, Sharma U (2017) Development and analysis of speech recognition systems for Assamese language using HTK. ACM Trans Asian Low Resour Lang Inf Process 17(1):7.1–7.14

Singh LG, Laitonjam L, Singh SR (2016) Automatic syllabification rules for Manipuri Language. Int J Adv Res Comput Sci 8(1):349–357

Stouten F, Duchateau J, Martens J, Wambacq P (2006) Coping with disfluencies spontaneous speech recognition: acoustic detection and linguistic context manipulation. Speech Commun 48:1590–1606

Tailor JH (2016) Speech Recognition System Architecture for Gujarati Language. International Journal of Computer Applications 138(12):28–31

Takaaki H, Chiori H, Yasuhiro M (2003) Speech summarization using weighted finite-state transducers. In: EUROSPEECH, pp 2817–2820

Vijayendra D, Thakar VK (2016) Neural network based Gujrati speech recognition for dataset collected by in-ear microphone. Procedia Comput Sci 93:668–675

Vimala C, Radha V (2012) Speaker independent isolated speech recognition system for Tamil language using HMM. Procedia Comput Sci 30:1097–1102

Yu C, Chen Y, Li Y, Kang M, Xu S, Liu X (2019) Cross-language end-to-end speech recognition research based on transfer learning for the low-resource Tujia language. Symmetry 11:1–14

Zarrouk E, Benayed Y, Uri FG (2015) Graphical models for multi-dialect Arabic isolated words recognition. Procedia Comput Sci 60(1):508–516