INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Development of Small Vocabulary Continuous Speech-to-Text System for Kannada Language/Dialects

G Thimmaraja Yadava[1]*, **B G Nagaraja**[2], **S Yogesh Kumaran**[3], **A C Ramachandra**[1], **N M Arun Kumar**[1]

1 Nitte Meenakshi Institute of Technology, Bengaluru, Karnataka, India
2 Vidyavardhaka College of Engineering, Mysuru, Karnataka, India
3 Faculty of Engineering and Technology, Jain deemed-to-be University, Kanakapura, Karnataka, India

OPEN ACCESS

Received: 19-09-2022

Accepted: 30-10-2022

Published: 07-12-2022

Citation: Yadava GT, Nagaraja BG, Kumaran SY, Ramachandra AC, Kumar NMA (2022) Development of Small Vocabulary Continuous Speech-to-Text System for Kannada Language/Dialects. Indian Journal of Science and Technology 15(45): 2476-2481. https://doi.org/10.17485/IJST/v15i45.1884

* **Corresponding author**.

thimrajyadav@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop a speech-to-text (STT) system using Kaldi speech recognition toolkit for continuous Kannada language/dialects. **Methods:** A continuous Kannada speech data is collected from 100 speakers/farmers of Karnataka state in field. The lexicon/dictionary and set of phonemes for Kannada language/dialects are created and transcribed the collected speech data using transcriber tool. The ASR models are developed at different phoneme levels using Kaldi. **Findings**: In this work, an effort is made to develop a robust small vocabulary STT system for continuous Kannada language using Kaldi. The various acoustic modelling techniques are used to develop a robust ASR model and achieved a word error rate (WER) of 0.23%. The performance of the developed ASR model is compared with existing works and analyzed by offline speech recognition. **Novelty**: Many STT systems have been developed for Indian and International languages/dialects, but not for Kannada language. This work is first of its kind using Kaldi in Kannada language under the constraints of limited data. The developed ASR model could be used further in the development of end-to-end ASR system for speech processing applications.

Keywords: Automatic Speech Recognition (ASR); Word Error Rate (WER); Continuous Kannada Speech Data; Kannada Language/Dialects; Lexicon

1 Introduction

The recognition of speech data by machines automatically under an uncontrolled environment is a challenging task and this is called automatic speech recognition (ASR). Two types of ASR systems can be categorized: speaker-based ASR and speech-based ASR systems[1]. To build an ASR system, it is necessary to create ASR models using speech recognition toolkits. The different toolkits for speech recognition are available to build robust ASR models. They are the hidden Markov model (HMM) toolkit, Julius, CMUSphinx, Kaldi, etc. A large lexicon/vocabulary continuous speech recognition model was developed in[2] using the HTK toolkit. The TIMIT speech dataset with 855 words was used for the conduction of experiments. Using transcription and dictionary, the obtained word error rates (WER) are 9%, 13% and 37% for the test set of size 25, 105 and 855 respectively. To overcome the drawbacks of HMM in the development of continuous speech recognition models, the authors in [3] have worked on template matching using the dynamic time wrapping (DTW) technique. The experiments were conducted using HMM and a combination of HMM and DTW techniques. The results revealed that the combined HMM+DTW method has given a better performance than the traditional HMM technique. Using the language resources and proposed technique, the achieved WER is 3.54% and 2.41% for HMM and HMM+DTW respectively. The importance of reservoir computing (RC) for acoustic continuous speech features modelling was explained in [4]. The WSJ0 speech corpora were exploited for the experiments. The speech dataset comprised of 7240 utterances collected by 84 speakers. The RC+HMM-based continuous ASR modelling technique outperforms other acoustic modeling techniques with a WER of 6.2% for the bigrams language model and 3.9% for the trigram model.

The control over the ASR model complexity is still a challenging task in many speech processing applications and this was described in [5]. To overcome the problem of ASR model complexity, the authors have implemented a technique for the generalized-variable-parameter HMMs (GVP-HMM). The claimed method has increased the efficacy over the conventional GVP-HMM. The experiments were conducted on the Aurora-2 speech database which comprised of 420 speech sentences. The experimental results have improved the efficiency of the entire system by minimizing the WER of 28%. The application of micro modulation features in the robust continuous ASR system was depicted in [6]. The Aurora- 4 speech corpus dataset was completely exploited for the conduction of the experiment. The experimental results showed that there is a better improvement in the efficiency of 8% using proposed micro modulation features compared to a single stream system. The performance of the ASR system was significantly reduced due to the addition of different types of noises in speech data. To overcome this problem, a method was proposed in [7] by considering the advantages of autoregressive (AR) modelling. The WSJ and Aurora-4 databases were considered for the experiments. The training set comprised 7138 sentences by 84 speakers and 330 sentences by 8 speakers considered for the testing. The ASR model was created using Kaldi. The ASR model was developed for both degraded and clean speech corpora. The obtained WER using the proposed technique for clean and degraded speech data were 3.1% and 12.8% respectively.

The recognition of sentences and phrases spoken by talking face without or with audio was demonstrated in [8]. The authors have focused on solving the problem of limited number of phrases or words recognition. They compared two important models for reading the lip by considering the CTC and sequence to sequence losses. These models were built above on self-attention architecture. The authors claimed that the proposed models have surpassed the performance of all most all previous works on lip reading. In [9] an overview on ASR systems using machine learning was neatly described. The problems involved in the development of real time robust ASR systems and their improved performance by modern machine learning technology was explained in detail. The recent advancements with respect to deep learning and sparse representations for speech recognition are also analyzed. The methods for speech to speech and speech to text summarization using speech unit concatenation and speech unit extraction was demonstrated in [10]. Two stage automatic summarization procedure was adopted for speech sentence extraction and sentence compaction. The proposed methods were evaluated using both objective and subjective measures and the investigated results revealed that the proposed methods are very effective than the state-of-the-art techniques.

In the history of Kannada language, for the first time, an end-to-end (E2E) ASR system was developed for accessing the real time commodity prices information and weather forecasting in Kannada language/dialects [11] and demonstrated the continuous advancements [12–14] in the performance of E2E ASR system in terms of speech recognition accuracy by proposing noise reduction algorithm [12]. Development of spoken-query-system (SQS) to recognize the Kannada continuous speech sentences was demonstrated in [15]. The authors have created a large vocabulary Kannada continuous speech data from 2400 speakers and developed ASR models using Kaldi at various phoneme levels. The authors have achieved an accuracy of 95.9% for noisy speech data and they have used this model in SQS. The authors further enhanced the recognition accuracy of continuous Kannada SQS by integrating the noise reduction algorithm before speech feature extraction [16]. An ensemble of proposed noise elimination algorithm [17] and time delay neural network (TDNN) acoustic modelling technique has given a better improvement and achieved an accuracy of 97.60% compared to earlier continuous Kannada SQS [15]. An overview for developing E2E ASR system was given in [18]. The authors were described the comprehensive study on the techniques, algorithms, projects, tools, recent contributions and future directions and scope in ASR systems using limited vocabulary. In [19], the authors have investigated the optimal timing window duration for cepstral features extraction in ASR context. An ASR system for Kannada language was considered for the aimed task and developed the same using HTK. The obtained experimental results were analyzed for different time window lengths with existing literature results. The development ASR system for large vocabulary continuous Kannada speech data was demonstrated in [20]. The collected speech data was transcribed and validated for system training and testing using Kaldi. The authors have achieved a WER of 4.64% using DNN-HMM. All our previous works [11–17] described the development of large vocabulary isolated and continuous E2E ASR systems for Kannada language/dialects. In this work, an attempt is made to develop a continuous speech-to-text system for small vocabulary for Kannada language/dialects.

The salient contributions made in the current work lies in:

- Creation of small vocabulary continuous speech Kannada speech data.
- Created the phoneme set and dictionary/lexicon for Kannada language/dialects.
- The collected speech data is transcribed, validated, and extracted the speech features.
- Development of small vocabulary continuous speech-to-text (SVCSTT) system for Kannada language/dialects.

The rest of the paper is summarized as follows: Section 2 gives the complete methodology for developing the of small vocabulary SVCSTT system. The results and discussion are given in Section 3. The conclusions are depicted in Section 4.

2 Methodology

To create a robust ASR model, a Kannada continuous speech data is collected from 100 (Size of the database is 2000 continuous Kannada speech sentences) farmers/speakers under field conditions. The IVRS call flow structure is designed to collect the speech data and it is shown in Figure 1. Some of the continuous speech sentences used for data gathering is shown Figure 2. Initially, the farmers/speakers need to dial the toll-free number to get connected with the speech data collection server. Once it is connected, the server will play out the pre-recorded prompt that "Welcome to continuous Kannada speech data collection center, please tell the names of continuous Kannada speech sentences after the beep sound". In a particular session, server will prompt 20 speech sentences and the farmer/speaker needs to repeat the same as server prompts. Once all the sentences have been completed, the server will ask the farmer/speaker to repeat the procedure for one more time if he/she is not satisfied with the speech data collection. If he/she says yes, then it will start with the initial step, otherwise the call will be terminated.

The dictionary/lexicon and phoneme set play a vital role in the recognition of speech under field conditions. Using the phoneme set, a dictionary/lexicon is created at word and phoneme level. The schematic representation of mapping of Kannada phonemes using the created phoneme set and dictionary/lexicon for the few Kannada continuous speech sentences are shown in Figure 3 and Figure 4 respectively. The transcription of collected speech data is done at word level using the created phoneme set. Some of the transcription of Kannada continuous speech sentences is shown in Figure 5.

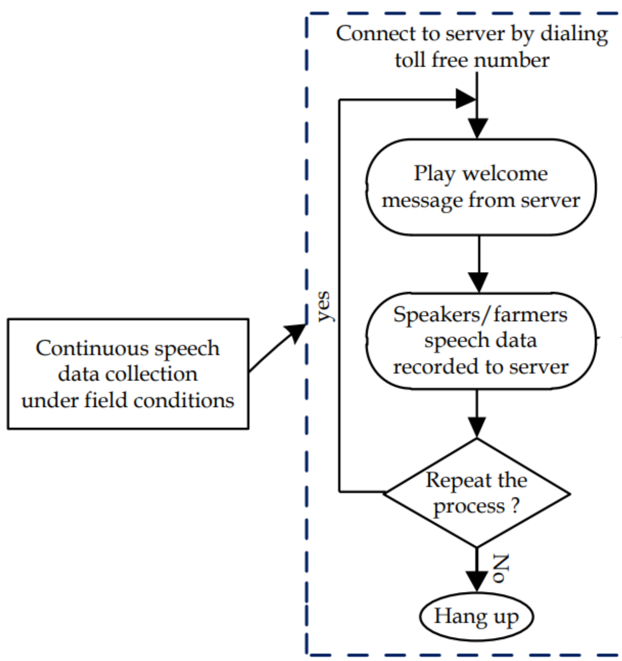

Fig 1. Speech data collection call flow structure

English version of continuous Kannada speech sentences	Kannada Version
Adda godeya mele deepa itta haage	ಅಡ್ಡ ಗೋಡೆಯ ಮೇಲೆ ದೀಪ ಇಟ್ಟ ಹಾಗೆ
Athi aase gathi gedu	ಅತಿ ಆಸೆ ಗತಿ ಗೇಡು
Bandaddella barali govindana daye irali	ಬಂದದೆಲ್ಲ ಬರಲಿ ಗೋವಿಂದನ ದಯೆ ಇರಲಿ
Akki mele aase nentara mele preeti	ಅಕ್ಕಿ ಮೇಲೆ ಆಸೆ ನೆಂಟರ ಮೇಲೆ ಪ್ರೀತಿ
Beline eddu hola meyitante	ಬೇಲಿನೆ ಎದ್ದು ಹೊಲ ಮೇಯಿತಂತೆ
Haavu saayalilla kolu muriyalilla	ಹಾವು ಸಾಯಲಿಲ್ಲ ಕೋಲು ಮುರಿಯಲಿಲ್ಲ
Veda sulladaru gaade sullagadu	ವೇದ ಸುಳ್ಳಾದರು ಗಾದೆ ಸುಳ್ಳಾಗದು
Adikege hoda maana aane kottaru baaradu	ಅಡಿಕೆಗೆ ಹೋದ ಮಾನ ಆನೆ ಕೊಟ್ಟರು ಬಾರದು
Haagalakaayige bevinakaayi saakshi	ಹಾಗಲಕಾಯಿಗೆ ಬೇವಿನಕಾಯಿ ಸಾಕ್ಷಿ
Enne bandaga kannu mucchhikondante	ಎಣ್ಣೆ ಬಂದಾಗ ಕಣ್ಣು ಮುಚ್ಚಿಕೊಂಡಂತೆ
Hiriyakkana chaali mane mandigella	ಹಿರಿಯಕ್ಕನ ಚಾಳಿ ಮನೆ ಮಂದಿಗೆಲ್ಲ
Hoovininda naaru swarga seritu	ಹೂವಿನಿಂದ ನಾರು ಸ್ವರ್ಗ ಸೇರಿತು
Bekkige chellata ilige praana sankata	ಬೆಕ್ಕಿಗೆ ಚೆಲ್ಲಾಟ ಇಲಿಗೆ ಪ್ರಾಣ ಸಂಕಟ
Hana andre henavu baayi biduttade	ಹಣ ಅಂದ್ರೆ ಹೆಣವು ಬಾಯಿ ಬಿಡುತ್ತದೆ
Maatu belli mouna bangara	ಮಾತು ಬೆಳ್ಳಿ ಮೌನ ಬಂಗಾರ
Tumbida koda tulukuvudilla	ತುಂಬಿದ ಕೊಡ ತುಳುಕುವುದಿಲ್ಲ

Fig 2. Some of the continuous Kannada speech sentences used for data gathering

Label set using IT3: UTF-8				Corresponding Kannada phonemes			
a	oo	t:h	Ph	ಅ	ಓ	ಠ	ಫ
aa	au	d	b	ಆ	ಔ	ಡ	ಬ
i	k	d:h	bh	ಇ	ಕ	ಢ	ಭ
ii	kh	nd	m	ಈ	ಖ	ಣ	ಮ
u	g	t	y	ಉ	ಗ	ತ	ಯ
uu	gh	th	r	ಊ	ಘ	ಥ	ರ
e	c	d	l	ಎ	ಚ	ದ	ಲ
ee	ch	dh	v	ಏ	ಛ	ಧ	ವ

Label set using ILSL12				Corresponding Kannada phonemes			
a	oo	txh	ph	ಅ	ಓ	ಠ	ಫ
aa	au	d	b	ಆ	ಔ	ಡ	ಬ
i	k	dxh	bh	ಇ	ಕ	ಢ	ಭ
ii	kh	nx	m	ಈ	ಖ	ಣ	ಮ
u	g	t	y	ಉ	ಗ	ತ	ಯ
uu	gh	th	r	ಊ	ಘ	ಥ	ರ
e	c	d	l	ಎ	ಚ	ದ	ಲ
ee	ch	dh	w	ಏ	ಛ	ಧ	ವ
ai	j	n	sh	ಐ	ಜ	ನ	ಶ
o	tx	p	s	ಒ	ಟ	ಪ	ಸ

Fig 3. Phoneme set for Kannada language/dialects

Label set using IT3: UTF-8	Label set using ILSL12
hand~a an:dre hend~avuu baayi bid:uttade	hanxa_andare_henxawuu_baayi_biduttade
huuvini:n:da naaru swarga seiritu	huuwininda_naaru_swarga_seiritu
hiriyakkana chaal:i mane man:digella	hiriyakkan_chaali_mane_mandgella
hittala gid:a maddalla	hittala_gida_maddalla
hut:t:u gund~a sut:a:t:aruu hoogolla	huttu_guna_suttaruu_hoogolla
ollada gan:d:anige mosaralli kallu	ollada_gandanige_mosaralli_kallu
raatri kan:d:a baaviili hagalu biddan:ge	raati_kanda_baaviili_hagalu_biddange
vinaasha kaalakke vipariita buddhi	winaasha_kaalakke_wipariita_buddi
yaradoo dud:d:u yallammana jaatre	yaradoo_duddu_yellammana_jaatre

Fig 4. Dictionary for some of the continuous speech sentences

Fig 5. Transcription of few Kannada continuous speech sentences using transcriber tool

3 Results and Discussion

Kaldi is one of the most widely used speech recognition toolkits to develop robust ASR models for various types of languages/dialects. The transcribed speech data is validated to make sure that there are no errors in the transcribed data. The validated continuous Kannada speech data is subjected to extract the speech features using Mel frequency cepstral coefficients (MFCC) technique. The 600 senons, 4, 8 and 16 mixtures of Gaussians and different acoustic modelling techniques are used (monophone, triphone1 and triphone2, SGMM, DNN+SGMM and DNN+HMM) in this work. The performance of the ASR model is evaluated using WER. Using continuous Kannada speech data transcription, phoneme set, lexicon, silence phones and Kaldi recipe, the achieved WER are shown in Table 1. From the Table, it can be observed that the hybrid modeling technique, DNN+HMM is outperformed the other acoustic modelling techniques in terms of WER and has given a better accuracy of speech recognition compared to the existing works [15,16,20]. This is since the vocabulary used for the system training and testing is less and the outcome of the triphone 3 is given as input to the DNN. Once the DNN training is done, the decoding phase will start by extracting the speech features using MFCC. The test features of speech are compared with the trained models. If the features are matched with training and decoding, then the recognition of speech takes place. This procedure is continuously considered till the completion test dataset. Once the testing phase is done, the DNN generates the decoding results in a log file. The decoding results show the recognition of continuous Kannada speech sentences which is represented inFigure 6.

Table 1. Performance comparison of ASR models for continuous Kannada speech data developed by proposed methodology with the existing work in terms ofWERs.

Acoustic modelling techniques	**WER achieved in [15]**	**WER achieved in [16]**	**WER achieved in [20]**	**Achieved WER in % for small vocabulary: Proposed work**
Training and testing at monophone level	7.66	7.06	8.23	1.27
Training and testing at triphone1	5.09	5.05	6.24	0.16
Training and testing at triphone2	5.80	5.66	5.59	0.23
Training and testing at triphone3	5.03	4.99	5.12	0.22
SGMM	4.65	4.45	4.86	0.20
DNN+SGMM	4.21	2.96	4.59	0.19
DNN+HMM	4.10	2.91	4.56	**0.15**

Fig 6. Offline recognition of continuous Kannada speech sentences

4 Conclusion

A demonstration of continuous speech-to-text system for Kannada language/dialects is presented in this work. The speech data was gathered from 100 speakers for creating a SVCASR models. The transcription and validation procedures were adopted for extracting the speech features using MFCC. The system training and decoding using Kaldi was done by considering the three acoustic modelling techniques such as monophone, triphone1 and triphone2, triphone3, SGMM, DNN+SGMM and DNN+HMM. With the recipe of Kaldi and the resources of Kannada language, a least WER of 0.15% was achieved using DNN+HMM. The experimental evaluations revealed that the combination of DNN+HMM has outperformed the other acoustic modelling techniques in terms of WER and the results obtained using the proposed methodology has been compared with existing works. Further, the performance of SVCSTT system is evaluated by offline recognition of continuous Kannada speech sentences. In future, further effort will be made to recognize the speech sentences under real time conditions using the developed SVCASR models.

References

1) Rabiner LR. Applications of voice processing to telecommunications. *Proceedings of the IEEE*. 1994;82(2):199–228. Available from: https://doi:10.1109/5.265347.
2) Zhao Y. A speaker-independent continuous speech recognition system using continuous mixture Gaussian density HMM of phoneme-sized units. *IEEE Transactions on Speech and Audio Processing*. 1993;1(3):345–361. Available from: https://doi:10.1109/89.232618.
3) De Wachter M, Matton M, Demuynck K, Wambacq P, Cools R, Van Compernolle D. Template-Based Continuous Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing*. 2007;15(4):1377–1390. Available from: https://doi:10.1109/TASL.2007.894524.
4) Triefenbach F, Demuynck K, Martens JPP. Large Vocabulary Continuous Speech Recognition With Reservoir-Based Acoustic Models. *IEEE Signal Processing Letters*. 2014;21(3):311–315. Available from: https://doi:10.1109/LSP.2014.2302080.
5) Su R, Liu X, Wang L. Automatic Complexity Control of Generalized Variable Parameter HMMs for Noise Robust Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2015;23(1):1–1. Available from: https://doi:10.1109/TASLP.2014.2372901.
6) Dimitriadis D, Bocchieri E. Use of Micro-Modulation Features in Large Vocabulary Continuous Speech Recognition Tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2015;23(8):1348–1357. Available from: https://doi:10.1109/TASLP.2015.2430815.
7) Ganapathy S. Multivariate Autoregressive Spectrogram Modeling for Noisy Speech Recognition. *IEEE Signal Processing Letters*. 2017;24(9):1373–1377. Available from: https://doi:10.1109/LSP.2017.2724561.
8) Afouras T, Chung JS, Senior A, Vinyals O, Zisserman A. Deep Audio-Visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;44(12):8717–8727. Available from: https://doi:10.1109/TPAMI.2018.2889052.
9) Deng L, Li X. Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Transactions on Audio, Speech, and Language Processing*. 2013;21(5):1060–1089. Available from: https://doi:10.1109/TASL.2013.2244083.
10) Furui S, Kikuchi T, Shinnaka Y, Hori C. Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech. *IEEE Transactions on Speech and Audio Processing*. 2004;12(4):401–408. Available from: https://doi:10.1109/TSA.2004.828699.
11) Yadava TG, Jayanna HS. A spoken query system for the agricultural commodity prices and weather information access in Kannada language. *International Journal of Speech Technology*. 2017;20(3):635–644. Available from: https://doi.org/10.1007/s10772-017-9428-y.
12) Yadava TG, Jayanna HS. Speech enhancement by combining spectral subtraction and minimum mean square error-spectrum power estimator based on zero crossing. *International Journal of Speech Technology*. 2019;22(3):639–648. Available from: https://doi.org/10.1007/s10772-018-9506-9.
13) Yadava TG, Jayanna HS. Enhancements in automatic Kannada speech recognition system by background noise elimination and alternate acoustic modelling. *International Journal of Speech Technology*. 2020;23(1):149–167. Available from: https://doi.org/10.1007/s10772-020-09671-5.
14) Yadava TG, Jayanna HS. Improvements in Spoken Query System to Access the Agricultural Commodity Prices and Weather Information in Kannada Language/Dialects. *Journal of Intelligent Systems*. 2018;29(1):664–687. Available from: https://doi.org/10.1515/jisys-2018-0120.
15) Kumar P, Yadava PS, Jayanna TG, S H. *Continuous Kannada speech recognition system under degraded condition Circuits, Systems, and Signal Processing*. 2020;39:391–419. Available from: https://doi.org/10.1007/s00034-019-01189-9.
16) Yadava GT, Nagaraja BG, Jayanna HS. Enhancements in Continuous Kannada ASR System by Background Noise Elimination. *Circuits, Systems, and Signal Processing*. 2022;41(7):4041–4067. Available from: https://doi.org/10.1007/s00034-022-01973-0.
17) Yadava TG, Nagaraja BG, Jayanna HS. *A spatial procedure to spectral subtraction for speech enhancement Multimedia tools and applications*. 2022;81:23633–23647. Available from: https://doi.org/10.1007/s11042-022-12152-3.
18) Louis J, Fendji KE, Tala DCM, Blaise O, & Marcellin AY. Automatic Speech Recognition Using Limited Vocabulary: A Survey. *Applied Artificial Intelligence*. 2022;36(1). Available from: https://doi.org/10.1080/08839514.2022.2095039.
19) Thalengala A, Hoblidar A, Girisha S, Tumkur. Effect of time-domain windowing on isolated speech recognition system performance. *International Journal of Electronics and Telecommunications*;2022(1):161–166. Available from: https://doi.org/10.24425/ijet.2022.139856.
20) Kumar P, Jayanna HS. Development of Speaker-Independent Automatic Speech Recognition System for Kannada Language. *Indian Journal of Science and Technology*. 2022;15(8):333–342. Available from: https://doi.org/10.17485/IJST/v15i8.2322.