Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 25–28
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Malayalam Speech Corpus: Design and Development for Dravidian Language

Lekshmi.K.R, Jithesh V S, Elizabeth Sherly
Research Scholar, Senior Linguist, Senior Professor
Bharathiar University, IIITM-K, IIITM-K
lekshmi.kr@iiitmk.ac.in, jithesh.vs@iiitmk.ac.in, sherly@iiitmk.ac.in

Abstract

To overpass the disparity between theory and applications in language-related technology in the text as well as speech and several other areas, a well-designed and well-developed corpus is essential. Several problems and issues encountered while developing a corpus, especially for low resource languages. The Malayalam Speech Corpus (MSC) is one of the first open speech corpora for Automatic Speech Recognition (ASR) research to the best of our knowledge. It consists of 250 hours of Agricultural speech data. We are providing a transcription file, lexicon and annotated speech along with the audio segment. It is available in future for public use upon request at "www.iiitmk.ac.in/vrclc/utilities/ml_speechcorpus". This paper details the development and collection process in the domain of agricultural speech corpora in the Malayalam Language.

Keywords: Malayalam, ASR, Agricultural Speech corpus, Narrational and Interview Speech Corpora

1. Introduction

Malayalam is the official language of Kerala, Lakshadweep, and Mahe. From 1330 million people in India, 37 million people speak Malayalam ie; 2.88% of Indians. (Wikipedia contributors, 2020). Malayalam is the youngest of all languages in the Dravidian family. Four or five decades were taken for Malayalam to emerge from Tamil. The development of Malayalam is greatly influenced by Sanskrit also.

In the Automatic Speech Recognition (ASR) area many works are progressing in highly and low-resourced languages. The present speech recognition system has achieved a 'Natural' degree of accuracy mainly in Standard American English (Xiong et al., 2016). The accurate recognition of speech exists only for highly resourced languages. But it is still lagging for "non-native" speakers. To increase the accuracy of such an ASR system the speech data for low- resource language like Malayalam is to be increased.

To encourage the research on speech technology and its related applications in Malayalam, a collection of speech corpus is commissioned and named as Malayalam Speech Corpus (MSC). The corpus consists of the following parts.

- 200 hours of Narrational Speech named NS and
- 50 hours of Interview Speech named IS

The raw speech data is collected from *"Kissan Krishideepam"* an agriculture-based program in Malayalam by the Department of Agriculture, Government of Kerala. The NS is created by making a script during the post-production stage and dubbed with the help of people in different age groups and gender but they are amateur dubbing artists. The speech data is thoughtfully designed - for various applications like code mixed language analysis, Automatic Speech Recognition (ASR) related research, speaker recognition – by considering sociolinguistic variables.

This paper represents the development of Narrational and Interview Speech corpora (NS and IS) collected from native Malayalam speakers. The literature survey of different speech corpora creation is detailed in section 2. Section 3 describes the design and demographics of speech data. The section 4 continues with transcription and section 5 deals with lexicon of the speech data and paper concludes with section 6.

2. Literature Survey

Many languages have developed speech corpus and they are open source too. The English read speech corpus is freely available to download for research purposes (Koh et al., 2019) (Panayotov et al., 2015). Similarly, a database is made available with the collection of TED talks in the English language (Hernandez et al., 2018). Databases are available for Indian languages on free download and a payment basis also. For the Malayalam language-based emotion recognition, a database is available (Rajan et al., 2019).

The corpus collection of low resourced languages is a good initiative in the area of ASR. One of such work is done on Latvian language (Pinnis et al., 2014). They created 100 hours of orthographically transcribed audio data and annotated corpus also. In addition to that a four hours of phonetically transcribed audio data is also available. The authors presented the statistics of speech corpus along with criteria for design of speech corpus.

South Africa has eleven official languages. An attempt is made for the creation of speech corpora on these under resourced languages (Barnard et al., 2014). A collection of more than 50 hours of speech in each language is made available. They validated the corpora by building acoustic and language model using KALDI.

Similarly speech corpora for North-East Indian low-resourced languages is also created (Hernandez et al., 2018). The authors collected speech and text corpora on Assamese, Bengali and Nepali. They conducted a statisti-

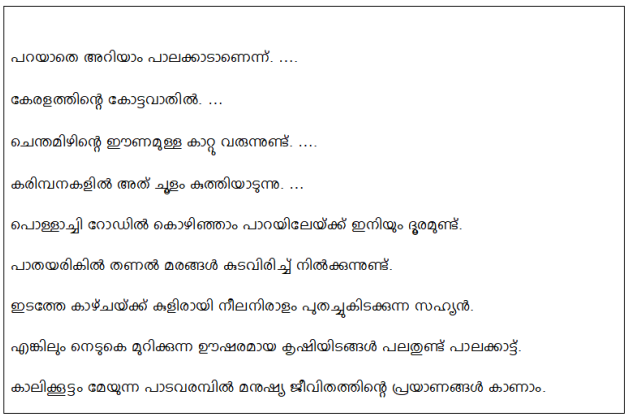
cal study of the corpora also.

3. The Speech Corpora

A recording studio is setup at our visual media lab with a quiet and sound proof room. A standing microphone is used for recording NS corpora. IS corpora is collected directly from the farmers using recording portable Mic at their place. Hundred speakers are involved in the recording of NS and IS corpora.

3.1. Narrational and Interview Speech Corpora

The written agricultural script, which is phonetically balanced and phonetically rich (up to triphone model), was given to the speakers to record the Narrational Speech. Scripts were different in content. An example script is provided in Fig:1. They were given enough time to record the data. If any recording issues happened, after rectification by the recording assistant it was rerecorded.


പറയാതെ അറിയാം പാലക്കാടാണെന്ന്. ....

കേരളത്തിന്റെ കോട്ടവാതിൽ. ...

ചെന്തമിഴിന്റെ ഈണമുള്ള കാറ്റ് വരുന്നുണ്ട്. ....

കരിമ്പനകളിൽ അത് ചൂളം കുത്തിയാടുന്നു. ...

പൊള്ളാച്ചി റോഡിൽ കൊഴിഞ്ഞാം പാറയിലേയ്ക്ക് ഇനിയും ദൂരമുണ്ട്.

പാതയരികിൽ തണൽ മരങ്ങൾ കടവിരിച്ച് നിൽക്കുന്നുണ്ട്.

ഇടത്തേ കാഴ്ചയ്ക്ക് കുളിരായി നീലനിരാളം പൂത്തുകിടക്കുന്ന സഹ്യൻ.

എങ്കിലും നെടുകെ മുറിക്കുന്ന ഊഷരമായ കൃഷിയിടങ്ങൾ പലതുണ്ട് പാലക്കാട്ട്.

കാലിക്കൂട്ടം മേയുന്ന പാടവരമ്പിൽ മനുഷ്യ ജീവിതത്തിന്റെ പ്രയാണങ്ങൾ കാണാം.

Figure 1: Example of script file for dubbing

The Narrational Speech is less expensive than Interview Speech because it is difficult to get data for the ASR system. The IS data is collected in a face-to-face interview style. The literacy and the way to communicate information fluently have given less focus. The interviewee with enough experience in his field of cultivation is asked to speak about his cultivation and its features. The interviewer should be preferably a subject expert in the area of cultivation. Both of them are given separate microphones for this purpose.

Few challenges were faced during the recording of the speech corpus. There were lot of background noise like sounds of vehicles, animals, birds, irrigation motor and wind. Another main issue that happened during post production is the difference in pronunciation styles in the Interview Speech corpora collection. This caused difficulty during validation of the corpus. The recording used to extend up to 5-6 hours depending on speakers. The recorded data is then given for post-production to clean unwanted information from that.

3.2. Speaker Criteria

We have set a few criteria for recording the Narrational Speech data.

- The speakers are at minimum age of 18
- They are citizens of India
- Speakers are residents of Kerala
- The mother tongue of the speaker should be Malayalam without any specific accents

3.3. Recording Specifications

Speech data is collected with two different microphones for NS and IS. For Narrational Speech, Shure SM58-LC cardioid vocal microphone without cable is used. For IS, we utilized Sennheiser XSW 1-ME2-wireless presentation microphone of range 548-572 MHz Steinberg Nuendo and Pro Tools are used for the audio post-production process.

The audio is recorded in 48 kHz sampling frequency and 16 bit sampling rate for broadcasting and the same is down sampled to 16 kHz sampling frequency and 16 bit sampling rate for speech-related research purposes. The recordings of speech corpora are saved in WAV files.

3.4. Demographics

MSC aims to present a good quality audio recording for speech related research. The NS and IS corpus have both male and female speakers. In NS, the male and female speakers are made up with 75% and 25% respectively. IS have more male speakers than females with 82% and 18% of total speakers.The other demographics available from the collected data are Community, Place of Cultivation and Type of Cultivation.

Category	NS (%)	IS (%)
Hindu	85	51
Christian	10	35
Muslim	05	14
Total	100	100

Table 1: Demographic details of speakers by community

Table 2 and 3 contains the details of the place of cultivation and the type of cultivation in Kerala.

Place of Cultivation (District wise)	IS(%)
Thiruvananthapuram	26
Kollam	21
Pathanamthitta	02
Ernakulam	07
Alappuzha	08
Kottayam	08
Idukki	09
Thrissur	12
Wayanad	03
Kozhikode	02
Kannur	02
Total	100

Table 2: Demographic details of speakers by place of cultivation

Type of Cultivation	IS (%)
Animal Husbandry	10
Apiculture	11
Diary	16
Fish and crab farming	05
Floriculture	07
Fruits and vegetables	22
Horticulture	04
Mixed farming	07
Organic farming	08
Poultry	07
Terrace farming	03
Total	100

Table 3: Demographic details of speakers by type of cultivation

4. Transcription

The NS and IS corpora are transcribed orthographically into Malayalam text. The transcribers are provided with the audio segments that the speaker read. Their task is to transcribe the content of the audio into Malayalam and into phonetic text. A sample of three transcribed data with demographic details is shown below and the annotated speech of first two sentences is depicted in Fig 2.

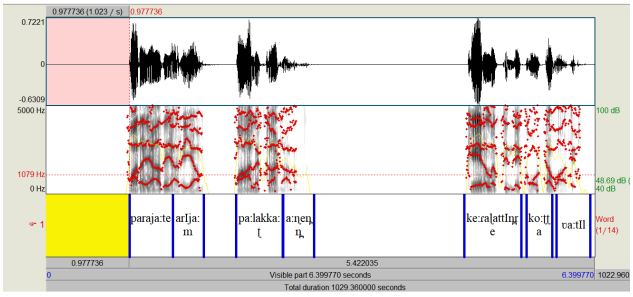


Figure 2: An example of Annotated Speech Corpora

Sample 1: **Record Entry No** : **180220_01_01**

In the first sample a Narrational Speech is detailed. The narrator is about 45 years old and he is describing the details about Palakkad a district in Kerala and a mango estate there. Few sentences are displayed below.

Sentence 1:

പറയാതെ അറിയാം പാലക്കാടാണെന്ന്

paraja:te arIja:m pa:lakka:ʈa:ɳeɳɳ

<Without saying we can understand that it is Palakkad>

Sentence 2:

കേരളത്തിന്റെ കോട്ട വാതിൽ

ke:ralattIɳre ko:ʈʈa ʋa:tIl

< Kerala's Castle door>

Sentence 3:

സേലവും ധർമ്മപുരിയും കൃഷ്ണഗിരിയുമൊക്കെയാണ് മൽഗോവയുടെ നാടുകൾ

se:laʋm dʰarmmapʊrIjʊm kɻʂɳagIrIjʊmokkeja:ɳ malgo:ʋajʊʈe ɳa:ʈʊkaɭ

<Selam dharmapuri and krishnagiri are the birthplaces of Malgova>

Sample 2: Record Entry 2: 180220_02_01

The sample shown below is an Interview Speech. The interviewer is an agriculture officer of age 50 and interviewee is the owner of farm about 55 years old.

Sentence 1:

മരുഭൂമിയിൽ നിന്ന് ആഗ്രഹിച്ചതുപോലെയുള്ള കാര്യങ്ങൾ ഇവിടെ ഈ കേരളത്തിലെ ഭൂമിയിൽ വന്നപ്പോൾ സാക്ഷാത്ക്കരിക്കാൻ പറ്റിയെന്ന തോന്നുണ്ടോ?

marʊbʰu:mIjIl ɳIɳɳ a:grahIʧʧatʊpo:lejʊlla ka:rjaŋŋaɭ IʊIte i: ke:ralattIle bʰu:mIjIl ʋaɳɳappo:ɭ sa:kʂa:tkkarIkka:ɳ parrIjeɳɳʊ to:ɳɳʊɳɳʊɳʈo:?

<Do you think you could fulfill what you have wished or envisioned from the desert, here in your homeland, Kerala?>

Sentence 2:

തീർച്ചയായിട്ടും, നമ്മൾ ഇവിടെ നമ്മുടെ കൈ കൊണ്ടു വെച്ച് അത് പൂത്ത് അതിന്റെ അകത്ത് നിന്ന് ഒരു മാങ്ങ പറിക്കുക അത് കഴിക്കുക അത് നമ്മുടെ ഏറ്റവും വേണ്ടപ്പെട്ടവർക്ക് കൊടുക്കുക എന്നുള്ളത് സാധിച്ചു.

ti:ɾʈʈaja:jIttʊm, n̪amma[IʊIte n̪ammʊte kəɪ koɳtʊ ʊeʈʈ at pu:tt atIn̪re akatt n̪In̪n̪ orʊ ma:ŋŋa parIkkʊka at kaɻIkkʊka at n̪ammʊte e:rrauʊm ʊe:ɳtappeʈʈauarkk koʈʊkkʊka en̪n̪ʊ[[at sa:dʰIʈʈʊ.

<Definitely we could. What we have planted here by ourselves blossomed, bore fruit, relished it and shared it with our dear ones>

5. Lexicon

The pronunciation dictionary, called Lexicon contains a collection of unique 4925 words. The audio collection process is still going on which will increase the lexicon size. The lexicon consists of word and its corresponding phonemic and syllabic representation as in the example shown in Fig 3.

Word	Phoneme	Syllable
അത് /at/	a t	a t
ഇവിടെ /lvlte/	l v l t e	l vl te
നാടുകൾ /n̪a:tʊkal/	n̪ a: t ʊ k a l	n̪a: tʊ ka l
നമ്മുടെ /n̪ammʊte/	n̪ a m m ʊ t e	n̪a mmʊ te
കഴിക്കുക /ka ɻ lkkʊka/	K a ɻ l k k ʊ k a	Ka ɻ l kkʊ ka
പൂത്ത് /pu:tt/	p u: t t	pu: tt

Figure 3: Example of the lexicon

6. Conclusion

Speech is the primary and natural mode of communication than writing. It is possible to extract more linguistic information from speech than text like emotions and accent. Speech related applications are more useful for illiterate and old people. The articulatory and acoustic information can be obtained from a good audio recording environment. One of the important features of speech data is that, there is less interference from a second party compared to textual data.

To encourage the academic research in speech related applications, a good number of multilingual and multipurpose speech corpora for Indian languages is required. The responsibility to develop such corpora still lies on the shoulder of the concerned researcher. Also the role of language corpora is very significant to preserve and maintain the linguistic heritage of our country.

The release of MSC will be one of the first speech corpora of Malayalam, contributing 200 hours of Narrational Speech and 50 hours of Interview Speech data for public use. The lexicon and annotated speech is also made available with the data. Future work includes creation of corpora related to tourism and entertainment domains and enhancement of quality of speech by building an ASR using KALDI toolkit. The updates on corpus will be accessible through "www.iiitmk.ac.in/vrclc/utilities/ml_speechcorpus".

Acknowledgements

This research is supported by the Kerala State Council for Science, Technology and Environment (KSCSTE). I thank KSCSTE for funding the project under the Back-to-lab scheme. I also thank Agri team, Indian Institute of Information Technology and Management-Kerala, Kissan Project for collecting the audio data.

Bibliographical References

Barnard, E., Davel, M. H., van Heerden, C., De Wet, F., and Badenhorst, J. (2014). The nchlt speech corpus of the south african languages. In *Workshop Spoken Language Technologies for Under-resourced Languages (SLTU)*.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., and Estève, Y. (2018). Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208. Springer.

Koh, J. X., Mislan, A., Khoo, K., Ang, B., Ang, W., Ng, C., and Tan, Y.-Y. (2019). Building the singapore english national speech corpus. *Malay*, 20(25.0):19–3.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Pinnis, M., Auzina, I., and Goba, K. (2014). Designing the latvian speech recognition corpus. In *LREC*, pages 1547–1553.

Rajan, R., Haritha, U., Sujitha, A., and Rejisha, T. (2019). Design and development of a multi-lingual speech corpora (tamar-emodb) for emotion analysis. *Proc. Interspeech 2019*, pages 3267–3271.

Wikipedia contributors. (2020). Malayalam — Wikipedia, the free encyclopedia. [Online; accessed 21-February-2020].

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.