Malayalam Speech Corpus: Design and Development for Dravidian Language

Lekshmi.K.R, Jithesh.V.S & Elizabeth Sherly

24 MAY 2019

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.
- 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.
- This work focuses on 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".

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.[7]
- 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.

Introduction

- In the Automatic Speech Recognition (ASR) area many works are progressing in low-resourced languages.
- 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

Introduction

- 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 air and web 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.

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.[3] [4]
- Similarly, a database is made available with the collection of TED talks in the English language.[2]
- For the Malayalam language-based emotion recognition,a database is available.[6]
- Another work is done on Latvian language.They created 100 hours of orthographically transcribed audio data and annotated corpus also.[5]

Literature Survey

- In addition to that a four hours of phonetically transcribed audio data is also available.
- South Africa has eleven official languages. An attempt is made for the creation of speech corpora on these under resourced languages.[1]
- A collection of more than 50 hours of speech in each language is made available.
- Similarly speech corpora for North-East Indian low resourced languages is also created.[2]

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.
- They were given enough time to record the data.
- If any recording issues happened, after rectification by the recording assistant it was rerecorded.

Narrational and Interview Speech Corpora

പറയാതെ അറിയാം പാലക്കാടാണെന്ന്. ....

കേരളത്തിന്റെ കോട്ടവാതിൽ. ...

ചെന്തമിഴിന്റെ ഈണമുള്ള കാറ്റു വരുന്നുണ്ട്. ....

കരിമ്പനകളിൽ അത് ചൂളം കുത്തിയാടുന്നു. ...

പൊള്ളാച്ചി റോഡിൽ കൊഴിഞ്ഞാം പാറയിലേയ്ക്ക് ഇനിയും ദൂരമുണ്ട്.

പാതയരികിൽ തണൽ മരങ്ങൾ കുടവിരിച്ച് നിൽക്കുന്നുണ്ട്.

ഇടത്തേ കാഴ്ചയ്ക്ക് കുളിരായി നീലനിരാളം പുതച്ചുകിടക്കുന്ന സഹ്യൻ.

എങ്കിലും നെടുകെ മുറിക്കുന്ന ഊഷരമായ കൃഷിയിടങ്ങൾ പലതുണ്ട് പാലക്കാട്ട്.

കാലിക്കൂട്ടം മേയുന്ന പാടവരമ്പിൽ മനുഷ്യ ജീവിതത്തിന്റെ പ്രയാണങ്ങൾ കാണാം.

Figure: Example of script file for dubbing

Narrational and Interview Speech Corpora

- 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 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.

Challenges

- 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.
- The difference in pronunciation styles in the Interview Speech corpora collection.
- The recording used to extend up to 5-6 hours depending on speakers.

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.

Recording Specifications

- A standing microphone is used for recording NS corpora.
- IS corpora is collected directly from the farmers using recording portable Mic at their place.
- 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 Pro Tools are used for the audio post-production process

Recording Specifications

- 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.

Demographics

- 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 are shown in tables.

Demographics

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: Demographic details of speakers by place of cultivation

Demographics

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: Demographic details of speakers by type of cultivation

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.

Transcription

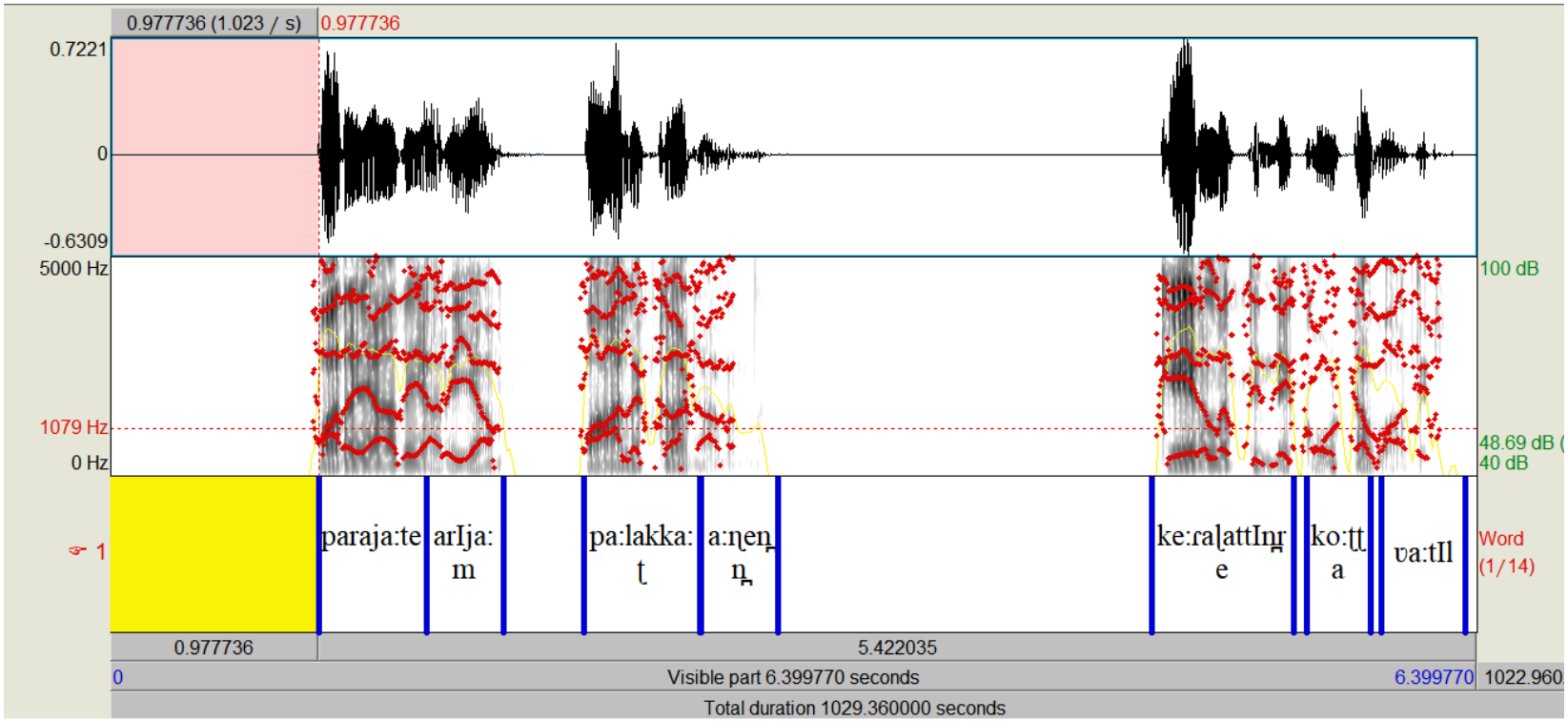

Figure: An example of Annotated Speech Corpora

Transcription

പറയാതെ അറിയാം പാലക്കാടാണെന്ന്

paraja:te arIja:m pa:lakka:ʈa:ɳeɳɳ

<Without saying we can understand that it is Palakkad>

കേരളത്തിന്റെ കോട്ട വാതിൽ

ke:ɾalattIɳɳre ko:ʈʈa ʋa:tIl

< Kerala's Castle door>

സേലവും ധർമ്മപുരിയും കൃഷ്ണഗിരിയുമൊക്കെയാണ്
മൽഗോവയുടെ നാടുകൾ

se:laʊm dʰarmmapʊrIjʊm kɻʂɳagIrIjʊmokkeja:ɳ
malgo:ʋajʊʈe ɳa:ʈʊkal

<Selam dharmapuri and krishnagiri are the birthplaces of Malgova>

Transcription

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

marʊbʰu:mIjIl n̪In̪n̪ a:grahIr̪r̪t̪aʊpo:lejʊlla ka:rjaŋŋal̪ IʊIte i:
ke:ralattIle bʰu:mIjIl ʊan̪n̪appo:l̪ sa:kʂa:tkkarIkka:n̪ parrIjen̪n̪ʊ
to:n̪n̪ʊn̪n̪ʊɳo:?

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

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:rr̪t̪aja:jIttʊm, n̪ammal̪ IʊIte n̪ammʊte kəɪ koɳtʊ
ʊer̪t̪ at pu:tt atIn̪re akatt n̪In̪n̪ orʊ ma:ŋŋa parIkkʊka
at kaɻIkkʊka at n̪ammʊte e:rraʊm ʊe:ɳtappet̪t̪aʊarkk
koɽʊkkʊka en̪n̪ʊllat sa:dʰIr̪t̪t̪ʊ.

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

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.

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 /kaɻlkkʊka/	K a ɻ l k k ʊ k a	Ka ɻl kkʊ ka
പൂത്ത് /pu:tt/	p u: t t	pu: tt

Figure: Example of the lexicon

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.
- To encourage the academic research in speech related applications, a good number of multilingual and multipurpose speech corpora for Indian languages is required.
- The role of language corpora is very significant to preserve and maintain the linguistic heritage of our country.

Conclusion

- 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.
- The updates on corpus will be accessible through "www.iiitmk.ac.in/vrclc/utilities/ml_speechcorpus".

References I

[1] Etienne Barnard et al. "The NCHLT speech corpus of the South African languages". In: *Workshop Spoken Language Technologies for Under-resourced Languages (SLTU)*. 2014.

[2] François Hernandez et al. "TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation". In: *International Conference on Speech and Computer*. Springer. 2018, pp. 198–208.

[3] Jia Xin Koh et al. "Building the singapore english national speech corpus". In: *Malay* 20.25.0 (2019), pp. 19–3.

[4] Vassil Panayotov et al. "Librispeech: an asr corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 5206–5210.

References II

[5] Marcis Pinnis, Ilze Auzina, and Karlis Goba. "Designing the Latvian Speech Recognition Corpus.". In: *LREC*. 2014, pp. 1547–1553.

[6] Rajeev Rajan et al. "Design and Development of a Multi-lingual Speech Corpora (TaMaR-EmoDB) for Emotion Analysis". In: *Proc. Interspeech 2019* (2019), pp. 3267–3271.

[7] Wikipedia contributors. *Malayalam — Wikipedia, The Free Encyclopedia*. [Online; accessed 21-February-2020]. 2020. URL: https://en.wikipedia.org/w/index.php?title=Malayalam&oldid=941882964.

Thank You !!!