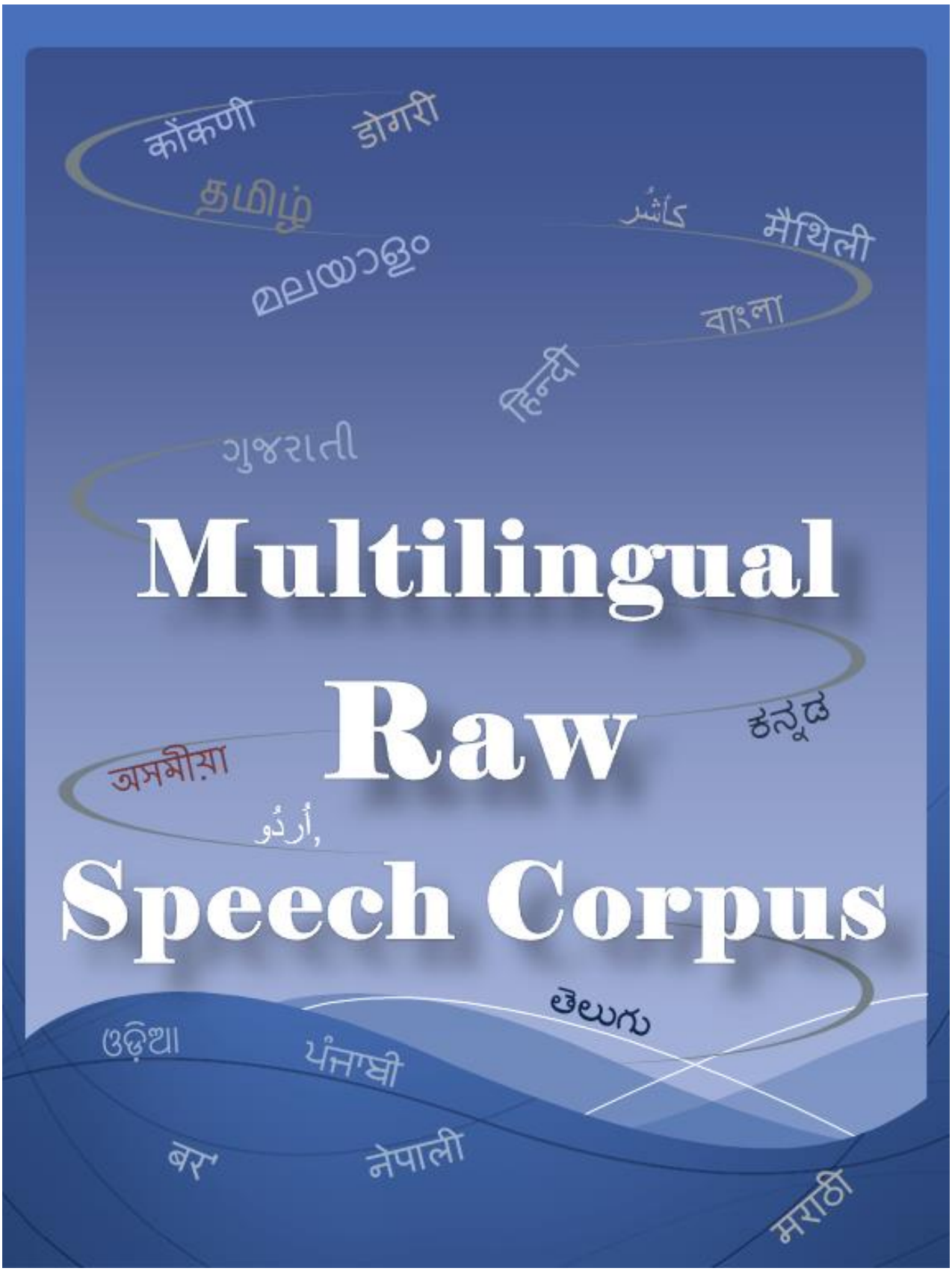
कोंकणी
डोगरी
தமிழ்
كأشُر
मैथिली
മലയാളം
বাংলা
हिन्दी
ગુજરાતી
Multilingual
Raw
Speech Corpus
ಕನ್ನಡ
অসমীয়া
اُردُو,
తెలుగు
ଓଡ଼ିଆ
ਪੰਜਾਬੀ
बर'
नेपाली
मराठी

Multi-Lingual Raw Speech Corpus

Authors:
Rajesha N.
Manasa G.
Narayan Choudhary

Annotated, quality language data (both-text & speech) and tools in Indian Languages to Individuals, Institutions and Industry for Research & Development - Created in-house, through outsourcing and acquisition.

Linguistic Data Consortium for Indian Languages
Central Institute of Indian Language
Mysore, India-570006

CENTRAL INSTITUTE OF INDIAN LANGUAGES
Manasagangotri, Mysuru, Karnataka, India, 570006
www.ciil.org

Title : Multilingual Raw Speech Corpus
Authors: Narayan Kumar Choudhary, Rajesha N., Manasa G.

e-ISBN: 978-81-948885-3-6
CIIL Publication No.: 1281

First published: AD 2021 May
Vaisakha 1943 Saka

© *Central Institute of Indian Languages, Mysuru 2021*

Publisher: C.G. Venkatesha Murthy, Director, CIIL

Production Team
Head, Publication Unit: Umarani Pappuswamy
Officer-in-Charge, Publication Unit: Aleendra Brahma
Artist: H. Manohara
Staff-in-charge: R. Nandeesh
Compositor: M.N. Chandrashekar
Cover design:N. Rajesha

Contents

Table of Figures

1 LDC-IL MULTI-LINGUAL RAW SPEECH CORPUS

1.1 INTRODUCTION

The LDC-IL Multi-Lingual Raw Speech Corpus dataset is extracted from the raw speech corpora published by LDC-IL in various Indian languages. The Multi-Lingual speech dataset sampling is taken from the content type of 'Creative Text-T2' There are three age groups selected from the LDC-IL datasets. They are, '16 to 20 years', '21 to 50 years' and 'above 50 years'. For more details about how the data is collected from the field, coverage, etc., please refer the overview of speech corpora (Choudhary, et.al, 2019) and specific language documentation available at the LDC-IL Data distribution portal (https://data.ldcil.org). This dataset is built to address the needs of some applications like language identifier modules where multiple language samples are a requirement, to explore cross-linguistic variations and diatopic comparison to determine what generalizations are possible about the types of variable features, to build multilingual phoneme set and models etc.

1.2 CONTENT TYPE

This Multi-Lingual speech dataset sampling is taken from the content type of 'Creative Text-T2'. 'Creative Text-T2' is extracted mainly from literary sources. It is used to capture literary terms. Creative Texts are Stories or Essays collected from books. It exhibits the language style of the period from which the text is taken.

The creative text of the LDC-IL Speech dataset comprises of essays or short stories. One of these essays or short stories, selected randomly from a data set, is assigned to a speaker for reading out. The same story may be read out by multiple speakers.

1.3 TECHNICAL SPECIFICATIONS

The LDC-IL data is recorded using Roland EDIROL Recorder. It is a 24-bit Linear PCM (R-09) Recorder, at the sample rate is 48.0 KH, with 16 bit WAV recording mode. The audio segments are recorded using rechargeable batteries or alkaline batteries.

1.4 TEXT - SPEECH MAPPING AND NAMING CONVENTIONS

The collected data is segmented and mapped with its corresponding text and metadata. Each recording is named in accordance with its metadata information like Language Name, Speaker id, Content id, Gender, Age, Content type etc.

The Naming convention of LDC-IL Multi-Lingual Raw Speech Corpus dataset is as follows:
LDC_IL_Scheduled_<Language>_<Gender>_<Age Group>_<Content Type>_<Speaker ID>_<ContentID>

A Typical LDC-IL naming convention for Speech corpus is shown bellow.

LDC-IL_Scheduled_Bodo_Female_16To20_Creative Text-T2_SP-0021_T2-0004.wav
LDC-IL_Scheduled_Bodo_Female_16To20_Creative Text-T2_SP-0021_T2-0004.txt

Where .wav extension represents the audio and the .txt represents corresponding metadata file

2 METADATA

The value of speech data can be determined according to the quality of metadata obtained. It is imperative to maintain metadata of the entire data collection for linguistic analysis. A brief of each of these 25 fields/legends is given in the table below:

SL	Legend	Description
1	Language	Name of the Language
2	SpeakerID	Each speaker has a unique identity language. However, this is within the language. If one is working on speech corpus from more than one language, the IDs may get repeated.
3	ContentType	This corresponds to the notation of the content types noted above.
4	ContentID	This corresponds to the ID of the text being read out.
5	Gender	Notes gender, whether it is male, female or other.
6	AgeGroup	Three age groups of 16 to 20, 21 to 50, and50+
7	Dialect	Notes the dialect of the language. An attempt has been made to cover all the dialects of the language as agreed upon in the academia of the language experts.
8	ReadInScript	The script in which the content has been provided to read in.
9	RecordingEnvironment	A brief info on the environment in which the recording has been done.
10	PowerSource	The source of the power using which the recording was done. It may be Li-ion, NiCd or Alkaline batteries.
11	RecorderManufacturer	Manufacturer of the recorder.
12	RecorderType	Type of the recorder. It is mostly 24-bit Linear PCM (R-09).
13	SamplingFrequency	Sampling frequency. It's mostly 48.
14	BitPerSample	Bit per sample. It is mostly 16-bit.
15	Channel	How many channels. All of LDC-IL data is stereo. Only data set is mono which is segregated and constitutes a separate dataset of its own.
16	State	Name of the Indian state/province to which the speaker belongs to.
17	District	Name of the Indian district to which the speaker belongs to.
18	Place	Name of the place to which the speaker belongs to.
19	MotherTongue	Mother tongue of the speaker. It is note that data has been taken from people who professo to speak the language. However, it may be that the speaker uses the target language as a second or third language. However, as long as the speaker confidently says (and it is also verified by other speakers of the community), some samples have been taken from these types of users as well. This adds to the variety of the speech data collected.
20	EducationalQualification	Highest educational qualification of the speaker.
21	PlaceOfElementaryEducation	Place of the elementary education. This usually corresponds to the early childhood experiences which happens to more than often affect the way a language spoken.
22	RecordingDate	Date when the recording took place.
23	Investigator	Name of the Investigator.
24	RecordedText	Text of the recorded speech (in the script of the language).
25	TextInRoman	Text of the recorded speech (in the Roman transliteration as per the LDC-IL transliteration scheme. If the text is long (as is the case with T1 and T2 content types), a reference of the corresponding file is given.)

Table 1: Metadata Legends and their Description

2.1 TRANSLITERATION SCHEME

The recorded text is provided in the native language as well as in the transliterated format of Roman (except for Kashmiri and Urdu) The transliteration schema for each language are given in the LDC-IL transliteration schema document available at https://ldcil.org/Tools/CorporaToolsPackage/LDC-IL%20Transliteration%20Scheme.pdf.

3 SUMMARY OF THE CORPUS

The LDC-IL multi-lingual raw speech has these figures may be helpful in tuning the corpus for various purposes of training, testing and evaluating various algorithms as well as provide useful insights into the dataset. The data size is of 62.2 GB with the total duration 97:43:54 (hh:mm:ss) comprising 1,916 Speakers.

The table below shows the distribution of each language in terms of total number of Speakers, Size and Duration in LDC-IL Multi-Lingual Raw Speech Corpus

Language	Female			Male			Total		
	Duration (hh:mm:ss)	Speakers	Size (in GB)	Duration (hh:mm:ss)	Speakers	Size (in GB)	Duration (hh:mm:ss)	Speakers	Size (in GB)
Assamese	2:33:40	68	1.64	2:34:33	64	1.65	5:08:13	132	3.30
Bengali	2:38:34	56	1.59	2:47:32	61	1.69	5:26:06	117	3.29
Bodo	2:30:39	42	1.61	2:41:04	40	1.72	5:11:43	82	3.34
Dogri	1:16:44	30	0.84	1:35:00	31	1.01	2:51:44	61	1.84
Gujarati	2:32:10	45	1.63	2:30:40	42	1.61	5:02:50	87	3.25
Hindi	2:37:28	44	1.66	2:30:18	44	1.57	5:07:46	88	3.23
Kannada	2:37:06	45	1.68	2:32:50	48	1.63	5:09:56	93	3.32
Kashmiri	2:32:26	30	1.63	2:39:46	29	1.71	5:12:12	59	3.34
Konkani	2:50:24	62	1.82	2:41:25	62	1.74	5:31:49	124	3.57
Maithili	2:46:28	54	1.71	2:53:31	50	2.00	5:39:59	104	3.48
Malayalam	2:38:16	68	1.69	2:28:17	61	1.59	5:06:33	129	3.29
Manipuri	2:15:42	29	1.45	2:44:43	32	1.76	5:00:25	61	3.22
Marathi	2:38:26	56	1.70	2:41:57	58	1.73	5:20:23	114	3.43
Nepali	2:51:09	44	1.83	2:58:41	52	1.91	5:49:50	96	3.75
Odia	2:38:24	63	1.70	2:32:10	60	1.63	5:10:34	123	3.33
Punjabi	2:41:13	67	1.72	2:35:40	62	1.66	5:16:53	129	3.40
Tamil	2:35:24	78	1.57	2:45:20	70	1.66	5:20:44	148	3.24
Telugu	2:06:18	24	1.33	3:00:40	38	1.93	5:06:58	62	3.27
Urdu	2:20:22	53	1.50	2:48:54	54	1.81	5:09:16	107	3.31

Table 2: Summary of Multi-Lingual Raw Speech Corpus