
Indian English - Kannada Variant
Sentence Aligned Speech Corpus

Indian English - Kannada Variant Sentence Aligned Speech Corpus

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: Indian English-Kannada Variant Sentence Aligned Speech Corpus
Authors: Rejitha K. S., Vijayalaxmi F. Patil, Rajesha N., Manasa G., Srikanth D., Nithin S.,Narayan Kumar Choudhary, Shailendra Mohan

ISBN: 978-81-19411-35-1
CIIL Publication No.: 1433

First published: AD 2023 September
Bhadrapada 1945 Śaka

Publisher: Prof. Shailendra Mohan, 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: K.S. Rejitha

Contents

List of Figures

List of Table

1 Speech Annotation

1.1 Introduction

Linguistic Data Consortium for Indian Languages (LDC-IL) has been acting as the repository of Indian language datasets since 2008. It has so far released 42 datasets covering 20 scheduled languages. The released datasets include raw text corpora and raw speech corpora. More details about the scheme and the datasets are available in Choudhary, 2021 and Choudhary and Rao, 2020.

This paper presents a documentation of sentence aligned speech corpora in various languages. This is in continuation of the earlier raw speech corpora in various languages. The difference between the raw speech corpora and the sentence aligned corpora can be easily guessed. As the title of the dataset suggests, the sentence aligned corpus is split at the 'sentence' or a meaningful 'utterance' boundary while the raw speech corpus had speech segments that were usually larger in size.

1.2 LDC-IL Speech Annotation

LDC-IL speech corpora are collected using two methods (complete details about the raw speech corpus and the process of its creation are included in the raw speech corpus documentation of respective languages).

a. Read Speech: A vast majority of the speech dataset in LDC-IL is created by reading texts from various sources. These sources are usually part of the corresponding language taw text corpus.

b. Spontaneous Conversation: These have been used for data collection in some of the languages and conversational data for other languages will be collected in subsequent phases.

We have manually transcribed both of these kinds of data and aligned the transcriptions at the sentence level. Even though the read speech data is created by reading the texts, a manual transcription was necessitated because a lot of times speakers do not reproduce the texts verbatim and so the original texts do not necessarily correspond to the speech recordings. Moreover, the speech transcriptions represent the different variant forms in speech as well as other speech properties (viz false start, mispronunciation, etc) and non-speech properties (viz background noise) - all of these are meticulously marked during the annotation process.

LDC-IL speech data is annotated with the official script of a language, wherever such a script is specified. For example, Malayalam speech data is annotated in Malayalam Script in UTF-8 encoding, Hindi is annotated in Devanagari script and so forth. The languages which do not have

an official script are annotated using the most commonly used script for writing that language. All annotations and transcriptions are carried out using the Praat software.

1.2.1 Speech Annotation

Speech is transcribed according to what the speaker says. This annotation layer is prepared on the basis of pronunciation and it is strictly following the LDC-IL speech annotation guideline. If the speaker articulates a word which is not in the prompt sheet then the uttered word is transcribed in the annotation text.

3.1.1.1 Guidelines for Sentence Aligned Speech Annotation

- Initial phase starts with aligning the audio files with the corresponding text file
- Annotation of this aligned data is carried out as per the pronunciation of the speaker in the audio. The wrong pronunciation, that is, deviation from the text is transcribed accordingly.
- Following are the instances which should be marked in the annotation

1. Use of '#'

While annotating the audio data, '#' symbol is used when the audio file has unnecessary sounds which need to be removed. Boundary is marked in the unwanted portion with '#' symbol. After the whole audio file is marked in the similar manner, it is processed which removes the portion marked with '#' along with boundary lines and merges the file.

2. Use of '0'

The use of '0' is done in the audio file in case of long silences or unwanted speech. Boundaries are made in the beginning and the end of the unwanted portion and marked as 0. Once the file processes the portion which is labelled as 0 will be removed from the audio file and the boundary is created automatically. i.e. '0' has to be used in the annotation only when the particular place needs a boundary. For e.g. A sentence ends and silence or flipping sound of paper before the next sentence starts. 0 is used in this case. It is also used in cases where long silence or unwanted speech is present in the beginning or at the end of the audio file. In such cases, the boundary is determined and marked as 0. After processing the audio file, this portion will be removed.

3. Silences are removed.

A period of relative quietness occurs in the audio files where the speaker stops to think or hesitates before saying a word. The expression relative quietness is used here because there is no actual silence in the speech signal due to line and environmental noise. Any silence longer than 50 ms is marked using #.

4. Cut off speech and intended speech is marked.

E.g.: a. [mini]*ster —shows that the speaker intended to speak minister but spoke mini in an unclear fashion and ster clearly.

5. Annotation of speech disfluency:

Restarts/false starts are marked. For example, the speaker intends to speak “Mysuru” but speaks “my Mysuru”. Then it is marked as ‘my Mysuru’.

6. Numbers:

All number sequences are spelled out. Years are transcribed in spoken format. E.g.: 1983 is transcribed as spoken—“nineteen eighty three.”

7. The text has been transcribed as per the wave. For e.g.: If the wave is ‘They is going’ then it is transcribed as ‘They is going’. No grammatical correction is done.

8. Mispronunciations

If a speaker mispronounces a word and the mispronunciation is not an actual word, transcription is done as the word is spoken.

9. Symbols and Abbreviations:

Punctuation that is part of the English orthographies, like Hyphenated words and apostrophes, is retained as such but punctuation like full stop, comma, semicolon, etc. is not used. Abbreviations are written as pronounced.

10. Utterances should be no longer than 30 secs. So the annotator should find a long silence around 500 ms and split the sentence appropriately.

1.3 Speech Annotation quality assurance

Transcriptions are susceptible to human errors, such as typos or spelling mistakes. To ensure data accuracy, both phonetically normalized, and orthographically normalized textual layers undergo third party evaluation. The third-party evaluator must check the transcription against its corresponding audio to rectify any mistakes. The linguist may accept or reject the corrections by means of matching the original and evaluated transcriptions against the audio. The linguist's specialized knowledge is essential in bringing the transcriptions back in line with accuracy. It is ensured that the mistakes encountered by third party evaluators are also corrected.

2 INDIAN ENGLISH - KANNADA VARIANT SPEECH ANNOTATION

2.1 Overview of Sentence Aligned Speech Corpus

Indian English – Kannada variant Sentence Aligned Speech Corpus is created by annotating the speech data (Ramamoorthy L., et. Al, 2021) collected by LDC-IL. A detailed explanation of the Indian English – Kannada variant Speech Corpus will be available in the Indian English - Kannada Speech Data Documentation (Rejitha K.S., et. Al, 2021). LDC-IL Indian English – Kannada Variant Sentence Aligned Speech files contains an audio file and a textual layer in Roman script. Each File is named in accordance with its metadata information like language name, speaker id, content id, gender, age, content type etc.

A Typical LDC-IL naming convention for Sentence Aligned Speech data is
'EnglishKan_Female_Above51_Creative_Text-T2_SP-0046_T2-0002-038.wav'

The speech is annotated on the basis of English spelling convention. The words are labelled manually to the corresponding wave. LDC-IL Sentence Aligned Speech corpus contains four content types such as contemporary text, creative text, sentences and date format. The contemporary text and creative text are recordings of news and essays/novels. Each speaker has uttered typically 25 sentences randomly selected from phonetically balanced sentences list of LDC-IL speech data set. Date format is kept as uttered by the speaker. The corpus consists of an audio file for each recording and corresponding textual layer consisting of the orthographically normalised annotation.

2.2 Observations

LDC-IL sentence level speech annotation strictly follows the Standard English spellings. Deviations in the phonological features are evident in Indian English because English is an inter-language in India. Indians have their own mother tongue and those language features are drastically influencing their English pronunciation. Therefore, speakers from different regions speak the same word in different ways. Most of the Indian-English speakers do not know the Received Pronunciation properly. Stress, accent, intonation pattern of Received Pronunciation is difficult to follow for a common Indian speaker. (Rejitha K.S., N. Rajesha, 2020).

The reading speed differs from reader to reader. Fast reading informants pose difficulty in annotation. Since news items contain sports news, it includes the informant reading all types of numbers. Speakers sometimes wrongly utter large digit numbers like thousands or lakhs, decimal numbers, fractions etc. It is observed that speakers read Cricket score, Tennis score etc. in their own way and very few speakers read it properly. Most of the speakers show difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely used words interrupt the reader's fluency. All these factors contribute to the complexity in speech which makes it a rather difficult task. Since the dialect of the annotator can differ from that of the

informant, the annotation process may need repetitive hearing in some cases. The annotation has to discard the data in particular places where the investigator has communicated with the informant. Some background noise like the sound of a bell, bus horn can be heard in the recording. Since this can be heard along with the voice of the informant, they have to be retained. This slows down the annotation process. Vocal noise of informants like coughing, sneezing etc. can also be observed.

2.2.1 Phonetic Alternation in English Speech Data

Read speech has disfluency like unwanted pauses, elongated syllables, word fragments, self-corrections, and repeated words. When speakers notice what they utter then they suspend their speech and add, delete, or replace words they have already produced. Some fluctuated occurrences were detailed as follows:

English is a non-phonetic language and there is no one-to-one correspondence between the letters and the sounds. Since Indian languages are almost phonetic, people have difficulty in the production of English sounds. Stress, intonation and pitch are essential factors to produce English words and sentences which are difficult for a foreign speaker. These factors vary word by word. Such specific language oriented characteristics make trouble in language learning.

- **Deviation in Vowels**

Kannada speakers equate the English vowels /a, ʌ, ɜ: and ə/ equate with / ə or a /. South Karnataka people use /a/ whereas north Karnataka people use /ə/. /ɪ/ become /e/ and /ɔ/ become /ʌ/ in most of the observations. It is a common phenomenon among Kannada speakers that the English short vowels tend to be long vowels and diphthong becoming monophthong and vice versa. Pillar /pɪlə/ as /pailar/. The mother tongue influence introduces vowels at the end of words like school /sku:l/ become /sku:lu/. This is a common phenomena even in loan words. For e.g.: [bus] as [basu], [car] as [caru]

- **Deviation in Consonants**

Several Kannada speakers have difficulty to articulate specifically /s/ and /ʃ/, /f/ and /ph/, /w/ and /v/ etc. Some of the Kannada speakers pronounce words in different way such as so /so/ as /ʃo/ and /bats /bæts/ as /bætz/. Alveolar fricative /s/ is spread across post-alveolar, retroflex and palatal positions. Kannada speakers are not properly distinguishing /s/, /z/, /ʒ/ and /ʤ/. Non-standard English speakers in Karnataka pronounce/morning /ˈmɔːnɪŋ/ as/ˈmoːrnɪŋ/ and learn /lɜːn/ as /lɜːrn/. Most of the Kannada Speakers are rhotic. They utter 'r' in middle and final positions even when it is silent.

Words ending with l, m, n are pronounced as əl, əm, ən. It is noticed in southern Karnataka, the ‘h’ insertion or deletion in Kannada words' pronunciation in large masses. Gemination is

common in Dravidian languages so it is a habit to utter the double consonants. For e.g. summer /sʌmə(r)/ as /səmmər/ and cindrella /sɪndərelə/ as /sɪndərellə/ or /sɪndərella/

- **Other Observation**

When there are unfamiliar words in the prompt sheet the speakers try to utter according to the spelling. For eg: 'aisle' /aɪl/ is pronounced in many ways such as /aɪl/, /asəɪl/, /əɪsel/, /asel/; daining room /daɪnɪŋ ruːm/ as /dɪnɪŋ ruːm/. The speaker usually interchanges the articles 'a, an, the', pronouns 'he and she', article 'the' with the pronoun 'that'. Moreover they miss or alter the prepositions. The speakers assume the forthcoming words and utter words which might not be in the prompt sheet. It is extensively observed that the plural marker 's' is omitted or added where it is not needed. Indian English speakers' Suprasegmental features like intonation, stress and tone are different from Received Pronunciation.

2.3 Summary of the Corpus

The total duration of Indian English – Kannada variant Sentence Aligned Speech Corpus is 11:17:40 (hh:mm:ss) comprising 6,166 audio segments from 53 speakers. Figures 1, 2 and 3 show the distribution of the corpus with respect to gender, age and content type, respectively. Figures 4 and 5 are showing gender and age distributions for each content type respectively. Table 1 gives a break-up of the corpus in terms of recordings obtained from different kinds of texts and also other demographic details. Table 2 shows the age and gender-wise distribution of all the speakers.

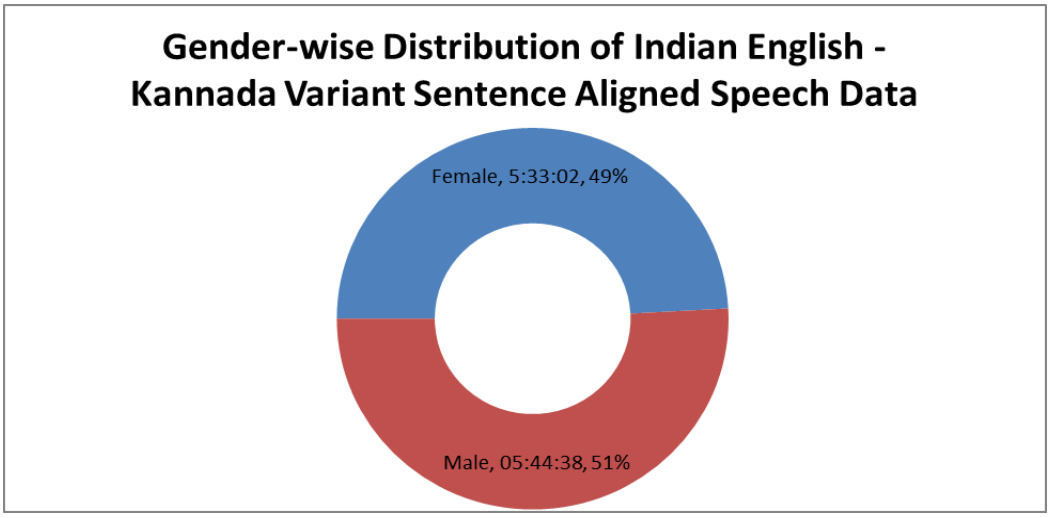

Figure 1:Gender-wise Distribution of Indian English - Kannada Variant Sentence Aligned Speech Data

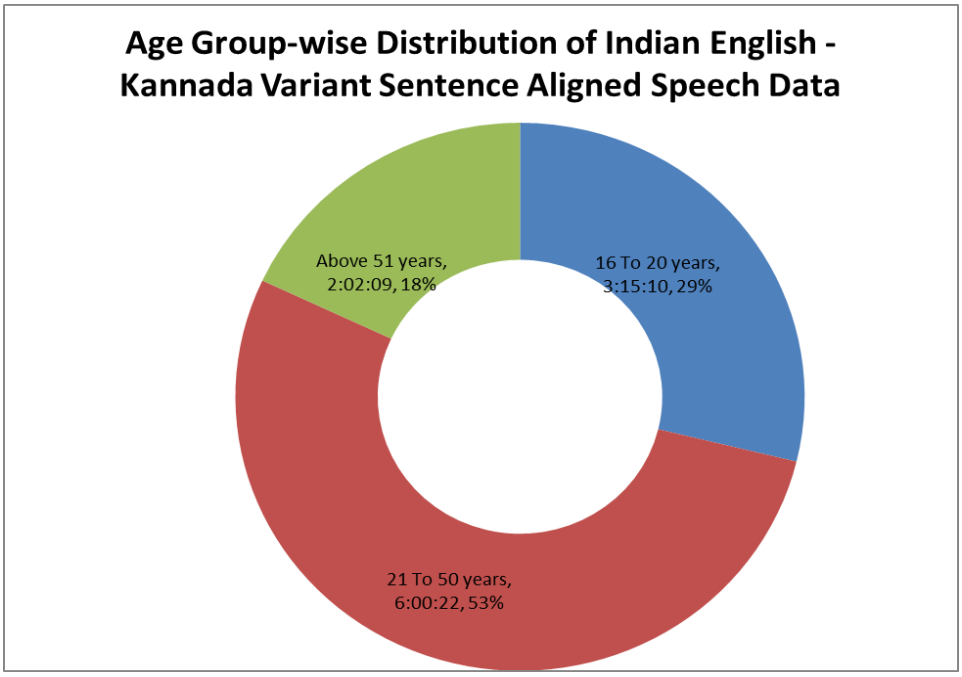

Figure 2: Age Group-wise Distribution of Indian English - Kannada Variant Sentence Aligned Speech Data

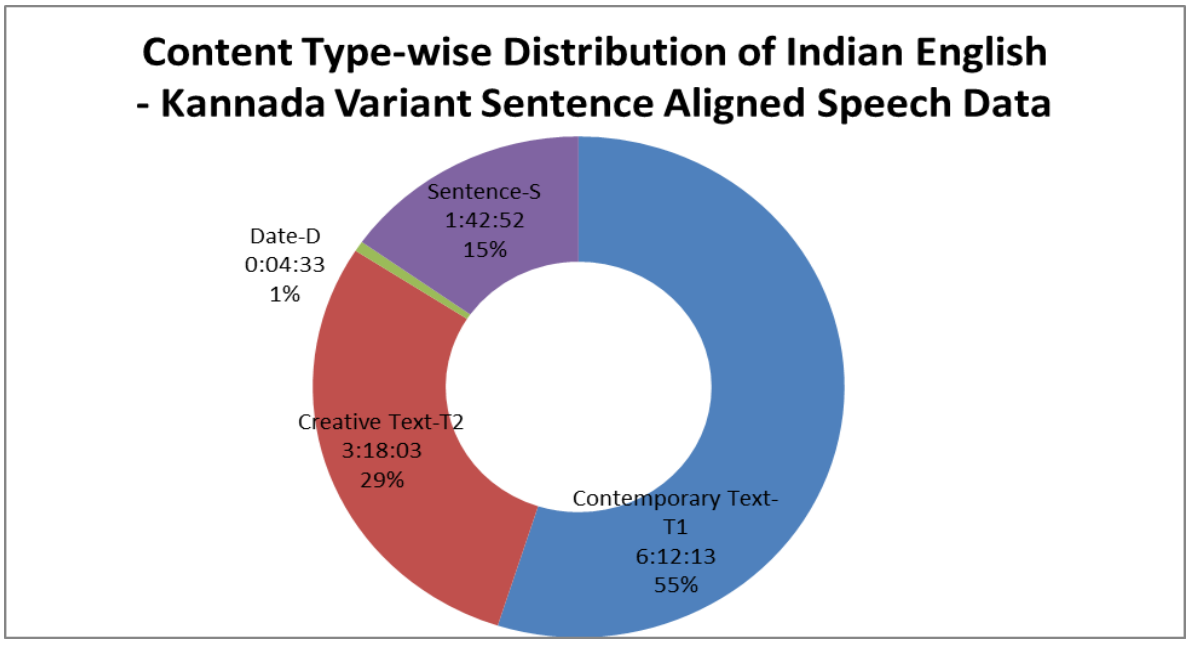

Figure 3: Content Type-wise Distribution of Indian English - Kannada Variant Sentence Aligned Speech Data

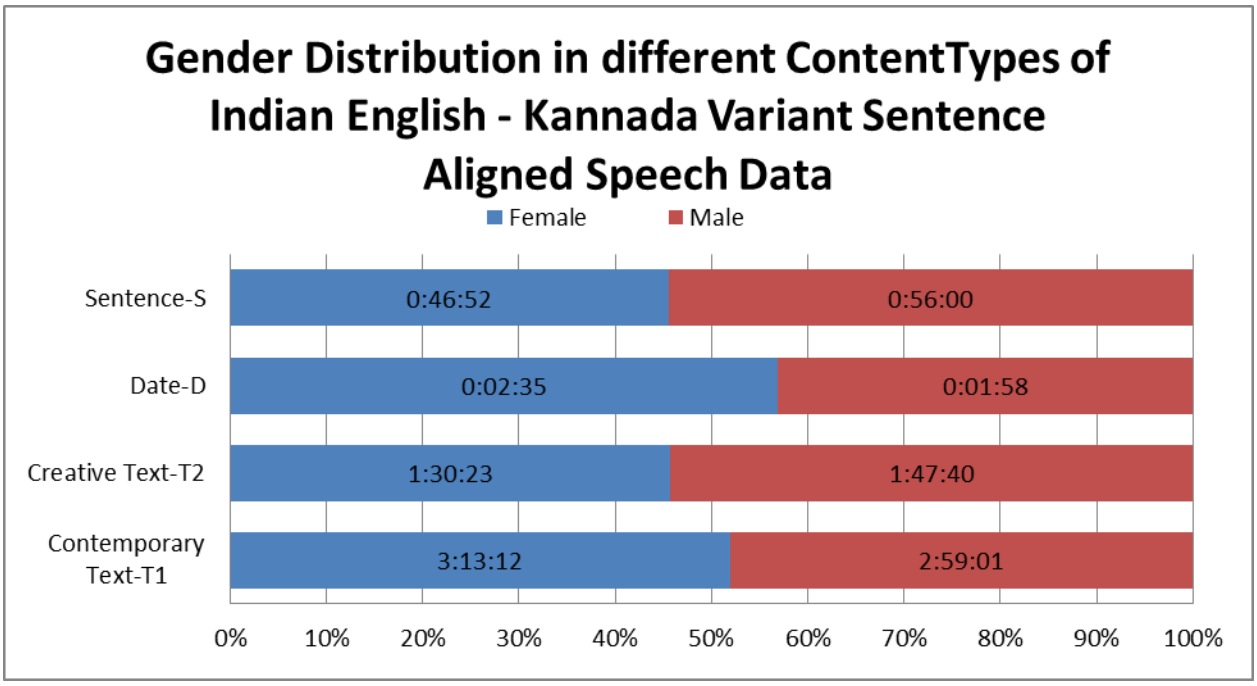

Figure 4: Gender Distribution in different Content Types of Indian English - Kannada Variant Sentence Aligned Speech Data

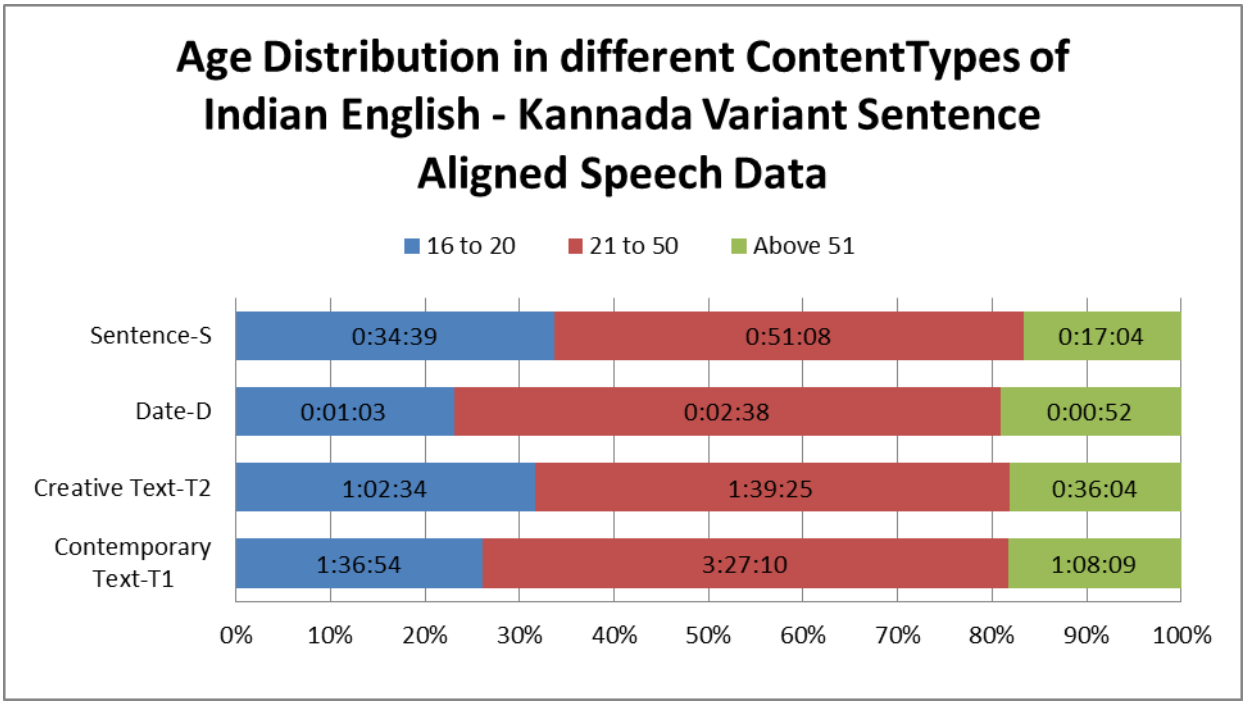

Figure 5: Age Distribution in different Content Types of Indian English - Kannada Variant Sentence Aligned Speech Data

2.3.1 Duration of the Indian English - Kannada Variant Sentence Aligned Speech Data

The table below shows the duration of each of the content types and their distribution across a few factors in Indian English – Kannada Variant Sentence Aligned Speech Data.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	01:02:28.741077	03:13:11.782010	06:12:13.027237
		21To50	01:40:24.956010		
		Above51	00:30:18.084923		
	Male	16To20	00:34:25.205666	02:59:01.245227	
		21To50	01:46:45.262068		
		Above51	00:37:50.777493		
Creative Text-T2	Female	16To20	00:23:16.523228	01:30:23.163705	03:18:02.845505
		21To50	00:50:25.904379		
		Above51	00:16:40.736099		
	Male	16To20	00:39:17.552623	01:47:39.681799	
		21To50	00:48:59.361319		
		Above51	00:19:22.767856		
Date-D	Female	16To20	00:00:36.554232	00:02:35.074273	00:04:32.600275
		21To50	00:01:27.910021		
		Above51	00:00:30.610020		
	Male	16To20	00:00:26.264155	00:01:57.526002	
		21To50	00:01:09.751868		
		Above51	00:00:21.509979		
Sentence-S	Female	16To20	00:11:32.067080	00:46:52.231114	01:42:51.771055
		21To50	00:26:25.565244		
		Above51	00:08:54.598789		
	Male	16To20	00:23:06.898949	00:55:59.539942	
		21To50	00:24:42.836948		
		Above51	00:08:09.804044		

Table 1: Indian English – Kannada Variant Sentence Aligned Speech Data Duration

2.4 Summary of Speakers

The table below shows the total number of speakers and their distribution in the Indian English – Kannada Variant Sentence Aligned Speech Data.

Age Group	Female	Male	Total
16To20	7	6	13
21To50	14	16	30
Above51	5	5	10
Total	26	27	53

Table 2: No. of Speakers of Indian English – Kannada Variant Sentence Aligned Speech Data

3 REFERENCES

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora. In Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
2. Choudhary, N. 2021. LDC-IL: The Indian Repository of Resources for Language Technology. Language Resources & Evaluation. Springer, Vol. 55, Issue 1. doi: https://doi.org/10.1007/s10579-020-09523-3
3. Choudhary, Narayan, Rajesha N., Manasa G. & L. Ramamoorthy. 2019. “LDC-IL Raw Speech Corpora: An Overview” in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 160-174.
4. K.S. Rejitha & N.Rajesha A Corpus Analysis of English Pronunciation Deviations among Kannada Speakers. Working Papers on Linguistics and Literature, Department Of Linguistics, Bharathiyar University, Coimbatore. Volume: XIV 2020. pp. 367-373. ISSN 2349-8420.
5. Rejitha K.S., Rajesha N., Manasa G., Narayan Choudhary. 2021. "Indian English Raw Speech Corpus - Kannada Variant" in Compendium of Linguistic Resources in Indian Languages, Central Institute of Indian Languages, Mysore. pp. 58-65.
6. Ramamoorthy L., Narayan Kumar Choudhary, Bharatha Raju A., Rejitha KS, Rajesha N., Manasa, G. 2021 Indian English Raw Speech Corpus - Kannada Variant Central Institute of Indian Languages, Mysore. 978-81-948885-9-8.