मैथिली
अनुवाचिका कॉर्पस

Maithili
Text to Speech Corpus

MAITHILI TEXT TO 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
Mysuru, India-570006

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

Maithili Text to Speech Corpus

Shantanu Kumar, Stephen Fernandes, Nithin S., Roopashri M. R., Dr. Narayan Kumar Choudhary,
Prof. Shailendra Mohan

e-ISBN: 978-93-48633-36-1
CIIL Publication No.: 1515

First published: AD 2025 March :: *Phalguna 1946 Śaka*

Publisher: Prof. Shailendra Mohan, Director, CIIL

Printing & Publication Unit
Aleendra Brahma, Officer-in-Charge, Printing Press & Publication Unit
B Manju Bhargavi, Unit-in-Charge, Printing Press & Publication Unit
M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

Cover design: Ankita Tiwari

TABLE OF CONTENTS

Table of Contents iv
Figures iv
Tables iv
1 Text to Speech Corpus 1
1.1 Introduction 1
1.2 TTS Prompt Preparation 1
1.3 Data Preparation 1
1.4 quality assurance 2
1.5 References 2
2 Maithili Text to Speech Corpus 3
2.1 Introduction 3
2.2 Overview 3
2.3 References 4

FIGURES

Figure 1: Gender-Sentence Category Chart 4
Figure 2: Category-Text Chart 4

TABLES

Table 1: Summary of the Corpus 3

1 TEXT TO SPEECH CORPUS

Dr. Narayan Choudhary, Rejitha K. S.

1.1 INTRODUCTION

The Linguistic Data Consortium for Indian Languages (LDC-IL), established by the Ministry of Education, Government of India, serves as a national repository of linguistic data at the Central Institute of Indian Languages, Mysore. It develops and distributes qualitative linguistic resources and software technologies to support and enhance Indian languages. Details of previously published linguistic datasets by LDC-IL are available in [1], [2] and [3].

A text-to-speech (TTS) is a speech synthesis technology that converts written text to a natural-sounding speech. Text-to-Speech (TTS) applications utilize speech synthesis technology to convert textual content into audible speech, making them a valuable tool across various domains. These applications analyze text to determine proper pronunciation and generate corresponding speech output. By enabling users to listen to digital content, TTS technology enhances accessibility and usability. It is widely utilized in navigation systems, customer service interfaces, and assistive technologies, particularly benefiting individuals with visual impairments by providing digital accessibility.

LDC-IL developed a high quality Text-to-Speech (TTS) dataset contains speech recordings and the corresponding transcriptions along with the metadata. The dataset can be used in research, development, and evaluation of Text-to-Speech systems.

1.2 TTS PROMPT PREPARATION

The text prompts were majorly prepared from LDC-IL text corpus and ensured it covered a wide range of phonetic features. The corpus includes all kinds of sentences with varying stress and pitch patterns. It is ensured that, no word of any sentence has more than 4 syllables. Sentences that have an overlap of more than 50% words are avoided.

1.3 DATA PREPARATION

The prepared prompts were recorded in studio environment by professional voice artists. LDC-IL preferred one male and female from 21-50 age group. The specification of audio is as follows:

a. **Sampling Rate** : 48 kHz
b. **Bit Rate** : 16000 Mbps
c. **Channels** : Stereo
d. **File Format** : WAV

1.4 QUALITY ASSURANCE

Once the data is recorded, it undergoes transcription and subsequent evaluation by a third-party evaluator. The evaluator is responsible for verifying the transcription against the corresponding audio to identify and correct any inaccuracies. It is ensured that the mistakes encountered by a third party evaluator are also corrected by arbitrating it further by a third linguist.

1.5 REFERENCES

1. Choudhary, Narayan (ed.). 2019. Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-295-8.
2. Choudhary, Narayan (ed.). 2021. Compendium of Linguistic Resources in Indian Languages. Central Institute of Indian Languages, Mysore. ISBN: 978-81-948885-6-7.
3. Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2023. Compendium of LDC-IL Sentence Aligned Speech Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-19411-34-4.

2 MAITHILI TEXT TO SPEECH CORPUS

Shantanu Kumar, Narayan Kumar Choudhary

2.1 INTRODUCTION

Maithili is a language spoken in India and Nepal and has six major varieties: Standard Maithili, Southern Standard Maithili, Western Maithili, Eastern Maithili, Chhika-chhiki, and Jolha Boli[1]. Based on the geographical area provided in his work covering the varieties, four major varieties, namely, the Bajjika, Sotipura, Angika, and Surjapuri varieties of modern Maithili correspond to Grierson's Western, Standard, Chhika-Chhiki, and Eastern varieties [2] [3]. [4] is the dataset available for Maithili. Maithili Text to Speech Corpus provides a high-quality and phonetically balanced dataset of Maithili speech recordings along with their corresponding transcriptions. It is specifically designed for training, testing, and evaluating Maithili TTS models while also facilitating linguistic analysis of the language's phonetic, prosodic, and intonational characteristics.

2.2 OVERVIEW

The corpus includes assertative (statements), interrogative (questions), imperative (commands), and exclamatory sentences, as well as sentences with different stress and pitch patterns. The text prompts were majorly prepared from [4]. A few sentences were manually added to ensure the command and exclamation words.

LDCIL follows a standard naming convention for storing audio files. Each audio file name consists language notation, gender id, sentence id, and the audio format name.
A typical audio file name is given below:
MT-TTS-M01-S-000001.wav
MT-TTS-F01-S-000001.wav

The summary of the Maithili TTS corpus is as follows:

Category	Sentence Count	Word Count	Male (Duration)	Female (Duration)	Total Duration
Statement	11,563	79,823	11:10:19	11:29:08	22:39:26
Question	1,500	15,756	01:03:15	01:43:10	02:46:25
Command	1,586	8,194	00:52:53	01:02:19	01:55:11
Exclaimation	1,500	15,766	01:35:30	02:02:44	03:38:15
Total	16,149	1,19,539	14:41:57	16:17:21	30:59:17

Table 1: Summary of the Corpus

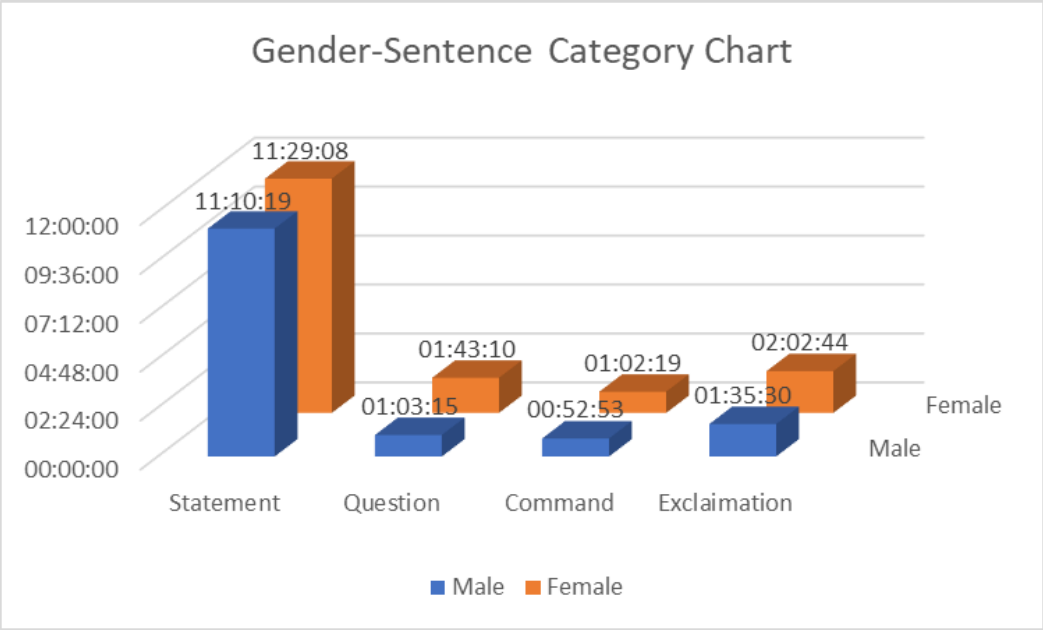

Figure 1: Gender-Sentence Category Chart

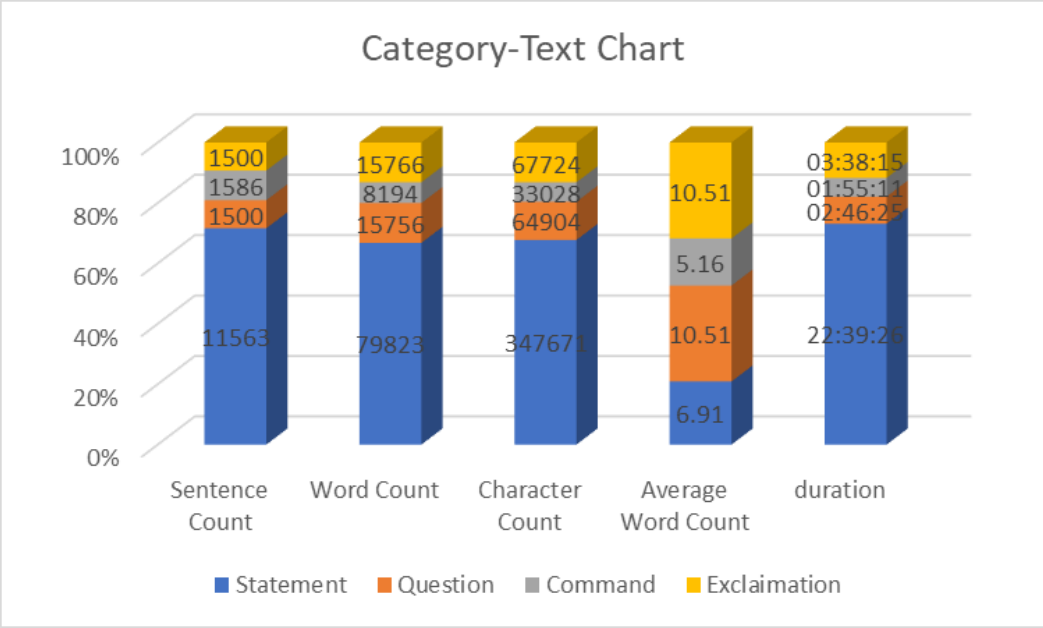

Figure 2: Category-Text Chart

2.3 REFERENCES

1. Grierson, George Abraham. Linguistic survey of India. Vol. 1. Office of the superintendent of government printing, India, 1928.
2. Jha, Subhadra. "The formation of the Maithili language." PhD diss., Luzac London, 1958.
3. Jha, Govinda. Uchchatar Máithilī Vyākaran. Patna: Maithili Akadami, 1979.

4. Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh & Dinesh Mishra. 2019. A Gold Standard Maithili Raw Text Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-236-1.