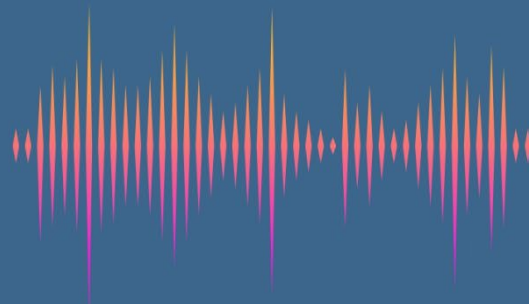


অসমীয়া অনূৰাচিকা কৰ্পাচ

ক খ গ



**ASSAMESE TEXT TO SPEECH
CORPUS**

ASSAMESE TEXT TO SPEECH CORPUS



34

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

Assamese Text to Speech Corpus

Syeda Mustafiza Tamim, Prangshu Manjul, Stephen Fernandes, Nithin S., Roopashri M. R.,
Dr. Narayan Kumar Choudhary, Prof. Shailendra Mohan

e-ISBN: 978-93-48633-45-3

CIIL Publication No.: 1514

First published: AD 2025 March :: Phalguna 1946 Śaka

© Central Institute of Indian Languages, Mysuru 2025

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: Syeda Mustafiza Tamim

TABLE OF CONTENTS

Table of Contents	4
Figures.....	4
Tables	4
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 Assamese Text to Speech Corpus.....	3
2.1 Introduction	3
2.2 Overview	3
2.3 References	5

FIGURES

Figure 1: Gender-Wise Word Count Category Chart	4
Figure 2: Gender-wise Duration Catgory Chart.....	4

TABLES

Table 1: Summary of the Corpus	3
--------------------------------------	---

1 TEXT TO SPEECH CORPUS

Dr. Narayan Choudhary

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 ASSAMESE TEXT TO SPEECH CORPUS

Syeda Mustafiza Tamim, Narayan Kumar Choudhary

2.1 INTRODUCTION

Assamese, which is primarily spoken in Assam and parts of Northeast India. Assamese has a rich literary heritage and a unique phonetic structure derived from the Eastern Nagari script[1]. According to some recent linguistic studies, the two dialectal groups can be further divided into four different varieties namely Eastern also known as Xiboxagoria, Central, Kamrupi and Goalparia group. Here, the Central group is spoken in Nagaon, Sonitpur, Morigaon and other adjoining areas; the Kamrupi variety is spoken in Kamrup, Barpeta, Nalbari, etc.; and the Goalparia group is spoken the Goalpara region. [2] and [3] are the datasets available for Assamese. Assamese Text to Speech Corpus provides a high-quality and phonetically balanced dataset of Assamese speech recordings along with their corresponding transcriptions. It is specifically designed for training, testing, and evaluating Assamese 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 [3]. 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:

AS-TTS-F01-S-000001.wav

AS-TTS-M01-S-000001.wav

The summary of the Assamese TTS corpus is as follows:

Category	Sentence Count	Word Count	Male (Duration)	Female (Duration)	Total Duration
Statement	11,313	105,430	16:15:07	16:50:06	33:05:13
Question	1,682	15,308	02:30:12	02:03:00	04:33:12
Command	1,605	144,509	01:50:41	01:29:36	03:20:17
Exclamation	1,697	10,967	01:57:59	01:51:53	03:49:52
Total	16,297	2,76,214	22:33:59	22:14:35	44:48:34

Table 1: Summary of the Corpus

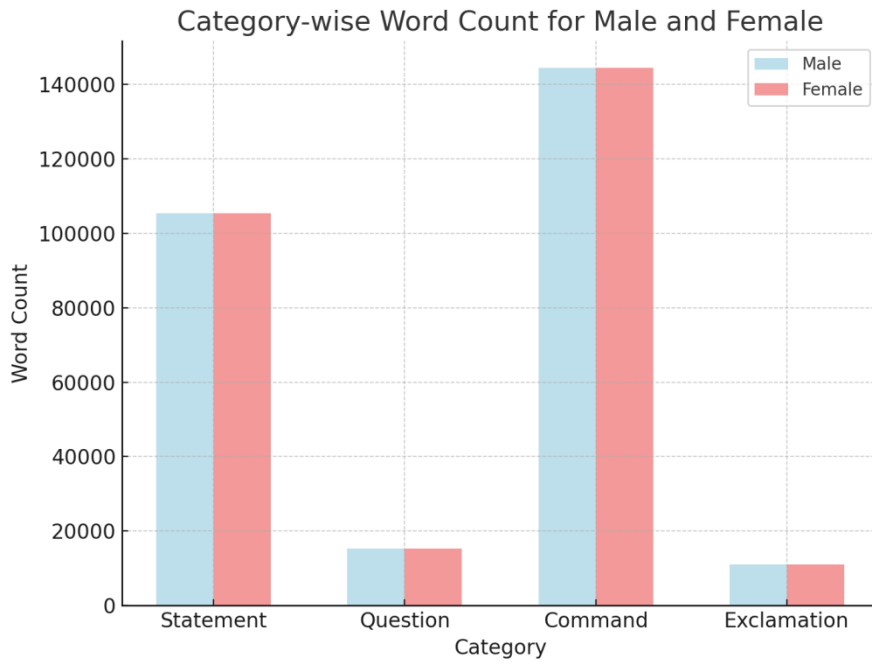


Figure 1: Gender-Wise Word Count Category Chart

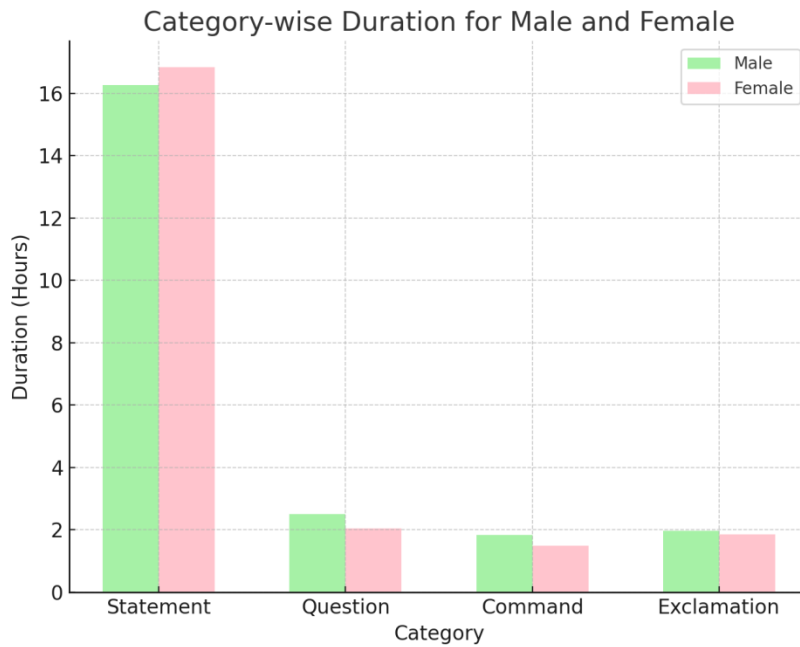


Figure 2: Gender-wise Duration Category Chart

2.3 REFERENCES

1. Tamim, S. M., Adhyapak, P., Rajesha, N., Manasa, G., Srikanth, D., Choudhary, N. K., & Mohan, S. (2023). Assamese Sentence Aligned Speech Corpus. Central Institute of Indian Languages. e-ISBN: 978-81-948885-4-3.
2. Ramamoorthy, L., Choudhary, N. K., Sharma, A., Gogoi, A., Kalita, J., Bharadwaj, S., Bora, P., Adhyapak, P., Tamim, M., Rajesha, N., & Manasa, G. (2021). Assamese Raw Speech Corpus. Central Institute of Indian Languages. e-ISBN: 978-81-948885-5-0.
3. Hussain, T., Adhyapak, P., Tamim, S. M., Rajesha, N., Manasa, G., Choudhary, N., & Ramamoorthy, L. (n.d.). Assamese Raw Text Corpus. Linguistic Data Consortium for Indian Languages, Central Institute of Indian Languages.