2023 9th International Conference on Smart Computing and Communications (ICSCC)

Development of Automatic Speech Recognition System for Kannada Language/Dialects

Thimmaraja Yadava G
Dept. of Electronics & Communication Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, Karnataka, India
thimrajyadav@gmail.com

Nagaraja B G
Dept. of Electronics & Communication Engineering
Vidyavardhaka College of Engineering
Mysuru, Karnataka, India
nagarajbg@gmail.com

Sanjay B D
Dept. of Electronics & Communication Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, Karnataka, India
sanjaybd.pd@gmail.com

Mohamed A S
Dept. of Electronics & Communication Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, Karnataka, India
siddiquetvl2010@gmail.com

Prajwal S D
Dept. of Electronics & Communication Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, Karnataka, India
dukandarprajwal@gmail.com

Pankaja K
Dept. of Electronics & Communication Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, Karnataka, India
pankajamurthy2001@gmail.com

2023 9th International Conference on Smart Computing and Communications (ICSCC) | 979-8-3503-1409-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSCC59169.2023.10334997

***Abstract*—Many end-to-end automatic speech recognition (ASR) systems have been developed for Indian and International languages, but no efficient ASR system is developed for Kannada language. This paper attempts to demonstrate the development of end-to-end (E2E) ASR system for Kannada language/dialects under noisy conditions. We develop an ASR system by combining speech-to-text (STT) and text-to-speech (TTS) systems under noisy conditions. To develop an ASR system, a task specific speech data plays a vital role, and it is collected from the speakers. The dictionary and phoneme set are much indeed to transcribe the collected speech data. The language and acoustic models for the validated speech data are developed using Kaldi speech recognition toolkit. The text output of an STT system is given as input to TTS system. The texts are normalized, converted to phonemes, and compared with the original speech database. The phoneme that suites the best is concatenated to words and waveforms are generated to get back speech as the output.**

***Index Terms*—End-to-end (E2E), automatic speech recognition (ASR), Phoneme, speech-to-text (STT), text-to-speech (TTS), Kaldi, Mel frequency cepstral coefficient (MFCC)**

I. Introduction

Speech is the fundamental way of interpersonal communication. Speech recognition is the operation of turning sound/voice signals recorded by a telephone or microphone into a collection of words. The action of converting a human sound into instructions or words is called speech recognition. Recognizing the speech is one of the rapidly growing engineering technologies. Speech recognition is essential because language hurdles become a barrier for many individuals from communicating with one another. One of the motives of speech recognition is also to reduce the language barriers [1]. Even though speaking is the best and most natural way for people to communicate, merely capturing speech as an audio signal presents challenges in efficiently examining, retrieving, and re-utilizing speech documents. Speech processing involves the analysis of speech signals and the implementation of signal processing techniques. It encompasses a wide range of operations and manipulations performed on the speech signal, like speech synthesis, speech recognition, speaker identification/verification, enhancement of the speech quality. Speech recognition takes the speech as an input and process the input to generate the text as an output. The action of converting the text input into speech output is referred as speech synthesis. Speaker verification/identification is concerned with identifying the speaker whether it's a male or a female speaker. When a raw speech data collected from speaker is fed to an ASR system, an ASR system may fail to recognize the spoken words due to the noise in the speech data. Therefore, the degraded speech data has to go through the speech enhancement process to improve the speech quality [2].

Speech recognition processes an acoustic signal that represents a conversation or spoken word collected by a microphone or telephone and converts it to a collection of words. Speech recognition is further classified as speaker dependent and speaker independent also, based on the manner of speech like isolated speech recognition, continuous speech recognition, spontaneous speech recognition. Speaker dependent systems are trained by a selected speaker who is going to be using the system. These systems achieve an accuracy of above 95% of word recognition. The disadvantage of this method is that

979-8-3503-1409-0/23/$31.00 ©2023 IEEE

it can respond accurately only to the speaker who trained the system. Whereas Speaker independent could be a system used to recognize anyone's voice. The accuracy isn't high as compared to speaker dependent since the speaker independent system must recognize various patterns in several voices. These systems are commonly found in telephone applications [3-5]. The Mel frequency cepstral coefficients (MFCC) is one among the standard methods for feature extraction [6,7]. To represent the statistical properties of speech signal, the modern speech recognition systems use acoustic and language models. An acoustic model represents the connection between an audio signal and phonemes whereas the language model distinguishes the words and phrases that sound phonetically similar. The sound features of a spoken audio file are compared to the spoken word text to build this model. There are several acoustic modeling techniques can be used, viz., Hidden Markov Model (HMM), Gaussian mixture model (GMM), deep neural network (DNN), convolutional neural network (CNN) and so on [8,9].

The framework for ASR system is broadly classified into two types, namely, speaker dependent framework and speaker independent framework. The ASR system developed in this work is of speaker independent type. The developed ASR system must be able to recognize the speech data or the audio recordings irrespective of the speaker. The speech output from speech-to-text (STT) system is fed into text-to-speech (TTS) system to get back the speech as an output. The complete process is referred to as an end-to-end (E2E) ASR system. The literature revealed that ASR models have been developed for many foreign languages like Arabic, Brazilian, Chinese, English, French, Spanish, Turkish and also for many of the Indian languages like Malayalam, Hindi, Urdu, Tamil etc., but not for Kannada language. In this work, an effort is made to develop an ASR system by combining STT and TTS systems. The STT System is the method for turning speech input into text output. The general architecture of STT system is shown in Fig. 1. The STT relies on an acoustic model, capturing sound waves with precision and resolve. Then is passes through a language model, deciphering context and grammar, ensuring words evolve. Next, the system employs a decoder, mapping phonemes to written characters, a task it endeavors. Together, these components create a seamless fusion, enabling speech to be transcribed with utmost precision.

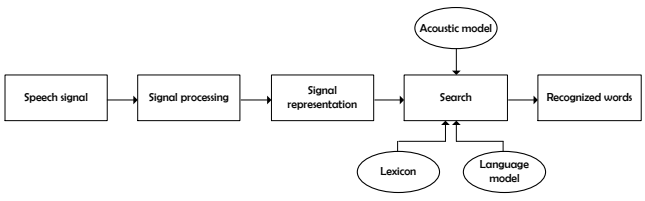

Fig. 1. Basic structure of STT system.

The basic architecture of TTS system is shown in Fig. 2. Within a TTS, a linguistic journey takes place, transforming written words into an auditory embrace. The process begins with a linguistic analysis, dissecting the text's structure and phonetic basis. A synthesis engine then crafts the vocal output, generating speech that resonates with clarity. Finally, the synthesized voice emerges, delivering speech that gracefully surges. The text output from the STT system is fed into

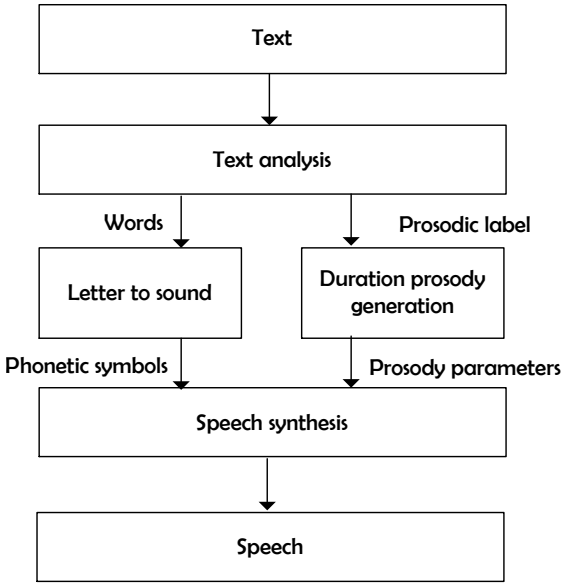

Fig. 2. Basic structure of TTS system.

the TTS system where the input texts are normalized to the standards and are converted to the phonemes then these phonemes are compared with speech database to get back the speech as an output.

II. Speech Data Collection

In order to encompass a wide range of pronunciations, the speech dataset is compiled from speakers residing in various dialect regions within the state of Karnataka. The collection of speech samples occurs in real-time settings. The dataset consists of recordings from 200 speakers, with each speaker uttering 10 continuous Kannada speech sentences, resulting in a total of 2000 sentences. The gathered speech data is transcribed at the word level, predominantly focusing on formal transcription. Table 1 illustrates the speech sentences used for data collection. Wavesurfer tool is used to record the speech data. The Fig. 3 describes the pictorial representation of data collection using wavesurfer tool. 80% of the collected speech data is dedicated to system training, while the remaining 20% is utilized for testing to develop ASR models.

TABLE I
SPEECH SENTENCES USED FOR DATA COLLECTION.

No.	Speech sentences collected
1.	ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯವು ಬೆಂಗಳೂರಿನ ಪ್ರತಿಷ್ಠಿತ ಸಂಸ್ಥೆಗಳಲ್ಲಿ ಒಂದಾಗಿದೆ .
2.	ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದ ವಾರ್ಷಿಕೊತ್ಸವದ ಹೆಸರು ಅನಾಧ್ಯಂತ.
3.	ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದಲ್ಲಿ ಶಿಕ್ಷಣದ ಜೊತೆಗೆ ಇತರ ಚಟುವಟಿಕೆಗಳು ಸಹ ಆಧ್ಯತೆ ನೀಡಲಾಗುತದೆ
4.	ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದ ಅಂಗಳದಲ್ಲಿ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ವಸತಿ ನಿಲಯವನ್ನು ಕಲ್ಪಿಸಿ ಕೊಡಲಾಗಿದೆ.
5.	ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದಲ್ಲಿ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಉತ್ತಮ ಶಿಕ್ಷಣವನ್ನು ನೀಡಲಾಗುತ್ತದೆ.
6.	ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದಲ್ಲಿ ಉದ್ಯೋಗಾವಕಾಶಗಳಿಗೆ ಹೆಚ್ಚಿನ ಮಹತ್ವವನ್ನು ನೀಡಲಾಗುತ್ತದೆ.
7.	ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದಲ್ಲಿ ಹಲವು ರಾಜ್ಯಗಳ ವಿದ್ಯಾರ್ಥಿಗಳು ತಮ್ಮ ವಿದ್ಯಾಭ್ಯಾಸವನ್ನು ಪೂರ್ಣಗೊಳಿಸುತಿದ್ದರೇ.
8.	ವಿದ್ಯುನ್ಮಾನ ಮತ್ತು ಸಂವಹನ ವಿಭಾಗವು ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದ ವಿಭಾಗಗಳಲ್ಲೇ ಒಂದಾಗಿದೆ.
9.	ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದಲ್ಲಿ ಪ್ರಸ್ತುತ ತಂತ್ರಜ್ಞಾನಗಳನು ಪರಿಣಾಮಕಾರಿಯಾಗಿ ಭೋದಿಸಲಾಗುತ್ತದೆ.
10.	ಐರಿಸ್, ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದ ಸಂಘಟನೆ ಸಮಿತಿಗಳಲ್ಲಿ ಒಂದಾಗಿದೆ.

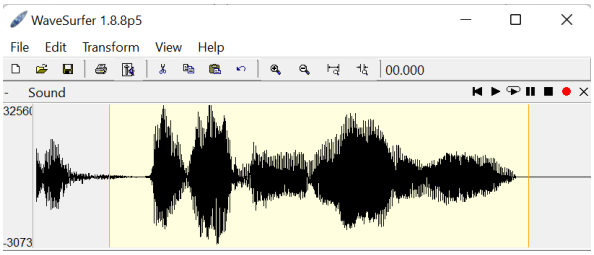

Fig. 3. Speech data collection using Wavesurfer tool.

III. PROPOSED METHODOLOGY

The architecture of the proposed system is shown in Fig. 4. It has mainly two parts, namely, STT and TTS system. The STT converts the raw speech data into its equivalent text information and the vice versa is known as TTS.

A. Creation of Dictionary and Phoneme Set

The dictionary and phoneme set are much indeed to transcribe the collected speech data. The lexicon and phoneme set are created for all the possible pronunciations. The Table 2 provides the phonemes for the continuous Kannada speech sentences.

B. Transcription

Transcription is a process representing the speech signal information into its equivalent text. The tool used for transcribing the speech data is shown in Fig. 5.

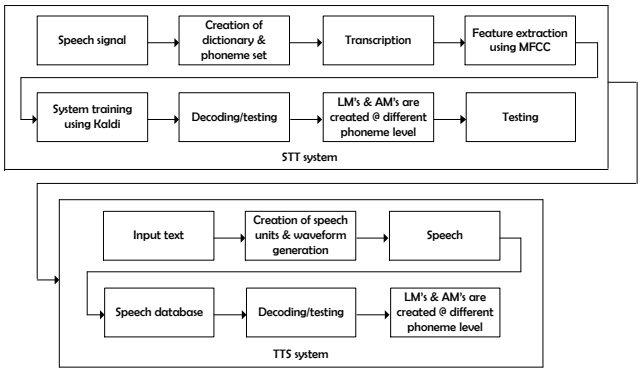

Fig. 4. Proposed E2E ASR system.

TABLE II
PHONEMES FOR THE CORRESPONDING SPEECH SENTENCES.

Word level	Phoneme level
ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯವು ಬೆಂಗಳೂರಿನ ಪ್ರತಿಷ್ಠಿತ ಸಂಸ್ಥೆಗಳಲ್ಲಿ ಒಂದಾಗಿದೆ .	ನ್ +ಇ +ಟ್ +ಟ್ +ಎ ಮ್ +ಈ +ನ್ +ಎ +ಕ್ +ಷ್ +ಇ ತ್ +ಎ + ನ್ +ತ್ +ಇ +ರ್ +ಇ +ಕ್ +ಎ ಮ್ +ಎ +ಹ್ +ಎ ವ್ +ಇ +ದ್ +ಎ +ಯ್ +ಎ +ಲ್ +ಎ +ವ್ +ಉ ಬ್ +ಎ+ನ್ +ಗ್ +ಎ +ಳ್ +ಉ೫ +ರ್ +ಇ +ನ್ +ಎ ಪ್ +ಎ +ರ್ +ಎ +ತ್ +ಇ +ಷ್ +ಇ +ಟ್ +ಇ +ತ್ +ಎ ಸ್ +ಎ+ ಮ್ +ಎ+ನ್+ಎ+ಥ್ +ಎ+ಗ್ +ಎ +ಳ್ +ಎ +ಲ್ +ಎ+ಲ್ +ಇ ಒ +ನ್+ಎ +ದ್+ಎ +ಗ್ +ಇ +ದ್ + ಎ.
ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದ ವಾರ್ಷಿಕೊತ್ಸವದ ಹೆಸರು ಅನಾಧ್ಯಂತ .	ನ್ +ಇ +ಟ್ +ಟ್ +ಎ ಮ್ +ಈ +ನ್ +ಎ +ಕ್ +ಷ್ +ಇ ತ್ +ಎ + ನ್ +ತ್ +ಇ +ರ್ +ಇ +ಕ್ +ಎ ಮ್ +ಎ +ಹ್ +ಎ ವ್ +ಇ +ದ್ +ಎ +ಯ್ +ಎ +ಲ್ +ಎ +ವ್ +ಎ +ಡ್ +ಎ ಹ್ +ಎ+ನ್ +ಎ +ರ್ +ಉ ಆ +ನ್ +ಎ +ಧ್ +ಎ +ಯ್ +ಎ +ನ್ +ಎ +ತ್ +ಎ.
ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದಲ್ಲಿ ಶಿಕ್ಷಣದ ಜೊತೆಗೆ ಇತರ ಚಟುವಟಿಕೆಗಳು ಸಹ ಆಧ್ಯತೆ ನೀಡಲಾಗುತದೆ .	ನ್ +ಇ +ಟ್ +ಟ್ +ಎ ಮ್ +ಈ +ನ್ +ಎ +ಕ್ +ಷ್ +ಇ ತ್ +ಎ + ನ್ +ತ್ +ಇ +ರ್ +ಇ +ಕ್ +ಎ ಮ್ +ಎ +ಹ್ +ಎ ವ್ +ಇ +ದ್ +ಎ +ಯ್ +ಎ +ಲ್ +ಎ +ವ್ +ಎ +ಡ್+ಎ ಎ +ನ್ +ಎ +ಗ್ +ಎ +ಳ್ +ಎ +ದ್ +ಎ +ಲ್+ಲ +ಇ ವ್ +ಇ +ದ್ +ಎ +ಯ್ +ಎ +ರ್ +ಥ್ +ಇ +ಗ್ +ಎ +ಳ್ +ಇ+ಗ್ + ಎ ವ್ +ಎ +ಸ್ +ಎ +ತ್ +ಇ ನ್ +ಇ +ಲ್ +ಎ +ಯ್ +ಎ +ವ್ +ಎ +ನ್+ ನ+ಉ ಕ್ +ಎ +ಲ್ +ಪ್ +ಇ +ಸ್ +ಇ +ಕ್ + ಓ +ಓ+ಡ್ +ಎ +ಲ್ +ಎ +ಗ್ +ಇ +ದ್ +ಎ.
ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದಲ್ಲಿ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ವಸತಿ ನಿಲಯವನ್ನು ಕಲ್ಪಿಸಿ ಕೊಡಲಾಗಿದೆ .	ನ್ +ಇ +ಟ್ +ಟ್ +ಎ ಮ್ +ಈ +ನ್ +ಎ +ಕ್ +ಷ್ +ಇ ತ್ +ಎ + ನ್ +ತ್ +ಇ +ರ್ +ಇ +ಕ್ +ಎ ಮ್ +ಎ +ಹ್ +ಎ ವ್ +ಇ +ದ್ +ಎ +ಯ್ +ಎ +ಲ್ +ಎ +ವ್ +ಎ +ಡ್+ಎ ಎ +ನ್ +ಎ +ಗ್ +ಎ +ಳ್ +ಎ +ದ್ +ಎ +ಲ್+ಲ +ಇ ವ್ +ಇ +ದ್ +ಎ +ಯ್ +ಎ +ರ್ +ಥ್ +ಇ +ಗ್ +ಎ +ಳ್ +ಇ+ಗ್ + ಎ ವ್ +ಎ +ಸ್ +ಎ +ತ್ +ಇ ನ್ +ಇ +ಲ್ +ಎ +ಯ್ +ಎ +ವ್ +ಎ +ನ್+ ನ+ಉ ಕ್ +ಎ +ಲ್ +ಪ್ +ಇ +ಸ್ +ಇ +ಕ್ + ಓ +ಓ+ಡ್ +ಎ +ಲ್ +ಎ +ಗ್ +ಇ +ದ್ +ಎ.
ನಿಟ್ಟೆ ಮೀನಾಕ್ಷಿ ತಾಂತ್ರಿಕ ಮಹಾ ವಿದ್ಯಾಲಯದಲ್ಲಿ ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ ಉತ್ತಮ ಶಿಕ್ಷಣವನ್ನು ನೀಡಲಾಗುತ್ತದೆ.	ನ್ +ಇ +ಟ್ +ಟ್ +ಎ ಮ್ +ಈ +ನ್ +ಎ +ಕ್ +ಷ್ +ಇ ತ್ +ಎ + ನ್ +ತ್ +ಇ +ರ್ +ಇ +ಕ್ +ಎ ಮ್ +ಎ +ಹ್ +ಎ ವ್ +ಇ +ದ್ +ಎ +ಯ್ +ಎ +ಲ್ +ಎ +ವ್ +ಎ +ಡ್+ಎ +ಎ+ಲ +ಲ +ಇ ವ್ +ಇ +ದ್ +ಎ +ಯ್ +ಎ +ರ್ +ಥ್ +ಇ +ಗ್ +ಎ +ಳ್ +ಇ+ಗ್ + ಎ ಉ೫ +ತ್+ತ್ +ಎ +ಮ್+ಎ ಶ್ +ಇ +ಷ್ +ಎ +ಣ್ +ಎ +ವ್ +ಎ +ನ್ +ನ್+ಉ ನ್ +ಈ +ಡ್ +ಎ +ಲ್ +ಎ +ಗ್ +ಉ +ತ್ +ಎ +ದ್ +ಎ.

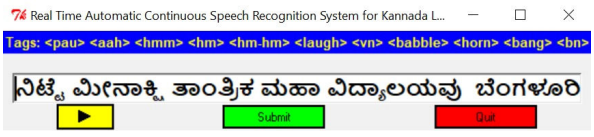

Fig. 5. Transcriber tool.

C. MFCC Feature Extraction

Feature extraction includes extracting various parameters of speech signals such as, pitch, frequency, amplitude of the signal and etc,. The MFCC [10-14] technique is used to extract the speech features and extracted features are used to develop ASR models.

D. TTS Normalization and Phonetic Analysis

It is a technology that converts written text into spoken words. TTS [15,16] systems analyze the input text and generate a synthesized voice output that can be listened to by humans. These systems are commonly used in applications such as voice assistants, navigation systems, accessibility tools for individuals with visual impairments, and many more. TTS technology has significantly advanced in recent years, with more natural and expressive voices being developed [17-19]. It enables computers and other devices to communicate with users through spoken language, enhancing accessibility and improving user experience in various domains.

IV. Result Analysis

The WER is used as a performance metric to evaluate the efficacy of the developed continuous Kannada ASR models. The WER is mathematically represented as follows:

$$\text{WER} = \frac{\text{NS} + \text{ND} + \text{NI}}{\text{NW}} \quad (1)$$

where NS represent the count of substitutions, ND represent the count of deletions, NI represent the count of insertions and NW denote the total number of words in the reference.

TABLE III
THE DESCRIPTION OF WERS AT DIFFERENT PHONEME LEVELS.

Phoneme level	WER - 1	WER - 2	WER - 3	WER - 4	WER - 5	WER - 6	WER - 7
Monophone	2.19	2.19	2.19	2.19	2.19	2.19	1.88
Triphone 1	1.25	1.56	1.56	1.56	1.56	1.56	1.56
Triphone 2	0.94	0.94	0.94	0.94	0.94	0.94	0.94
Triphone 3	0.94	0.94	0.94	0.94	0.94	0.94	0.94

KNKKK0101
ನಿಟ್ಟೆ_ಮೀನಾಕ್ಷಿ_ತಾಂತ್ರಿಕ_ಮಹಾ_ವಿದ್ಯಾಲಯವು_ಬೆಂಗಳೂರಿನ_ಪ್ರತಿಷ್ಠಿತ_ಸಂಸ್ಥೆಗಳಲ್ಲಿ_ಒಂದಾಗಿದೆ
LOG (gmm-latgen-faster[5.5.1057~1546-be222]:DecodeUtteranceLatticeFaster():decoder-wrappers.cc:375) Log-like per frame for utterance KNKKK0101 is -3.5138 over 640 frames.
KNKKK0102 ನಿಟ್ಟೆ_ಮೀನಾಕ್ಷಿ_ತಾಂತ್ರಿಕ_ಮಹಾ_ವಿದ್ಯಾಲಯದ_ವಾರ್ಷಿಕೊತ್ಸವದ_ಹೆಸರು_ಅನಾಧ್ಯಂತ
LOG (gmm-latgen-faster[5.5.1057~1546-be222]:DecodeUtteranceLatticeFaster():decoder-wrappers.cc:375)
KNKKK0103
ನಿಟ್ಟೆ_ಮೀನಾಕ್ಷಿ_ತಾಂತ್ರಿಕ_ಮಹಾ_ವಿದ್ಯಾಲಯದಲ್ಲಿ_ಶಿಕ್ಷಣದ_ಜೊತೆಗೆ_ಇತರ_ಚಟುವಟಿಕೆಗಳು_ಸಹ_ಆಧ್ಯತೆ_ನೀಡಲಾಗುತದೆ
LOG (gmm-latgen-faster[5.5.1057~1546-be222]:DecodeUtteranceLatticeFaster():decoder-wrappers.cc:375)
KNKKK0104
ನಿಟ್ಟೆ_ಮೀನಾಕ್ಷಿ_ತಾಂತ್ರಿಕ_ಮಹಾ_ವಿದ್ಯಾಲಯದ_ಅಂಗಳದಲ್ಲಿ_ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ_ವಸತಿ_ನಿಲಯವನ್ನು_ಕಲ್ಪಿಸಿ_ಕೊಡಲಾಗಿದೆ
LOG (gmm-latgen-faster[5.5.1057~1546-be222]:DecodeUtteranceLatticeFaster():decoder-wrappers.cc:375)
KNKKK0105
ನಿಟ್ಟೆ_ಮೀನಾಕ್ಷಿ_ತಾಂತ್ರಿಕ_ಮಹಾ_ವಿದ್ಯಾಲಯದಲ್ಲಿ_ವಿದ್ಯಾರ್ಥಿಗಳಿಗೆ_ಉತ್ತಮ_ಶಿಕ್ಷಣವನ್ನು_ನೀಡಲಾಗುತ್ತದೆ
LOG (gmm-latgen-faster[5.5.1057~1546-be222]:DecodeUtteranceLatticeFaster():decoder-wrappers.cc:375)
KNKKK0106
ನಿಟ್ಟೆ_ಮೀನಾಕ್ಷಿ_ತಾಂತ್ರಿಕ_ಮಹಾ_ವಿದ್ಯಾಲಯದಲ್ಲಿ_ಉದ್ಯೋಗಾ_ಅವಕಾಶಗಳಿಗೆ_ಹೆಚ್ಚಿನ_ಮಹತ್ವವನ್ನು_ನೀಡಲಾಗುತ್ತದೆ
LOG (gmm-latgen-faster[5.5.1057~1546-be222]:DecodeUtteranceLatticeFaster():decoder-wrappers.cc:375)
KNKKK0107
ನಿಟ್ಟೆ_ಮೀನಾಕ್ಷಿ_ತಾಂತ್ರಿಕ_ಮಹಾ_ವಿದ್ಯಾಲಯದಲ್ಲಿ_ಹಲವು_ರಾಜ್ಯಗಳ_ವಿದ್ಯಾರ್ಥಿಗಳು_ತಮ್ಮ_ವಿದ್ಯಾಭ್ಯಾಸವನ್ನು_ಪೂರ್ಣಗೊಳಿಸುತಿದ್ದರೇ
LOG (gmm-latgen-faster[5.5.1057~1546-be222]:DecodeUtteranceLatticeFaster():decoder-wrappers.cc:375)

Fig. 6. The output of STT system.

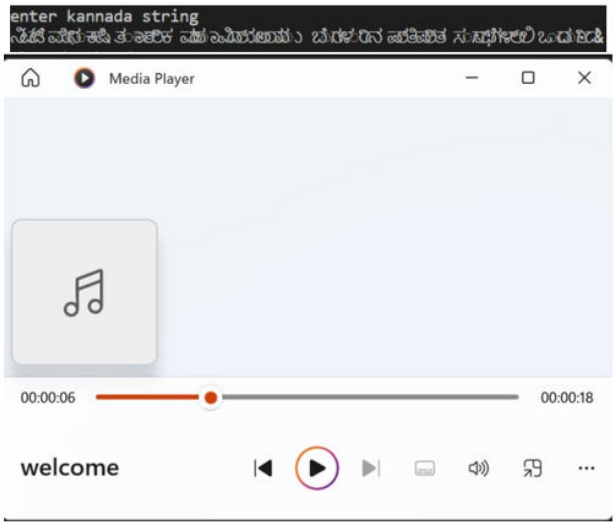

Fig. 7. The output of the TTS system.

20% and 80% of the validated speech data is used for testing and training respectively. With the recipe of Kaldi, resources of Kannada language/dialects and validated speech data, we achieved a least WER of 0.94% at triphone 2 and triphone 3 acoustic levels. The detailed description of obtained WER at various phoneme levels is shown in Table 3. The least WER ASR model is used for online recognition. The online recognition of continuous Kannada speech sentence (STT) is shown in Fig. 6. The output of STT is given as input to the TTS system which converts the text information into speech as shown in Fig. 7. From Fig. 6, it can be observed that, the continuous Kannada speech sentence file name is represented as KNKKK0101 and it is decoded correctly. The decoded results are checked with the test speech data. The text based output is given as input to the TTS system which further converts into the speech.

V. Conclusion

This study exemplifies the development of dependable end-to-end ASR models for Kannada language/dialects in real-world conditions. A total of 200 speakers contributed task-specific speech data in a noisy environment, encompassing various pronunciations. To transcribe the speech data, a comprehensive dictionary and phoneme set were utilized. The speech data underwent feature extraction using the MFCC technique. Training and testing of the system were conducted using the Kaldi toolkit. Finally, the resulting text output from the STT system was employed as input for the TTS system.

References

[1] Vineet Vashisht, Aditya Kumar Pandey, and Satya Prakash Yadav, "Speech recognition using machine learning," in IEIE Transactions on Smart Processing and Computing, vol. 10, no. 3, June 2021.

[2] M. Kobayashi, M. Sakamoto, T. Saito, Y. Hashimoto, M. Nishimura and K. Suzuki, "Wavelet analysis used in text-to-speech synthesis," IEEE Transactions on Circuits and Systems, vol. 45, no. 8, pp. 1125-1129, 1998.
[3] Gokul G Nair, C Santhosh Kumar, "Speech Enhancement System for Automatic Speech Recognition in Automotive Environment," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 198-209, 2021.
[4] N. Mahadev and RR. Malagi Harkudeand, "Automotive noise and vibration sources prediction and control", Proceeding of NCRIET-2015 Indian J.Sci.Res, vol. 12, no. 1, pp. 001-006, 2015.
[5] S. E. Bou-Ghazale and J. H. L. Hansen, "HMM-based stressed speech modelling with application to improved synthesis and recognition of isolated speech under stress," IEEE Transactions on Speech and Audio Processing, vol. 6, no. 3, pp. 201-216, May 1998.
[6] H. A. Murthy, F. Beaufays, L. P. Heck and M. Weintraub, "Robust text-independent speaker identification over telephone channels," IEEE Transactions on Speech and Audio Processing, vol. 7, no. 5, pp. 554-568, Sept. 1999.
[7] I. Varga, "ASR in mobile phones - an industrial approach," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 8, pp. 562-569, Nov. 2002.
[8] D. O'Shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis," Proceedings of the IEEE, vol. 91, no. 9, pp. 1272-1305, Sept. 2003.
[9] S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," IEEE Transactions on Speech and Audio Processing, vol. 12, no. 4, pp. 401-408, July 2004.
[10] Jainar, S.J., Sale, P.L. and Nagaraja, B.G., "VAD, feature extraction and modelling techniques for speaker recognition: a review," International Journal of Signal and Imaging Systems Engineering, vol. 12(1-2), pp. 1-18, 2020.
[11] Nagaraja, B.G. and Jayanna, H.S., "Kannada language parameters for speaker identification with the constraint of limited data," International Journal of Image, Graphics and Signal Processing, vol. 5(9), p.14, 2013.
[12] Nagaraja, B.G. and Jayanna, H.S., "Combination of features for crosslingual speaker identification with the constraint of limited data. In Proceedings of the Fourth International Conference on Signal and Image Processing 2012, vol. 1, pp. 143-148, 2013.
[13] Ayadi, M, Kamel, MS, Karray, F. (2011). "Survey on speech emotion recognition: features, classification schemes, and databases," Patt. Recognit., 44(3), 572–587.
[14] Ververidis, D, Kotropoulos, C. (2006). "Emotional speech recognition: resources, features, and methods," Speech Comm., 48(9), 1162–1181.
[15] Sheikhan, M, Gharavian, D, Ashoftedel, F. (2012). "Using DTW neural–based MFCC warping to improve emotional speech recognition," Neural Comput. Appl., 21(7), 1765–1773.
[16] Cid, Felipe, Cintas, R, Manso, Luis, Calderita, Luis, Sánchez, A, Núñez, Pedro. (2011). "A real-time synchronization algorithm between Text-To-Speech (TTS) system and Robot Mouth for Social Robotic Applications".
[17] Oh, Kyunggeune, Jung, Chan-Yul, Lee, Yong-Gyu, Kim, Seung-Jong "Real-time lip synchronization between text-to-speech (TTS) system and robot mouth. Proceedings", IEEE International Workshop on Robot and Human Interactive Communication.
[18] Page, J., Breen, "The Laureate text-to-speech system - Architecture and applications," BT Technology Journal. 14,1987.
[19] Sunitha, C., Chandra, Evania. (2015). "Speaker Recognition using MFCC and Improved Weighted Vector Quantization Algorithm," International Journal of Engineering and Technology. 7. 1685-169.