Spectrogram Analysis and Text Conversion of Sound Signal for Query Generation to Give Input to Audio Input Device

Kavita Sharma(✉) and S. R. N. Reddy

Indira Gandhi Delhi Technical University for Women, Delhi 11006, India
{kavita004phd20,srnreddy}@igdtuw.ac.in

Abstract. The world is being reshaped by Natural Language Processing. Audio inputs are used in modern electronics. Different types of people supply input to the system in their native language. The system accepts the person's speech, processes it, and responds accordingly. Cooking is a huge problem for a variety of people, including the elderly, those who are confined to their beds, and those who have a specific sort of handicap, such as those who are unable to use their hands and require assistance at all times. To help these people reach their full potential, an audio input device for giving cooking instructions to a cooking system has been proposed in this paper. The gadget takes the user's spoken English language as input, converts it to text using deep learning algorithms, and generates instructions with the help of context-aware words extracted from the recorded audios to send the instruction to the cooking device. To analyses, the audio signal for user authentication is a challenging task due to gaps and pauses between spoken characters, and existing noise in the environment. As a result, the audio input device developed for kitchen systems must analyze the audio input signal to create a more secure environment for authenticated users. As a result, the objective of this paper is to analyze the audio input signal captured in real-time and process the accepted signal to convert into the text to generate instructions for a larger system. The sound signals captured in the real environment are analyzed with Mel spectrogram, MFCC spectrogram, and PRAAT software. The sound signal is processed with the help of a natural language toolkit to generate instructions.

Keywords: Mel-spectrogram · Audio signal · MFCC · Tokenization

1 Introduction

Processing of the sound signals received from the user has much popularity in the modern world as it fills the gap between machine understanding, learning, and human spoken language. It saves a lot of time for the user of the technology instead of giving commands to the machine with the help of written text in a particular machine-readable language. The advancement in speech technology has several challenges for the technology developer e.g.: recording of the speech spoken audio signals, analyzing then, removal of noise elements, and extracting the desired signal with maximum accuracy. Using machine

A. Dev et al. (Eds.): AIST 2021, CCIS 1546, pp. 171–181, 2022.
https://doi.org/10.1007/978-3-030-95711-7_15

learning and deep learning models, the ASR (Automatic Speech Recognizer) assists in recording audio speech signals and converting them to text. The display of various spectrograms aids in the study and classification of audio signals. This technology's implementation in the kitchen offers a lot of potential for increasing the efficiency of the cooking process. The study consists of attempting to synthesize a specific speaker's voice with a few lines to leverage a vast amount of labeled speech recorded via the audio input device and text data received after conversion of the recorded signal, from which labels are to be extracted to generate a query that is to be given to the culinary system. The most popular platform e.g.: Amazon Echo, Google Home, Cortona, Siri, etc. extracts the words from the recorded clips of the spoken audio in some language and text conversion. This extracted text is used for sending queries to the system working in the background and the user desired output is provided in the audio form again. A normal human being has no difficulty in operating a smart device using any input mode eg: text input, sign input, or audio input. However, a certain group of people, such as the elderly or those with a specific sort of impairment, face unique challenges when using text or sign input. With audio input, they feel at ease. People who are unfamiliar with technology, on the other hand, find auditory input to be comforting. As a result, the goal of this study is to take audio input from a variety of users.

The audio accepted by the user can be represented in different ways based on the application. The analysis of audio input is based on various parameters e.g.: spectral centroid, spectral shape, zero-crossing statistics, harmony, fundamental frequency and temporal envelop frequency, etc. For classification of the sound different representations are popular in literature. The magnitude representation of raw audio [2] is done with the help of Mel-scale and Mel Frequency Cepstral Coefficients. Mel spectrum [4, 6] is used to map characters using recurrent sequence to sequence feature prediction. The log-Mel spectrogram and MFCC characteristics help recognize Alzheimer's dementia and classify environmental sounds [7]. CycleGAN-VC2 (Cycle Consistent Adversarial Network Voice Conversion 2) [9] have mappings without using parallel corpus between source and destination speech. As a result, classifying the users' sounds is a vital duty for the system's security. As a result, the paper's major goal is to record user input to classify different audio inputs given by the user with the help of Mel-spectrogram and MFCC classification tools.

The remaining paper is organized as follows: The second section is a review of the literature on the tools used to classify audio signals. The suggested audio input device is described in Section three. The fourth section focuses on various audio signal formats. Section five contains the results and discussion of several audio signal representations, as well as the suggested model's potential reach. The overall paper is concluded in Section six.

2 Literature Survey

The audio signals are accepted for various types of applications to provide input based on the context of the application. When an audio signal is captured from the open environment, it has a different type of interference and noise data embedded with it. Thus, literature provides various methods for analysis and classification tasks of recorded

audio data. Mel spectrogram and MFCC are the most popular signal classification tools of capturing the low-level shape of modulation spectra e.g.: Spectrograms [1] are used to generate audio using neural network single-channel STFT (Short-Time Fourier Transform), style transfer, The analysis of SNR (signal to noise ratio) [2] using the Mel-scale and Mel-Frequency Cepstral Coefficients spectrograms techniques increases recognition accuracy of an audio signal, spectrograms [3] are normalized into grayscale to extracts the block to extract features, character embedding [4] is mapped with Mel-scale spectrogram with the help recurrent sequence to sequence feature prediction. Mel spectrogram [5] extracts the appropriate hyperparameters of augmentation, parallel neural vocoder [6] integrate linear predictive synthesis filter in the model, the efficacy of log-Mel spectrogram and MFCC [7] are effective techniques of recognizing Alzheimer's dementia(AD), the environmental sound [8] classification is based on spectrogram to strain less informative and irrelevant frequencies from the signal, low dimensional features of an audio signal are defined by spectrograms [10] to capture the shape of modulation spectra, CycleGAN-VC3 [14] is enhanced version of CycleGAN-VC2, Acoustic Scene Classification [11, 12] gives scene's comprehensive with the help of spectrograms representation. Natural language toolkit [9, 13] is used for text mining to generate queries from the text extracted from the audio signal. Thus, in a nutshell, we can say that:

Spectrograms are the better tool for audio signal analysis, classification, and representation.

Natural language toolkit helps in text mining to extract the words from the text received from audio signal conversion.

3 Proposed Model: Audio Input Device

A typical guy can easily operate in the kitchen, but a person with a unique kind of impairment, such as someone who is unable to leave their home due to a health issue or is of advanced age, finds it difficult to perform everyday tasks without assistance. The kitchen is the most important place in the house, and people spend 8 to 10 h a day there preparing food. The smart kitchen is popular in this day when everything is going to be smart. Smart kitchens necessitate smart gadgets capable of controlling smart kitchen objects. Figure 1 depicts the proposed model in this paper, which is an audio input device. The audio input device is designed using a microcontroller and accepts audio signals from a human being for ASR (Automatic Speech Recognizer) The ASR recognizes the input signals and sends the input to the machine learning model where the audio signal is converted into text to extract the context-dependent features. When the system accepts the audio signal from the user, the signal is recorded as a wave file.

The wave file is recorded and given as input to the ASR system stored on the cloud. The designed device is a part of a system and resource-constrained device; thus, it stores only audio files and puts them through the same process. The remaining background programs are kept in the cloud and launched on the server as needed. When the user wishes to give the system instructions, he turns on the audio input device and begins speaking. The audio signal varies from user to user due to differences in pitch, phonemes, utterance level, spectrum, amplitude, and other factors. Sounds and noises make up the most basic audio data. Human speech is an example of this. Speech is more difficult to

encode since it encodes spoken words. The first step in solving an audio classification problem is to listen to a sound sample and determine the class it belongs to from a list of available classes in the database.

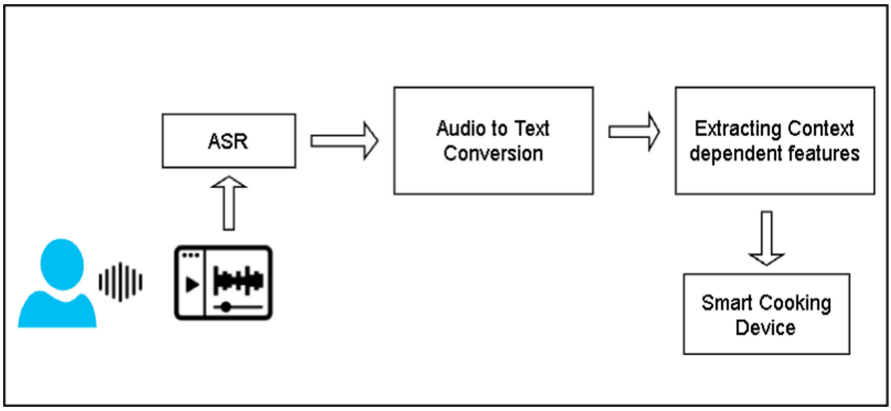

Fig. 1. Block diagram of audio input system working

The Speech-to-Text conversion training data is classified as follows:

Audio clips of spoken words (X) are input features.
Labels for the target (y): a text extracted from the transcript of the recorded audio of the user (Fig. 2).

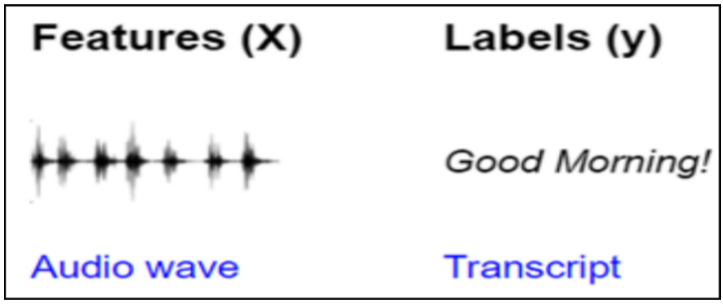

Fig. 2. Input and output of the device

Audio waves are used as input characteristics and text transcripts extracted from audio signals are used in analyzing and classifying the audio signal and target text for query generation in an automatic speech recognition system. The model is created to learn, interpret the audio input to make predictions based on text content extracted from the recorded audio clip and the extracted words and phrases. The audio data is then processed as seen in Fig. 3 and a Mel Spectrogram is generated to classify different signals. Classification of the sound signal by converting raw audio waves to Mel spectrogram images for extracting user sound signal and minimizing noise, deep learning popular Python library librosa is used here.

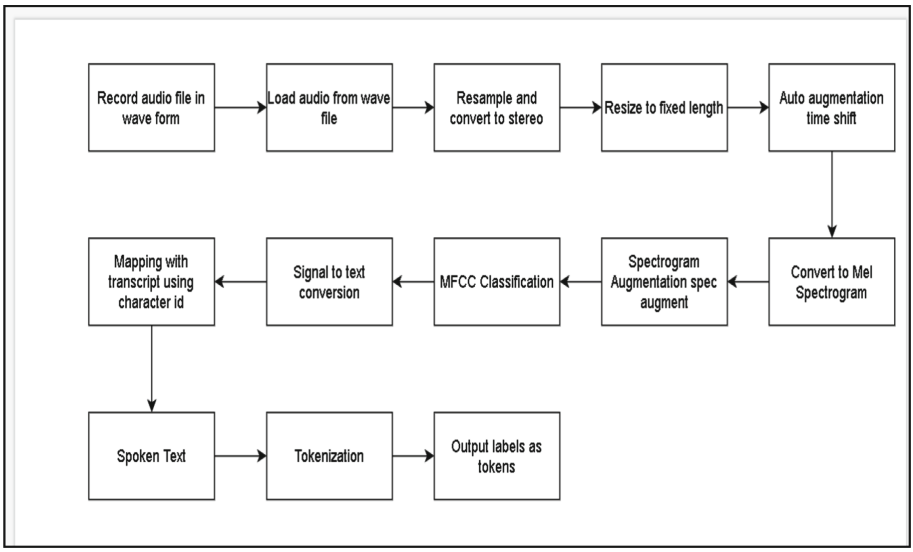

Fig. 3. Processing of audio data

4 Different Representations of Audio Signals

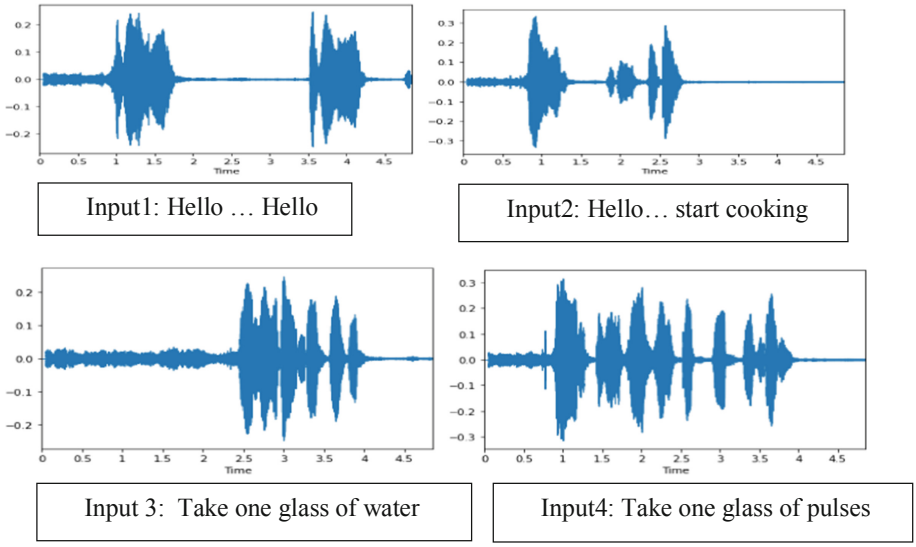

Fig. 4. Wave plotting of Audio sound signal for time

4.1 Recording of Sound

The sound signal to accept the input from the user is recorded in real-time. The wave plot of the recorded file is represented in Fig. 4.

4.2 Loading and Processing Audio Files for Analysis

Load the above input audio signals for further processing in an audio format such as ".wav" corresponding to the above four inputs. The short-time Fourier transform (STFT)is represented in Fig. 5 has the above audio signal as input.

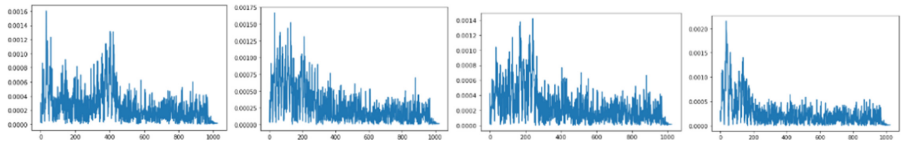

Fig. 5. STFT representation of Audio input data

The recorded wave audio data is sent to an array of 2D NumPy that contains an array of numbers for measurement of the amplitude and intensity of the sound signal at a specific movement of time.

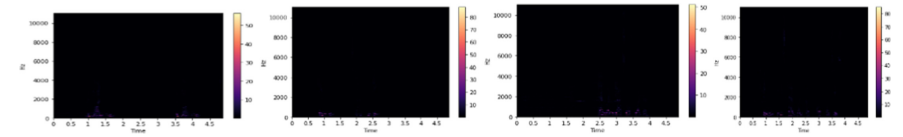

Fig. 6. STFT representation of the sound signal

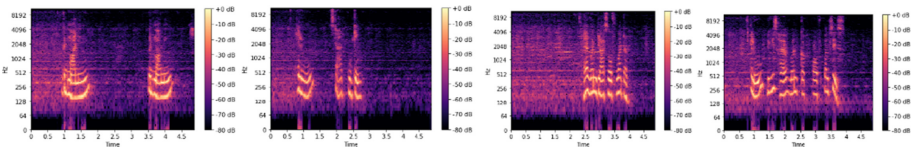

Fig. 7. Amplitude to dB representation of the sound signal

The sample rate reveals the number of measurements taken for consideration and the sample rate used in this paper to record the audio is 44.1 kHz thus, the array in the NumPy library has a single row of 44,100 values for each second of recorded audio. Mono and stereo are the two channels that are commonly used by audio signals. The recorded four audio uses two audio channels where the second channel has a similar sequence of amplitude numbers. Thus, the NumPy array has 3D with a depth of 2. As a result, the sample rate, channels, and duration are all converted to uniform dimensions (Fig. 6).

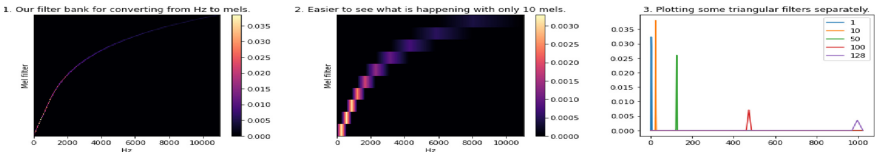

Fig. 8. Conversion of the sound signal in Mel and representation with 10 Mel of audio input1

Clips may be sampled at different speeds in real-time for different types of commands, or they may have a varying number of channels in separate audio files. As seen in Fig. 7, 8, 9 the clips will most likely have varying durations and dimensions.

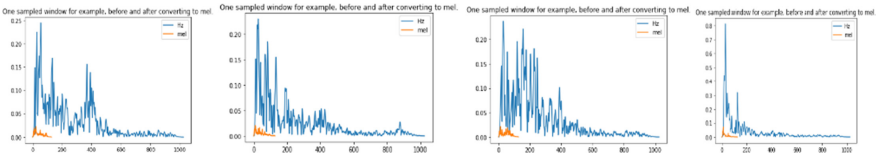

Fig. 9. Comprehensive analysis of original audio signal and Mel signal

The Mel Spectrogram has now been changed to MFCC (Mel Frequency Cepstral Coefficients) to extract only the most context-aware frequency coefficients corresponding to the human being's frequency ranges of talking to the surroundings. MFCCs provide a compressed form of the Mel Spectrogram (Fig. 10, 11).

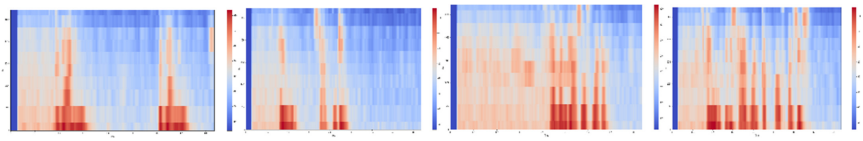

Fig. 10. Log Mel spectrogram representation of the sound signal

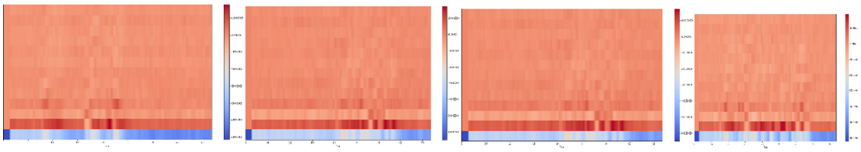

Fig. 11. MFCC representation of a sound signal with order 1

Data cleaning helps in standardizing the audio data dimensions as deep learning models need all inputs of similar size to process. As a result, the audio file is resampled into the same number of channels at the same sampling rate by padding the shorter sequences and truncating the longer sequences (Fig. 12).

Background noise is removed using a noise removal algorithm to improve audio quality. As a result, in Fig. 7, we get a better sound signal than in Fig. 4.

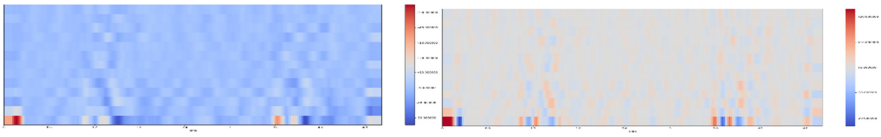

Fig. 12. Delta1 and Delta2 MFCC representation of the sound signal with order 2

By hyper-parameter and data augmentation, this stage improves the features of spectrograms for optimal performance. To extend the learning of the model to an accurate range of inputs, data augmentation techniques are used due to the diversity in the input data. It is done by randomly shifting the audio signal in the left or right direction by a small proportion based on user pitch or speed of the audio.

This raw audio is now decomposed into the set of frequencies to recognize the nature of the signal for more having more accurate signal and to recognize and classify the signals with the help of Mel spectrogram and MFCC delta 1 and delta 2 versions.

Audio to text conversion: the audio signal captured from the user is converted into text using speech to text "Speech Recognition" python library for further text mining process.

5 Results and Discussion

As we have analyzed from Fig. 4 to Fig. 12 with the help of code written in the python language and the sound signals are recorded manually with the help of code written in python. The sound signals contain the sentences used in an input device designed for the cooking device. The sound signals are accepted by the same user. Now the same sound signals are analyzed with the PRAAT audio analysis tool (Fig. 13) which provide the following measurements (Table 1) w.r.t spectrogram:

Table 1. Mean F1 and F2 of the audio signal in PRAAT audio analysis

Input	Mean F1	Mean F2
Audio Input1	430.2049424322493 Hz	1751.7334002196676 Hz
Audio Input2	405.8863713297039 Hz	1751.5263508130217 Hz
Audio Input3	476.0224521852952 Hz	1584.1387055004666 Hz
Audio Input4	491.3848052596775 Hz	1832.477938977869 Hz

We have analyzed the sound signals, having parameters shown in Table 2, with the help of Mel spectrogram, MFCC, and PRAAT tool representation which give a better representation than original sound and extract the sound features e.g.: pitch, intensity, pulses, etc. It also helps in the visualization of environmental noise embedded with the sound to clean the sound signal for further processing. After analyzing the audio signal is converted into text and processed with the nltk python library to extract the

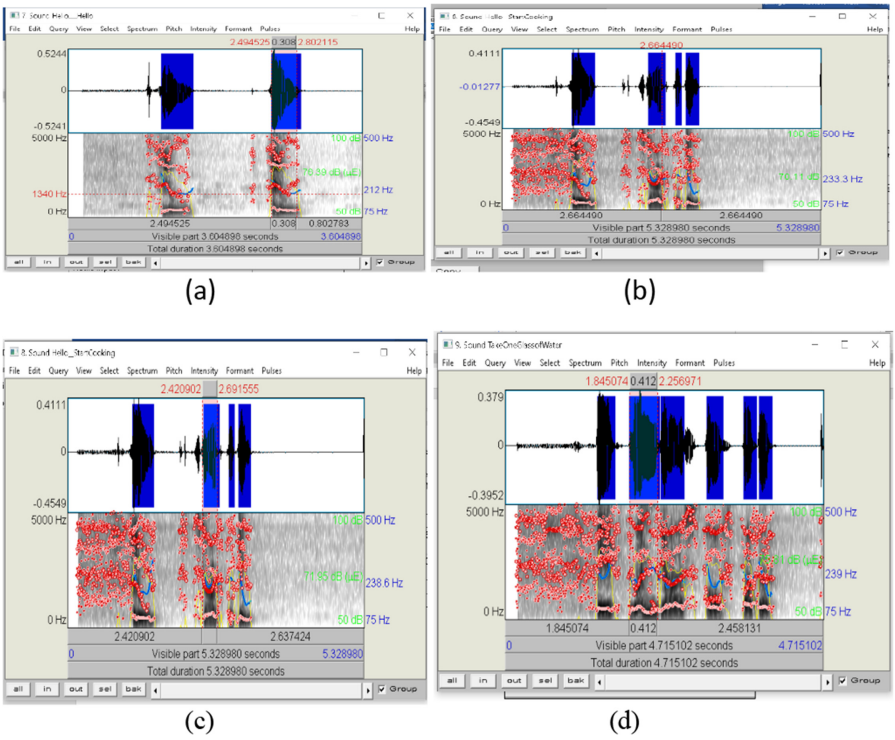

Fig. 13. Visualization of spectrogram, pitch, intensity, formants, and pulses w.r.t. sound signals in PRAAT software

context-aware text and generate instructions. The speech recognition library provides the following output corresponding to four audio inputs in Fig. 4.

After converting audio text into text, the tokenization process provides the following output (Figs. 14 and 15):

```
Converting audio transcripts into text ...
good morning
start cooking
Heaven glass of water
hello please have one cup of pulses
```

Fig. 14. Audio to text conversion of input data

Table 2. Different parameters used in audio signal representation

Property	Value
Channels	2
Rate of sampling	44100 Hz
Recording time of the audio waveform	5 s
Hop length	512
n_mel	10
n_fft	2048
n_mfcc	13

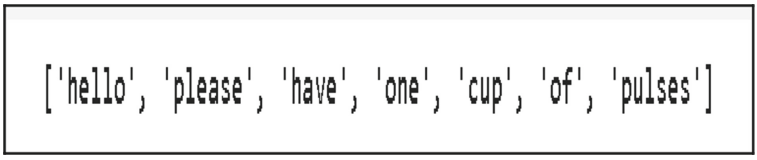

Fig. 15. Token generated from tokenization process

6 Conclusion

This paper has a novel framework for accepting the audio signal from the user using the microphone. The features of the audio signals are extracted for analysis with the help of a low-level spectrogram using python audio libraries and visualization tools. The same audio signals having sentences related to input to be provided to the cooking device are processed in PRAAT software and different features of the sound signals with environmental noise are analyzed. The audio signal is processed by a speech recognition library to convert the audio signal into text. The converted text is processed by the standard python library NLTK to generate the instruction for the larger system. In this paper, we have analyzed the audio signals to classify the inputs provided by the used and converted the audio signal into text, and tokenized the text. In future, the further analysis on audio input will be done to clean the sound signal and classify the sound signal to identify authenticated users to provide security to the authorized user and the more valid input signal to the text processing tool. In the future, the tokenized text will be used to generate instruction for the large system.

References

1. Wyse, L.: Audio spectrogram representations for processing with convolutional neural networks. arXiv preprint arXiv:1706.09559 (2017)
2. Papadimitriou, I., et al.: Audio-based event detection at different SNR settings using two-dimensional spectrogram magnitude representations. Electronics **9**(10), 1593 (2020)
3. Dennis, J., Tran, H., Li, H.: Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Process. Lett. **18**(2), 130–133 (2011)

4. Shen, J., et al.: Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018)
5. Hwang, Y., et al.: Mel-spectrogram augmentation for sequence-to-sequence voice conversion. arXiv preprint arXiv:2001.01401 (2020)
6. Juvela, L., et al.: GELP: GAN-Excited linear prediction for speech synthesis from mel-spectrogram. arXiv preprint arXiv:1904.03976 (2019)
7. Meghanani, A., Anoop, C.S., Ramakrishnan, A.G.: An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. In: 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE (2021)
8. Khunarsal, P., Lursinsap, C., Raicharoen, T.: Very short time environmental sound classification based on spectrogram pattern matching. Inf. Sci. **243**, 57–74 (2013)
9. Jha, N.K.: An approach towards text to emoticon conversion and vice-versa using NLTK and WordNet. In: 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA). IEEE (2018)
10. Kinnunen, T., Lee, K.A., Li, H.: Dimension reduction of the modulation spectrogram for speaker verification. In: Odyssey (2008)
11. Ngo, D., et al.: Sound context classification basing on join learning model and multi-spectrogram features. arXiv preprint arXiv:2005.12779 (2020)
12. Zheng, W., et al.: CNNs-based acoustic scene classification using multi-spectrogram fusion and label expansions. arXiv preprint arXiv:1809.01543 (2018)
13. Contreras, J.O., Hilles, S., Abubakar, Z.B.: Automated essay scoring with ontology based on text mining and nltk tools. In: 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE). IEEE (2018)
14. Kaneko, T., et al.: CycleGAN-VC3: examining and improving CycleGAN-VCs for mel-spectrogram conversion. arXiv preprint arXiv:2010.11672 (2020)