ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 7, Issue 6, June 2020

Development of Speech Interface for Challan Application using Speech Recognition

K.Tamilselvi[1], V.Sujitha[1], R.Kirubamanohari[2], S.Gayathiri[2], M.Nivetha[2], K.Abinaya[2]

Assistant professor, Computer Science and Engineering, P.A. College of Engineering and Technology, Coimbatore, India[1]

Student, Computer Science and Engineering, P.A. College of Engineering and Technology, Coimbatore, India [2]

Abstract: Communicating with technological devices via voice has become so popular and natural in the engineering world. The plebeians found difficulty in form filling. The approach aims to overcome the challenges faced by the plebeians while filling the forms. The objective of this project is to develop a flexible structure for development of speech interface for form filling application. The speech interface provides an integrated framework for developing STS and TTS system across the languages and dictionaries specific to various forms and domains. The accurate working of speech interface is done with the help of dynamic down sampling and de-noising. The speech to text works in the speech interface using speech recognition module in python. Specific scope of this project includes the speech interface enabled challan application form filling.

Keywords: Text-to-speech; Speech-to-Text; Speech Recognition; Python libraries; Perturbation elimination

I. Introduction

Voice is the future. The world's technology giants are making a beeline for voice based applications. A Speech Interface makes natter human interaction with computers possible, using Speech Recognition to grasp spoken words and typically text to speech for more interactive session. Voice Recognition software have been added to automobiles , home automation systems, computer operating systems , home appliances like washing machines , microwave ovens and television remote controls.

II. Related works

In 2007, a CNN business article reported that voice command was over a billion dollar industry and that companies like Google and Apple were trying to create speech recognition features. It has been years since the article was published, and since then the world has witnessed a variety of speech recognition devices. As well, Google created a speech recognition engine called pico TTS and Apple has released siri. The increased use of smart speakers has opened up new application areas and they are put to practical use in many real-world fields such as computerized driving directions and instructions in hotel rooms or public spaces. Voice control technology devices are becoming more widely popular, and innovative ways for using the human voice are always being created. Therefore, there is a need to strengthen the security measures for smart speakers.

Threats arising from inadequate speech recognition applications may cause many real world problems. If speech recognition is vulnerable and open to the public, unauthorized access to personal information, illegal computer access, and unauthorized falsification may occur through smart speakers. An adversarial example [2] is an instance with small, intentional feature perturbations that cause machine learning model to make a false prediction. In 2018, CarLini et al., proposed an audio adversarial example [3]. Their target was speech-to- text systems based on speech-to-text transcription neural networks. An audio adversarial example can be created by adding perturbations to the input voice.

III. Scope of Research

In this paper we propose a flexible structure for development of speech interface form filling application. In 2018 Ananya Paul et al., proposed an GUI for Text-to-Speech Recognition using Natural Language Processing [1].They developed a software application to read the file and convert the text into audio format. In [4] , involves a typical speech to text by sequence – to- sequence voice conversion , which doesn't suits for limited audio data samples .Therefore we use python speech recognition library with deep speech for more appropriate speech to text conversion. The following are the main contributions of this paper:

- A modularized framework for development of speech interface for form filling application.
- A speech interface form filling application that provides a audio instruction , based on the users form filling context ; the system use python audio libraries for the conversion of users speech to the form of text in the appropriate text box.

IV. Overview of Speech Interface

Speech interface is one of the artificial computation of providing a audio instruction, based on the users form filling context and converting the users speech into text in the particular text field of the form using the module speech recognition in python .The audio instruction is used to help the user to know the content to be filled in the text area near to the context. The audio files are stored in the mp3 format. The user's speech is converted to ext using the speech recognition module in python. To capture the microphone input and down-sample the audio input, the PyAudio package is imported. The PyAudio is a open source, cross platform package which allows to play audio and record audio. It provides a python bindings for port audio. The PyAudio recognize the frequency, filter the low frequency audio samples.

Recognizing speech requires audio input and speech recognition makes retrieving text from the audio input really easy. The speech recognition library acts as a wrapper for several popular speech API's. The speech recognition happens with the recognizer class. The primary purpose of a recognizer instance is to recognize speech. The recognize_*() method will throw a Speech Recognition request error exception when the audio input can't be recognized by the recognizer class. In this way the user's input is filled in the appropriate text field with more accurate conversion of speech to text. The completion of all the text fields in the form, user can submit the details. The details will be stored in the database.

A STT system consists of one interface as front end and the other as back end. The first interface divides the single audio input into several fragments and then undergoes the process of dynamic down sampling for each fragmented audio input. The second interface divides the output of the first interface into small fragments. It removes the low frequency sound. Therefore the perturbations are eliminated and the conversion of speech to text will be accurate.

V. Speech Interface Synthesis Model

The speech interface provides the audio instruction to the user first and it gets a speech as an input from the user. The conversion of speech to text takes place by several steps.

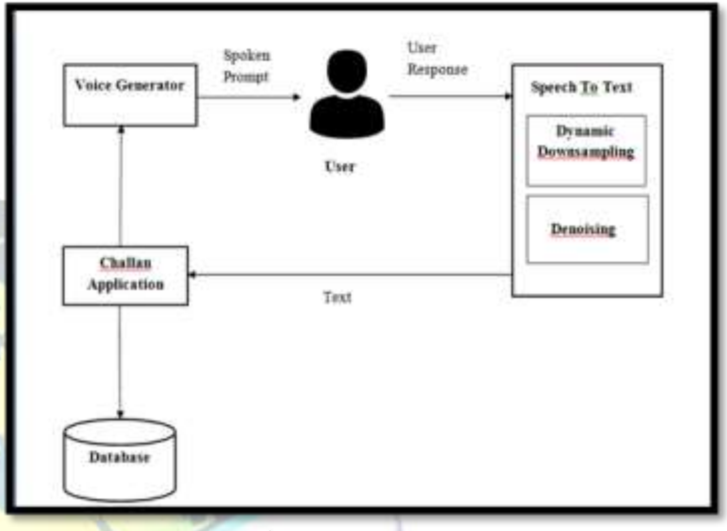


Fig. 1.Speech Interface challan application overview

A. *User*: The plebeians or the person those have difficulty in filling the form needs a human assistant or the speech interface.

B. *Challan application*: The challan application has three pages namely, home page, challan form, admin page.

1) *Home page*: The home page is designed with the html, php, css. The page displays the welcome greeting to work with the speech interface.
2) *Challan form*: The challan form is a front end shown to the user in the clearly understandable format. The challan form consists of the fields like Account holder name : the person name who accepts legal responsibility for handling the account, Account number : It is a unique string of numbers, letters and characters that defines the owner of a service and permits access to it , Date: Indicates the date in which the person depositing or remitting the amount, Branch: It is the Physical location of a banking corporation, Phone number : It is the sequence of digits helps to communicate with the people, Amount : The cash to be deposited, Remitters name : A person name who sends the payment .
3) *Admin login*: The admin login has a set of credentials used to authenticate a user. It consists of

user name and password. The Admin will register for username and password. A username is a name that uniquely identifies someone on a computer system. The username is almost paired with a password. While logging in, the username may appear on the screen and the password is kept secret. By keeping the password private, the details of the users are secured. The Admin has the access to login for checking of the user details who used the speech interface challan application.

C. *Voice generator*: The man-machine communication in the speech interface challan application is initiated by providing the audio instruction to the user. The voice generator helps in generating the appropriate audio instruction to the particular field.

D. *Speech to Text (STT)*: he user response was taken as input to the STT. The audio input was converted to text by undergoing two processes with the help of python libraries accurate conversion. The two processes are namely dynamic downsampling and denoising.

1) *Dynamic downsampling*: dynamic downsampling is the process of making the digital audio signal smaller by lowering its sampling rate or sample size (bits per sample).
   - i. The audio input waveform is considered as ‘x’, window frame size is considered as ‘n’.
   - ii. The input ‘x’ is divided into smaller fragments based on window size. Sx = len(x)/n.
   - iii. The fragment audio sample are subjected to downsampling. The downsampled audio fragments are denoted as dSx_i .

The output of the dynamic downsampling is obtained by the operation XORin the downsampled audio fragments. $Y=dSx_1 \oplus dSx_2 \oplus dSx_3 \oplus \ldots\ldots \oplus dSx_n$

2) *Denoising*: Denoising is any signal processing method which reconstructs a signal from a noisy one. Its goal is to remove noise and preserve useful information. In this process we divide the downsampled audio fragments into two parts. The first audio part is taken as $c(r_1)$ and the second audio part is taken as $c(r_2)$, compare $c(r_1)$ with $c(r_2)$, if the length of $c(r_1)$ is longer than $c(r_2)$ then the audio of $c(r_1)$ is again divided into two parts. The division step will be repeated until the downsampled audio is divided into many small fragments. In all the tiny fragments, the low frequency sounds are eliminated. Thus the noise in the audio gets eliminated; the tiny fragments are fused together. The downsampled and denoised audio is converted into text. These processes are done by the pyAudio and SpeechRecognition packages in python.

E. *Database*: A database is an organised collection of data. Python, PHP, MySQL , is used to design the database in the speech interface challan application. The required tables with input fields are created for managing the records efficiently.

VI. DESIGN AND IMPLEMENTATION

Our designed web application is called the speech interface challan application, with the speech to text functionality. The system was developed using PHP, Python 3.6.2, HTML, CSS, Java script.

The application is divided into two main modules. The first module which includes the basic GUI components which handles the basic operations of the application such as generating audio instruction, receiving user speech (audio) as input. The second module, the main conversion engine which integrated into the main module is for the acceptance of audio input, hence the conversion.

Speech interface (STT) converts speech to text by receiving the audio input from the user’s speech. The python library packages start the conversion of speech to text. The recognition of speech takes 20 to 30 seconds and finally the text will be displayed in the appropriate text field. STT shows an exceptional error when the speech cannot be recognized by the STT engine.

The following figure depicts the working procedure of the speech interface challan application,

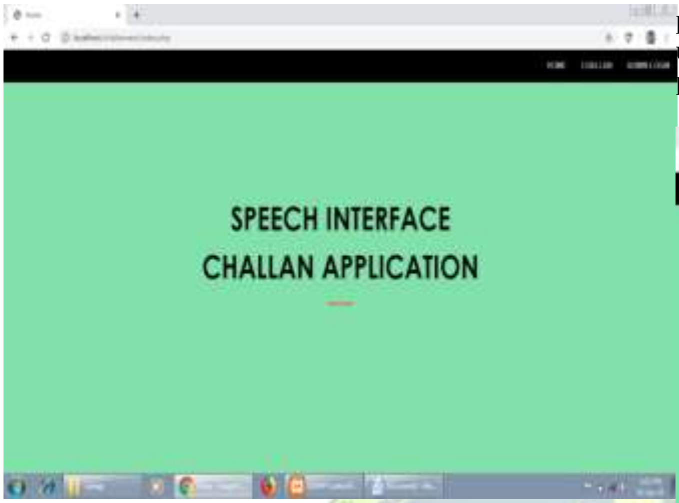


Fig. 2. Screenshot of the home page of speech interface challan application

This the first page of the web application. This screen appears in the full screen mode when the application is launched. As we can see home, challan , admin button present in the top right corner of application window , each having different functions. Click on the challan to move to the page challan form. Let's see working of audio instruction in the challan form,

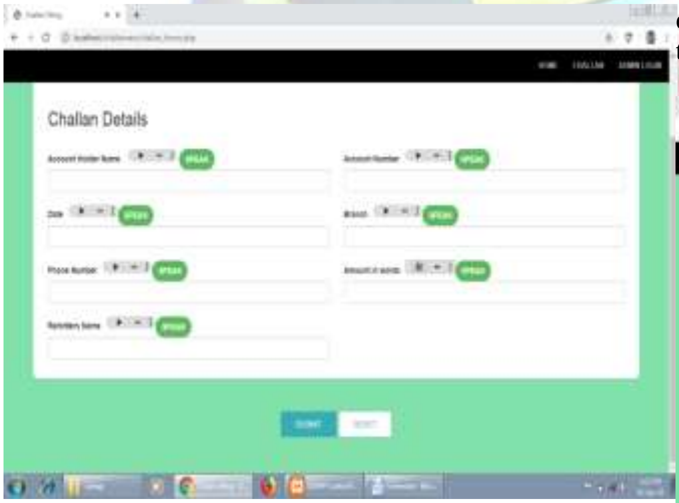


Fig. 3. Working of audio instruction in the speech interface challan application

The user clicks the play button to hear the audio instruction. The volume button is used to adjust the volume of the audio played. After listening to the audio instruction, the user understands what to be filled in the field and the user provides the audio/speech as input,

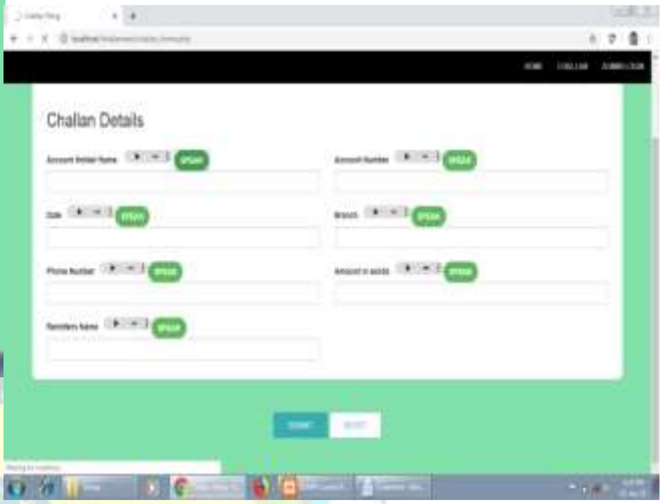


Fig. 4. The recognition of speech

The audio input is provided to the STT by clicking the speak button once. The python program recognizes the speech in 20 to 30 seconds and displays the text in the appropriate text field. In this way all the text box can be filled. Fig.5. displays the challan form is filled with user details in all the text boxes.

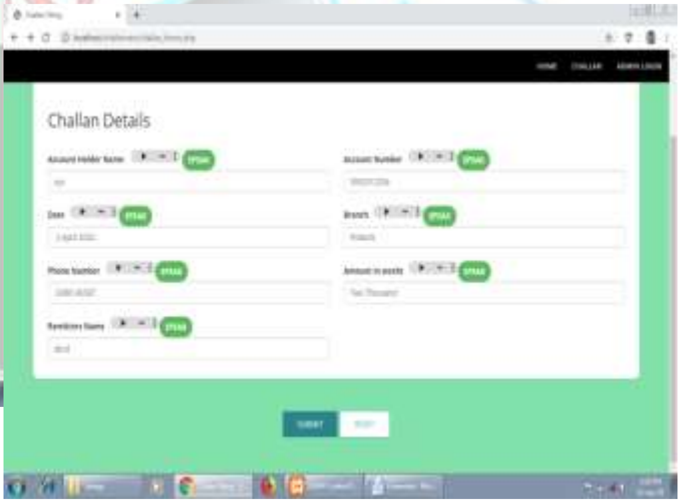
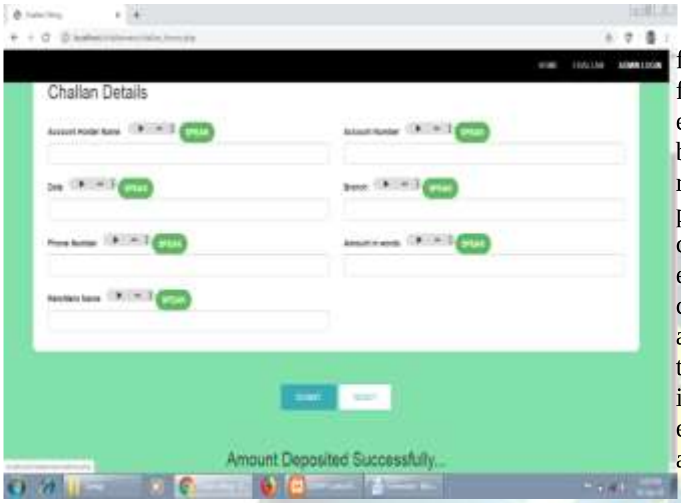


Fig. 5. Challan form is filled with user details

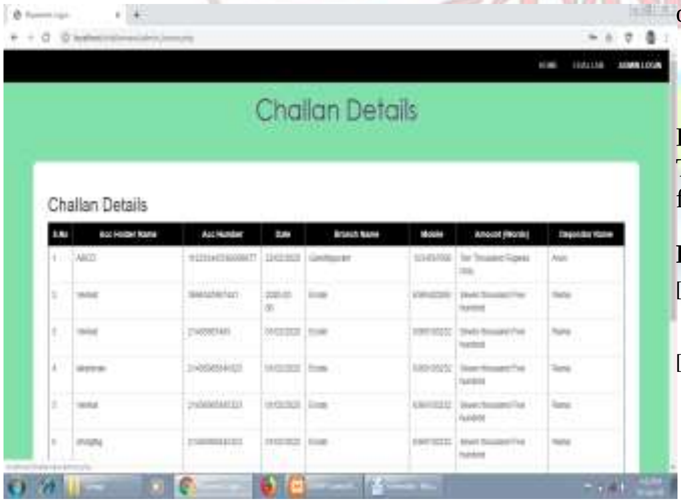
After filling the details, the submit button is clicked by the user. Fig.6. shows the “amount deposited successfully” will be displayed at the bottom off the page.


Challan Details

Amount Deposited Successfully...

Fig. 6. The acknowledgement for the amount deposit

When the amount is deposited successfully the acknowledgement is shown to the user and the details get automatically stored in the database. To know the details of the user who accessed the speech interface challan application, the admin can login to the database to view the records.


Challan Details

Challan Details

S.No	Acc Holder Name	Acc Number	Date	Branch Name	Mobile	Amount (Words)	Depositor Name
1	ABCD	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Fig. 7. The records of the users

The Fig. 7. Table displays the details of the user who accessed the speech interface challan application.

VII. Conclusion

With the world’s technology giants are making a beeline for voice based applications, the speech interface form filling will be more useful for plebeians. Here the exploration of speech recognition accuracy came to known by reviewing the existing literature survey. Thus the speech recognition in speech interface is done by the python packages. Accuracy of the software is excellent in the conversion to text of its ability to work in the real life environment. The speech interface is implemented in the challan application form filling. The speech interface challan application helps the plebeians to fill their form by themselves in bank sectors, without sharing their personal information to third persons. Further discussed about the expansion of speech interface in the field of other application.

Future enhancement: The following section discusses some limitations identified in the speech recognition. The major limitation in the speech interface is designed for one particular language. The future work planned to carry out the proposed limitations by expanding voice based speech interface applications in multi languages or indigenous languages. The addressed speech interface systems have great potential in developing the speech based applications in wide variety of service sectors for the aadhar card, PAN card, Gas connection, education, health care, tourism and other public and private sectors.

Acknowledgment

The authors thank the Computer Science and Engineering department, P.A. College of Engineering and Technology, Coimbatore, India for granting laboratory facilities to carry out the project work.

References

[1]. Ananya Paul et al “Development of GUI for Text-to-Speech Recognition using Natural Language Processing”, IEEE 2018.

[2]. Keii chi Tamura et al “Novel Defense Method against Audio Adversarial Example for Speech-to-Text Transcription Neural Networks”. In proceeding of IEEE 11th International Workshop on Computational Intelligence and Applications November 9-10, 2019, Hiroshima, Japan.

[3]. N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in 2018 IEEE Security and Privacy Workshops (SPW), May 2018, pp. 1–7.

[4]. Yuan Jiang et al "Improving Sequence-to-Sequence voice conversation by adding text-supervision". At National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China in 2019.

[5]. Snezhanapleshkova et al "Reduced Database for Voice Commands Recognition Using Cloud Technologies, Artificial Intelligence and Deep Learning". In proceedings of XVI-th International Conference on Electrical Machines, Drives and Power Systems ELMA 2019, 6-8 June 2019, Varna, Bulgaria.

[6]. Atma Prakash Singh et al "A Survey: Speech Recognition Approaches And Techniques" In proceeding of IEEE Uttar Pradesh Section International Conference On Electrical, Electronics and Computer Engineering 2018.

[7]. Farhan Khan et al. "Voice to Text transcription using CMU Sphinx" . Journal of IEEE transactionon human machine system, VOL. 47, NO. 6, DECEMBER 2017.

[8]. Jiangtao Wang et al "CAPFF: A Context Aware Assistant For Paper Form Filling". Journal of IEEE Transactions on Human-Machine Systems vol. 47, no. 6, December 2017.

[9]. Dr. JayashriVajpai et al "Industrial Applications of Automatic Speech Recognition Systems". Journal of IEEE Engineering Research and Applications, Vol. 6, Issue 3, (Part - 1) March 2016, pp.88-95.

[10]. Yogita H. Ghadage "Speech to Text Conversion for Multilingual Languages". In proceeding of International Conference on Communication and Signal Processing, April 6-8, 2016, India.

[11]. Vishnudas Raveendran et al "An Approach to File Manager Design for Speech Interfaces", in 2016, India.

[12]. YoungJae Song et al "Classifying Speech Related vs. Idle State towards Onset Detection in Brain-Computer Interfaces", in 2014.

[13]. Lie deng et al "New types of deep neural network learning for Speech recognition and related applications: An overview", IEEE 2013.

[14]. Lei Xie et al "Speech and Auditory Interfaces for Ubiquitous, Immersive and Personalized Applications". In proceeding of Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing 2010.

[15]. Sunil Issar "A Speech Interface for Forms on WWW". At Carnegie Mellon University in 2009.

[16]. Shih-Jung Peng et al "A Generic Interface Methodology for Bridging Application Systems and Speech Recognizers" IEEE 2007.

[17]. Bernhard Suhm "Interactive recovery from Speech Recognition errors in speech user interfaces" IEEE 2006.

[18]. Sadaoki Furui "Automatic speech recognition and its application to information Extraction" 2005.

[19]. Nobuo Hataoka et al "Robust Speech Dialog Interface for Car Telematics Service", IEEE 2004.