VIRTUAL ENQUIRY SYSTEM USING NATURAL LANGUAGE PROCESSING AND VISUAL GRAPHICS

Nair Aishwarya[1] Student[1]
Nakhwa Shumayla[2] Student[2]
Shaikh Mohammed Shafiq[3] Student[3]
Ansari Habbibur Rehman[4] Student[4]
Shaikh Mohammed Ashfaque[5] Professor[5]

[1]Computer Engineering,
[1]Rizvi College of Engineering, Mumbai, Maharashtra, India

ABSTRACT : In many infrastructural organizations or at different venues we see help desk, boards or receptions that are placed to help visitors or newcomers navigate into the building or enquire about any service provided by them. We are virtualizing this help desk so that it is available 24X7 to the users. We propose a system that interacts with the user in audio form with visual graphics to enhance the whole experience. This system would several APIs and services to take a voice input which is would be converted into text and processed with the help Natural language processing and for animating the user experience. The output would be given in audio form in sync with the graphics used. The design is created in a form to make the system in a more natural way. The system can use computer vision and sensors to enhance the whole interaction with the user.

Keywords: Chat-bot, Automatic Speech Recognition (ASR), Natural Language Processing (NLP), Natural Language Understanding (NLU), Natural Language Generation (NLG), Visual Graphics, Lip Sync.

I. INTRODUCTION

The need of enquiry is a must for every individual on a daily basis whereas the type of enquiry may differ with the context of location, designation, situation and some other factors. Most visitors or guests are unaware of many services and location to be approached in an infrastructural organization or at various venues. The major reason for proposing this system is that it would provide the user with available information.

The Virtual Enquiry System (VES) is an interactive agent that would use chatbot technology to help guests to navigate their way inside an infrastructure. Its input will be the questions of the user that will be naturally spoken and will provide an answer within the restricted domain. This respective conversation will be conducted till the user finishes enquiring. This system will also use visual graphics to enhance the user experience while the interaction.

The idea behind the VES is to take the traditional concept of having a receptionist or help desk and virtualizing the user experience. This interaction will be more interesting due to the graphics. This system makes the help desk be available for 24X7 at a cheaper cost.

II. LITERATURE REVIEW

The VES system consists of different mechanism blocks which accepts the users voice as an input, analyses it, processes it and synthesizes a proper text output and converts it into voice output which is synced with the visual graphics.

1. AUTOMATIC SPEECH RECOGNIZATION (ASR)

To transform speech into text, a computer has to perform several complicated steps. We create vibrations in the air while speaking. The **analog-to-digital converter (ADC)** translates these vibrations which is an analog wave into digital data that is machine understandable. VES then filters the digitized sound to cut out unnecessary noise and to separate it into different bands of **frequency**. It also normalizes the sound to a constant volume level. The speed of sound of the user varies from person to person, so the sound has to be adjusted in order to match the speed of the sound samples that are already stored in the system's memory. The program then figures what the user was possibly speaking and outputs it as a computer command [1].

2. NATURAL LANGUAGE UNDERSTANDING (NLU)

The objective of Natural Language Understanding (SLU), is to extract all the information from an audio-based input that is related to a particular task.

NLU performs two main roles

(a) Parsing the Automatic Speech Recognition (ASR) output into meaningful segments that result in accurate search

(b) Understanding user's intention.

Stages of NLU

1. **RELATIONSHIP EXTRACTION**
2. **SEMANTIC PARSING**
3. **PARAPHASE AND NATURAL LANGUAGE INFERENCE**
4. **SENTIMENT ANALYSIS**

3. NATURAL LANGUAGE GENERATION (NLG)

Natural Language Generation is the production of Natural Language from a given dataset. It turns statistical data and facts into meaningful English. In short NLG means translating the computerized data into natural language which can be displayed as text or spoken by the VES system.

Stages of Natural Language Generation:

1. **CONTENT DETERMINATION**
2. **DOCUMENT STRUCTURING**
3. **AGGREGATION**
4. **LEXICAL CHOICE**
5. **REFERRING EXPRESSION GENERATION**
6. **REALISATION**
7. **SPEECH SYNTHESIS**
8. **TEXT TO WORDS**
9. **WORDS TO PHONEMES**
10. **PHONEMES TO SOUND**

The Natural Language Processing Toolkit available in Python programming language is used in the project. These are some of NLP Pipeline used. These are as follows [6]:

- ***ACCEPTING THE INPUT (Raw Text, Sourcing and Normalization)***
- ***PRE-PROCESSING***
- ***REGULAR EXPRESSIONS***
- ***POS TAGGING & GRAMMARS***
- ***CHUNKING***
- ***INFORMATION EXTRACTION***
- ***OUTPUTTING THE VOICE OUTCOME***

➢ ACCEPTING THE INPUT

The input is accepted using Google's API or any other APIs like RealSense, iSpeech, Dragon Speech and many more. The input data for the database purpose can also be accepted in DOC and PDF format or can read contents from an RSS feed.

➢ PRE-PROCESSING

In Pre-Processing, Tokenisation, Stemming, Lemmatization and other concepts are involved.

➢ REGULAR EXPRESSION

Regular expressions are one of the most simple and basic, yet most important and powerful, tools to be used. More commonly known as regex, that are used to match patterns in text. Basically used for pattern matching purpose as a way to do text analysis like text match, text search or text extraction operation.

➢ POS TAGGING & GRAMMARS

Tagging is the process of classifying the words in a given sentence using parts of speech (POS). Software that helps achieve this is called tagger. There are different types of taggers like In-built tagger, Default tagger, Regular tagger & Lookup tagger. CFG describes a set of rules that can be applied to text in a formal language specification to generate newer sets of text. CFG in a language comprises the following things: Starting token, ending symbols, non-ending symbols, rules or production.

➢ CHUNKING

Chunking is the process of extracting short phrases from text. POS tagging algorithms can be leveraged to do chunking. The main point to be noted is that the tokens (words) produced by chunking do not overlap.

➢ INFORMATION EXTRACTION

The natural step after these processes is to identify the Interested Entities in a given piece of text. When processing large amounts of data, the main interest is to find out whether any famous personalities, places, products, and so on are mentioned or not. And

these things are called named entities in NLP. These help us understand more about what is being referred to in a given text so that we can further classify the data. Since named entity comprise more than word, it is sometimes difficult to find these from the text.

4. VISUAL GRAPHICS

We are providing an avatar with our VES so as to increase the personal touch of humans and also you need a face to match the voice. After the formation of response, the processor examines the response input given to the avatar letter by letter.

The avatar would be speaking the sentence produced by the NLP hence it has to match the facial expressions to the sentence it has to speak. The program would have different expressions and the way face shapes when a letter is pronounced already in it. The avatar will properly open its mouth when letters like 'O' are to be pronounced or closed for letters like 'F'. The avatar would also have to know when pause at certain spaces and punctuation in the sentence. There are different workspaces created for visual graphics like Tupad and softwares form Adobe.

Traditionally animation worked when a lot of similar drawings where run together very fast. This is now done using code wherein one drawing is used and section of it is changed and instructions like move, turn etc makes it an animation.

After the speech synthesis process, the processed output i.e. the final text answer is then converted into speech (audio format). This output is served as an input to the graphics system. Where the phonemes of the text are checked with their respective lip stimuli and displayed on the screen. This works in such a way that it first checks the phoneme of the text and accordingly the reaction for that particular text phoneme is the lip movement of the animated character or the avatar on the screen.

Let's understand this in more detail with an example, if the output of the speech synthesis is let say, 'Check on second floor'. When the word 'check' is inputted to the graphics system, when check is pronounced usually the lips spreads horizontally wide and teeth are visible so in this way the movement of avatar's lips are already coded with different scenarios and conditions, also when a word 'floor' has to be read the lips basically becomes round and this movement of the character's lips takes place.

III. METHODOLOGY

The design of the system is shown in the figure. The user will ask something to the system and the system would recognise the input. Any services can be used to convert the audio input into text, some of the speech recognizers than can be used are Google Speech Recognition API or Watson services on IBM. The output text would be a structured sentence.

The Text will be further analysed using Natural Language Understanding (NLU).During this the text is tagged for entities which are classified into categories. Pattern recognition is applied on the keywords thus giving us a raw output. This raw output is then synthesized into a proper sentence with the help of NLG. Another API will be applied to this text so that a voice file which contains an appropriate response to the original input is generated. The final step would be to sync this audio with the visual character. This syncing will be done by checking for all conditions i.e. which phoneme pronunciation creates what kind of respective movement of the lips so that the character would copy it.

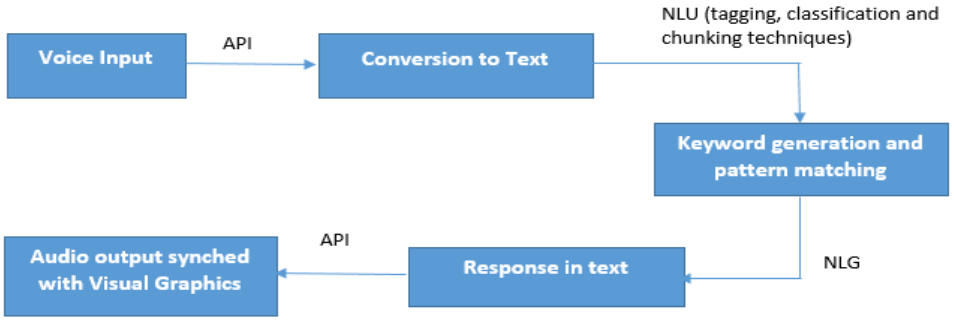

Figure 1: Block Diagram of VES

IV. CONCLUSION

The Virtual Enquiry System can be a very good alternative of the present human helpdesks. All the queries related to anything in a restricted domain can be easily answered and in the best way. Also in market, there can be many devices or kits that can speak or interact acoustically but what makes VES distinct from others is its graphical interface that would create a very real human-like experience and also its boundary that is restricted to an organization or an institution or another infrastructural complex makes it easier to install without large complexity in terms of memory, data and its processing.

V. REFERENCES

[1] https://electronics.howstuffworks.com/gadgets/high-tech-gadgets/speech-recognition1.htm

[2] https://sflscientific.com/data-science-blog/2015/12/11/natural-language-processing-information-extraction

[3] https://en.wikipedia.org/wiki/Semantic_parsing

[4] https://towardsdatascience.com/convolutional-attention-model-for-natural-language-inference-a754834c0d83

[5] https://en.wikipedia.org/wiki/Sentiment_analysis8

[6] Natural Language Processing with Python by Krishna Bhavsar, Naresh Kumar and Pratap Dangeti, Packt.

[7] https://www.geeksforgeeks.org/artificial-intelligence-natural-language-generation/

[8] https://www.explainthatstuff.com/how-speech-synthesis-works.html