Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

Speech Oriented Virtual Restaurant Clerk using Web Speech API and Natural Language Processing

Gautam R
Sir MVIT
Bengaluru, India

Akahay G J
Sir MVIT
Bengaluru, India

Dhavan R
Sir MVIT
Bengaluru, India

Amika Kumawat
Sir MVIT
Bengaluru, India

A. Ajina
Sir MVIT
Bengaluru, India

***Abstract*—Voice interfaces and artificial speech synthesis is revolutionizing the human computer interaction in a multitude of unimaginable ways. Predominantly gaining popularity due to their similarity to human conversation. This technology has also extended its arms to a larger neglected crowd. From placing phone calls and texting their caretakers to unlocking front doors and ordering groceries, virtual assistants are making important steps towards accessible UIs for the masses. Our project explores a specific dimension overcoming the challenges and improving the flexibility with restaurant ordering workflow. We have experimented with a system that acts as a restaurant-clerk and ameliorates the overall customer experience. Through this paper, we have researched various possible speech recognition techniques and picked the most suitable one for our application. The system implements face biometrics, innate food recommenders, and speech recognition algorithms that are deployed through python and web front-end.**

Keywords*— *Face Recognition, Machine Learning, Collaborative filtering, Convolution Neural etworks, Speech UI, Natural Language Processing.

I. INTRODUCTION

Speech recognition technology entered the public consciousness rather recently, with the glossy launch events from the tech giants making worldwide headlines. The earliest advances in speech recognition focused mainly on the creation of vowel sounds, as the basis of a system that might also learn to interpret phonemes (the building blocks of speech) from nearby interlocutors. Machine learning, as in so many fields of scientific discovery, has provided the majority of speech recognition breakthroughs in this century. Google combined the latest technology with the power of cloud-based computing to share data and improve the accuracy of machine learning algorithms. Speech recognition system has the edge over other systems as it is more user-friendly and caters to a broader audience (including people with hearing ailments). This is also the reason for the popularity of this system.

The goal of our project was to simulate restaurant-clerk behavior. It must be able to provide information and ask client questions Similarly to how a human clerk does. The earlier dialogue sub-system uses several kinds of knowledge which are represented as frames, rules and class instances. Preliminary researchers used frames to represent client interaction. Each frame represents a class of elements, and is a compound of a slot set. The linguistic knowledge necessary for clients' natural language analysis is represented as syntactic and semantic rules, which are stored in the Output Interface Knowledge Base.

Siri. Amazon's Alexa. Google's Home Hub. Facebook's Portal. Apple's HomePod. Samsung's Bixby. Microsoft's Cortana. Today's assortment of voice assistants and smart speakers have integrated into the lives of everyday people — and the restaurant industry is taking note.

When 72% of smart-speaker owners claim their device is an integral part of their daily routine, restaurant owners and managers have a pivotal opportunity to use voice search to their business' benefit.

II. SPEECH INTERFACE

A. Voice Automated Restaurent System

The primary approach into the development of a restaurant menu speech-interface was attempted in the form of predefined set of Dialogues between users and systems as entities. Frames [9][11] to interpret the client interactions and restrictions, coupled with rules and class instances to address the linguistic natural language analysis while hosting a knowledge base for the items on the menu respectively. The NLG [12] (natural language generator) has two functions – deep generator and surface generator. The former is responsible for generating what to say i.e. context of speech interaction. While latter

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

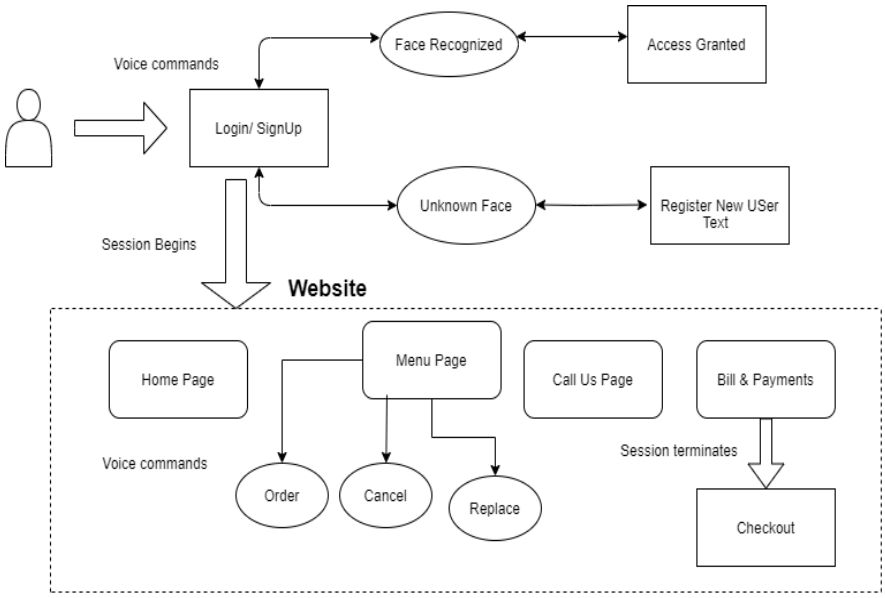

Fig. 1. Schematic View of the Experimental Model

concerns with two criteria: delivering accurate information to avoid client confusion along with enhanced naturalness of the conveying system.

B. Convolution Neural Networks for Speech

Speech recognitions in ASR's has a long history with GMM (Gaussian Mixture models) and HMM (Hidden Markov Models). However, recent experiments displayed a significant spike in the accuracy, promoting the exploration towards the CNN's and their combinatorial versions to predict speech inputs. [13] Demonstrated experiments with all possible CNN combinations and concludes the TF-CNN yields the best results with a reasonably high metrics. Spectrogram noise smoothing effect has augmented this performance boost to CNN's. For the first time, we have seen a relatively lower WER (Word Error Rate). Deep Neural Nets have extrapolated the potential CNN's can achieve with appropriate [15][16] training data and diverse [13] [14] inherent patterns.

C. Web Speech API

The API encompasses both speech synthesis and analysis. Speech-to-text (STT) and Text-to-speech (TTS) are both supported functions. In addition, the API is purely coded in JavaScript, extending the flexibility into the front-end development and integration.[1] Event based architecture asynchronously processes speech and reports intermediate speech recognition results. Intermediate and final recognition results are represented as candidate sentences with associated confidence values. Highest confidence score pertains to the most likely sentence recognized. Word accuracy and semantic metrics validate against the TSP database with 1400 recordings of Harvard Sentences [7][8]. API supports currently just the Chromium web browser (V.25+).

III. NATURAL LANGUAGE PROCESSING

NLTK stands for Natural Language Toolkit. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. This is one of the most usable and mother of all NLP libraries.

spaCy, is a completely optimized and highly accurate library widely used in deep learning.

A. Natural Language Toolkit

NLTK is undoubtedly the best open source module that offers the best of tools for Natural Language Processing. Tools are designed considering factors such as usability, simplicity, extensibility, consistency and simplicity. A vast community and precise documentation are what makes this library the most desirable. For our application, we have taken advantage of the Tagger modules and Dependency parser that identifies Parts of speech and sentence structures with a decent accuracy. Visualization modules deliver good insights in processing the correlations between sentence semantics. Tree structure visualizations have enabled in depth understanding for our project.

B. spaCy

spaCy visualization tools and chunk tagging enabled us to view the relationship strengths amongst the words forming the sentence. The statistical models spacy offers prove highly beneficial in understanding sentence semantics and neighboring word associations.

IV. FACE RECOGNITION

Face Recognition has transformed into the state-of-the-art biometric techniques globally. Ranging from face indexing to HCI (Human Computer Interaction) this technology has mimicked the human pattern to recognize individuals in real-life. [2] [3] The basic principles stem from loading an image, detecting the face, validating against dataset and finally recognizing the target. [2][3] This research compared several algorithms to accurately detect and verify faces, tested against Indian database, Yale Face Database B, AT&T (ORL database) and lastly the FERRET database. A compound study marks the benefits [17][18] of curating the best practices and sustainable

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

challenges leading us to make wiser choices for our system. Hence, we were convinced to apply the prolific Deep Neural nets in our case.

V. METHODOLOGY

The core focus in our project was to improve the efficiency

Fig. 2. Face Detection and Recognition

of restaurant management while simultaneously reduce the human errors (waiters confusing the orders). This would serve as a boon to the growing food industry to better handle customer relationships with minimized overheads and concentrate only on their workflow. The transformation of this idea into a reality is what our project emphasizes. Some of the challenges that we solved include - placing orders directly by the customers, tracking status for the foods in transit, accurate recommendations based on customers taste interests, seamless handling of billing and payments. The tour into our system reveals two phases: Firstly, customer authentication through face-biometric i.e. login/signup into the system. Secondly, order placing followed by billing and payment modes. The system design is a dramatic shift in UI with Hands-free voice commands for navigation across all components that constitute. Our system also extends its accessibility to the less favored (disabled individuals) instilling a sense of independent dining experience.

A. Voice Automated Restaurent System

The login page retains the traditional familiar outlook, but the additional features create a dynamic shift in user experience. The signup/login page renders a facial recognition system that identifies the user (customer), authenticating them into the restaurant menu system. The face recognition biometrics implements the CNN [19] modules to detect and assign facial encodings to the face-images. The output encodings from CNN modules are compared against the user template images in the file-system, applying CNN encodings to the live video stream. This offers us the convenience to provide single user-image registration and accurate recognition at the time of login. In case of a new user the absence of facial encodings directs the system to identify the user an unknown customer. Following which the user is redirected to the signup page for registration. The feature extraction and classification of the facial encodings [11] [12] run through grey scale conversion of the training images and SVM classifiers respectively. The accurate distinguishing ability of the system is crucial and pivotal in user identification and further recommendations. To further boost the system performance and curb compute overhead, we have resized the training images to quarter the resolution while maintaining accuracy.

B. Food Recommendation System

Diagnosing the Menu page implicates the presence of foods and drinks sorted based on cuisines, nature of foods, taste preferences, and their ratings. [5] Proposes matrix factorization for food recommender system that fuses rating information and user tags to achieve significantly better prediction accuracy. Collaborative Filtering forms the essence of our recommender. Largely due to the fact of real-time updates, finite and maintainable database, and lower compute footprints. To begin with, mapping of users and items into a matrix take place through MF (matrix factorization). The ratings by corresponding user's to items occupy the respective cells of the matrix. This data processes into the nth dimensional space to map users-item data points. The application of KNN (K Nearest Neighbors) congregates similar users based on ratings. The recommendations tailored for user1 are similar to 10 (K=10) other users with similar ratings. The top-rated foods yet to be tried by user1, but rated by similar users will be recommended to user1. [5] [6] Hybrid system are to be a part of our future endeavors based on the induced flexibility and performance.

C. Integrating Web Speech API

Fundamental differentiating factor of our system is the speech interaction-controlled environment the entire speech handling applications are monitored and processes for STT (speech to text) and TTS (text to speech) via web speech API. The motive was to keep the client-side code light weighted in order for synchrony amongst other services. The web speech API inclusiveness also stems from the consolidated efforts of different companies like Google and OpenStream to process recognition tasks accurately while monitoring periodic updates. The vast training data on which the models are trained are easily accessed [1] just by the mere integration of the API. The input from the client microphone is carried instantly into the API and the consequent processing involves analysis of the spectrograms and producing intermediate and final candidate outputs. Every candidate is tagged with a confidence value. The most likely confidence is returned as suitable transcript. This API performance validates against the TSP dataset [5] consisting of 12 male and 12 female individuals. Diagnosing the Menu page implicates the presence of foods and drinks sorted based on cuisines, nature of foods, taste preferences, and their ratings.

D. Combining Speech Processing and spaCy tools

For the most important aspect of our system's functionality i.e. NLP (Natural Language Processing) for the recognized speech. We have provisioned modules from spaCy libraries and NLTK (Natural Language Toolkit). We trained a custom dependency parser on the entire Menu and navigation commands that a user could potentially inject. Using the POS tagging (parts of speech) we could extract valuable keywords like - item names and their quantities. This enabled improved

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

system accuracy and navigation across the site smooth. Once the processed items and quantities appeared onto the order screen, we could track the status of the food in preparation. Red indicates the option for cancellation, yellow indicates the food is in transit and green implies the served orders as shown in Fig. 3. Later on, the system allows the users to pay the auto-generated bills using UPI, cash, and card payment modes shown in Fig. 4.

Other options include calling the manager in case of system failure or login issues.

E. Design and Integration

- **Front-end design**

The Whole Project is voice based Web Application. Therefore, we used a HTML, CSS and JavaScript for developing the front-end of our web application. HTML is used for basic website(html) components, CSS for styling the HTML page and JavaScript for the basic Functionalities such as clicking a button, reloading a page, call specific functions etc.,

- **Database**

We have used a Structured Database –SQLITE3 for storing of data and to perform queries on the data stored depending upon our need. We used Python with SQLite package for accessing the Database contents through SQL queries.

- **System Integration and Function calls**

Frontend Inputs communicate to Server-side python script using Ajax and HTTP Requests. These HTTP Requests and operational data renders in JSON format.

The facial python script integration into the web template incurred a few challenges. To eliminate them we took the needful steps:

a) Image Frames to be displayed on the Browser

For this we had built a channel that ensured every fame that was processed at the Backend had to be sent to the AJAX for it to be displayed as a video i.e. sequence of image frames at constant rate. Python 'yield' function returns every processed image to the front-end.

b) Retrieving User Credentials

For this, we used a python's global variable concept i.e. we used a global variable that would be accessed by both Face Recognition Function and the Flask function. Hence, enabling us to get the 'user id' recognized through Face Recognition Function and perform the further processing at Flask.

VI. RESULTS AND DISCUSSIONS

This project has given us an opportunity to learn and experiment building systems with modern speech interfaces and face recognition technologies. The outcome reveals interesting applications to embed such technologies in future business domains. We have successful digitized the traditional restaurant service such as manually taking orders, keeping track of the item status and also delivering the accurate bills to customers. Our results have managed to imagine how the restaurants in the future would possibly operate. The results that our face recognition model delivers is reasonably fair and the quick access nature into the system boost the customer-system experience.

Fig. 3. Menu Page

The Following Images display the Menu and other navigation pages within the website. We can easily notice the recommendations and ratings of customers. The recommendations are solely a by-product of customer ratings. We also observe the crucial aspect to a restaurant's profitability i.e. the billing and payment info. The accuracy of the facial recognition system stands at 89% while the speech recognition is assigned triggers to begin the speech recognition. This trigger in our case 'Ram' quickly activates the recognition of speech and starts to take input. Adding triggers to our system has significantly improved the recognizing capabilities of the system. For the payments, we accept all three modes of cash, card and UPI. Enabling customers to seamlessly dine and walk out of the restaurant after payments.

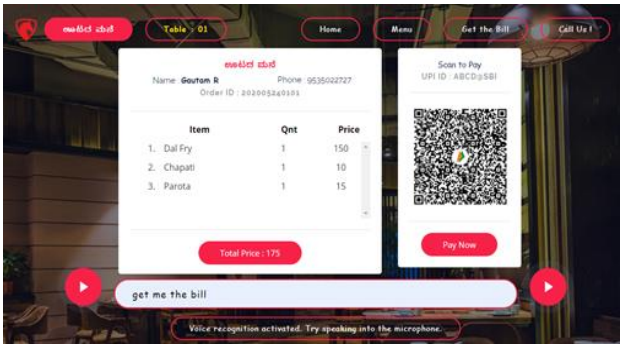

Fig. 4. Billing and Payments

VII. CONCLUSION AND FUTURE ENHANCEMENTS

In conclusion, our attempt aims at addressing the emerging Foodie community across the country. With new concepts such as Zomato gold and Dineout providing attractive deals on several restaurant partners. Massive crowds are flocking to these outlets regularly. Such environments call for an extreme organization in order tracking, customer support and accurate bill generations. To aid this cause our system deliver tools that ease the workflow of the restaurant and build customer experience. For future endeavors, we plan to integrate a native ASR (automatic speech recognition), preferably TF-CNN's (Time-Frequency convolution neural nets) for quicker speech processing and proprietorship benefits. We also intend to transform the web interface into mobile platforms like Android and iOS. Another area of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

possibilities we look forward into is reconciling with virtual assistants like Amazon Alexa and Google Homes. This system also delivers an important replication advantage. Which can adapt to multiple restaurant menus and listings.

ACKNOWLEDGMENT

Sir M Visvesvaraya Institute of Technology supports this project under the department of Computer Science. We would like to acknowledge our sincere gratitude to our HOD (head of the department) Dr. Banu Prakash and Project guide Dr. A. Ajina for the immense support and guidance that led to the successful completion of this project.

REFERENCES

[1] J. Adolf, "Web Speech API", 2013.
[2] Shakir F. Kak, Firas Mahmood Mustafa, and Pedro Valente "A Review of Person Recognition Based on Face Model ", Eurasian Journal of Science & Engineering, August 2018, doi:10.23918/eajse.v4i1sip157
[3] Kavita , Ms. Manjeet Kaur, "A Survey paper for Face Recognition Technologies", International Journal of Scientific and Research Publications, July 2016.
[4] Christoph Trattner, David Elsweiler, "Food Recommender Systems Important Contributions, Challenges and Future Research Directions", 2017.
[5] Edward Loper, and Steven Bird "NLTK: The Natural Language Toolkit", Department of Computer and Information Science University of Pennsylvania, Philadelphia, 2002.
[6] Freyne, J. & Berkovsky, S. (2010). Intelligent food planning: Personalized recipe recommendation. In Proceedings of the 15th international conference on intelligent user interfaces (pp. 321–324). New York, NY.
[7] P. Kabal,'TSP Speech Database," tech. rep., McGill University, 2002.
[8] H. R. Silbiger and J. L. Sullivan, \IEEE Recommended Practice for Speech Quality Measurements," IEEE Transactions on Audio and Electroacoustics, vol. 17, no. 3, pp. 225{246, 1969}.
[9] "Inteligencia Artificial, Segunda edicion". Mc-Graw Hill, 1994.
[10] J. Allen "Natural language understanding". The Benjamin/Cummings Publishing Company Inc., 1995.
[11] M. Nagala, T. Morimoto. "First steps towards statistical modelling of dialogue to predict the speech act type of the next utterance". Speech Communication 15 (1994) 377-378.
[12] R. López-Cózar, P. Garcia, J. Diaz and A.J. Rubio," Voice activated dialogue system for fast food restaurant applications"
[13] V. Mitra, W. Wang, and H. Franco, "deep convolutional nets and robust features for reverberation-robust speech recognition," in Proc. of SLT, pp. 548–553, 2014
[14] T. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural network for LVCSR," Proc. Of ICASSP, 2013.
[15] . G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task," ETSI STQ-Aurora DSR Working Group, June 4, 2001.
[16] G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task," ETSI STQ-Aurora DSR Working Group, June 4, 2001.
[17] Barnouti, N. H. (2016). Improve Face Recognition Rate Using Different Image Pre-Processing Techniques. American Journal of Engineering Research, 5(4), 46-53.
[18] Barnouti, N. H., Mahmood, S. S., & Matti, W. E." Face Recognition: A Literature", (2016, September).
[19] Bheleet, S. G., & Mankar, V. H. ,"A Review Paper on Face Recognition Techniques", International Journal of Advanced Research in Computer Engineering & Technology, 1(8) (2012, October).
[20] T. Hain, A. el Hannani, S. Wrigley, and V. Wan, "Automatic speech recognition for scientific purposes - webASR," in Interspeech'08, 2008, pp. 504–507.