ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse101922020.pdf
https://doi.org/10.30534/ijatcse/2020/101922020

Speech Recognition-based Billing System: A multi-model design and implementation

Aditya Dhinavahi[1], Prashant R Nair [2], Sundeep V V S Akella[3], Aneeswar K S S[4]
[1]UG-Scholar, Department of Mathematics, Amrita School of Engineering Coimbatore
Amrita Vishwa Vidyapeetham, India.[1]dhinavahiaditya@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering, Amrita School of Engineering Coimbatore, Amrita Vishwa Vidyapeetham, India.[2] prashant@amrita.edu.
[3,4]UG-Scholar,[3,4]Department of Computer Science and Engineering,
Amrita School of Engineering Coimbatore,
Amrita Vishwa Vidyapeetham,India.[3]sundeep.akella1999@gmail.com[4]aneeshkalisapudi@gmail.com.

ABSTRACT

Speech recognition has emerged as a cutting-edge area of research with wide-ranging applications. In this work, we propose a billing system, which has the capability to generate a bill by taking the customer's voice as the input source. The system aims to implement a multi-model approach in order to avoid conflicts in recognition i.e., having individual models (also considered as 'subjects') that recognizes the necessary speech. Also, the proposed system is incorporated with Google's speech to text model as it also has two custom-trained models that were trained using 1 Dimensional CNN. This system is also successful in recognizing different dialects of global pronunciations. To make it more interactive with the customer or the user, the system has a 'text-to-speech' technology that can speak or produce the given sentence(s) in the form of speech.

Key words : Billing, Google-speech-recognition, Google-text-to-speech, Multi-model, Python, TensorFlow.

1. INTRODUCTION

Speech recognition has emerged as a cutting-edge area of research with wide-ranging applications. One such system is the 'Google Assistant', which can be a nice friend to a person and can help the user do his search. Google Assistant has rendered convenience and added value to mobile phones and hand-held devices. Another application is 'Alexa' which helps the users to control his/her home appliances. The foundation of speech-recognition lies in the concepts of Natural Language Processing. These concepts are gainfully deployed in Google's Speech Recognition. For development and education, Google also provides a Speech-Recognition API. When one uses this API, it basically takes our input and leverages its knowledge based on the model which was being trained publicly for many years since its inception. The system recognizes several dialects and voices. Initially, users train their voice in Google Assistant and this serves as the foundation for the speech recognition system.For example, take the popular Food item in India called 'dosa', which is basically a pancake. There are many pronunciations in vogue such as 'dosha', 'dosai',' dosa' and many more. When training the recognition of this item in a single way of pronunciation or in a limited number of pronunciations, then it will be difficult to recognize other types of pronunciations. Keeping this in mind, Google has trained its model with many pronunciations and voices. Also, when considering Google's text-to-speech application, it uses techniques the same as that of Natural Language Processing to provide the voice output of the given text. It has a huge corpus of data on which this model was trained on. This data includes the text and the voice, which is further processed to get it into a sentence. This is because it is very difficult to train a model with different combinations of words; the existing text needs to be processed such that it can be recognized by the model. Google offers the tool in multiple languages apart from English such as various Indian languages like 'Hindi', 'Telugu', 'Tamil' and Malayalam. It offers different formats of English such as 'English-US', 'English-IN' and 'English-UK'.

The cutting edge area of technology implemented in recent times of research work, include many applications based on machine learning [6], neural networks [7] and deep learning [8].

2. LITERATURE SURVEY

In the research work presented by "Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Ryback, Pedro Moreno" have presented their work on "Bringing Contextual Information to Google Speech Recognition", where they have introduced a system which helps in improving the accuracy of speech recognition.

Sentence recognition is challenging because it usually has out-of-vocabulary words or not properly represented in a Language model. Hence we use the on-the-fly rescoring mechanism and perform n-gram biasing for the given language model. This method shows significant accuracy in various datasets in different contexts.[1]

In the research work presented by "Masanobu Nakamura, Koji Iwano and Sadaoki Furui" on "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance" the researchers have explained that when it comes to speech derived from spontaneous speeches, the recognition does not perform well whereas from reading texts, news broadcasts and other reads aloud speeches it has got high accuracy, This is only because spontaneous speeches and read-aloud speeches are completely different from each other, both acoustically and linguistically. In this paper, the system compares read the speech and spontaneous, their acoustic feature aspects. Experiments have shown that spontaneous speeches and read Speeches could be differentiated by reduced spectral space. Spontaneous speeches have more spectral space shrinks. In this execution, the researchers also talk about the spectral space which reduces the phoneme recognition accuracy. Results indicate that spectral reduction is one of the major reasons for the decrease of recognition for spontaneous speeches[2]. In the research presented by "Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani" on "Generation of large-scale simulated utterances in virtual rooms to train deep neural networks for far-field speech recognition in Google Home", the researchers state that they have trained the model with various room dimensions and cross-sections, a wide distribution of reverberation time and signal-to-noise ratios, even recognizing the elevates sound from the noise is also considered and a range of microphone or the amplifier and sound source locations. They have artificially generated and reproduced data by randomly sampling a noise configuration for every new training example model[3]. In the research presented by "Jasmine Bhaskar, Sruthi K, Prema Nedungadi" on "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining" the researchers have explained that Speaker's emotion cannot be recognized by the common speech recognition techniques, Hence the system mainly concentrates on emotions of the speaker from speech and texts. Researchers come to a conclusion that this system proposed creates a single unique feature vector for classification. The method used has Conventional approaches such as Natural Language Processing, Support Vector Machines, SentiWordNet and WordNet effect Dataset was outfitted by Semeval -2007 and enterface05 emotion Database[4]. In the research work presented by "Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, Noah A. Smith" on "Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters", the researchers state the problems of part-of-speech tagging for an informal, these online conversational or communicating texts that are generally being done in the social media. This is done by assessing large range unsupervised information or word clustering and new lexical features to improve tagging efficiency. In this genre presented by the researchers, conventional methods such as NLP and linguistic phenomena are also included. It also improved in twitter tagging from 90% to 93% (3% accuracy gain). It also helped in parts of speech annotation guidelines which are used to get a new dataset for English language twitter tweets It also improved in twitter tagging from 90% to 93% (3% accuracy gain). It also helped in parts of speech annotation guidelines which is used to get a new dataset for English language twitter tweets.[5]

3. PROPOSED SYSTEM

Typically in restaurants, hotels and food kiosks, the food counters have people who take our order, deliver the food ordered and help us complete the transaction. Our idea is to automate this process of 'billing' in food counters. In order to achieve that, we had an idea of making our application 'Speech-based'.

Our application aims at making billing more convenient to the customer by taking his/her voice and generating the bill. This makes two jobs easier i.e., there will be no need for human support in the counters as also the customer can avoid queues. Our model also aims for secure billing while keeping the process as simple as possible.

This proposed application is expected to take the item input and quantity input from the customer in the form of the customer's voice. Now, this input for item and quantity is separate, and, in order to avoid confusion or mistakes in this billing process, the application validates the process and checks with the customer to ensure that his/her order is correct.

In order to avoid conflicts in recognition, we have divided our Speech-recognition system into three models i.e., item, quantity, decision. The first model is based on Google's speech recognition module. It recognizes the customer's speech and passes it further to check the existence of the item. The second and third models are custom-trained using TensorFlow and Keras. These models recognize numbers and the decision of the customer. The audio is taken from 'Kaggle'(an open-source website that provides datasets). After analysis, the audio signals are resampled to ensure good accuracy in the model. We have also trained our model using some of the local dialect pronunciations, which made it easier for the application to recognize people's voices efficiently. The custom-trained models can recognize single words, whereas Google's speech recognition model can recognize sentences. To make the billing process even more interactive, we have used Google's text-to-speech module. This serves as a virtual biller who actually talks to the customer, while billing is happening and hence lowers the chance of confusion of the customer.

3.1 main modules

1. Speech Recognition
2. Text-to-Speech
3. Bill Generation

3.1.1 Speech Recognition

This module is where we have applied our multi-model strategy. Basically, the application uses the appropriate model according to the process and its attributes. Firstly, Google's Speech-to-Text model predicts the item asked by the user. This is done by using its machine-learning models. The second model is built with the help of Tensor-Flow and it predicts the quantity (ranging from 1-9). Since it was trained manually, some errors are evident in the prediction. The third model has a very simple job of predicting the words 'yes' and 'no'. Since it has to predict

only two words, it's accuracy is much better than the second model. The multi-model approach seems to be a better option since there is better accuracy in the overall application run. The models trained using TensorFlow use a basic concept of predicting audio waves using a 1-dimensional Convolution Neural Network.

3.1.2 Text-to-Speech

To make our application more interactive with the user, the system has Google's Text-to-Speech. This basically processes the text and gives the output in the form of audio. It basically analyzes the text, and then sets the linguistic parameters for the utterances and then produces the audio waves which will be further converted into the audio file, and the file will be played. This requires internet connection as the system uses an instance of Google's Text-to-Speech and the analysis is done on the basis of the knowledge the model has. The model resides in the cloud.

3.1.3 Bill Generation

The process of Bill Generation is very simple. The items and its rates are stored in a dictionary. Whenever an item input is given, it checks for it in the dictionary and if the item is there, then it's rate is taken and it will be multiplied with the quantity and then added to the total bill rate. Any new items can be easily appended to the dictionary. Also, the application is intelligent enough that if any customer asks for the item which is not present in the menu, then it will inform the customer that 'the item is not present' and then ask if the customer wants any other item to be ordered. This makes the work of the bill generation easier and hence we also have nullified the chances of ordering something that is not there. At There was no use of any of the tokenization so as to keep the application simple by using python dictionaries. This implementation using dictionaries significantly improves the performance of the application.

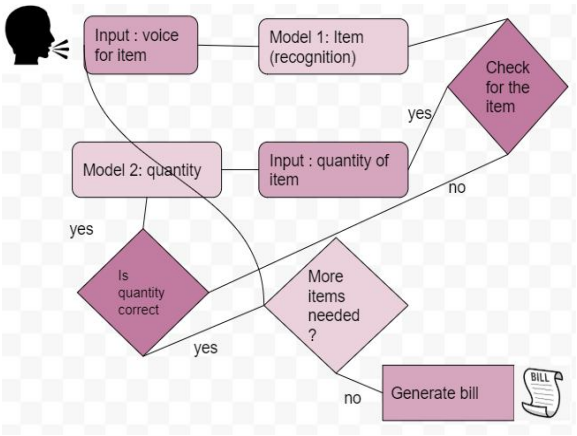

Figure 1: shows a flow diagram of the proposed system.

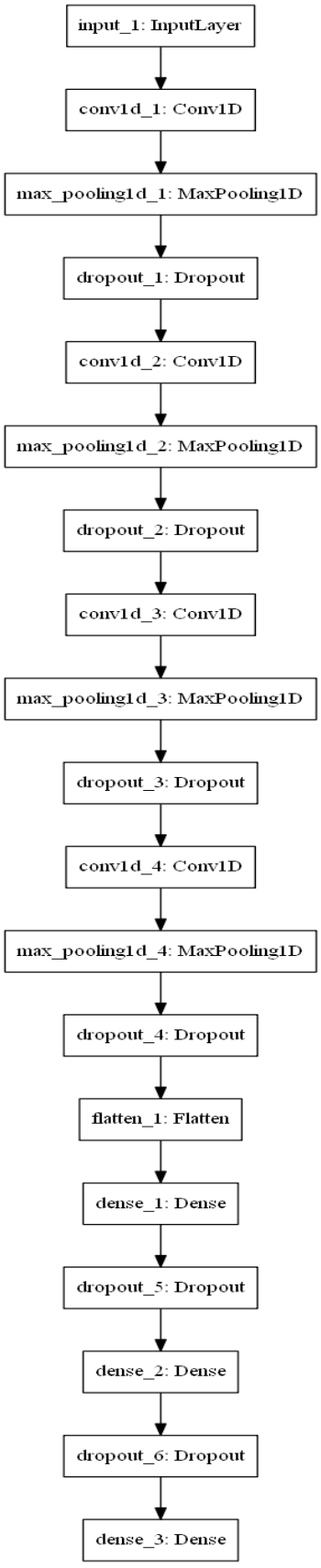

Figure 2: shows a proposed model plot.

4. ACCURACY

According to the training and validation statistics, the second model displays the accuracy of 93.69% and for the third model, it is 95.18% since there are only two words to predict. Note that the same configuration was used to train the models.

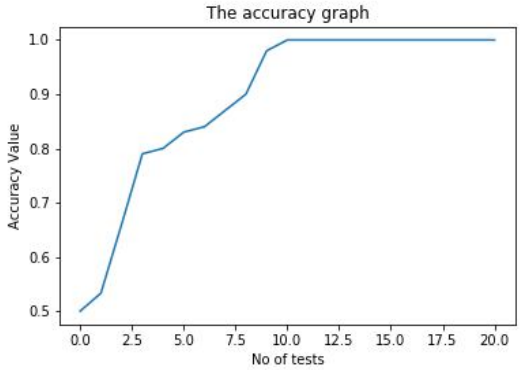

Figure 3: shows the accuracy obtained by application. The x-axis shows the abundance of tests made on the application. The y-axis indicates the accuracy of the application. The accuracy of the application is the average of the prediction of the model(1 for correct prediction and 0 for the wrong prediction).

4.1 The accuracy graph of testing

The proposed application is tested in different environments and the accuracy obtained from the model is approximately 89 percent. This application was tested with different types of microphones such as the default Laptop Microphone, a Professional Conference Microphone, a normal Bluetooth Microphone. The application was tested in a crowded environment to bring in the question of the practical usage of the application. The above graph is a sample of the tests. Epochs are the time step which is incremented after every time it has gone through each and every sample in the training dataset.

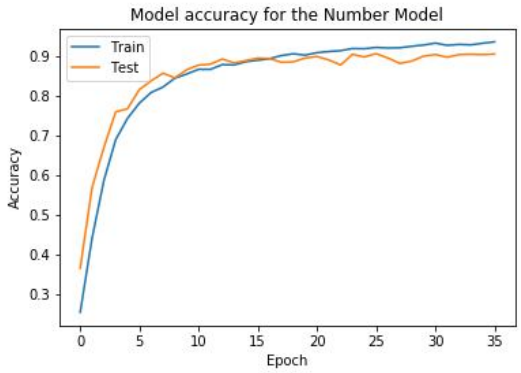

Figure 4: shows the accuracy graph of the second model(number model) during the training and validation process. The blue line indicates the accuracy of the training dataset, it shows an increase when plotted against epochs The orange line indicates the accuracy of the test dataset/validation dataset .it also shows an increase when plotted against the epochs.

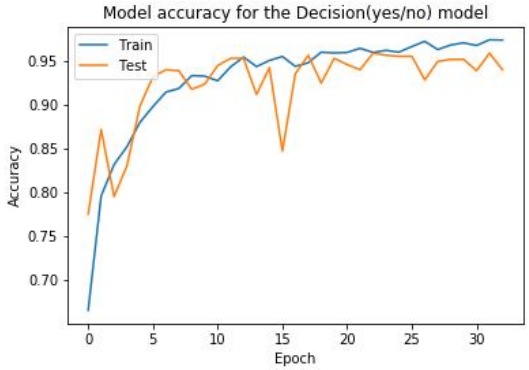

Figure 5: shows the accuracy graph of the third(decision model) during the training and validation process. The x-axis depicts the number of epochs and the y-axis shows the accuracy for each epoch. We have the blue line that shows the accuracy for the training dataset and the orange line shows the accuracy for the validation which is done after each epoch. The validation had a decrement in the accuracy since the validation loss was more. However, the training accuracy did not have this effect, due to the huge variation in the size of the training and validation datasets. At the final epoch, the difference between the accuracies of the training and validation is less as compared to the previous epochs. Since the idea is to save the model which has the best accuracy in the validation, we have considered the accuracy of the validation whose accuracy is the maximum in this process.

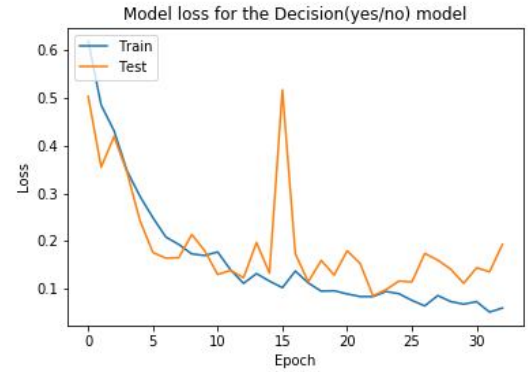

Figure 6: shows a Plot explaining the loss of values during training and validation for prediction of decision (yes / no). The x-axis is the epochs and the y-axis shows the loss for each epoch. The blue line shows the loss during the training and similarly, the orange line shows the loss during validation after each epoch. At the 15th epoch, since there is a drop in the accuracy, it, therefore, implies that the loss for that validation was more. Gradually, this has come down to a good extent at the final epoch.

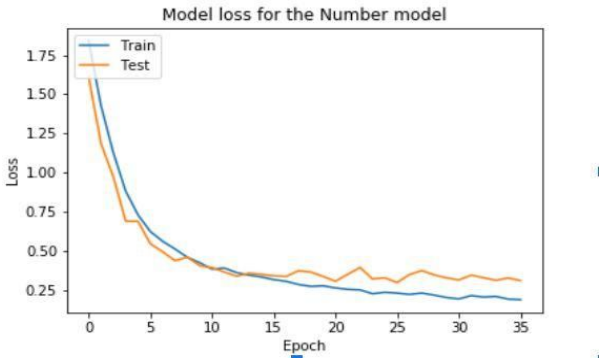

Figure 7: shows a Plot explaining the loss of values during training and validation for prediction of number. The x-axis has the epochs and the y-axis shows the loss for each epoch. The blue line shows the loss during the training and the orange line indicates the loss during validation after each epoch.

5. PERFORMANCE

The application shows extreme promise and delivers excellent results provided the following are satisfied:

1. The system must be having good configuration
2. It must be connected to the internet. Although we don't recommend super-speed broadband networks for this application. A decent network with a good speed is enough for giving good results.
3. From our tests, it is recommended that a concentrated input source like a separate microphone is used, as we see them in the ticket counters in multiplexes, and in some food counters too.

According to the observations made, the application works well in a crowded environment, and on average a customer can order an item within 15 seconds when the application is running with recommended expectations as stated above.

Some failure cases arose mainly due to poor microphone or too much disturbance in the environment such that the microphone was unable to capture our voice properly and hence the recognition fails.

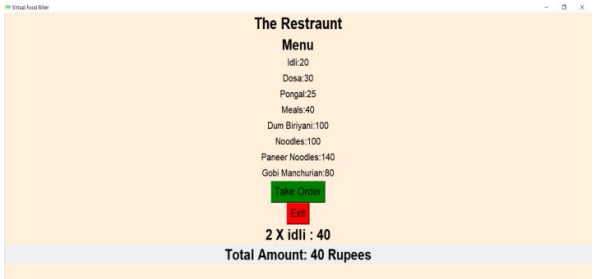

Figure 8: shows a sample GUI explaining how the system works.

6. CONCLUSION

In the current scenario, most fast food services, either depend on manual billing(i.e., billing with the help of a person) or kiosk billing wherein the customers use screens to order their food, which is more like ordering the food online. Billing with the help of an assistant which is actually virtual can improve the customer experience, allowing him/her to order food easily by speaking. In this work presented, the application aims to achieve billing through speech which can make the job of the customer easier and also make it secure for the service provider as the customer can only give his/her speech as the input and cannot access any other part of this billing system, keeping the transaction completely secure. After various tests on the application with different scenarios, an internal microphone fails to recognize if there is a lot of noise present in the background, hence it is highly recommended that a proper external microphone is used, and an internet connection with a decent speed is suggested for a better performance of the application.

REFERENCES

1. Aleksic, Petar S., Mohammadreza Ghodsi, Assaf Hurwitz Michaely, Cyril Allauzen, Keith B. Hall, Brian Roark, David Rybach and Pedro J. Moreno. "**Bringing contextual information to google speech recognition.**" *INTERSPEECH* (2015).
2. M. Nakamura, K. Iwano and S. Furui. **Differences between acoustic characteristics of spontaneous and read a speech and their effects on speech recognition performance,** *Computer Speech and Language*, vol.22, no. 2, pp. 171-184, April 2008. https://doi.org/10.1016/j.csl.2007.07.003
3. C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani. **Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home** in *Proc. Interspeech*, Stockholm, 2017. https://doi.org/10.21437/Interspeech.2017-1510
4. J. Bhaskar, K. Sruthi and P. Nedungadi. **Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining,** *Procedia Computer Science*, vol. 46, pp. 635-643, 2015 https://doi.org/10.1016/j.procs.2015.02.112
5. O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider and N. A. Smith. **Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters** in *Proc. NAACL-HLT 2013*, Atlanta, Georgia, June 2013, pp. 380-390
6. S. More. **Machine Learning Techniques with IoT in Agriculture,** *International Journal of Advanced Trends in Computer Science and Engineering*, vol.8, no.3, pp. 742-747, May-June 2019. https://doi.org/10.30534/ijatcse/2019/63832019
7. L. Teng, K. Kuok & M. Imteaz, W. Y. Lai and K. Derrick. **Development of Whale Optimization Neural Network for Daily Water Level Forecasting,** *International Journal of Advanced Trends in Computer*

Science and Engineering, vol.8, no.3, pp. 354-762, May-June 2019.
https://doi.org/10.30534/ijatcse/2019/04832019

8. V. Bharat and N. Malik,. **Study of Detection of Various types of Cancers by using Deep Learning: A Survey.** *International Journal of Advanced Trends in Computer Science and Engineering*. vol. 8, no.4, pp. 1228-1233, August 2019.
https://doi.org/10.30534/ijatcse/2019/31842019