Conversational Interactive Voice Response System with few-shot intent classifier

Abhinav Parag Mishra[1], Rishabh Shukla[2], Manish Pandey[3], and Gyanendra Nath Dwivedi[4]
[1,2,3,4] *UG student, Raj Kumar Goel Institute of Technology/AKTU*

***Abstract* - Interactive Voice Response (IVR) systems have been helping businesses in providing support to their customers over a telephone network. Modern IVRs not only have speech recognition facility, but also have an AI-powered conversational system with an intent classifier. Such IVRs improve user experience but building such a system is a resource-intensive and labor-intensive process. Resources are required to fine-tune and train the intent classifier every time intent is added. The labor-intensive part of the process is creating a domain-specific labeled dataset for training the intent classifier. In this paper, we propose a smart IVR system with an intent classifier based on dual sentence encoders. The intent classifier used not only requires very few resources for training but also works better than models like fine-tuned BERT in few-shot setups (when examples per intent are less). The cost of building conversational IVR is brought down even more when such an intent classifier is built in conjunction with open-source technologies like Asterik (for voice server) and Festival (for text-to-speech conversion).**

***Index Terms* - Dual sentence encoder, Intent classifier, Interactive Voice Response (IVR), VoiceXML.**

I.INTRODUCTION

IVR systems are widely used by businesses for automating customer support. An IVR system is a telephony technology that automates the task of gathering information and the intent of the caller. This information is used to forward the caller to a customer support agent of the appropriate department. To gather the information, IVR system uses either DTMF signals from keypad presses or the caller's voice. Apart from an efficient call routing facility, IVRs can also provide self-service options. Self-service options enable customers to retrieve basic information, perform simple tasks and solve issues without any involvement of customer support agents. This helps in reducing operational costs and improving customer satisfaction. Out of top 50 companies in Fortune250 companies list, 53% [1] still use DTMF-only IVRs. In such IVRs, a pre-recorded voice prompts the caller to press a specific key for choosing the required department or getting a specific issue resolved. This implementation of IVR systems is easier and straightforward, but in some cases, it results in a poor user experience. For example, if menus are overcomplicated or the required option comes last in the menu, then an impatient user will become frustrated. With recent advancements in NLP (Natural Language Processing), IVRs nowadays come with Natural Language Understanding (NLU) capabilities. Such IVRs allow users to interact with the system using speech, thus giving a more natural user experience. Users can directly speak and convey what they want without having to go through a multi-level menu.

When it comes to response time, 90% [2] of consumers consider immediate response very important for customer service questions. Modern IVRs help in achieving immediate response but developing conversational IVR that is customized for a particular business domain is difficult. In this paper, we propose an IVR system that uses speech recognition for taking input from the caller and identifies the intent of the caller using USE+ConveRT model. Fine-tuned BERT model is commonly used in IVRs to achieve intent classification, but the process of fine-tuning is quite resource intensive. In few-shot scenarios, for example, where only 10 examples per intent are available, fine-tuning BERT might cause overfitting [3]. Dual Sentence Encoders (USE + ConveRT) not only outperform BERT in intent detection, but they are also compact and require very few resources. Since USE+ConveRT gives better performance [3] in few-shot scenarios as well, small businesses can easily implement the proposed IVR, even when manual annotation of data is required.

II. RELATED WORK

In [4] authors presented a voice-based mobile application for tackling the problem of medication and prescription errors in health care services. Users had to just dial an appropriate number for accessing the application for their prescription. The architecture was based on a 3-tier client and server model, wherein voice response was rendered to the phone interface using VoiceXML. Automatic speech recognition (ASR) and text-to-speech (TTS) systems present in the voice server made the voice-based interaction possible. The proposed system not only helped patients by providing correct prescriptions, but also the health care workers who would get patient's medical records in real-time.

In [5] authors proposed an intelligent e-Learning system that was aimed at helping visually impaired and dyslexic students. The voice user interface (VUI) in this system was developed using VoiceXML and for providing intelligent services, Case-Based Reasoning (CBR) AI paradigm was used. A cost-effective IVR system was proposed in [6] using open-source tools. While the architecture suggested in it was for DTMF-only IVRs, open-source tools like Asterik and Festival could definitely be employed while implementing the IVR system presented in this paper.

The efficient dual sentence encoders based intent classifier was proposed in the paper [3]. The authors introduced methods of intent detection using USE and ConverRT, which are pre-trained encoders. In the paper, they demonstrated how USE+ConveRT outperforms fixed BERT and fine-tuned BERT in intent detection. The research shows how easily and quickly the model could be trained on a single CPU. They also tested the model on 3 popular datasets: HWU64, CLINC150, and BANKING77. The benefits were more pronounced in few-shot setups, which is why we use USE+ConveRT as the intent classifier in the proposed architecture.

III. PROPOSED SYSTEM

Figure 1 illustrates the system architecture of the proposed IVR system. Users dial a particular number using a mobile phone or landline and the call reaches the voice gateway through Public Switched Telephone Network (PSTN). The caller is greeted with a welcome message and asked to speak what he/she wants. The voice from the caller is then processed by VoiceXML interpreter using Automatic Speech Recognition (ASR), which is part of the voice browser. The voice browser then transmits the gathered input from the caller to a web server using HTTP (if sensitive information is being transferred then HTTPS should be used). The web server then makes a system call to the trained intent classifier model which is present on the same machine. The output from the intent classifier suggests what the caller wants and necessary actions are taken. The action could either be replying to the voice server with some information or making API calls to internal application servers and then confirming the user through the voice server. The voice server receives a reply from the web server and converts it into speech using its text-to-speech (TTS) module. Thus, the user gets back a reply in form of speech and the conversation moves further.

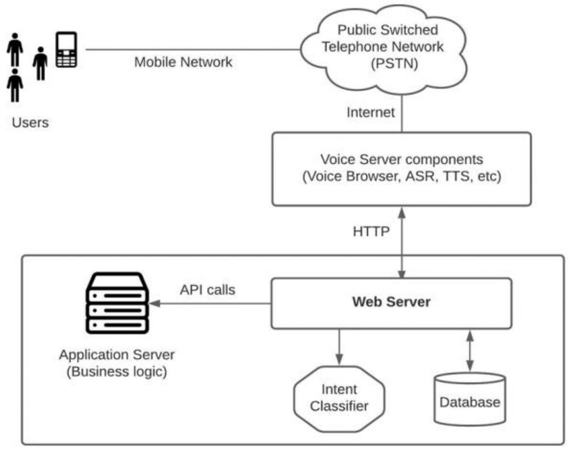
Users
Mobile Network
Public Switched Telephone Network (PSTN)
Internet
Voice Server components (Voice Browser, ASR, TTS, etc)
HTTP
API calls
Web Server
Application Server (Business logic)
Intent Classifier
Database

Fig. 1. Proposed IVR system architecture

The components of the architecture that an organisation has to setup are as follows:

A. Voice server components:

The main components of the voice server are voice browser, automatic speech recognition (ASR), and text-to-speech (TTS) module. VoiceXML is the technology that makes interaction with the web through telephone possible. A VoiceXML document is rendered by voice browser, which contains VoiceXML interpreter. VoiceXML interpreter makes calls to ASR and TTS to recognize voice and convert text to speech respectively. An organization could implement the voice server using Asterik. Asterik is a free to use (open source) technology, which cuts down the cost of implementation.

B. Web Server and Database:

The web server is responsible for communicating with the voice server to get user input and then give back a reply. The web server communicates with a database and an intent classifier to perform required tasks. The database stores general information that users might ask for, for example, information about a new offer, future product launches, etc. The database can also store useful caller information that could be helpful in improving user experience, for example, ticket id for an issue raised could be stored for future references. The intent classifier takes the caller's dialog as an input and returns what the corresponding intent is. With the help of this intent, the web server decides a particular action. The action could be either making an API call to the internal application server, for example, when the caller wants to change his profile information, or it could be replying with information that is stored in the database.

C. Intent classifier:

The intent classifier is designed using dual sentence encoders (USE+ConveRT). This design of intent classifier is especially helpful in new domains in which large datasets are not present and manual labeling is required. USE+ConveRT performs better [3] than its popular alternative BERT in case of low data availability (few-shot setups). Training it requires very less resource and it can be trained on a single CPU machine within few minutes. To use BERT in a new domain, fine-tuning is required which is very resource-intensive and time taking. To explore the working of USE+ConveRT model, readers are encouraged to go through the original paper [3] in which it was proposed. The authors have discussed the implementation of this model which should be referenced while creating the model.

As specified in the original paper, the dataset can be prepared with as low as 10 examples per intent. Once trained, the model is deployed on the same machine where the web server is running. The web server then communicates with the intent classifier through system calls. On Linux, a bash script could be executed with caller dialog as a command-line argument. The bash script (Batch script in case of Windows) will invoke the model, provide input to it and then return the output (intent) to the web server program to process it and proceed further.

IV. CONCLUSION

Interactive voice response systems are an excellent way of adding value to a business. Often IVRs are the first point of contact by consumers for support. DTMF-based IVRs are helpful on their own, but their shortcomings could be easily overcome by giving speech-based input facilities to the caller. When IVR is built using open-source technologies like Asterik and Festival, and the intent classifier is built using dual sentence encoders, then implementation cost is reduced significantly. The cost also comes down because the suggested intent classifier works great in few-shot scenarios as well (when only 10-30 examples per intent are available). Thus, implementing IVR for low capital business becomes easier.

REFERENCES

[1] Omni-channel customer engagement Infographic - The current state of the IVR, Nuance Communications, Inc [online]. Available: https: //www.nuance.com/omni-channel-customer-engagement/infographic/state-of-the-ivr-infographic

[2] John Dick - Live Chat Exposes a Fatal Flaw in Your Go-to-Market, HubSpot, sec. The evolution of communication: The slow march toward speed [online]. Available: https://blog.hubspot.com /sales/live-chat-go-to-market-flaw

[3] I.Casanueva, T.Temčinas D.Gerz, M.Henderson, and I.Vulić ― Efficient Intent Detection with Dual Sentence Encoders ,‖ Association for Computational Linguistics, vol. Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI [online], pp.38-45,July 2020. Available: https://www.aclweb.org /anthology/2020.nlp4convai-1.5

[4] N.A.Omoregbe and A.A.Azeta, - A Voice-based Mobile Prescription Application for Healthcare Services (VBMOPA) ,‖ International Journal of Electrical and Computer Sciences [online], 10 (2). pp. 73-78, December 2010. Available: http:// eprints.covenantuniversity.edu.ng/id/eprint/12

[5] A. A. Azeta, C. K. Ayo, A. A. Atayero and N. A. Ikhu-Omoregbe - A Case-Based Reasoning approach for speech-enabled e-Learning system, 2nd International Conference on Adaptive Science & Technology (ICAST), January 2009,

pp. 211-217, DOI: 10.1109/ICASTECH. 2009.5409721.

[6] Anil Kumar and S. Niranjan - Design, Development and Implementation of an Automated IVR System with feature-based TTS using Open-Source Tools, International Journal of Engineering Research & Technology, vol. 1 issue 3, May 2012.