Automatic Telephone Interview Survey System

Che-Wen Chen, Yao-Hsiang Cheng, Shu-Wei Chuang,
Pin-Yang Hsu, Shih-Pang Tseng and Jhing-Fa Wang

November 6, 2021

Automatic Telephone Interview Survey System

Che-Wen Chen
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
kfcmax300@gmail.com

Yao-Hsiang Cheng
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
shawn2200696@gmail.com

Shu-Wei Chuang
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
n26091576@gs.ncku.edu.tw

Pin-Yang Hsu
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
pinyang870606@gmail.com

Shih-Pang Tseng
School of Software and Big Data
Changzhou College of Information Technology
Changzhou, China
tsengshihpang@ccit.js.cn

Jhing-Fa Wang
Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan
wangjf@mail.ncku.edu.tw

***Abstract*—Nowadays, governments or enterprises often use questionnaires to investigate people's views on issues or products, and telephone interviews are one of the most commonly used methods for questionnaire surveys. The proposed system is mainly divided into four parts. The first part is pre-processing, according to the questionnaire to be interviewed to generate the data needed in the follow-up dialogue in advance. The second part is to build a VoIP soft phone, the purpose is to realize the function of making calls on the software side with multiple SIP lines at the same time through the SIP architecture. In addition, the system also connects the front-end softphone module with the back-end dialogue module through a dialogue interface to automatically conduct the telephone interview. The third part is multi-turn question answering dialogue. If the call is answered, the dialogue module will actively ask the respondent according to the questions in the questionnaire, and then through hybrid answer matching to determine the option matched by the respondent's answer. Finally, the dialogue module decides the question to be asked in the next turn to achieve the effect of multi-turn dialogue. The fourth part is call detail record and table generation, which is responsible for recording the call status and answering situation during the telephone interview. In addition, the system performs post-processing according to the telephone interview results of all lines after the execution, so that we can quickly understand the situation of the telephone interview survey.**

Keywords*—*Questionnaire survey, Telephone interview, VoIP, SIP, Multi-turn question answering

I. Introduction

Questionnaire survey is an important way to collect information or opinions. The type of questions for the questionnaire can be divided into two types [1], namely open-ended and closed-ended. Open-ended questions allow respondents to express their opinions freely. There are no pre-designed answers to choose from. Closed-ended questions have an appropriate number of options and allow the respondent to choose one of the options as the answer. In order to make statistical analysis after the telephone interview more conveniently, the telephone interview usually uses closed-ended questions for the survey, such as the work done by [2].

In recent years, due to the vigorous development of computer systems and information technology, a survey method using the Computer-Assisted Telephone Interview (CATI) system for telephone interviews have been further developed. The CATI system is a software system mainly used to assist telephone interviewers to conduct telephone interviews, which can be used with physical phones or phones developed by VoIP [3]. There are many different protocols for VoIP, among which the most widely used is the Session Initiation Protocol (SIP) established by the IETF (Internet Engineering Task Force) [4].

Since communication through dialogue can effectively shorten the distance between humans and bots, Spoken Dialogue System (SDS) [5] has become a hot research topic and application. SDS can be divided into task-oriented and chat-oriented. In general, task-oriented SDS [6] is more common and is used to deal with dialogue tasks in a specific field. For a telephone interview survey, the content of the dialogue will be determined through a pre-designed questionnaire. The conversation is then done through a multi-turn question answering dialogue between the telephone interviewer and the respondent. Such a highly repetitive operation gives us the idea of combining a question answering dialogue system with the telephone system to replace telephone interviewers for dialogue.

Therefore, we hope to combine the question answering dialogue system with telecommunications technology to automatically conduct telephone interviews. In this paper, we propose an automatic telephone interview survey system, the remaining organization is shown as follows. Section II introduces our proposed system architecture. In Section III, we show the experimental results. Section IV presents the conclusions of our system.

II. Proposed System

A. System Overview

The system architecture of the proposed approach is shown in Fig. 1, which has four parts: <1> Pre-processing, <2> Multi-line SIP Softphone, <3> Multi-turn Question Answering Dialogue, and <4> Call Detail Record and Table Generation. There is also a <5> Post-processing part which is the additional processing part of the system after the operation ends.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

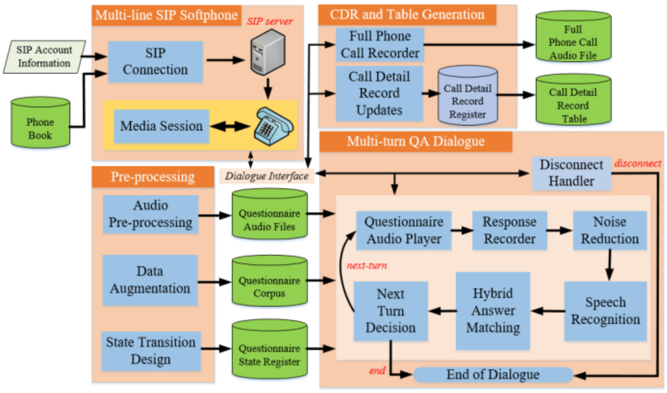

Fig. 1. System Architecture

B. Pre-processing Module

Before conducting a telephone interview survey, we need to prepare the questionnaire for the interview. This questionnaire consists of closed-ended questions. We use the pre-processing module to generate the data needed by the multi-turn question answering dialogue module in the system. These data include the question audio files of the questionnaire, the corpus of the questionnaire, and the jump logic of the questionnaire.

Audio pre-processing generates audio files for the questions in the questionnaire by pre-recording. Data augmentation is used to generate synonyms for the options in the question. We use the method of word2vec [7] to vectorize the word for word similarity calculation. Besides, we also consider the professional experience of the questionnaire designer. The example of data augmentation is shown in TABLE I. The state transition is designed to store the jump logic of the questionnaire. The pre-processing module establishes a process to generate necessary questionnaire data, even for different questionnaires.

TABLE I. THE EXAMPLE OF DATA AUGMENTATION

Q：請問您認為現任市長做得最好的是哪一項？請回答: (1)福利、(2)經濟、(3)文化、(4)治安、(5)衛生	
Original option	After data augmentation
福利	[福利, 社福, 托育, 保障]
經濟	[經濟, 財政, 貿易]
文化	[文化, 人文, 藝術, 教育]
治安	[治安, 秩序, 警務, 公共安全]
衛生	[衛生, 環境, 醫療]

C. Multi-line SIP Softphone Module

We use the multi-line SIP softphone module to achieve the function of making calls from the SIP softphone to telephones. In this module, we connect to the SIP server of the telecommunication service provider. At the beginning of the module for single-line case, the phone book and the SIP account are configured first. Next, the module dials the SIP outbound call and connects to the telephone for the establishment of a session. After the media session is established, the two parties can start the conversation.

However, the execution of a single line is not efficient enough for the system. Therefore, we use the method of running multiple processes at the same time by the operating system to achieve the effect of simultaneous multi-line telephone interview, which can be regarded as every process executes single-line telephone interview at the same time. As far as we observe the situation of the multi-line telephone interview, twenty is the maximum number of lines for the system to perform stable multi-line execution under our software and hardware environment. The flow diagram of multi-line SIP softphone module is shown in Fig. 2.

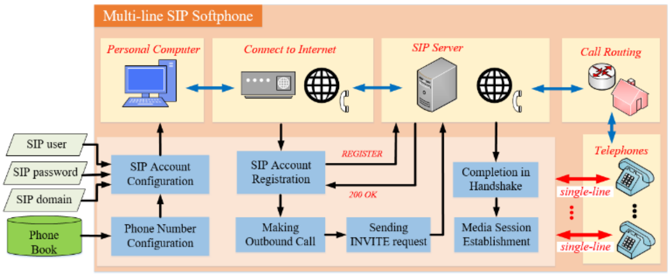

Fig. 2. The flow diagram of multi-line SIP softphone module

D. Multi-turn Question Answering Dialogue Module

After establishing the media path between the SIP softphone and the telephone, the multi-turn question answering dialogue module reads the questionnaire data generated by the pre-processing module to start the conversation through the dialogue interface. The disconnect handler handles the disconnection of the session to avoid audio transmission problems after the conversation starts. During the conversation, the questionnaire audio player actively plays the audio file of the question to ask the respondent. When the respondent starts to answer it after listening to the question. At this time, response recorder records the answer and automatically ends the recording. Finally, response recorder outputs the recording result as an audio file. This audio file is then converted into text after noise reduction and speech recognition by Google Speech API [8].

We match the options of the question based on the text result of respondent's answer. In addition to the option itself, the answer may also be a synonym for the option, so we need to match the option when the answer is different from the option but has the same meaning. We first delete some stop words, such as interjections or pronouns from the text result through stop word deletion after the respondent’s answer is converted into text. Then we perform string matching with the text result based on the options and its synonyms in the questionnaire corpus. The example of answer matching for synonym is shown in TABLE II.

TABLE II. THE EXAMPLE OF ANSWER MATCHING FOR SYNONYM

Q : 請問您對台南市的環境整潔滿意嗎？請回答: (1)很不滿意、(2)不滿意、(3)普通、(4)滿意、(5)很滿意		
Original Answer	After Deletion	After String Matching
很滿意	很滿意	**Matching Result:** match with "choice005"
啊我很滿意啦	很滿意	**Matching Result:** match with "choice005"
啊我非常滿意啦	非常滿意	**Matching Result:** match with "choice005" (by synonym)

Furthermore, because the respondent's answer is based on the options, the answer is usually very short. The word-level short answer may cause ASR to recognize the result as a near-homophone of the answer, which is a word that sounds similar to the answer. In this case, relying only on synonyms for answer matching is obviously not enough. However, since the questions are closed-ended questions, we can convert the text into Mandarin Phonetic Symbols [9]. Then through phonetic symbols processing and phonetic symbols matching to compare whether the respondent's answer (after ASR) and options are phonetically similar. In other words, we use the hybrid method of answer matching for synonym and near-homophone to determine whether the respondent's answer matches one of the options of the question. The example of answer matching for near-homophone is shown in TABLE III.

TABLE III. THE EXAMPLE OF ANSWER MATCHING FOR NEAR-HOMOPHONE

<table>
<tr><td colspan="3">Q：請問您認為現任市長做得最好的是哪一項？請回答:
(1)福利、(2)經濟、(3)文化、(4)治安、(5)衛生</td></tr>
<tr><td>Original Answer</td><td>ASR Result (Near-Homophone)</td><td>Phonetic Symbols Matching</td></tr>
<tr><td>經濟</td><td>星際</td><td>both are ["ㄧㄥ", "ㄧ"]
<b>Matching Result:</b>
match with "choice002"</td></tr>
</table>

After the end of hybrid answer matching, the next turn of question answering dialogue is determined by the next turn decision. We complete the transition to the next state through the matching result and the questionnaire state register generated in the pre-processing module. After the state transition is completed, when the state represents the end, the module will end the question answering dialogue, which is the end of the multi-turn dialogue between the two parties. If there is a jump in the transition of the state, the module will jump to ask that question according to the state result, if not, the module will continue to ask the next question.

E. Call Detail Record and Table Generation Module

During the telephone interview, the status of each call and the answers of the respondents need to be recorded for subsequent analysis. We update these call detail records at any time by grasping the situation of session state and conversation. The module then stores this information in the call detail record register. The module will wait until all the calls for the telephone interview are over and generate a table that record this information. In addition, the full phone call recorder also records the content of the conversation between the two parties.

F. Post-processing Module

Since the simultaneous multi-line telephone interview makes each line generate its own table result, we need to merge these table results and filter out the successful interview cases. Then we can perform statistical analysis of the answers to the questionnaire based on these results or inquire the call status of each line to quickly understand the results of the telephone interview. The flow diagram of post-processing module is shown in Fig. 3.

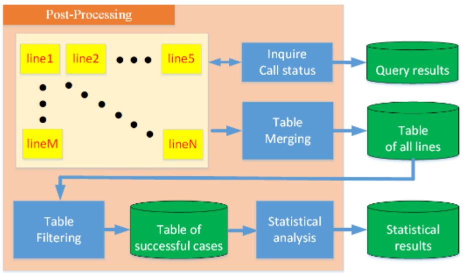

Fig. 3. The flow diagram of post-processing module

III. Experimental Results

A. Experiment of Multi-turn QA Dialogue for Proposed System

In order to understand the effect of the system in conducting telephone interviews, we conduct tests under different numbers of lines. Because there will be cases where the call is not answered or no one responds to the questions after making a call, we define the "Valid Call Rate", which represents the percentage of calls that are answered and someone responds (including accepting or not accepting interviews). The formula is shown as follows:

$$\frac{number\ of\ valid\ calls}{number\ of\ total\ calls} \times 100\% \tag{1}$$

The focus of this system is whether the dialogue is completed after the call is answered and multi-turn question answering dialogue have been started with the respondent. This method can objectively evaluate the dialogue effect of our system. We define the "Dialogue Completion Rate" for multi-turn question answering, which represents the proportion of respondents who have actually completed multi-turn question answering dialogue with the system, that is, the entire questionnaire interview survey has been completed. The formula is shown as follows:

$$\frac{number\ of\ calls\ completing\ multi-turn\ QA\ dialogue}{number\ of\ calls\ \ answering\ more\ than\ one\ turn} \times 100\% \tag{2}$$

Although the focus of the system is the dialogue effect, we still define the "Total Response Rate" for the whole telephone interview survey. The formula is shown as follows:

$$\frac{number\ of\ calls\ completing\ multi-turn\ QA\ dialogue}{number\ of\ total\ calls} \times 100\% \tag{3}$$

For this experiment, we conducted the experiment with a questionnaire consisting of 10 closed-ended questions for multi-turn QA dialogue. In addition, we also tested the "Dialogue Completion Rate" with different number of lines. The experimental results are shown in TABLE IV.

TABLE IV. THE EXPERIMENTAL RESULTS OF MULTI-TURN QA DIALOGUE

	10 lines	15 lines	20 lines
Calls *(per line)*	100	200	200
Calls *(total)*	1000	3000	4000
Valid Call Rate	11.3%	10.2%	12%
Total Response Rate	3.2%	2.0%	2.4%
Dialogue Completion Rate	88.89%	83.33%	85.84%

The results show that as long as the respondent is willing to conduct multi-turn question answering dialogue with our system, there is a high probability that the entire questionnaire interview survey can be completed.

B. Experiment of MOS for Proposed System

To understand the practicality, fluency and user experience of our proposed system, we use Mean Opinion Score (MOS) to evaluate the subjective assessment of users. MOS is typically expressed as a single rational number in the range of 1 to 5, where 1 is regarded as the lowest perceived quality, and 5 is regarded as the highest perceived quality.

The experiment of MOS was carried out with thirteen participants to evaluate the overall performance of the proposed system. Every participant was asked to fulfill the MOS evaluation after interacting with our proposed system. We use three different aspects to evaluate our system and the experimental results are shown in TABLE V.

TABLE V. THE MOS SCORE FOR THE PROPOSED SYSTEM

Description	Average Score
The accuracy of the system	4.43
The adaptability of the system	4.38
The fluency of the system	4.68
	Average: 4.49

The average MOS score we get is 4.49, representing that our system can give users a good experience in practical applications.

IV. CONCLUSIONS

This paper proposes an automatic telephone interview survey system. This system connects the SIP softphone at the front end of the system and the multi-turn question answering dialogue module at the back end of the system through a dialogue interface. This method achieves the effect of automating the process of telephone interviews, and saves the labor cost originally required for telephone interviews by telephone interviewers. The system also improves the efficiency of telephone interview through multi-line execution. In addition, the system proposes hybrid answer matching for the respondents' answers to deal with answers other than options, or situations where ASR recognizes the answer as a near-homophone. The experimental results show that the system has a high and stable dialogue completion rate. Finally, the average MOS score of our system is 4.49, which means that the functions of the proposed system are practically acceptable.

REFERENCES

[1] A. M. Kazi and W. Khalid, "Questionnaire designing and validation," Journal of the Pakistan Medical Association, vol. 62, no. 5, p. 514, 2012.

[2] S. Brar et al., "Perinatal care for South Asian immigrant women and women born in Canada: telephone survey of users," Journal of Obstetrics and Gynaecology Canada, vol. 31, no. 8, pp. 708-716, 2009.

[3] S. Jalendry and S. Verma, "A detail review on voice over internet protocol (voip)," International Journal of Engineering Trends and Technology (IJETT), vol. 23, no. 4, pp. 161-166, 2015.

[4] J. Rosenberg et al., "RFC3261: SIP: session initiation protocol," ed: RFC Editor, 2002.

[5] R. López-Cózar, Z. Callejas, D. Griol, and J. F. Quesada, "Review of spoken dialogue systems," Loquens, vol. 1, no. 2, p. 012, 2014.

[6] Z. Wei et al., "Task-oriented dialogue system for automatic diagnosis," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 201-207.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[8] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx)," Int. J. Eng. Res. Appl, vol. 7, no. 03, pp. 20-24, 2017.

[9] 注音符號 -Wikipedia. Available: https://zh.wikipedia.org/wiki/%E6%B3%A8%E9%9F%B3%E7%AC%A6%E8%99%9F.