INTERSPEECH 2021: Show & Tell Contribution
30 August – 3 September, 2021, Brno, Czechia

Duplex Conversation In Outbound Agent System

Chunxiang Jin[1], *Minghui Yang*[2], *Zujie Wen*[3]

Ant Group
HangZhou, China
{chunxiang.jcx, minghui.ymh, zujie.wzj}@antgroup.com

Abstract

Intelligent outbound is a popular way to contact customers. The traditional outbound agents communicate with users in a simplex way. The user and the agent cannot speak at the same time, and the user cannot actively interrupt the conversation while the agent is playing audio generated by TTS. The traditional solution is based on the output of the VAD module, once the user voice is detected, the agent will immediately stop talking. However, the user sometimes expresses the short answer at will, not to interrupt the agent, and it will cause the agent to be frequently interrupted. In addition, when users say named entity nouns(numbers, locations, company names, etc), their speech speed is slow and the pause time between words is longer, and they may be interrupted by the agent unreasonably. We propose a method to identify user's interruption requests and discontinuous expressions by analyzing the semantic information of the user's utterance. As a result, fluency of the dialogue is improved.

Index Terms: duplex conversation, automatic speech recognition, dialogue flow, interruption requests, discontinuous expressions, natural language understanding, voice activity detection

1. Introduction

In traditional simplex conversation process of the outbound agent, the voice stream of the user is converted to text with the ASR service, and then the user's utterance is passed to the dialogue agent as a parameter for Natural Language Understanding(NLU) module. The agent recognizes the user's intention according to the text and selects proper dialogue strategies and scripts, and broadcast the scripts with the TTS service finally, as show in Fig. 1.

Since outbound calls are initiated by agents, users' willingness for conversation is much lower than online text agents, and customers' tolerance for bad conversation flow is low. In the scene for simplex conversation, intermittent conversation proportion accounts for about 7% to 1%), the user may directly hang up the phone, which cause failure of communication task.

In the traditional simplex conversation, the agent cannot respond to the user's interrupt request when the agent is talking. Because the dialogue system rely on VAD[1] to detect whether the user is talking. One of the conversation policy is that the agent stops talking and waits for the user to finish if positive signal is detected by the VAD module. This will lead to intermittent dialog flow. For example, the user just echoes without any semantic meaning, this will cause the agent to be frequently interrupted and affect the fluency of the dialogue. Some common practices in natural conversations have been studied to generate natural conversation [2]

Another problem is the user's discontinuous expression. When the user speaks entity nouns, he may think while speaking, and the interval between words will become longer, which cause the agent to start talking before the user has finished speaking. So the user feels offended and the conversation is unnatural.

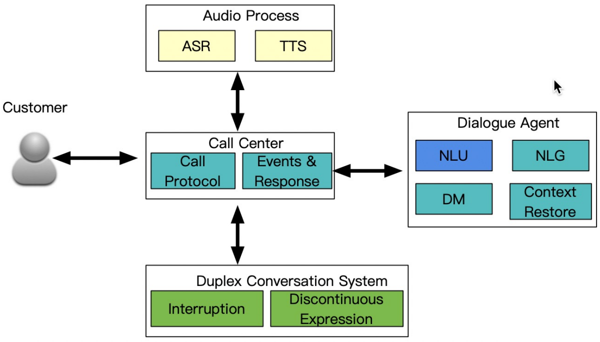

Figure 1: *Dialogue Agent System.*

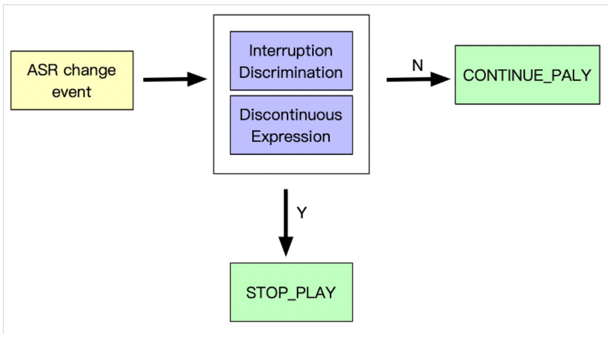

Figure 2: *Framework of Duplex Conversation*

We propose a natural duplex conversation scheme based on semantic understanding, which is mainly composed of full-duplex interaction state machine, interruption discrimination, and discontinuous expression detection modules to solve the above problems, improve the fluency of the dialogue, and increase the completion rate of the dialogue.

2. System Description

We propose a way to use the events in the ASR output results to convert the intermittent ASR into a streaming ASR, and use the event mechanism of ASR and TTS to build a natural duplex interactive system(show in Fig. 2).

2.1. Duplex Conversation State Machine

We abstracted out six major types of events from ASR and TTS services(Table 1) and responses from dialogue agent system (shown in Table 2). The duplex conversation state machine judges whether the interruption should be performed according to the speaking state of user's utterance in the IVR system. According to the ASR segment every 300ms, the previous recognition results are superimposed, and the text is sent to the interruption detection module to determine whether the interruption should be executed at the current moment.

Table 1: *ASR and TTS events*

event description	event code
agent starts to play	PRE_PLAYBACK_SATRT
agent finishes playing	PLAYBACK_END
silent of user timeout	SILENT_TIMEOUT
ASR of user starts	ASR_START
ASR of user changed	ASR_CHANGED
user finish speaking(VAD)	ASR_RESULT

Table 2: *response from dialogue agent*

response description	response code
start ASR recognize	START_RECOGNIZE
stop ASR recognize	STOP_RECOGNIZE
start timer for user	START_INPUT_TIMER
start playing TTS	START_PLAYBACK
stop playing TTS	STOP_PLAYBACK
call agent TTS	CALL_NLU

2.2. Interruption Discrimination

The traditional simplex interruption strategy is based on the VAD module to determine whether the user is speaking. The agent will be interrupted when the user says some meaningless answers like 'hmm'.

We propose a method that uses the streaming ASR result to continuously detect the semantic information of the user's utterance to determine whether to perform interruption. The timing of triggering the interruption is not as early as possible, because the first few words of the user's utterance is often meaningless without any specific intention, and the output result of the downstream NLU model is unknown. If the interruption is triggered at this time, there is not any proper script for the agent. Therefore, the strategy we adopt is as follows: from the first word that the user says, the interruption discrimination module is called continuously until the content that the user says has clear semantics, then the dialogue agent executes the interruption.

The interruption model is essentially a binary text classification task, which determines whether the user's utterance is meaningful or not. The method of constructing corpus is to sample user scripts in conversation logs. From the first word to the last word clause, NLU model is called to identify user intention in turn. If it is classified to a meaningful tag, it is a positive sample, otherwise it is a negative sample(shown in Fig. 3). Then a text classification model with the pretrained BERT[3] is trained from the corpus as the interruption discrimination module.

Because of the streaming input, the call frequency is very high. In order to ensure the real-time response of the agent, it is necessary to distill the model, with soft target and hard target. The accuracy of the interruption discrimination model is 90%, and the reasonable rate of interruption is 76%. After excluding the background noise samples, the reasonable rate of interruption is 86%, which is increased by 10%.

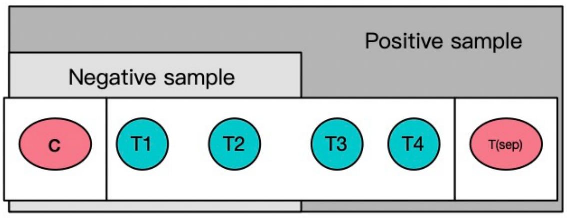

Figure 3: *corpus of interruption model*

2.3. Discontinuous Expression Detection

When the user speaks entity nouns, he may think while speaking, and the interval between words will become longer. This may cause the VAD module fail to recognize the complete user's utterance. Therefore, we propose a discontinuous expression detection module to judge whether the content of the current user is complete. We select the clauses with complete entity nouns as positive samples, and take the sentences with partial entity nouns as negative samples. We use these samples to train the binary classification model[3] as discontinuous expression detection module.

2.4. Dialogue Agent Performance Improvement

With the duplex conversation, the dialogue agent gets more complete user utterance, so the intention recognition rate is increased by 5.2% and the dialogue completion rate is increased by 2.3%(Table. 3)

Table 3: *Conversation Performance*

metrics	turn on	turn off	improve
intention recognition rate	67.1%	70.6%	+5.2%
dialogue completion rate	87.5%	89.5%	+2.3%

3. Conclusions and Future Work

With the use of semantic information to identify user's interruption requests and discontinuous expressions, we can improve the fluency of the conversation and dialogue completion rate, and ultimately improve the achievement of the goal of the conversation. In the next step, we will introduce the speech signal into the interruption recognition model to detect the background noise, separate the human voice from the background noise, and improve the accuracy of interruption.

4. References

[1] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *2009 17th European Signal Processing Conference*. IEEE, 2009, pp. 2549–2553.

[2] Y. Leviathan and Y. Matias, "Google duplex: an ai system for accomplishing real-world tasks over the phone," 2018.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.