Silence Detection and the Effect in a Spoken Dialogue System

Yuto Maetosa[1], Ryo Saegusa[2, *]

1. Graduate School of Informatics and Engineering, the University of Electro-Communications, Japan
2. Graduate School of Engineering, Kanagawa Institute of Technology, Japan

*corresponding author: ryo.saegusa@syblab.org

Keywords: Silence Detection, Spoken Dialogue System, Voice Processing

1. Introduction

Conversation requires the roles of the speaker and listener who converse according to the rules of turn taking in speech(1). Usually, the listener starts speaking after the listener recognizes that the present speaker has finished speaking ((i) turn taking). However, the listener prefers to wait for the present speaker to speak again ((ii) turn keeping), if the pause in speaking is short.

Many spoken dialogue systems incorporate only the function (i), while a few systems incorporate the function (ii) as a backup response(2). However, the silence detection of the previous studies ignores the turn keeping, and therefore, the response manner gets unnatural and looks different from that of humans. This paper proposes a spoken dialogue system that incorporates silence detection to allow turn taking and turn keeping, and evaluates the function in the conversation with users.

2. Framework of Silence in a Dialog

Fig. 1 illustrates a framework of silence in a dialog. In the proposed system, the moment when the system has finished speaking is regarded as the starting point of the process. The time duration of user's silence from the starting point to the moment when the system starts speaking again is denoted as "the mute" of the user. The duration from the starting point to the moment when the user starts to speak is denoted as "the pause before speaking". The time duration of a temporal pause in the user's speaking is denoted as "the pause in speaking".

The silence is modeled based on statistical distributions. The waiting time for "the mute" is determined by a normal distribution with the mean as 2.8 and deviation as 6.14. When the waiting time is over, the system interrupts the voice detection and expresses the user's response with the symbol "..." in the system buffer, where the number of the symbols to be printed is set at random. The waiting time of "the pause before speaking" is determined by a uniform distribution with the range from 1[s] to 4[s], and if the waiting time is over, two symbols of "..." are added into the buffer. The pause shorter than 0.5[s] in the user's speaking is detected as "the pause in speaking". If the pause is longer, the system takes turn of speaking from the user.

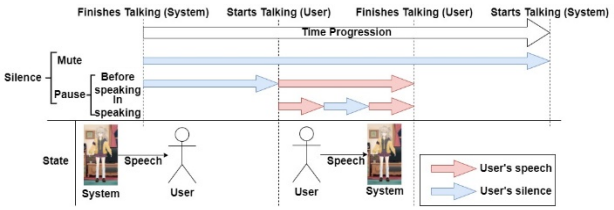


Fig. 1. Framework of silence in a dialog

3. Response for Speaking and Silence

The responses of the proposed system is based on the preset scenarios and generative contexts (Fig. 2). After the user's mute, the system responds as follows: In the first time, the system picks up a sentence from the preset scenario texts and use it. In the second time and thereafter, the system generates a sentence that continues from the previous system's speech. In the other case of silences

such as “the pause before/in speaking,” the system responds normally to what the user has spoken with variations generated by random modification.

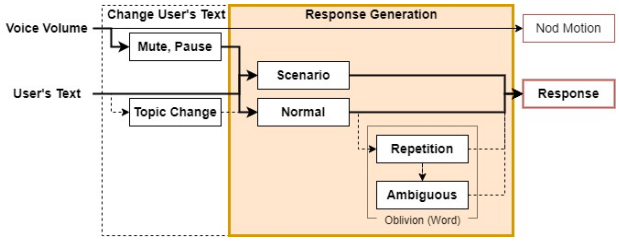


Fig. 2. Response model for speaking and silence

4. Experiments

The experiments of conversation were conducted with two conditions regarding the treatment of silence. In condition A, the system ignored silence and waited for the users to speak, while in condition B, the system detected silence and dealt with it. After the trials of conversations, the subjects of the experiments were to answer for a five-point questionnaire. The subjects were seven mixed genders between the age of about 10 to 40.

Tab. 1 shows results of the questionnaire, where values indicate average scores of the items. Comparing condition B to condition A, “system response” and “continuity (of the whole conversation)” are about 0.2 higher. The result seems to suggest that the system's silence detection reduces discomfort of silence in the flow of conversation, though the reduction rate is small. This is due to the fact that the silence detection is limited to temporal features. Multi-modal features such as gestures and face expressions are expected to be incorporated.

5. Conclusion

A spoken dialogue system based on silence detection was proposed and the effect for the man-machine conversation was evaluated. The experimental results suggested that the silence detection on mute and pause, and corresponding responses contributed to reduce discomfort in conversing with the system. On the other hand, the proposed silence detection is limited to the temporal feature based on statistical distributions. In the future, the implementation of silence detection and response will be improved to allow the contexts of past conversations and emotions.

Question Items	A	B
System Response	3.57	3.71
Naturalness	3.43	3.43
Connection	3.71	4.14
Friendly	4.43	4
Enjoyment	3.86	3.43
Continuity	4	4
Reaction	4.57	4.43
Ease of Speech	4	3.71
Humanity	3.57	3.14

Tab. 1. Average scores in conversations

Reference

(1) Harvey Sacks, et al., “A simplest systematics for the organization of turn taking for conversation,” Studies in the organization of conversational interaction, 7-55, 1978.

(2) Koji Inoue, et al., “An Attentive Listening System with Android ERICA: Comparison of Autonomous and WOZ Interactions,” Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), 118–127, 2020.

Biography



Yuto Maetosa is a master course student at the University of Electro-Communications. His research interest includes human-computer interaction.



Ryo Saegusa is an associate professor at Kanagawa Institute of Technology. His research interest includes symbiotic interaction between human and machines towards coevolution.