Interspeech 2022
18-22 September 2022, Incheon, Korea

Confusion Detection for Adaptive Conversational Strategies of An Oral Proficiency Assessment Interview Agent

Mao Saeki[1], *Kotoka Miyagi*[1], *Shinya Fujie*[2], *Shungo Suzuki*[1], *Tetsuji Ogawa*[1], *Tetsunori Kobayashi*[1], *Yoichi Matsuyama*[1]

[1]Waseda University, Japan
[2]Chiba Institute of Technology, Japan
{saeki, miyagi, fujie, ssuzuki, ogawa, matsuyama}@pcl.cs.waseda.ac.jp

Abstract

In this study, we present a model to detect user confusion in an online interview dialogue using conversational agents. Conversational agents have gained attention for reliable assessment of language learners' oral skills in interviews. Learners often face confusion, where they fail to understand what the system has said, and may end up unable to respond, leading to a conversational breakdown. It is thus crucial for the system to detect such a state and keep the interview going forward by repeating or rephrasing the previous system utterance. To this end, we first collected a dataset of user confusion using a psycholinguistic experimental approach and identified seven multimodal signs of confusion, some of which were unique to an online conversation. With the corresponding features, we trained a classification model of user confusion. An ablation study showed that the features related to self-talk and gaze direction were most predictive. We discuss how this model can assist a conversational agent to detect and resolve user confusion in real-time.
Index Terms: conversational agents, oral proficiency interview, computational paralinguistics, confusion detection

1. Introduction

Assessment is a crucial step in language learning; however, assessment of oral proficiency currently has relied heavily on human-led interviews, which is costly and possibly biased without extensive rater training [1]. In recent years, conversational agents have gained attention for delivering a low-cost and reliable speaking test [2, 3]. One key challenge for such agents is to adaptively change their conversational strategies in response to the users' affective state. Especially, users with low oral proficiency will often fail to understand the system utterance and become confused.

When confused, the user might explicitly request help, which can be detected relatively easily using methods such as pattern matching. However, in many cases, users would become lost in thought or wait for the system to assist them, failing to give a verbal response. If the user is left in such a state for too long, the conversation might break down, or users may lose engagement, both of which are critical for an assessment. A simple timeout strategy is not viable, as it may cut off users who are not confused but simply formulating what to say. Confused users tend to show signs such as producing fillers and frowning. Therefore, it is essential for the system to promptly detect these multimodal signs and give assistance. Previous studies have attempted to detect user confusion in task-oriented [4, 5] or chat-oriented dialogue with language learners [6]. Given our work focuses on the confusion in a chat-oriented online conversation, the difference in conversational settings may question the applicability of previous findings. For instance, since there are no task-related objects present that the user may stare or point at, signs of confusion can be delivered differently. Meanwhile, online conversation provides different causes of confusion, such as audio loss due to a bad network connection. Additionally, little has been examined about how the models proposed in previous studies can be applied in dialogue systems to resolve confusion in real-time.

In this study, we present a model for predicting the confusion of language learners in an online interview dialogue. One of the main challenges is the low availability of data. Confusion does not frequently occur in a real conversation, even with language learners. Moreover, given that the signs of confusion can significantly vary across users, even with a large amount of interactional data, it is not feasible to observe and learn valuable features end-to-end. Considering these characteristics of confusion and its signs, previous works have attempted to identify relevant signals and device predictive features [7, 8]. To better understand the phenomenon of confusion and its realization in conversation, we adopted a psycholinguistic approach to elicit learner confusion through linguistically manipulated interview questions (e.g., pseudo-words, silence insertion) [6]. Using the dataset of confusion, we identified characteristic signs of confusion and crafted feature extractors. We then trained a classification model to detect confusion and conducted an ablation study to identify the features with the optimal level of predictability. Lastly, we conclude our paper by discussing how this model can assist a conversational agent in detecting and resolving user confusion in real-time.

2. Related work

Confusion has been studied concerning the application of tutoring systems. Features in various modalities have been used to predict confusion, such as prosody [9], gaze [10], facial expression [11], and head movement [5]. Confusion has also been examined in task-oriented dialogues [4], where the user is required to solve some problems through the interaction with an interlocutor. However, confusion in these works occurs from various difficulties in the task, such as requiring a novel perspective or contradicting information being present. Signs of confusion subsequently could have differed from our setting, that is, a more chat-oriented dialogue, where there are no task-related objects the user may point or gaze at.

In most related work, Cumbal et al. [6] detected confusion among language learners in a conversation practice with a social robot, controlled by a WoZ operator. However, the applicability of their model to real-time confusion detection should be carefully discussed because the detection of confusion takes place

10.21437/Interspeech.2022-10075

when the operator selected the next system utterance, which is not obtainable in reality. From the perspective of the current study, another difference lies in the mode of conversation. Cumbal et al.'s model was trained based on the face-to-face interaction data, whereas our target context is an online conversation.

Taken together, we focus on identifying characteristic signs of confusion in an online interview conversation and handcraft features so that detection can be achieved using a small amount of data. We also show how this model can be used to detect and resolve confusion in real-time.

3. Data collection

3.1. Confusion elicitation procedure

As confusion does not occur frequently, we decided to efficiently elicit confusion by expanding on an experiment design by [6]. In the original experiment by Cumbal et al., a social robot asked various questions to the user. Some of the questions were manipulated by increasing lexical complexity, and increasing the speaking rate of the text-to-speech (TTS). These manipulations, however, do not cover other possible causes of confusion.

According to Levelt [12], speech comprehension entails different processes, and they proceed in the following order: speech sound recognition, spoken word recognition, retrieval of word meanings, and sentence parsing and establishing the mental representation of the text. From the perspective of speech comprehension mechanisms, the aforementioned manipulation in Cumbal et al.'s work can only enhance the demands on the second and third processes of speech comprehension. To better understand learner confusion and its possible signs comprehensively, it is essential to elicit confusion, shedding light on the other processes of speech comprehension. More specifically. it can be hypothesized that the reaction to the breakdown in the course of speech comprehension can be different according to which processes experience breakdown. Consequently, the system may be able to detect confusion regarding different observable cues. For example, breakdowns in the recognition of words can be better dealt with by repeating the previous question slower, whereas breakdowns in the parsing of sentences can be addressed by rephrasing with a simpler sentence structure.

Therefore, in our data collection, we added two additional manipulations, focusing on the processes of speech recognition and sentence parsing. The first manipulation was done by silencing some part of the system utterance, resembling the situation in an online conversation where the speaker's utterances can be silent due to the poor internet connection. The second was operationalized by the grammatical complexity of utterances, which was achieved by increasing the sentence length in words and adding subordinate clauses without changing the meaning of the utterances. We also slightly adapted the original manipulation of lexical complexity by inserting pseudowords, instead of using existing infrequent lexical items, so that the user would be likely to experience breakdowns in the lexical processes (i.e., spoken word recognition, lexical retrieval). A list of the cause of confusion and its elicitation method is shown in Table 1. Note that we did not include the manipulation based on the final process of establishing the mental representation. This process is achieved regarding the user's background knowledge, meaning that the demands on this process cannot be manipulated systematically.

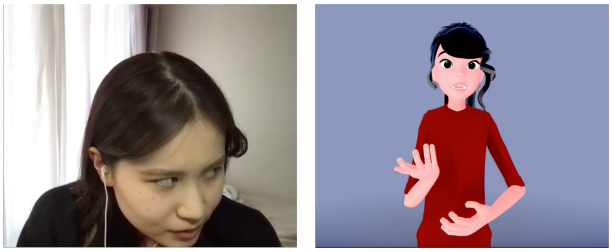

Figure 1: *A user (left) showing signs of confusion by averting gaze from screen and self-talking when the agent (right) asked a question.*

3.2. WoZ data collection

Forty-seven Japanese learners of English participated in the data collection. All participants were university students at a Japanese private university, with varying English oral proficiency.

With the aforementioned manipulations, we prepared a scenario for an online interview for language assessment. We used the InteLLA virtual agent[13] and controlled it in a Wizard-of-Oz (WoZ) style, where an operator selects the system utterance from a list of possible actions. The agent was able to repeat or rephrase the utterance to assist users in the case of confusion. The interview was designed to be completed in approximately six minutes. Four of the system utterances were swapped with each type of manipulation. The manipulation was spread throughout the conversation so that participants did not notice that the whole interview was set up to elicit confusion, and lose motivation.

Users were questioned on multiple topics, such as "favorite season" and "pros and cons of social media". Users joined the interview with the virtual agent in a video chat setting via their personal devices, and the entire conversation was video-recorded. Figure 1 shows a screenshot of the recording with the user showing signs of confusion.

3.3. Annotation

Signs of confusion may appear while the system is still speaking. We therefore clipped five seconds of user recording, starting from two seconds before the end of system utterance as a single data sample. Videos were clipped every second in a sliding window to augment data samples until either the system or user took the turn. If the user began speaking within three seconds of the system's end of utterance, the system can simply wait without any consequence, therefore the video clip was removed from the subsequent data analysis.

Next, we labeled each data as either "confused" or "not-confused". Labels for each data were assigned automatically based on the following action of the user or system. If the user asked for clarifications, or if the system assisted the user through repetition or rephrasing, the data was labeled as confused. The remaining data were labeled as not-confused. A total of 372 confused, and 155 not-confused samples were obtained.

Initially, the confused data were categorized into four subcategories based on the four conditions of confusion elicitation. This is because in our precursor project where we collected interview data by human interviewers, the interviewers were able to detect the source of confusion and give appropriate assistance. Assistance included slowing the speed of deliv-

Cause of Confusion	Elicitation Method	Examples
Failed to recognize sound	Silencing part of utterance	"What is your -"
Failed to recognize word	Increasing speech rate	Doubling the speech rate of the synthesizer
Failed to retrieve word meaning	Mixing non-existing words	"Can you evice me your deningham?"
Failed to parse and sentence	Increasing grammatical complexity	"Could you tell me about your friends you feel closest to?"

Table 1: *Causes of confusion and its elicitation method.*

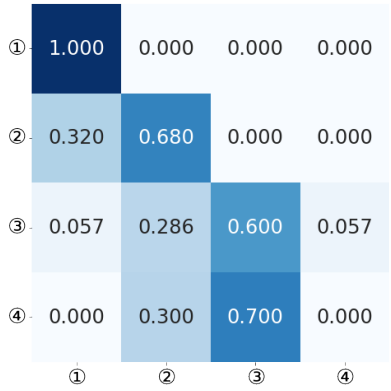

Figure 2: *Confusion matrix for predicting the cause of user confusion by a human annotator. The possible causes are ①Failed to recognize sound ②Failed to recognize word ③Failed to retrieve word meaning ④Failed to parse and sentence.*

ery and rephrasing to simpler sentences. Despite this fact, our preliminary result based on the current data revealed the difficulty of predicting the cause of confusion, as shown in Figure 2. Many cases were misclassified as confusion caused by an error in earlier stages of the comprehension process. For instance, the confusion that was elicited by the failure to recognize words (②) was commonly misclassified as the failure to recognize the sound (①). This is because the error in the previous step propagates to the next step. If a user fails to recognize the sound, they will certainly not be able to recognize the word, and it is hard to predict the initial cause of the error. Another key insight is that it is hard to discern the cause of confusion using *only* user image and audio information. In reality, human interviewers would consider additional information, such as the linguistic difficulty of their own utterance and the listener's oral proficiency and knowledge level, to deduce the cause of the error. Using such information is out of the scope of this study, therefore we resulted in the binary prediction of confusion, and leave the classification of causes for future studies.

4. Confusion detection

4.1. Feature extraction

Through analysis of the data samples, we identified seven characteristic signs of confusion, many of which often appeared in combination. We explain each sign and the feature extraction method below.

- **Increased blinking**: The frequency of blinking often increased. We extracted the activity of action unit (AU) 45 which corresponds to blinking.
- **Averting gaze from screen**: A common visual cue was to stare away from the screen. Since it is sufficient to know that the user is staring away from the screen, and not the direction, we calculated the absolute distance between the current gaze direction and the direction of the screen. The direction of the screen was estimated by calculating the average gaze direction during the first 10 seconds of the dialogue when the user is likely staring at the agent.
- **Rapid head movement**: Some users rapidly rotated their heads from left to right. We extracted absolute rotation of the head from the position when they are staring at the screen, similar to gaze direction.
- **Rapid eye movement**: Similar to the previous sign, but some users fixed their heads while their gaze moved from left to right. We used the same features as "averting gaze from screen".
- **Moving the face towards screen**: Some users moved their faces closer to the screen to inexplicitly indicate "I could not catch what you said, so please repeat". This is a feature rather unique to an online conversation. We calculated the head rotation and horizontal distance between the screen and head.
- **Silence**: One simple but common reaction was complete silence. we used voice activity detection (VAD) to detect the absence of user utterance.
- **Self-talk**: Self-talks are utterances that are not directed at the interlocutor, and users would often repeat words or parts of a sentence to try and make sense of what they heard. For human listeners it is easily distinguishable with utterance directed towards the interlocutor, however, there has been little work on it's automatic detection. The main difference is it being very quiet and uttered while the user is looking away from the screen. We calculated the relative loudness by dividing the current user loudness by the mean loudness of all previous utterances.

Voice activity was detected using the Python WebRTC-VAD library[1], and Head position, gaze direction, and AUs were extracted using OpenFace [14]. All features were extracted for every 40 milliseconds. Since some signs shared the same feature, a total of six features were extracted every frame; voice activity, relative loudness, AU 45 intensity, gaze distance from the screen, head rotation, and head distance from the screen.

4.2. Model Training and Evaluation

To capture the temporal dependencies of the features extracted in the previous section, we built a neural network classifier using Long Short-Term Memory (LSTM). We used 70% of the data for training, and the remaining 30% for evaluation. The not-confused class was oversampled to reduce the effect of data imbalance. To identify the most predictive features, we also performed an ablation study by excluding one feature at a time, as well s using only audio or visual features.

[1]https://github.com/wiseman/py-webrtcvad

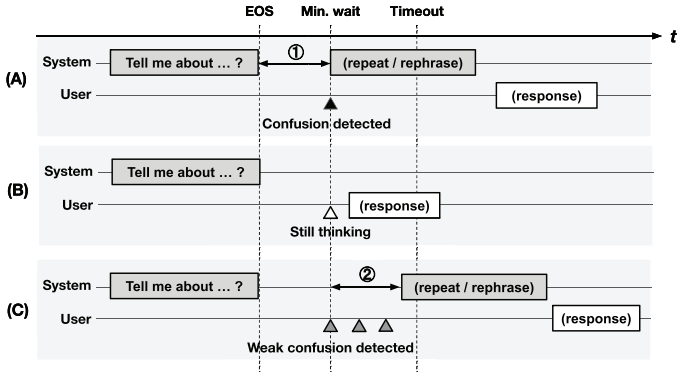

Figure 3: *Examples of adaptive dialogue flows by detecting and resolving confusion in real-time. Right after the end of the system's sentence (EOS), the system starts monitoring the user's behaviors until the minimum waiting time (①). In case (A), the system detected the user's confusion and instantly repeated the previous utterance. In case (B), the system did not detect any confusion, meaning the user was still thinking and waited for a small time interval until the user started speaking. In case (C), confusion was initially not detected, and the system extended the waiting time (②).*

Model	Acuracy	Precision	Recall	F1
Full Model	**0.808**	**0.694**	0.641	**0.667**
w/o head rotation	0.663	0.517	0.769	0.619
w/o head distance	0.767	0.579	0.564	0.571
w/o gaze distance	0.735	0.500	0.564	0.530
w/o relative loudness	0.792	0.490	0.590	0.525
Visual only	0.776	0.470	**0.800**	0.591
Audio only	0.668	0.413	0.795	0.544

Table 2: *Confusion classification result for model using all features, and ablation study.*

Classification results of the full model and ablation study are shown in table 2. For the ablation study, only the results for the top four contributing features are shown. First, we observe that using all features results in the highest score for all measures, surpassing the majority baseline accuracy of 0.706. Some signs required a combination of features to predict, and this is likely the reason for this result. From the ablation study, we see that removing relative loudness reduces the F1 score the most. While other features dynamically change even for the case of "not-confused", self-talk which is related to relative loudness only occurs when the user is confused. Therefore it is plausible that this was the most predictive feature of all. Finally, the combination of visual features was more predictive than the combination of audio features, which agrees with the result in [6], showing that visual features tend to be more informative for the detection of states related to confusion.

5. Adaptive interview scenario

In this section, we explain how the model in section 4.2 can be used in a dialogue system to detect and resolve confusion in real-time while avoiding interruption. As shown in Figure 3, the system will initially wait for the minimum wait time (e.g. 3.0 sec.). If the user replied to the question or asked for help, such as repetition, there is no need for detecting confusion. However, if no response was registered, the probability of confusion will be calculated using information up to that point. If the probability was above the threshold, the user will be assisted by either a repetition or rephrasing (A). If the probability was below the threshold, the system would wait for a small time interval (e.g. 1.0 sec.). If the user responded within that interval, confusion is automatically resolved (B). However, if no response was registered, confusion probability will be calculated again, and the system would either decide to help (C) or wait for another interval. This process will be repeated until the max wait time is reached, at which point the system will assist the user. This allows the system to assist the user as early as possible while avoiding interruption.

6. Conclusion

Detecting and resolving user confusion is crucial for avoiding conversational breakdown, and keeping the user engaged in the interview. In this paper, we presented a model to detect user confusion in a online interview dialogue. Firstly, we built a robust confusion detection model by identifying characteristic signs of confusion. We identified seven multimodal signs, some of which were unique to online conversation, and extracted six related features. Through feature-engineering, we were able to extract relevant information and train a LSTM model showing good classification result, given the small amount of data. Self-talk and gaze direction were shown to be the most predictive features. Secondly, we presented a model for detecting and resolving confusion in real time. The model can incrementally detect user confusion, enabling conversational agents to assist the user in a timely manner, while avoiding interruption.

One limitation of our work is that we only collected data from Japanese English-learners, and it is unclear to what extent the signs of confusion are the same with users of other background. Detecting the cause of confusion will enable the system to assist users in a more appropriate manner, which is likely to improve engagement. However our preliminary result showed that more context, such as the difficulty of the system utterance, and user knowledge were necessary to achieve this. We would like to explore such directions in future work.

7. Acknowledgements

This paper is based on results obtained from a project, JPNP20006 ("Online Language Learning AI Assistant that Grows with People"), subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

8. References

[1] A. Brown, "Interviewer variation and the co-construction of speaking proficiency," *Language Testing*, vol. 20, no. 1, pp. 1–25, 2003.

[2] D. Litman, S. Young, M. Gales, K. Knill, K. Ottewell, R. van Dalen, and D. Vandyke, "Towards Using Conversations with Spoken Dialogue Systems in the Automated Assessment of Non-Native Speakers of English," in *SIGdial*, no. September, 2016, pp. 270–275.

[3] K. Zechner and K. Evanini, *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. Routledge, 2020.

[4] D. Kontogiorgos, A. Pereira, and J. Gustafson, "Estimating uncertainty in task-oriented dialogue," in *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction*. Association for Computing Machinery, Inc, oct 2019, pp. 414–418.

[5] N. Li, J. D. Kelleher, and R. Ross, "Detecting Interlocutor Confusion in Situated Human-Avatar Dialogue : A Pilot Study," in *SEMDIAL 2021*, 2021.

[6] R. Cumbal, J. Lopes, and O. Engwall, "Detection of Listener Uncertainty in Robot-Led Second Language Conversation Practice," in *ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction*. Association for Computing Machinery, Inc, oct 2020, pp. 625–629.

[7] S. K. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 147–187, 2010.

[8] J. F. Grafsgaard, K. E. Boyer, and J. C. Lester, "Predicting facial indicators of confusion with hidden markov models," in *ACII*, 2011.

[9] R. Kumar, C. P. Rosé, and D. J. Litman, "Identification of confusion and surprise in spoken dialog using prosodic features," *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP*, vol. 4, no. March, pp. 1842–1845, 2006.

[10] M. Pachman, A. Arguel, L. Lockyer, G. Kennedy, and J. Lodge, "Eye tracking and early detection of confusion in digital learning environments: Proof of concept," *Australasian Journal of Educational Technology*, vol. 32, no. 6, Dec. 2016. [Online]. Available: https://ajet.org.au/index.php/AJET/article/view/3060

[11] N. Bosch, Y. Chen, and S. D'Mello, "It's written on your face: Detecting affective states from facial expressions while learning computer programming," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8474 LNCS, pp. 39–44, 2014.

[12] W. J. M. Levelt, *Speaking: From intention to articulation*. Cambridge, Mass: MIT Press., 1989.

[13] M. Saeki, R. Matsuura, S. Suzuki, K. Miyagi, T. Kobayashi, and Y. Matsuyama, "InteLLA: A speaking proficiency assessment conversational agent," *The 12th Dialog System Symposium*, pp. 15–20, 2021.

[14] A. Zadeh, Y. Chong Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit tadas baltrušaitis," *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.