Is Corpus Suitable for Human Perception?: Quality Assessment of Voice Response Timing in Conversational Corpus through Timing Replacement

Sadahiro Yoshikawa* Ryo Ishii† Shogo Okada*
* Japan Advanced Institute of Science and Technology, Japan
E-mail: {s2230033, okada-s}@jaist.ac.jp
† Human Informatics Laboratories, NTT Corporation, Japan
E-mail: {ryo.ishii.ct}@hco.ntt.co.jp

***Abstract*—The timing estimation models in spoken dialogue systems (SDSs) have typically been trained by human responses in order to achieve the appropriate response timing. However, human response timings are not always appropriate: in a previous experiment in which annotators listened to responses with the timings replaced by fixed values, some responses with the mode value sounded more realistic than actual human responses. Since this previous experiment was a small-scale preliminary one that only showed that some speakers tended to be significantly preferred, in the current study, we conducted an experiment on about 1,700 human responses, and scored whether they could be replaced with the mode value. The results showed that the annotators tended to feel that mode (or perhaps from 0 ms to 400 ms) responses are more appropriate than actual overlappings. We determined the responses that could and could not be replaced with the mode value by a chi-square test and then formulated a detection task to predict them from the scores. The evaluation results showed that our proposed simple model outperformed random selection with the AUC of 0.650. On the basis of these results, we present examples of SDSs, using the score to predict which responses or response timings are appropriate for the SDS users. Our findings may suggest a more efficient way to determine the appropriate response timing for SDSs compared to training models by corpus data.**

I. Introduction

To achieve natural responses in spoken dialogue systems (SDSs), it is crucial to predict the response time that humans perceive as appropriate and then have the system respond at the appropriate timing. However, in daily conversations, the appropriate response time for humans is not always clear, depending as it does on human perception. In addition, the speaker who responds to the interlocutor's utterances does not always pay attention to the response time, so some responses may sound inappropriate (early or late) to the observer. A previous study reported that some responses with the timing replaced by a fixed value sound more like real conversations than actual human responses [1]. This implies that even if a response generation model is trained using response times recorded in corpus data (i.e., actual conversations), it is not always possible to respond with the appropriate timing. Therefore, to achieve the appropriate response timing, it is crucial to not only train models by corpus data, but also to analyze what response times humans perceive as appropriate for each response.

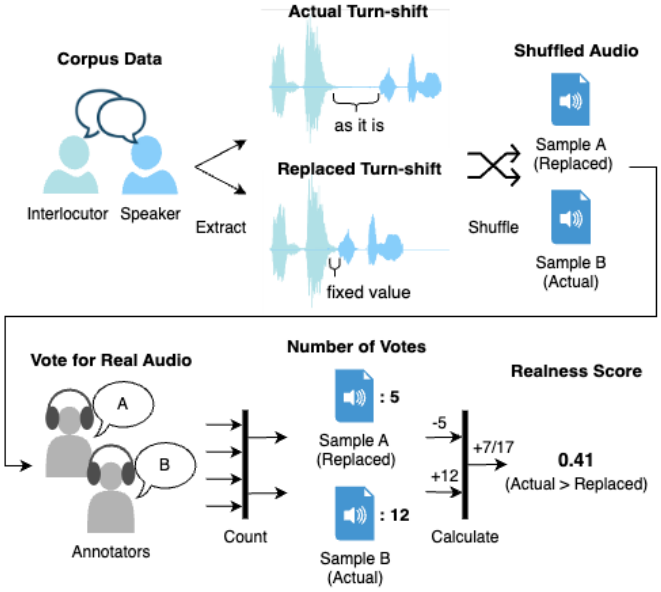

Fig. 1: Procedure of listening test and calculation of realness score (RS). Extraction of Corpus Data is explained in Section III-A, and Vote for Real Audio and calculation of RS is explained in Section III-B.

In the current study, we expanded on previous research [1] with two methodologies. First, since the previous study only showed that some speakers tended to be significantly preferred between an actual response and a response in which the timing was replaced by a fixed value, we utilized more data to analyze the differences for each response to explore the common characteristics of response timing across speakers in Section III. Second, we built a model based on the acquired data and explored how well the model could predict whether response times could be replaced by a fixed value in Section IV. Next, on the basis of these results, we suggested applications to SDSs, using the model to predict which responses or response timings are appropriate for the users in Section VI.

First, we conducted a similar listening test to the previous study to analyze which response times are considered appro-

priate for humans. The procedure of the listening test is shown in Fig. 1. For the corpus data of the listening test, as we aim to explore the diversity of responses for each speaker, we chose a corpus of non-task conversations in Japanese between a speaker and multiple interlocutors. Through several filter conditions, we made a dataset of about 1,700 turn-changes. For voting for the dataset, we gathered 17 annotators for each speaker, and asked them to vote which response voice in the audio samples sounded more realistic: the actual human response (called *actual*) or the response with the timing replaced by a fixed value (called *replaced*). We then defined a realness score (RS) as the average of the votes of the annotators for each response and utilized it to assess the extent to which humans perceived each response time as a real conversation. Since a medium RS contains some randomness, we also performed a statistical test to determine which response timings have statistically significant preferences (SSPs) between *actual* and *replaced*: in other words, which responses can be replaced by a fixed value and which cannot.

As a prediction task for RS, we evaluated how well a prediction model based on neural networks detected which turn-changes have SSPs. The recall rate of SSPs for *actual* indicates the ability of the model to detect the responses for which timing estimation is needed, and that for *replaced* reflects its ability to detect when it is not needed. We evaluated the model using speakers/interlocutor's voice to explore which feature is useful for the prediction, and based on the results, we present examples of SDSs using the model.

II. Related Work

One of the most representative human assessment scores in voice is the mean opinion score (MOS). The annotators of MOS evaluate the quality of voice subjectively and absolutely on a five-point scale ranging from 1 (bad) to 5 (excellent). In the VoiceMOS Challenge [2], a contest to automatically evaluate synthesized speech by building MOS prediction models, there is an application study that utilizes these models to learn only the audio, which is useful for speech synthesis from dark data such as YouTube [3]. Our listening test and RS also aim to achieve high-quality results through automatic evaluation.

There has been some prior work on response time estimation, including an experiment with the users of an automated call system for bus information [4], the development of a estimation model using LSTM [1], and efforts to reduce the effect of speech recognition delays [5], all of which featured models trained by human response timings. Interestingly, it has been reported that some responses replaced by the mode value of the corpus (called *mode*) sound more realistic than actual human responses. In one study [1], the authors defined several filter conditions to extract early responses (called *early*) and late responses (called *late*) from responses across corpus data and then replaced the response times with the mean of the entire *late* if responses were *early* and vice versa (called *opposite*) or *mode*. Then, annotators listened to both the *actual* (called *true* in the paper) and the replaced responses and answered the question "Which response timing sounds like it was produced in a real conversation?". Finally, they performed a statistical test based on the answers to confirm which speakers have SSPs between *actual* and the replaced responses (*opposite* or *mode*) at the $p<0.05$ level.

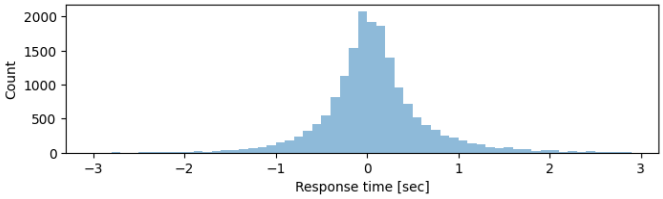

Fig. 2: Number of responses by response time in the target speakers in the corpus of this study.

The test results showed that ten out of 16 speakers of *true* were significantly preferred to *opposite*, while three out of 16 speakers of *actual* and three out of 16 speakers of *mode* were significantly preferred to the comparison. Therefore, in the current study, we explored *true*(*actual*) vs *mode*(*replaced*), which was the controversial result of the previous study.

III. Listening Test

To analyze which response times are considered appropriate for humans, we collected voice data and conducted the listening test. The procedure of the listening test is shown in Fig. 1. We used the mode value of the dataset from the corpus data (0 ms) as a fixed value. Annotators listened to both the actual response (*actual*) and the response whose timing was replaced with the mode value (*replaced*) and then answered the same question as the previous study [1] to vote which one was closer to a real conversation. For analysis and model prediction, we defined the realness score (RS) as the average of the answers of annotators for each response. We also performed a statistical test to confirm which responses humans perceived as a real response timing.

A. Dataset for Listening Test

We used the corpus of Hayashi et al. [6], which includes dyadic interactions in face-to-face conversations in Japanese. A key features of this corpus is the variety of voice samples for each speaker. The speakers to be recorded consisted of groups of four friends, and each speaker in a group recorded one-on-one conversations with three friends and with three other persons the speaker was meeting for the first time (called strangers). The difference between friends and strangers is outside the scope of this study. Each conversation between a speaker and an interlocutor was 1 hour.

To make a dataset of responses, we extracted the audio of eight friends in two groups as speakers, and all (six) interlocutors for each speaker. The distribution of the responses is shown in Fig. 2. We first segmented each conversation audio into voices. These were segmented by silences longer than 200 ms as inter-pausal units (IPUs), following the previous study [1]. We utilized a voice activity detection (VAD) tool https://github.com/wiseman/py-webrtcvad and a low-amplitude

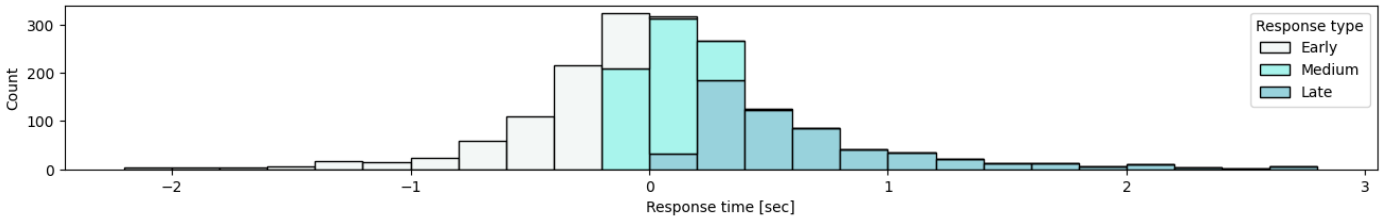

Fig. 3: Number of responses by response time for listening test.

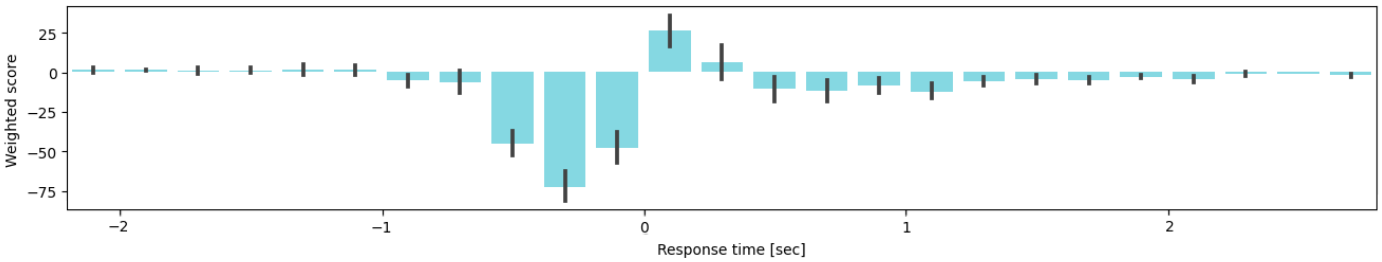

Fig. 4: Distribution of the weighted mean of realness score (RS) by response time multiplied by the number of responses. Vertical line on each bar represents the standard deviation from the mean.

detection script to divide audio into voice segments with 20-ms splits. We then extracted turn-changes from each speaker, defined as when an interlocutor speaks and a speaker responds. We also defined the response as when the speaker's voice is shorter than that of the interlocutor in a turn-change, and excluded those that did not meet this criterion. The response includes not only answers to interlocutors but also backchannels, laughter, reactions, and parts of a speaker's turn.

To make a dataset for the listening test, we extracted 36 turn-changes from the eight speakers and six interlocutors for each speaker. Due to errors during our experimental operations, some turn-changes were excluded. In total, 1,720 turn-changes were targeted in this study, which corresponds to about 10 % of the dataset of responses. The distribution of responses for the listening test is shown in Fig. 3. To extract the 1,720 turn-changes, we set the following filter conditions for this analysis.

Variety of Response Time: To collect a wide variety of response times for each speaker, we divided the responses into three types: *early*, *late*, and *medium*. We calculated response time statistics for each speaker and extracted those below the 30th percentile (as *early*), above the 70th percentile (as *late*), and between *early* and *late* (as *medium*). We collected four responses evenly on the basis of these types. There was a lot of overlap, but most of it occurred without interruption [7].

Extreme Value: We excluded responses later than –20 ms and earlier than 20 ms because we felt there would be almost no difference in such cases even if they were replaced with 0 ms. To avoid extreme values, we also excluded responses below the 0.1 percentile or above the 99.9 percentile.

Unnatural IPU: We excluded turn-changes in which the replacement of response time could be easily perceived by the factors other than voice quality, such as unnatural interruptions in IPUs.

As a result, the distribution of the response times ranged from –2140 ms to 3180 ms. The thresholds between *early* and *medium* for each speaker ranged from –280 ms to 60 ms, and those between *medium* and *late* ranged from 100 ms to 400 ms. The mean of the lengths of the IPUs was 2265 ms for the interlocutors and 908 ms for the speakers.

TABLE I: Statistics of realness score (RS) by response type.

Response type	Mean	SD	Minimum	Maximum
Early	–0.25	0.38	–1.00	0.88
Medium	0.02	0.32	–0.88	0.88
Late	–0.14	0.38	–1.00	0.77
Overall	–0.13	0.37	–1.00	0.88

B. Listening Test Settings

There were 17 annotators for each speaker in the listening test. All annotators were Japanese native speakers. Each used headphones to listen to all turn-changes of two speakers. We shuffled the order of the turn-changes to evaluate each turn-change without time series bias. For each response, annotators listened to both *actual* and *replaced*, and we asked them to answer "Which response timing sounds like it was produced in a real conversation?". We then calculated the RS from the answers, with +1 for answering *actual* and –1 for answering *replaced* and dividing by the number of annotators.

C. Listening Test Results

The results of the listening test are shown in Fig. 4 and Table I. As seen in Fig. 4, the mean of RS was positive from 0 ms to 400 ms and was negative from –1000 ms to 0 ms. From Table I, we found that the maximum RS was 0.88, indicating that there were no responses where all annotators chose *actual*. We also found that the mean RS of *early* and *late* responses was negative, indicating that the annotators preferred *replaced* in the two types.

D. Statistical Test

Since a medium RS contains some randomness, we performed a statistical test on the listening test results to determine which turn-changes had statistically significant preferences (SSPs) between *actual* and *replaced*. Specifically, a chi-square test was conducted to confirm whether there was any bias in the frequency of answers to *actual* or *replaced* for each turn-change at the $p<0.05$ level. The results revealed that there were SSPs in 52 turn-changes for *actual* and in 193 turn-changes for *replaced*. In the turn-changes with SSPs for *actual*, the response time closest to 0 ms was 60 ms, and the furthest was –2140 ms. For *replaced*, the closest was –20 ms, and the furthest was 2800 ms.

The turn-changes that had SSPs are shown in Table II. *Early* responses resulted in the least SSPs for *actual* and the most for *replaced*. The distribution by response time is shown in Fig. 5. In *medium*, all turn-changes that had SSPs for *actual* were positive, i.e. non-overlapping. For all the turn-changes examined in the listening test, in the responses earlier than 0 ms, i.e. overlap, there were fewer SSPs for *actual* than for non-overlap, and there were more SSPs for *replaced*. These results suggest that, in spoken dialogue, humans tend to feel that *mode* (or perhaps from 0 ms to 400 ms) responses are more appropriate than overlap.

TABLE II: Number of turn-changes that have statistically significant preferences (SSPs) between *actual* and *replaced*.

p <0.05	**Response type**			**Total (%)**
	Early	**Medium**	**Late**	
Actual	14	19	19	52 (3.0)
Replaced	107	12	74	193 (11.2)

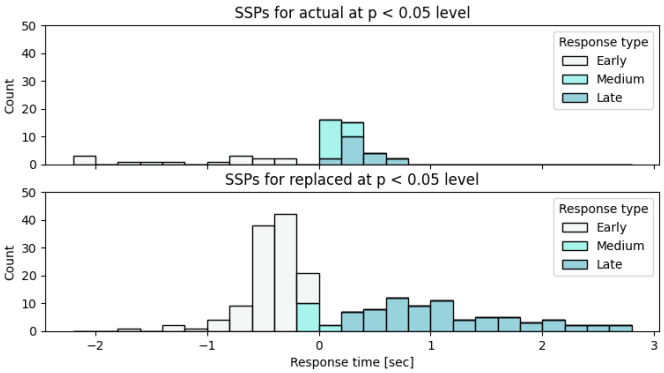

Fig. 5: Number of turn-changes that have SSPs between *actual* and *replaced* by response time.

IV. Realness Score Prediction

We built a model to predict realness score (RS) and explored how well the predicted scores could detect whether response timing estimates are needed or not. As a reference when applying RS to SDSs, we focused on evaluation using only the speaker's/interlocutor's voice. In this study, we define the turn-changes that have SSPs for *actual* as needing response time estimation, and those for *replaced* as not needing it.

A. Model

The VoiceMOS Challenge [8] has shown that self-supervised voice models are useful for predicting the quality of voice. The self-supervised learning framework learns a large amount of audio features, and we used it here as a audio feature extractor. We then implemented a simple model featuring a projection layer added to the self-supervised model as a baseline. For the projection layer, we utilized a 1536-dimensional linear layer that adopted ReLU [9] as the activation function and a linear layer for score output. As the self-supervised model, we used a public HuBERT [10] pretrained with Japanese speech https://huggingface.co/rinna/japanese-hubert-base.

B. Features

Following prior research [8], we input only audio waveforms to the model. We used the audio of the speakers and the interlocutors for a comparison. The number of interlocutors is different from the number of speakers because the speakers consist of friends whereas the interlocutors include both friends and strangers. Therefore, while not a strict comparison, it is sufficient for determining how effective a speaker's features are by using the features of another. We utilized RS normalized from 0 to 1 as the training label.

C. Experimental Settings

We used k-fold cross-validation and measured the out-of-fold prediction scores for each fold. We adopted k = 6 and equalized the number of speakers and interlocutors in each fold. Since we confirmed that strong regularization leads to excessive average results on training data, we set the learning rate to 1e–06 and the weight decay to 1e–05 for model fitting. The model was trained for 70 epochs using Adam [11] without scheduler. The loss function was L1 loss. To enhance reliability, we performed the cross validation with five seeds, and used the average of metrics calculated for each seed as the performance measures.

For evaluation, we defined three evaluation metrics: the recall of SSPs for *actual* (R_a), the recall of SSPs for *replaced* (R_r), and the balanced accuracy (BA). These were calculated as follows:

$$R_a = \frac{T_a}{T_a + F_r} \quad R_r = \frac{T_r}{T_r + F_a} \quad BA = \frac{1}{2}\left(R_a + R_r\right) \quad (1)$$

where:

- $T_{a/r}$ is the number of predicted responses that are necessary/unnecessary to estimate timings, which have SSPs for *actual*/*replaced*.
- $F_{a/r}$ is the number of predicted responses that are necessary/unnecessary to estimate timings, which have SSPs for *replaced*/*actual*.

For these evaluation metrics, we set the prediction score threshold as the mean of the score labels.

In addition, we defined a true actual rate (TAR) and a false actual rate (FAR) to evaluate the performance for each prediction score threshold using ROC curve and AUC:

$$TAR = \frac{T_a}{T_a + F_r} \qquad FAR = \frac{F_a}{F_a + T_r} \tag{2}$$

D. Experimental Results

The evaluation results are provided in Table III. R_a shows 0.669 and BA shows 0.618 using the speaker's voice as input. R_r is higher when learning using the interlocutor's voice. This result may be caused by using the features of the interlocutors or by the number of the interlocutors. The ROC curves are shown in Fig. 6. AUC was 0.650 using the speaker's voice, which indicates that the prediction model was more accurate than random selection.

As an ablation study, we also show the prediction results of each response type in Table IV. We found that the prediction of *early* responses had the highest AUC and the biggest difference of predicted RS between *actual* and *mode* of all response type.

V. Discussion and Future Work

In this study, there were some biases in the accuracy of the RS prediction to *early*. Towards the automatic evaluation of response timing, exploring how to predict the other patterns is required. Several aspects remain to be analyzed:

1) Difference between friends and strangers
2) Difference between each speaker
3) Dialogue act analysis (e.g., backchannel, laughter, etc.)
4) 5-point-scale evaluations gathered from annotators

It may be possible to uncover new patterns that can be detected as proper response timings for humans from the results of these analyses. To explore the characteristics of RS, experiments using the fixed values other than 0 ms are also required.

A. Modalities and Contexts

It is well known that the overlap frequency is different between video conferences and face-to-face meetings [7]. Since we conducted the listening test with a limited number of modalities (audio only) and contexts, conducting a listening test with multimodal perception or several contexts might reveal different results or model performances. In addition, there are differences in response times depending on the language [12], research in other languages is needed.

VI. Towards Developing Spoken Dialogue Systems with Appropriate Response Timing

We explored the prediction accuracy of realness score (RS) and found that it may be possible to achieve the automatic assessment of response timing quality using an RS prediction model. This section discusses the challenges in applying the RS model to SDSs. Since the realness of SDS response timing is determined by the SDS user, we need to explore how a user actually feels when interacting with an SDS. We therefore explore some of the topics relating to the development of SDSs using RS prediction models.

TABLE III: Comparison of evaluation metrics when using the mean of score labels as the prediction score threshold.

Audio feature	$\mathbf{R_a}$	$\mathbf{R_r}$	BA	AUC
Speaker's voice	**0.669**	0.567	**0.618**	**0.634**
Interlocutor's voice	0.446	**0.624**	0.535	0.559

Note: The number of interlocutors is more than that of speakers.

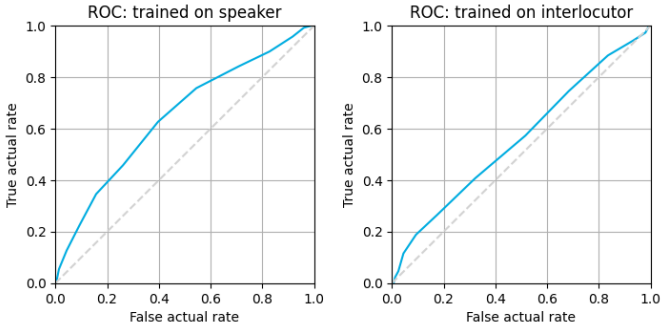

Fig. 6: ROC curves for realness score (RS) prediction model using the voice of speakers and interlocutors.

TABLE IV: Prediction results of each response type inputting the voice of speakers.

Response type	AUC	Mean of predicted RS	
		Actual	Replaced
Early	0.795	0.115	–0.181
Medium	0.540	–0.115	–0.153
Late	0.629	–0.023	–0.126
Overall	0.650	–0.019	–0.159

A. Fixed Value for Timing Replacement

We should first mention the fixed value for the timing replacement. SDSs and VAD tools incur latency when reacting to user utterances. As the fixed value is 0 ms in this study, we cannot utilize the current model directly unless we have prediction models of future conversational events [13]. However, the mode value was just an example of RS measurement and there were high RSs at the response timing from 0 ms to 400 ms; therefore, it would be interesting to evaluate RSs within ranges other than 0 ms.

B. Example of SDSs using RS: Natural Fixed Timing Filler

Under our experimental settings, we can predict RS using only system voices. If we extract system voices in advance, we would be able to predict the voice response timing quality before starting an interaction. In addition, as stated, a high RS response needs a response time estimation, while a low one can be replaced by a fixed value. Therefore, we can extract the responses that can be replaced by a fixed value before starting an interaction as low RS voices using an RS prediction model.

Fig. 7 shows an example of the turn-change using the model. The green zone indicates a low RS voice predicted by the model at a fixed timing, and SDSs can use it as a filler with natural response timing. This will make the quality of response timing stable and increase the time for generating responses.

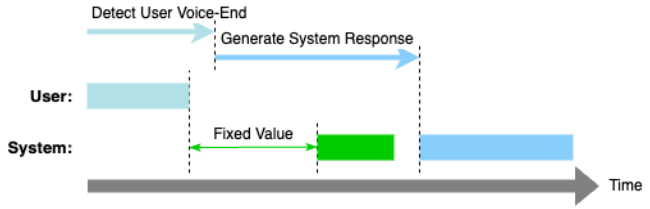

Fig. 7: Example of turn-change of SDSs using realness score (RS) prediction model. The model predicts natural voices at a fixed timing before starting the interaction. In the interaction, SDSs use them as fillers at the fixed timing (green zone).

C. Example of SDSs using RS: Multi-Timing Learning

We can use the model to perform inference during conversations. By predicting the RSs of multiple response timings by the same model, we can predict the most proper fixed timings, including human perceptions. We leave this for future work.

D. VAD Accuracy

Since the end timing detection of IPUs will be different depending on the annotators (including VAD tools), we recommend using the same VAD model not only in the training data but also in the applications. In addition, VAD models must deal with different cases such as phonemes, noises, and languages, so if the RS prediction model uses the interlocutor's features as input, we also have to deal with different cases in the model as well as VAD models. Even if using the system's ones, we need to pay attention to the accuracy of VAD models in tens to hundreds of milliseconds depending on the requirements of the SDS.

VII. Conclusion

Through an experiment on about 1,700 human responses, we clarified the distribution of RS and the possibility of detecting which responses can be replaced with fixed values for appropriate response timing by a simple model. Although there were some patterns that could not be observed due to the filtering process, the voice patterns of the speakers in this study were generally covered. The RS we proposed can also be utilized with previous timing estimation models or with other voice models. We believe the usage of this score will help in the construction of SDSs that are comfortable for humans.

References

[1] M. Roddy and N. Harte, "Neural Generation of Dialogue Response Timings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 2442–2452. DOI: 10.18653/v1/2020.acl-main.221.

[2] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023.

[3] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," *ArXiv*, vol. abs/2210.14850, 2022.

[4] A. Raux and M. Eskenazi, "Optimizing the turn-taking behavior of task-oriented spoken dialog systems," *ACM Transactions on Speech and Language Processing*, vol. 9, no. 1, 1:1–1:23, 2012, ISSN: 1550-4875. DOI: 10.1145/2168748.2168749.

[5] J. Sakuma, S. Fujie, and T. Kobayashi, "Response Timing Estimation for Spoken Dialog Systems Based on Syntactic Completeness Prediction," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 369–374. DOI: 10.1109/SLT54892.2023.10023458.

[6] T. Hayashi, C. O. Mawalim, R. Ishii, *et al.*, "A Ranking Model for Evaluation of Conversation Partners Based on Rapport Levels," en, *IEEE Access*, vol. 11, pp. 73 024–73 035, 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3287984.

[7] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," en, *Computer Speech & Language*, vol. 67, p. 101 178, May 2021, ISSN: 08852308. DOI: 10.1016/j.csl.2020.101178.

[8] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2021.

[9] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10, Madison, WI, USA: Omnipress, 2010, pp. 807–814, ISBN: 978-1-60558-907-7.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A.-r. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[12] T. Stivers, N. J. Enfield, P. Brown, *et al.*, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, Jun. 2009. DOI: 10.1073/pnas.0903616106.

[13] E. Ekstedt and G. Skantze, "Voice Activity Projection: Self-supervised Learning of Turn-taking Events," in *Proc. Interspeech 2022*, 2022, pp. 5190–5194. DOI: 10.21437/Interspeech.2022-10955.