Transition Relevance Point Detection for Spoken Dialogue Systems with Self-Attention Transformer

Kouki Miyazawa and Yoshinao Sato
Fairy Devices Inc.
{miyazawa,sato}@fairydevices.jp

Abstract

Most conventional spoken dialogue systems determine when to respond based on the elapsed time of silence following user speech utterances. This approach often results in failures of turn-taking, disrupting smooth communications with users. This study addresses the detection of when it is acceptable for the dialogue system to start speaking. Specifically, we aim to detect transition relevant points (TRPs) rather than predict whether the dialogue participants will actually start speaking. To achieve this, we employ a self-supervised speech representation using contrastive predictive coding and a self-attention transformer. The proposed model, TRPDformer, was trained and evaluated on the corpus of everyday Japanese conversation. TRPDformer outperformed a baseline model based on the elapsed time of silence. Furthermore, third-party listeners rated the timing of system responses determined using the proposed model as superior to that of the baseline in a preference test.

1 Introduction

Turn-taking is a fundamental aspect of speech communication, yet spoken dialogue systems often fail to take turns at the right moments. This awkward behavior arises from conventional systems that determine when to speak based solely on the elapsed time of silence after user utterances. Consequently, smooth communications between humans and machines become challenging (Skantze, 2021).

In this study, we aim to detect when a spoken dialogue system is permitted to start speaking, in contrast to conventional models that predict whether speaker shift will actually occur. Specifically, we investigate the transition relevant point (TRP) detection during a user utterance. To achieve this, we employ a self-supervised speech representation and a self-attention transformer. The proposed model receives a single-channel audio stream of user speech and infers whether it is preferable for the spoken dialogue system to take a turn at each time frame.

Most prior studies on turn-taking focused on batch inference conducted at the end of speech. Additionally, previous research commonly relied on linguistic features extracted from transcriptions. The proposed model is notable for performing consecutive inference without relying on linguistic features. We avoid transcription to eliminate any additional latency introduced by automatic speech recognition (ASR) and perform consecutive inference in the frame unit. These characteristics differentiate our approach from earlier studies on turn-taking (Skantze, 2017; Liu et al., 2017; Masumura et al., 2017, 2018; Roddy et al., 2018; Hara et al., 2018, 2019; Ekstedt and Skantze, 2020; Gervits et al., 2020; Yang et al., 2022; Threlkeld et al., 2022; Sakuma et al., 2023).

Notably, TRP detection differs from voice activity projection (VAP) (Ekstedt and Skantze, 2022a,b; Ekstedt et al., 2023). TRP detection infers whether a listener is allowed to start speaking without specifying who will actually take the turn. Conversely, VAP predicts whether each interlocutor will speak next, thereby identifying the likely next speaker. Hence, TRP detection and VAP address distinct tasks. Importantly, a TRP does not necessarily guarantee a speaker change; the present speaker may keep talking, or a listener may simply produce a backchannel response. Moreover, VAP does not distinguish whether the subsequent speech is a backchannel or turn. For the effective use of a VAP model for turn-taking, another model is required to predict whether it is appropriate to make a backchannel or take a turn. From the data perspective, a VAP model learns from whether each interlocutor actually started speaking in the conversational speech data, rendering it self-supervised. Conversely, a TRP detection model requires spoken dialogue data with TRP labels for training. Similar to VAP, most previous studies on turn-taking

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 268–274
Aug 25–27, 2025.

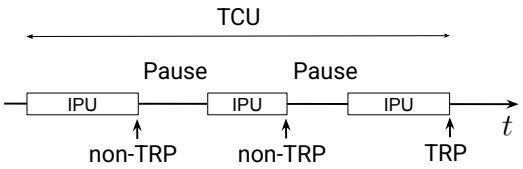

Figure 1: Structure of TCU. A TCU comprises IPUs and intermediate pauses. The end of an IPU is either a TRP or a non-TRP. t denotes the time.

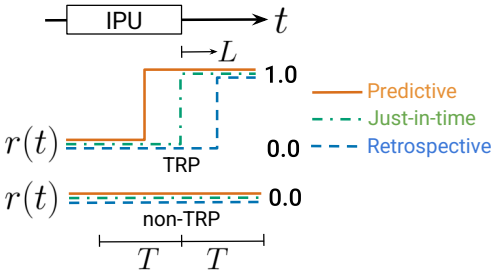

Figure 2: Transition relevance score $r(t)$. The model is trained to detect a TRP with a latency L relative to the IPU tail. The loss function is calculated within the time interval T before and after the IPU tail.

focused on predicting who would actually take the turn (Skantze, 2017; Liu et al., 2017; Masumura et al., 2017, 2018; Roddy et al., 2018; Hara et al., 2018; Ekstedt and Skantze, 2020; Yang et al., 2022; Sakuma et al., 2023).

Prior studies have used the term "TRP" to refer to different tasks. Threlkeld et al. (2022) analyzed the effect of pause duration on the probability of a speaker shift, assuming that the temporal position of the turn end is known. Gervits et al. (2020) focused on predicting the completion point of the current speech utterance from its partial transcription. Hara et al. (2019) used a TRP detection model as the first step in a two-step approach to actual turn-taking prediction. The definition of TRPs in (Hara et al., 2019) aligns with ours; however, their inference was performed in batch and relied on linguistic features, in contrast to our approach.

2 Method

We introduce fundamental elements and formalize TRP detection. An inter-pausal unit (IPU) is a continuous speech utterance without an intermediate pause. Furthermore, a pause (i.e., a period of time during which the speaker is not talking) may occur between adjacent IPUs. IPUs and pauses formulate a turn construction unit (TCU), i.e., a unit of speech that makes up a turn in the conversation, as illustrated in Figure 1. A TRP is located at the end of a TCU, where a turn shift (i.e., speaker change) is acceptable. Conversely, the end of an IPU in the middle of a TCU is referred to as a non-TRP.

2.1 Transition relevance score

As a supervisory target variable, we introduce a transition relevance score $r(t)$ that represents whether a spoken dialogue system is allowed to take a turn at a time frame t:

$$r(t) = \begin{cases} 1 & \text{when the system may take a turn} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

A TRP detection model is trained to predict $r(t)$ from an input audio stream in certain time intervals. Therefore, the TRP detection model should be causal. The predicted scores range from 0.0 to 1.0 (i.e., $r(t) \in [0, 1]$).

In this study, we explore three types of models: predictive, just-in-time, and retrospective. Predictive and retrospective models are trained to detect TRPs earlier and later than the actual point, respectively. The target score is shifted by L in time, as illustrated in Figure 2. L is referred to as latency. A positive value indicates a retrospective model, while a negative value signifies a predictive model. $L = 0$ represents a just-in-time model.

3 Model

This study proposes a TRP detection model with a self-attention transformer (TRPDformer). TRPDformer consists of a pre-trained contrastive predictive coding (CPC) model (Rivière et al., 2020), a causal self-attention transformer, and a detection head, as illustrated in Figure 3. The CPC model yields speech representations at a sampling rate of 100 Hz. A subsequent convolutional layer downsamples the representations to 20 Hz before passing them to the transformer. A causal attention mask is applied in the transformer. The detection head is a time-distributed dense layer with a logit function. The entire model yields the transition relevance score in intervals of 20 ms. The cross-entropy loss is employed for training.

This model structure is identical to that of the stereo VAP model, except for the inevitable modifications that arise from changes in inputs and outputs. Specifically, the input of the proposed model is single-channel audio, whereas the VAP model receives stereo audio. Consequently, the proposed models do not include cross-attention transformers

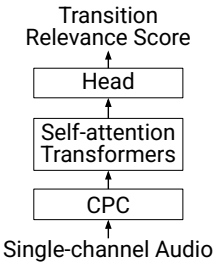

Figure 3: Model structure

between channels. The head is altered to yield a single score, namely $r(t)$, instead of stereo voice activities. Notably, the proposed and VAP models address different tasks: TRP detection and voice activity prediction, respectively. Therefore, the VAP model is not well-suited for TRP detection (see also Section 4.2).

We employ a model based on voice activity detection (VAD) as a baseline. This model detects a TRP when the pause duration exceeds a certain period of time after the IPU ends.

4 Data

This study utilizes a large-scale spontaneous conversational speech corpus known as the corpus of everyday Japanese conversation (CEJC) (Koiso et al., 2022). CEJC consists of 200 hours of speech data recorded during daily conversations involving 862 distinct speakers. The annotations contain long utterance units (LUUs).

An LUU is a basic unit for spoken dialogue that is determined considering syntactic, pragmatic, and prosodic aspects (Japanese Discourse Research Initiative, 2017). Notably, an LUU coincides with a TCU in most cases. In fact, turn-taking frequently occurs at LUU boundaries (Den et al., 2010). Even in the case where a turn-switch does not occur, the LUU endpoint can mostly be considered as a TRP (Enomoto et al., 2020). Therefore, considering the completion points of LUUs as surrogate labels for TRPs is reasonably valid. This approach is widely accepted in prior studies (Koiso and Den, 2011; Ishimoto et al., 2011; Ishimoto and Enomoto, 2017; Enomoto et al., 2020). A few exceptions to this identification of LUUs and TCUs were noted by Hara et al. (2019); Enomoto et al. (2020).

A key distinction between LUUs and TCUs lies in their use of backchannels. An LUU can function as a backchannel and therefore may not constitute a full turn. However, the LUU labels provide advantages for incorporating a TRP detection model into spoken dialogue systems; this is because the systems need to determine when to speak, regardless of whether the current user's speech is a turn or a backchannel. Hence, we treat the end of an LUU as a point to be detected and refer to it as a "TRP."

Table 1: Statistics of IPU transitions

Speaker	Overlap	TRP	non-TRP
Shift		39,827	16,084
Shift	✓	53,556	8,569
Hold		25,734	25,948

4.1 Preprocessing

For preprocessing, we applied speech enhancement to the CEJC data to suppress background noises and reverberations. Additionally, we filtered the data as follows: (1) Two-party dialogues were selected; (2) LUUs that were included in another LUU were excluded; (3) LUUs involving a non-speech sound, such as a laugh, breath, and cough, were eliminated; (5) For each IPU, the segment that overlapped with the subsequent IPU at the tail was cut off if it existed; (6) IPUs with a duration shorter than 500 ms were excluded. Consequently, we obtained 169,718 IPUs.

4.2 Statistics

Table 1 lists the statistics of the selected IPUs. By definition, an IPU with a tail overlap is always accompanied by a speaker shift. This statistical analysis indicates that the prediction of speaker change distinctly differs from the TRP detection. Even perfect speaker change prediction (i.e., perfect VAP) can achieve a recall of 0.60 and precision of 0.71 in the TRP detection task when focusing on non-overlap speech in CEJC. Notably, this theoretical performance is lower than that of the proposed model in our experiments (Fig. 6). The difference lies in the fact that a speaker shift does not necessarily signify a turn shift, as the subsequent IPU could be a backchannel. Furthermore, a TRP does not necessarily indicate a speaker shift, because the current speaker may keep talking.

5 Experiments

We conducted model training and evaluation in IPU units. For each IPU, all precedent IPUs within the same LUU were included in the data unit as the

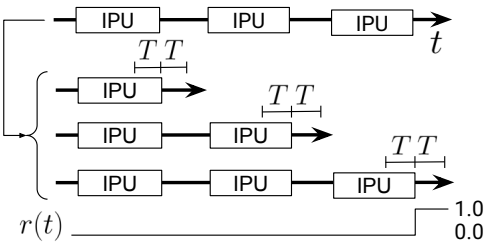

Figure 4: Units of data. Three data units are extracted from an LUU encompassing three IPUs. For each data unit, the time interval of T before and after the tail of the last IPU is considered by the loss function.

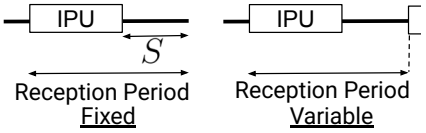

Figure 5: Reception period. A fixed reception period ends when a certain interval S elapses after the IPU offset, whereas a variable reception period ends at the subsequent IPU onset.

past context, as illustrated in Figure 4. The IPU-based data units were divided into three groups for training, validation, and testing with a ratio of 90:5:5. Consequently, we obtained 152,647, 8,545, and 8,526 data units, respectively.

5.1 Training method

To alleviate the imbalance of the score values, only the intervals of T before and after the end of each IPU are considered in the loss function, as illustrated in Figure 2 and 4. Note that the past context available for the model was not limited to this interval.

The embedding dimension of the self-attention was set to 256; the number of heads was set to four; the dropout probability was set to 0.1. In our experiments, we set T=1000 ms and examined different latencies, namely L=−240, 0, 240 ms.

5.2 Test method

The correctness of model inferences was determined based on whether a TRP was detected within a reception period. A reception period begins at the IPU onset, as illustrated in Figure 5. A fixed reception period ends when a certain interval S elapses after the IPU offset, disregarding any subsequent IPU. A variable reception period ends at the onset of the subsequent IPU. The IPUs with a tail overlap were excluded at the test time.

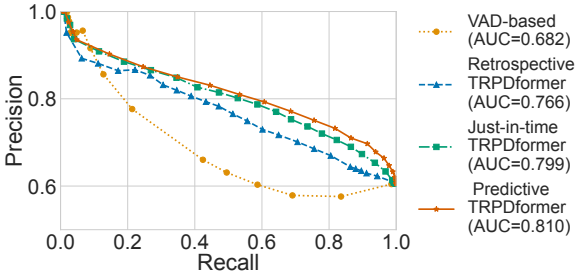

Figure 6: Precision-recall curve

5.3 Performance evaluation

Figure 6 presents the precision-recall curves and area under the curve (AUC) in the variable reception period condition. The proposed models (i.e., TRPDformers) outperformed the baseline model (i.e., the VAD-based model). The balance between the precision and recall can be adjusted by changing thresholds of the pause period and score for the VAD-based and proposed models, respectively. In the variable reception period condition, the models were required to detect a TRP (if applicable) by the onset of the subsequent IPU; hence, the predictive model was slightly advantageous. In the fixed reception period condition, all TRPDformer models achieved almost the same performance, namely an AUC of 0.83.

For the validation set, the retrospective, just-in-time, and predictive TRPDformers achieved AUC scores of 0.763, 0.807, and 0.801, respectively.

5.4 Detection delay analysis

Delays in the TRP detection by the proposed models were analyzed. This analysis was performed in the fixed reception field condition (S=1000 ms) to eliminate the influence of the pause period distribution in CEJC. Figure 7 depicts the detection delay distribution (independent of the detection correctness) with respect to the present IPU offset for score thresholds of 0.5 and 0.75. A negative value of the delay represents that the model detected a TRP earlier than the IPU tail.

Notably, the TRP detection delays increased as the score threshold was increased. This finding indicates that we can modify the system behavior regarding the response delay by adjusting the score threshold.

5.5 Preference test

We conducted a preference test to confirm that the differences between the VAD-based model and the proposed model were meaningful for humans. The

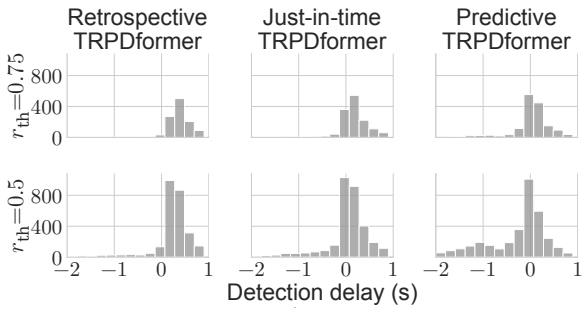

Figure 7: Distribution of the detection delay, where r_{th} denotes the score threshold.

test was designed to simulate a human interacting with a spoken dialogue system, where the user speech is real, while the system response is synthesized. Subjects were presented with a pair of short spoken dialogues: one attributed to the VAD-based model and the other to the proposed model. Each dialogue comprised two or three LUUs. The first and second LUUs were sampled from CEJC and were supposed to be user utterances. The first LUU was common between a pair. The second LUU was inserted only when the VAD-based model detected no TRP in the first LUU. The last LUU was a system response synthesized as follows: (1) The text was generated using Gemini 2.0 Flash by providing the dialogue history; (2) The speech was synthesized using Google Text-to-Speech by providing the generated text. The VAD-based and just-in-time TRPDformer models determined the system response timing. The IPUs in the first and second LUUs after the detected TRP were discarded.

The pause and score thresholds were set to 200 ms and 0.45, respectively, ensuring that both models achieved the same recall of 0.7. Consequently, we prepared 117 stimulus pairs where the two models disagreed on the TRP detection. Table 2 presents the number of stimulus pairs for each condition. Subjects rated which system response was more natural based on the timing on a scale of one to four. Each stimulus was rated by 40 subjects. Figure 8 depicts the distribution of the scores averaged over the subjects.

As expected, the subjects assigned higher ratings to the model that successfully detected TRPs compared to the model that failed to do so. This result indicates that differences in TRP detection are meaningful for humans. Interestingly, when both models incorrectly detected TRPs, the subjects preferred the proposed model over the VAD-based model. Given that the proposed model outperformed the VAD-based model AUC, the effectiveness of the proposed model when applied to spoken dialogue systems was suggested.

Table 2: Number of stimulus pairs

TRPDformer	VAD-based	Count
Correct	Incorrect	65
Incorrect	Correct	23
Incorrect	Incorrect	29

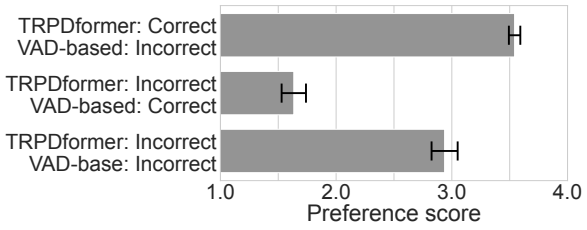

Figure 8: Results of the preference test. A higher score represents that the proposed model is better.

6 Conclusion

This study addressed the continual TRP detection from acoustic features to determine for spoken dialogue systems. The proposed model, TRPDformer, outperformed the VAD-based model when trained and evaluated on CEJC in units of IPUs. Furthermore, a preference test conducted with third-party listeners revealed that the superiority of the proposed model was perceptible to humans.

Finally, we discuss future direction. The computational delay of the proposed model in stream processing and its robustness against noise and reverberation should be evaluated in a manner similar to previous studies on VAP (Inoue et al., 2024, 2025). The proposed model should also be investigated in different domains other than everyday conversation, as well as in languages other than Japanese. Creating a new dataset of human-system spoken dialogues in quiet and anechoic conditions with TRP labels is a promising direction, though it would require considerable investment. Finally, evaluating a spoken dialogue system incorporating the proposed model remains an important direction for future research.

Acknowledgments

This work is supported by a NINJAL Collaborative Research Project: "A Comprehensive Study of Spoken Language Using a Multi-Generational Corpus of Japanese Conversation."

References

Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida. 2010. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2981–2990. Association for Computational Linguistics.

Erik Ekstedt and Gabriel Skantze. 2022a. How much does prosody help turn-taking? investigations using voice activity projection models. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 541–551. Association for Computational Linguistics.

Erik Ekstedt and Gabriel Skantze. 2022b. Voice activity projection: Self-supervised learning of turn-taking events. In *Interspeech*, pages 5190–5194.

Erik Ekstedt, Siyang Wang, Éva Székely, Joakim Gustafson, and Gabriel Skantze. 2023. Automatic evaluation of turn-taking cues in conversational speech synthesis. In *Interspeech*, pages 5481–5485.

Mika Enomoto, Yasuharu Den, and Yuichi Ishimoto. 2020. A conversation-analytic annotation of turn-taking behavior in Japanese multi-party conversation and its preliminary analysis. In *Language Resources and Evaluation Conference (LREC)*, pages 644–652. European Language Resources Association.

Felix Gervits, Ravenna Thielstrom, Antonio Roque, and Matthias Scheutz. 2020. It's about time: Turn-entry timing for situated human-robot dialogue. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 86-–96. Association for Computational Linguistics.

Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. In *Interspeech*, pages 991–995.

Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Turn-taking prediction based on detection of transition relevance place. In *Interspeech*, pages 4170–4174.

Koji Inoue, Bing'er Jiang, Erik Eksted, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and continuous turn-taking prediction using voice activity projection. *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Koji Inoue, Yuki Okafuji, Jun Baba, Yoshiki Ohira, Katsuya Hyodo, and Tatsuya Kawahara. 2025. A noise-robust turn-taking system for real-world dialogue robots: A field experiment. *arXiv preprint arXiv:2503.06241*.

Yuichi Ishimoto and Mika Enomoto. 2017. Prosodic changes leading to transition relevance place in spontaneous utterance. *Annual Conference of the Japanese Cognitive Science Society*, pages 31–37. (in Japanese).

Yuichi Ishimoto, Mika Enomoto, and Hitoshi Iida. 2011. Projectability of transition-relevance places using prosodic features in japanese spontaneous conversation. In *Interspeech*, pages 2061–2064.

Japanese Discourse Research Initiative. 2017. Utterance-unit labeling manual version 2.1. `https://www.jdri.org/resources/manuals/uu-doc-2.1.pdf`. Accessed: 2025-04-26.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the corpus of everyday Japanese conversation. In *Language Resources and Evaluation Conference (LREC)*, pages 5587–5594. European Language Resources Association.

Hanae Koiso and Yasuharu Den. 2011. A phonetic investigation of turn-taking cues at multiple unit-levels in japanese conversation. *International Congress of Phonetic Sciences (ICPhS)*, pages 1122–1125.

Chaoran Liu, Carlos Ishi, and Hiroshi Ishiguro. 2017. Turn-taking estimation model based on joint embedding of lexical and prosodic contents. In *Interspeech*, pages 1686–1690.

Ryo Masumura, Taichi Asami, Hirokazu Masataki, Ryo Ishii, and Ryuichiro Higashinaka. 2017. Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. In *Interspeech*, pages 1661–1665.

Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Ryuichiro Higashinaka, and Yushi Aono. 2018. Neural dialogue context online end-of-turn detection. In *Annual Meeting on the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 224–228.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Investigating speech features for continuous turn-taking prediction using LSTMs. In *Interspeech*, pages 586–590.

Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2023. Response timing estimation for spoken dialog systems based on syntactic completeness prediction. In *Spoken Language Technology Workshop (SLT)*, pages 369–374.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 220–230. Association for Computational Linguistics.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Charles Threlkeld, Muhammad Umair, and JP de Ruiter. 2022. Using transition duration to improve turn-taking in conversational agents. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 193–203. Association for Computational Linguistics.

Jiudong Yang, Peiying Wang, Yi Zhu, Mingchao Feng, Meng Chen, and Xiaodong He. 2022. Gated multimodal fusion with contrastive learning for turn-taking prediction in human-robot dialogue. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7747–7751.