リアルタイム音声対話システムのための応答タイミングと短文応答の同時予測

大中緋慧 [1,2]　河野誠也 [2,1]　大西一誉 [1,2]　吉野幸一郎 [1,2,3]
[1] 奈良先端科学技術大学院大学
[2] 理化学研究所ガーディアンロボットプロジェクト [3] 東京科学大学
{onaka.hien.oj5,onishi.kazuyo.oi5,kawano.seiya.kj0,koichiro}@naist.ac.jp

概要

対話における応答タイミングは発話者の意図を表現するための有用な手段である．この観点から応答タイミングを予測する手法の研究が進んでいる．他方で，このような手法を有効に活用するためには，音声合成などの生成モジュールの遅延を緩和する仕組みが必要となる．このような背景に基づき，応答タイミングと遅延緩和のための短文応答を同時予測するモデルを提案する．提案手法は応答すべきか否かを連続的に予測しながら，応答と判定した際に対照学習に基づくランキング付けにより適切な短文応答を選ぶ．二つのタスクでの客観評価を行い，応答タイミング，短文応答選択の両方で同条件の比較手法に対して優れた結果であることを確認した．

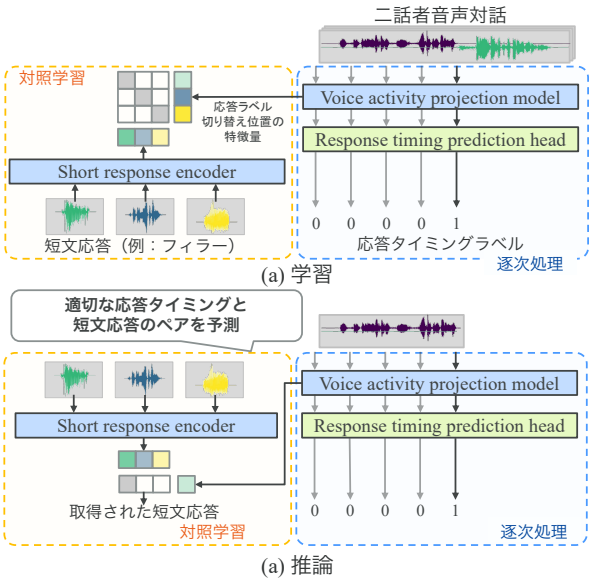

図 1　提案手法の概要図.

1　はじめに

人間は対話の中で，相互に発話権を調整しつつスムーズな情報の授受を実現するためにターンテイキングを行う．ここで，ターンテイキングとは発話権を調整する振る舞いの総称である．ターンテイキングは人とシステム間の自然な対話を実現するための重要な課題として研究されている．従来のターンテイキングモデル [1, 2] は，無音区間に基づいて高精度にユーザ発話の終端検出を行うことを目的としていた [3]．この設計では，システムはユーザの発話を邪魔しないように注意しつつ，発話が終わったことを確認してから応答を行う．

ユーザ発話を邪魔しないことを最重要とするモデル設計の中で，ターン交替時の交替時間（ポーズ長，間合いとも呼ばれる）は無視されがちな要素である．しかし，交替時間は視線 [4] や動作 [5] に並んで，発話者の意図や感情を表現する上での有用な手段である．例えば，応答の受け手は長い交替時間からネガティブな応答を予測すること [6] や，怒りパターン分類タスクにおける特徴量として交替時間が有用なこと [7] が報告されている．人とロボットの対話においても，人間が人間に相対する際と同様に，ロボットの交替時間から遅延以外の意味を汲み取ること [8] が報告されている．

近年では，より自然かつ円滑な対話システムの実現に向けた交替時間推定モデルの研究も盛んになっている．例えば，Roddy らはユーザ発話の音響・言語特徴量とシステム応答の言語特徴量からシステムの応答タイミングを 0.05 秒毎に逐次推定する手法である Response Timing Network [9] を提案した．また，Fujie らは視覚特徴を取り入れ，多人数対話にも対応可能な交替時間推定モデル [10] を提案した．交替時間推定のサブタスクとしてユーザ/システムの対話行為 [11] 推定を組み込む手法もある [12]．

これらは人間にとって許容可能な交替時間の中でベストな点を探索する研究である．他方で，極端に

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

長い交替時間は単なる遅延でありユーザ印象を低下させる [13]．そのため，応答タイミング予測を有効活用するためには，NLG や TTS の遅延を緩和する仕組みが必要となる．このような仕組みの一つとしてフィラー挿入による知覚的な遅延の緩和がある [13, 14]．これらは遅延を埋めるようにフィラーを出力することで，ロボットの応答に対する好ましさの低下 [13] や遅延に起因する気まずさを緩和 [14] できることを報告している．リアルタイム音声対話の実応用でも，フィラー戦略は遅延を緩和するための一般的なアプローチの一つである．一方，このアプローチの課題として，文脈に沿わないフィラーの出力や応答の多様さの乏しさが挙げられる．

本稿ではより現実的な設定を考慮し，応答タイミング予測と遅延緩和のための高度なフィラー戦略を実現するためのマルチタスク学習に基づく手法を検討する．図 1 に提案するモデルを示す．我々の手法では，自己教師ありターンテイキング (Voice Activity Projection; VAP) モデル [15, 16] をベースとして，交替時間予測と短文応答予測を解く．まず，交替時間予測では，0.5 秒後に応答すべきかどうかを二値分類により逐次的に予測する．推論時には逐次的に入力される音声に基づき応答タイミングを連続的かつリアルタイムに推定する．あらかじめ用意された短文応答の集合から適切な応答一つを選択するタスクである短文応答予測では対照学習ベースの方法を取り入れる．学習時には，ターン交替時の VAP モデル特徴量と短文応答エンコーダ出力に基づいて対照学習を行う．推論時には，交替時間予測が 0 から 1 に切り替わる（応答すべきと判定された）タイミングの特徴量に基づいて，短文応答候補から最も適切な一つを選択する．選択された短文応答は，そこまでの文脈と応答タイミングに合うものであることが期待される．評価実験では客観評価を用いて各タスクでの性能評価を行い，提案手法の有効性を述べる．

2 Voice activity projection (VAP)

VAP は Ekstedt らが提案したターンテイキングの自己教師あり学習 [15] である．これは二人の対話者の音声波形を入力として直後 2 秒の二話者の音声活動を予測するタスクであり，学習されたモデルは様々な下流タスクに Zero-shot で活用できる．

最新の VAP モデルのアーキテクチャを図 2 (a) に示す．まず，各話者毎に分かれた 2ch の音声を事前学習済みの CPC encoder [17] によってエンコードする．エンコードされた特徴量は各話者毎に固有の Transformer を通して，その後 Cross-attention Transformer によって統合される．最終的には，各タスクのための線形層を通して出力が得られる．ここで，現時点の音声活動を捉える Voice activity detection (VAD) と，直後 2 秒間の音声活動が $[0.2, 0.4, 0.6, 0.8]$ 秒で離散化された音声活動（$2^{2\text{speaker}\times 4\text{bin}} = 256$）256 クラスからどれに当たるかを捉える VAP がタスクとして採用されている．さらに，VAP モデルをベースモデルとして相槌タイミング予測に特化するように fine-tuning することの有効性も示されている [18]．

3 提案手法

提案手法では図 2 に示すように，VAP モデルベースのタイミング予測と，短文応答エンコーダを用いる対照学習ベースの応答選択によって実現される．この章では，それぞれのタスクについて説明する．

3.1 応答タイミング予測

応答タイミング予測は，応答を返すか否かという二値分類のタスクである．ここでの応答は，発話者がメインターンを有するような発話のみを指しバックチャネルについては考慮されない．正解ラベルの定義は，図 3 に示すように，speaker2 の発話を対象に [18] と同様に実際に応答を返したタイムスタンプを 500 秒前にずらしたものとした．

損失関数 L_{response} は正解ラベルと予測ラベルによる二値クロスエントロピー損失を採用した．また，予測ラベルは図 2 に示すように VAP モデル出力に追加の head を適用することで獲得するものとした．

3.2 短文応答予測

短文応答予測では，与えられた文脈に対してあらかじめ決められた応答の中から，適切な応答をランキング付けにより選択する．そのランキング付けのために対照学習ベースでタスクを解く．

データ抽出　二話者音声対話データセットからの短文応答抽出では，データセットに含まれるタイムスタンプに基づいてまず図 3 において Positive となるメインターン応答の先頭を抽出する．そのために，データセットのタイムスタンプとは別に追加で Silero VAD [19] 音声区間検出を適用する．これにより得られたタイムスタンプに基づいて，応答の先頭における更に細かいセグメントを短文応答として抽出する．ここで，図 2 における L は 2.5 秒とし，

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

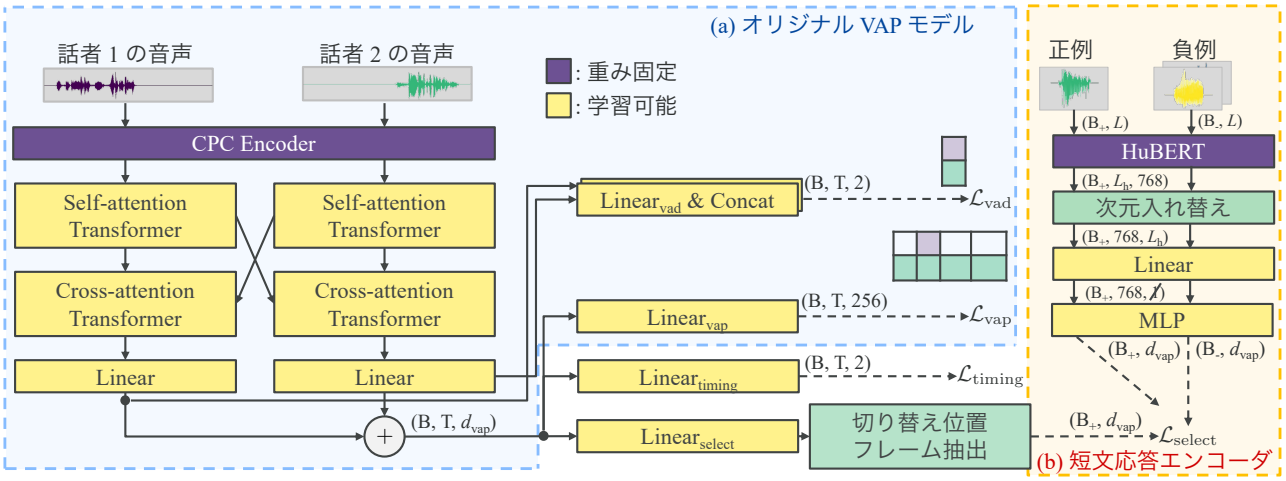

図 2　提案手法のシステム概要図．提案手法は大まかに VAP [15] モデルベースの応答タイミング予測と，short response encoder を用いた対照学習ベースの応答選択の二つから構成される．

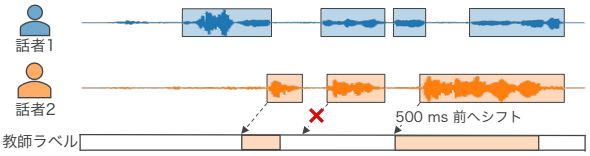

図 3　応答タイミング予測における正解ラベル抽出の例．上から，話者 1 の音声活動，話者 2 の音声活動，正解ラベルを示している．正解ラベルは，話者 2 の音声活動を 0.5 秒前に移すことで得られ，バックチャネルは無視される．

それに満たない短文応答はパディングを行い，超えるものは後方をトリミングした．また，負例のバリエーションを増やすために，同一セッション内のバックチャネルも短文応答の候補として抽出する．

モデル設計　短文応答のエンコーダには rinna 社の HuBERT [20] [1)] を採用する．その後時間方向に特徴量を圧縮する線形層と多層パーセプトロン (Multi-layer perceptron: MLP) を用いて，VAP モデル側の特徴量と同一次元にマッピングする．損失関数には InfoNCE [21] 損失を用いる．この時の正例と負例のサンプリングは次の通りである．まず，VAP モデル側では，Negative から Positive にラベルが切り替わる位置を抽出する．これに対応する短文応答のエンコーディング結果を正例として適用する．負例は，同一セッション内同一話者の短文応答の候補からランダムにサンプリングされる．

4　評価実験

以下の条件で評価実験を実施した．

データセット　データセットとして，高齢者傾聴音声対話コーパス [22] と CEJC コーパス [23] の二話者対話のセッションを用いた．それぞれのデータセットは二話者による音声対話コーパスであり，各話者の発話タイムスタンプがアノテーションされている．これを利用して，ターン交替時のラベルや短文応答の抽出を行った．また，各データセットを 0.8, 0.1, 0.1 の割合で学習，検証，評価用に分割した．

モデル学習　図 2 に示す L_{vad},L_{vap},L_{timing},L_{select} の総和を損失関数として，学習率 $3.63e{-}4$ の AdamW を用いて 20 エポック学習した．また，バッチサイズは 8，対照学習における負例数は 3 とした．各エポックでの検証セットに対する L_{timing}, L_{select} の和が最も小さいものを最良のモデルとして選択した．

4.1　応答タイミング予測精度

評価指標　評価指標として次の二つを用いる．一つはフレームレベルの F1/Recall/Precision である．また，応答の立ち上がりに注視した評価として，真の応答タイミングと推定応答タイミングの絶対誤差 x[s] を許容度とする F1/Recall/Precision を用いる．

比較手法　比較手法として，次の二つを用いる．

- **Random**: 常に Positive と予測（[18] と同様）
- **Pause-based(y)**: 無音が y[s] 続いた際に応答を返すように応答タイミング予測を行う典型的な無音区間に基づくモデル
- **VAP-zeroshot**: VAP モデル出力を変形して得られる 600–2000[ms] の音声活動確率に基づく応答タイミング予測

実験結果　フレームレベルのターンテイキング予測結果の F1/Recall/Precision による客観評価の結果を表 4.1 に示す．結果から，提案手法は二つの比較

1)　https://huggingface.co/rinna/japanese-hubert-base

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　フレームレベルの客観評価の結果

Methods	F1-score	Precision	Recall
Random	0.631	0.462	1.000
VAP-zeroshot	0.748	0.698	0.805
Proposed	**0.769**	**0.719**	**0.821**

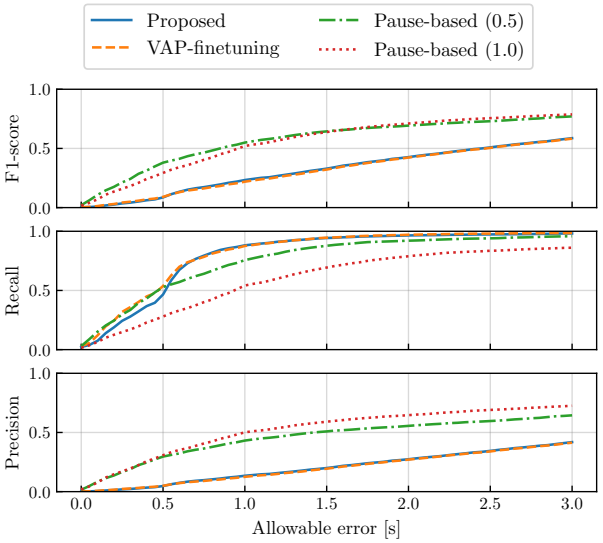

図 4　真の応答タイミングからの許容誤差 $x[s]$ に基づく F1/Recall/Precision .

表 2　短文応答予測の結果

Methods	top-1%	top-5%	top-10%	top-25%	top-50%
Random	1.0	5.0	10.0	25.0	50.0
Proposed	**1.6**	**20.6**	**34.9**	**69.8**	**93.7**

手法に対して優れたスコアを示すことが確認できた．この結果は，VAP モデルのみの fine-tuning よりも，関連するタスクを考慮するような固有の線形層を追加する方針がより良いことを示唆している．

次に，絶対誤差 x[s] を許容度として設定した F1/Recall/Precision による評価結果を図 4 に示す．こちらの観点では，提案手法と VAP-zeroshot がおおよそ同程度の性能であることが確認できる．Recall では VAP-zeroshot の方が僅かに高い精度であるが，これは VAP-zeroshot はメイン応答のタイミングだけでなく相槌のタイミングも暗に推定するため，適合率が高くなる傾向にあるのが要因である．他方で，Pause-based と比較すると提案手法は精度で劣ることが分かる．これは，提案手法は常に 0.5 秒先を予測する設定であるために，偽陽性が増えてしまう事が原因と考えられる．この問題を緩和するために，ユーザ発話の終端度予測などの補助タスク [24, 25] を導入することが今後の課題である．

4.2　短文応答予測

対話文脈からの対照学習に基づく相槌選択 [26] を参照し，同様の評価方法を用いる．

評価指標　応答の候補となる集合に対してランキング付けを適用し，真の応答が上位 k% に含まれるかを評価する top-k% を用いた．

比較手法　比較手法として，応答候補の集合からランダムに応答を選択する Random を用いた．この方法は，相槌やフィラーによる遅延緩和を実現する上で最も直感的なアプローチの一つである．

実験結果　実験結果を表 4.2 に示す．結果から，提案手法に基づく応答選択はランダムな選択と比較して大幅に精度が高いことがわかる．このことから，提案手法は VAP モデルにより得られる対話文脈に基づいて適切な応答を選択していると言える．また，今回の設定では 100 個の候補中実際にコーパスに現れた 1 個を正解としているが，実際はこの 1 個のみが文脈に対して許容される短文応答というわけではなく，やや厳しい設定となっていることに注意が必要である．

5　まとめと今後の展望

本稿では，より自然なリアルタイム音声対話の実現に向けた，応答タイミングと遅延緩和のための短文応答の同時予測モデルを提案した．提案手法は，VAP モデルと短文応答エンコーダに基づいて応答タイミングを連続的に予測し，返答時には適切な短文応答を選択する．評価実験では，各タスクに対して客観評価を実施し，同条件の比較手法に対して優れた性能を示すことを確認した．一方で，提案手法も応答タイミング予測では低い精度であり，短文応答においても難しい条件では精度は高くない．これらの原因の一つとして，過去の音声対話入力に基づく推定にとどまっている事が挙げられる．先行研究では，応答側の対話行為に従って応答タイミングが変化すること [27] や，ネガティブな応答はポジティブな応答よりも遅れる傾向にあること [6] が報告されている．そのため，提案手法の連続的な予測の枠組みの中で，応答側の状態（対話行為，感情など）も考慮したモデル構造に拡張することが今後の課題である．また，客観評価だけでなく，実際に対話システムに提案手法を組み込んだ主観的な対話評価についても進めていく必要がある．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究一部は JST ムーンショット型研究開発事業 JPMJMS2236 の支援を受けたものです.

参考文献

[1] Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa. Learning Decision Trees to Determine Turn-taking by Spoken Dialogue Systems. In **Proc. of ICSLP**, pp. 861–864, 2002.

[2] Antoine Raux and Maxine Eskenazi. A Finite-state Turn-taking Model for Spoken Dialog Systems. In **Proc. of NAACL**, pp. 629–637, 2009.

[3] Nigel G Ward, Anais G Rivera, Karen Ward, and David G Novick. Root Causes of Lost Time and User Stress in a Simple Dialog System. In **Proc. of Interspeech**, pp. 1565–1568, 2005.

[4] Akishige Yuguchi, Tetsuya Sano, Gustavo Alfonso Garcia Ricardez, Jun Takamatsu, Atsushi Nakazawa, and Tsukasa Ogasawara. Evaluating Imitation and Rule-based Behaviors of Eye Contact and Blinking Using An Android for Conversation. **Advanced Robotics**, Vol. 35, No. 15, pp. 907–918, 2021.

[5] Najmeh Sadoughi and Carlos Busso. Speech-driven Animation with Meaningful Behaviors. **Speech Communication**, Vol. 110, pp. 90–100, 2019.

[6] Sara Bögels, Kobin H Kendrick, and Stephen C Levinson. Conversational Expectations Get Revised as Response Latencies Unfold. **Language, Cognition and Neuroscience**, Vol. 35, No. 6, pp. 766–779, 2020.

[7] Narichika Nomoto, Hirokazu Masataki, Osamu Yoshioka, and Satoshi Takahashi. Detection of Anger Emotion in Dialog Speech Using Prosody Feature and Temporal Relation of Utterances. In **Proc. of Interspeech**, pp. 494–497, 2010.

[8] Kiyona Oto, Jianmei Feng, and Michita Imai. Investigating How People Deal with Silence in a Human-robot Conversation. In **Proc. of RO-MAN**, pp. 195–200, 2017.

[9] Matthew Roddy and Naomi Harte. Neural Generation of Dialogue Response Timings. In **Proc. of ACL**, pp. 2442–2452, 2020.

[10] Shinya Fujie, Hayato Katayama, Jin Sakuma, and Tetsunori Kobayashi. Timing Generating Networks: Neural Network Based Precise Turn-Taking Timing Prediction in Multiparty Conversation. In **Proc. of Interspeech**, pp. 3226–3230, 2021.

[11] Wu Mike, Nafziger Jonathan, Scodary Anthony, and Maas Andrew. HarperValleyBank: A Domain-specific Spoken Dialog Corpus. **arXiv 2010.13929**, 2021.

[12] Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. Response Timing Estimation for Spoken Dialog System Using Dialog Act Estimation. In **Proc. of Interspeech**, pp. 4486–4490, 2022.

[13] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. How Quickly Should Communication Robots Respond? In **Proc. of HRI**, pp. 153–160, 2008.

[14] Naoki Ohshima, Keita Kimijima, Junji Yamato, and Naoki Mukawa. A Conversational Robot with Vocal and Bodily Fillers for Recovering from Awkward Silence at Turn-Takings. In **Proc. of RO-MAN**, pp. 325–330, 2015.

[15] Erik Ekstedt and Gabriel Skantze. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In **Proc. of Interspeech**, pp. 5190–5194, 2022.

[16] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. Real-time and Continuous Turn-taking Prediction Using Voice Activity Projection. In **Proc. of IWSDS**, 2024.

[17] Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. Unsupervised Pretraining Transfers Well Across Languages. In **Proc. of ICASSP**, pp. 7414–7418. IEEE, 2020.

[18] Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection. **arXiv preprint arXiv:2410.15929**, 2024.

[19] Silero Team. Silero VAD: Pre-trained Enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad, 2024.

[20] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised Speech Representation Learning by Masked Prediction of Hidden Units. **IEEE/ACM Trans. ASLP**, Vol. 29, pp. 3451–3460, 2021.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. **arXiv preprint arXiv:1807.03748**, 2018.

[22] Koichiro Yoshino, Hiroki Tanaka, Kyoshiro Sugiyama, Makoto Kondo, and Satoshi Nakamura. Japanese Dialogue Corpus of Information Navigation and Attentive Listening Annotated with Extended ISO-24617-2 Dialogue Act Tags. In **Proc. of LREC**, pp. 2922–2927, 2018.

[23] Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Kenya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, et al. Design and Evaluation of the Corpus of Everyday Japanese Conversation. In **Proc. of LREC**, pp. 5587–5594, 2022.

[24] Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. Response Timing Estimation for Spoken Dialog Systems Based on Syntactic Completeness Prediction. In **Proc. of SLT**, pp. 369–374. IEEE, 2023.

[25] Jin Sakuma, Shinya Fujie, Huaibo Zhao, and Tetsunori Kobayashi. Improving the Response Timing Estimation for Spoken Dialogue Systems by Reducing the Effect of Speech Recognition Delay. In **Proc. of Interspeech**, Vol. 2023, pp. 2668–2672, 2023.

[26] Livia Qian and Gabriel Skantze. Joint Learning of Context and Feedback Embeddings in Spoken Dialogue. In **Proc. of Interspeech**, pp. 2955–2959, 2024.

[27] 大中緋慧, 河野誠也, 大西一誉, 吉野幸一郎. 情報案内・傾聴対話における対話行為とターン交替時間の関係分析. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 102, pp. 96–101, 2024.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).