日本語音声合成における対話行為情報の利用による韻律改善 *

☆佐藤匡紀, 高道慎之介, 猿渡洋 (東大院・情報理工)

1 はじめに

ディープラーニング技術の進歩により，テキスト音声合成 (text-to-speech: TTS) 技術は人間の音声に近い品質の音声を合成できるようになった [1]. しかしこのような研究で評価に使用される音声は主に独話の音声である．TTS を独話環境のみならず対話環境でも活用できるようにするため，対話音声コーパスを使用し対話文脈を考慮した音声対話システムのための TTS モデルの研究が進んでいる [2].

対話音声には対話を通して行われる行為や意図に関する情報が含まれる．これを対話行為 (dialogue-act: DA) 情報といい [3]，対話相手との意思疎通を円滑に進めるために重要な役割を持つ．例えば音声対話システムは，対話の中で適切なタイミングで発話ターンを授受する必要がある [4]．これには音声に含まれるターンテイキングの意思を示す情報が重要である．この情報は発話ターンの取得と維持・譲渡の 2 つの場面で利用することが考えられるが，本研究では後者に焦点を当てて TTS モデルがこの情報を表現することで音声対話システムがターンの維持・譲渡を伝達することを目標とする．

また，ターンの維持・譲渡によって句末境界音調 [5] の傾向が異なることが知られている [6, 7]．句末境界音調で TTS モデルを条件付けることで音声の自然性が改善する先行研究 [8] を踏まえ，ターンの維持・譲渡などの情報を表す対話行為情報ラベル (DA ラベル) と句末境界音調を組み合わせて利用する手法についても検討する．

本論文では，ターンテイキングと句末境界音調に焦点を当て，テキスト音声合成におけるターンテイキングの意思の表現とイントネーションの再現性を検証する．FastSpeech2 [1] をベースラインの TTS モデルとして，DA ラベルを用いて予測した句末境界音調を入力したモデルと，DA ラベルと句末境界音調の両方を直接入力したモデルを提案する．また，TTS モデルに対する DA ラベルの入力がターンテイキングにおける韻律の表現に有効であるかを検証する．

2 関連研究

2.1 ターンテイキングに関する情報の利用

対話相手と円滑にコミュニケーションをするには適切なタイミングで発話ターンを授受しなければならない．通常人間同士の対話では双方の協調によってターンの授受が成立しているが，これを音声対話システムが行うためには対話ターンの授受を示す手がかりとなる情報を利用することが求められる．例えば日本語においてはターンを維持する場合に音声の F0 とパワーが大きくなり，句末モーラが短くなる傾向がある [7].

Table 1 句末境界音調の種類と出現頻度

音調名	ラベル	CSJ [11] における出現頻度
下降調	L%	62.970 %
上昇調 1	L%H%	27.919 %
上昇下降調	L%HL%	8.770 %
上昇調 2	L%LH%	0.330 %
上昇下降上昇調	L%HLH%	0.012 %

このターンテイキングの情報を利用する場面としては，発話ターンの取得時と維持・譲渡時の 2 つが考えられる．前者では人間である対話相手の音声に含まれるターンテイキングの情報から音声対話システムが自然なターンの取得を行えるようにする．これを目的とした研究は多く，無音区間や強化学習を利用したものなどが研究されている [4]．一方で，音声対話システムがターンテイキングの情報を表現することでターンの維持・譲渡を対話相手に伝達しようとする研究は少ない．英語の TTS において発話がターンエンドであるかどうかで条件付けた先行研究 [9] があるが，アルゴリズムによって発話ターンを決定するのは容易ではない．そこで，主観的にターンの維持・譲渡のラベルを付与することが考えられる．

2.2 日本語における句末境界音調

句末境界音調 (phrase-final boundary tone: PBT) とは主に日本語のアクセント句末において生じるピッチの多様な変化であり，X-JToBI [5] などの韻律ラベリングスキームで付与されている．最も頻出する基本的な音調は下降調であり，他にも Table 1 に示すような 4 種類の音調がある．この音調はターンテイキングによって傾向が変化する [6, 7]．例えば，小磯 [7] によれば上昇調 1 と上昇調 2 はターンの譲渡が多く，下降調はターンの維持が多い．その他の音調のターンテイキングとの関係については議論があるが，これらの音調は PBT 全体の約 9 割を占めており PBT がターンテイキングの手がかりとなることがわかる．

また PBT は TTS の条件付けに利用されている．山下ら [8] は TTS モデルに対し X-JToBI ラベルで条件づけた場合の自発音声合成の自然性を検証し，PBT を含む複数種類のラベルが合成音声の自然性の改善に寄与できることを示している．さらに，佐藤ら [10] は TTS モデルの学習時には F0 から予測した PBT で条件付け，推論時にはテキストや品詞などのテキスト特徴量と事前学習済み BERT の出力から PBT を予測して使用することで，テキストからの予測誤差を含めて PBT の TTS への利用可能性を検証している．

3 対話行為情報を利用した TTS モデル

本研究ではターンテイキングの再現性を検証するため，会話コーパスを用いて音声データを作成し，各発話がターンの維持または譲渡のどちらに聞こえる

* Prosody Improvement by Using Dialog Action Information in Japanese Speech Synthesis. by Masaki Sato, Shinnosuke Takamichi, Hiroshi Saruwatari (University of Tokyo).

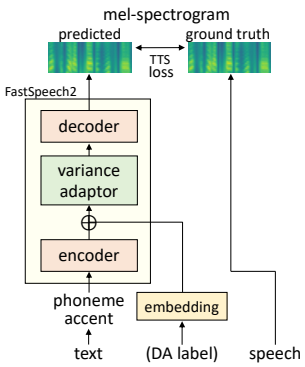

Fig. 1　DA ラベルを入力する TTS モデル．

かを表すラベルを作成する．このラベルに情報要求の対話行為を表す情報として発話末尾の疑問の有無を示すラベルを加えて DA ラベルとした．実験ではこのラベルの有効性と適切な入力方法を検証する．さらに，ターンテイキングや情報要求などの対話行為により出現する PBT の傾向が変化する [6, 7] ことを踏まえて，DA ラベルに加えて各アクセント句の PBT を示す PBT ラベルを TTS モデルに条件付ける．それに加え，PBT ラベルも使用した場合の有効性や PBT ラベルの予測に DA ラベルを使用することの有効性についても検証する．

3.1　DA ラベルの利用

本研究では，まず TTS モデルにおける DA ラベルの有効性を示すため FastSpeech2 [1] を DA ラベルで条件付けるモデルを作成する．このモデルを Fig. 1 に示す．このように DA ラベルを FastSpeech2 の variance adaptor の直前で加えることで DA ラベルで条件付けた音声を推論することができる．

3.2　PBT ラベルの利用

さらに，対話行為情報を TTS モデルで利用することが，PBT の TTS モデルでの利用に与える影響や利用方法を検証する．PBT ラベルを用いるモデルを Fig. 2 に示す．PBT ラベルを使用する場合は多段階の学習を行う．PBT ラベルは X-JToBI などのラベリングスキームで付与されているが，多くのコーパスではそのようなラベルを用いることができないことを踏まえ，まず人力で PBT ラベルが付与されているコーパスを学習データとして F0 から PBT ラベルを生成できるようにする PBT ラベル生成器を作成する．このモデルを Fig. 2(b) に示す．標準化対数 F0 を入力として線形層，Bi-LSTM [12] 層，Attentive Pooling [13] 層，線形層から成る PBT ラベル予測器を用いて出力したラベルを，人力で付与された PBT ラベルとのクロスエントロピー損失で学習する．FastSpeech2 の学習時には，この F0 から生成された PBT ラベルを FastSpeech2 の variance adaptor の直前で加える．

次に，Fig. 2(c) のように音調ラベルを言語特徴量から予測する．ここでは先行・当該・後続アクセント句に含まれる，音素・品詞・句の位置情報を含むテキスト特徴量の埋め込みを BERT の出力へ加える．次に，(c) に示すような PBT ラベル予測器を使用して音調ラベルを予測し，(b) で生成したラベルとのクロスエントロピー損失で学習する．こうすることで，テキストのみから PBT ラベルを予測する．音声を推論する際には，Fig. 2(e) のようにテキストから予測された PBT ラベルを TTS への入力とする．さらに，テキストに加えて対話行為情報も利用して PBT ラベルを予測することで，PBT ラベル予測における対話行為情報の利用の有効性を検証する．この場合は，テキスト特徴量の埋め込みと BERT [14] の出力の和に対して，さらに DA ラベルの埋め込みも加える．それ以降はテキスト特徴量と BERT の出力のみから予測をする場合と同様である．また，対話行為情報は PBT ラベルの予測だけでなく TTS モデルの条件付けにも利用可能である．そこで，TTS モデルをテキストと PBT ラベルだけでなく DA ラベルでも条件付けるモデルを考える．実験的評価ではこの有効性についても検証する．

4　実験

4.1　実験条件

4.1.1　データの処理

本研究では TTS モデルの学習データとして日常会話コーパスである CEJC (Corpus of Everyday Japanese Conversation) [15] を用いた．各音声データは 200ms 以上の pause で区切った単位を 1 発話とし，アクセント句ごとに標準化した標準化対数 F0 を作成した．また CEJC は雑音を含む日常会話コーパスであるため，以下のようにデータのスクリーニングを行った．まず他者との発話時間の被りがない 1 秒以上の発話に対して speechbrain [16] の事前学習モデル[1]を用いて音声強調を行った．次に，NISQA[2] [17] で各発話ごとに音声強調前後の音声品質を評価した．各発話について，当該発話の予測 MOS またはその発話が含まれる収録内発話の平均予測 MOS が 2.0 以下の発話と，音声強調により discontinuity が 0.3 以上減少した発話を除外した．結果として CEJC のうち 37 名による 4971 発話を用い，そのうち 4203 発話を訓練データ，256 発話を評価データ，512 発話をテストデータとした．

本実験では DA ラベルとしてターンの維持・譲渡のラベルと疑問の有無のラベルを用いた．このラベルを作成するため，クラウドソーシングサービスのランサーズ[3]を通して 510 名のラベラーに依頼して，発話音声とその書き起こしを元にその発話後に会話ターンを維持しそうに聞こえるか譲渡しそうに聞こえるかによってラベル付けを行った．ラベルの品質を保証するためにゴールドデータ法 [18] を用い，各ラベラー 50 問のラベル付けに加えて正解が既知である問題 10 問をランダムな位置に挿入し，この問題のうち 9 問以上を正答したラベラーによるラベルのみを採用した．各発話につき 5 つの回答を得てその多数決を最終的なラベルとした．

[1] https://huggingface.co/speechbrain/sepformer-wham16k-enhancement
[2] https://github.com/gabrielmittag/NISQA
[3] https://www.lancers.jp/

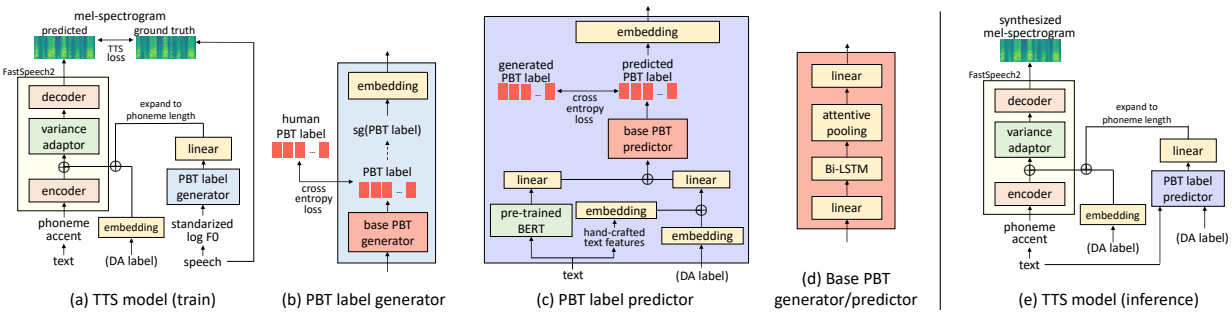

Fig. 2　DA ラベルと PBT ラベルを利用する TTS モデル．sg は勾配伝播の停止を意味する．

Table 2　比較する手法名の一覧．DA を入力する場合は，手法名の末尾に”&DA”を付ける．ただし PBT を入力せず DA のみを入力する手法は”DA”と表記する．

PBT 説明変数	手法名
PBT 入力なし	NoLabel
text	PBT(text)
text と DA	PBT(text&DA)
F0	PBT(F0)

Table 3　テキスト特徴量と DA ラベルからの PBT ラベルの予測精度．

Input	acc	rec	prec	F
Without DA labels	0.787	0.667	0.732	0.682
With DA labels	0.792	0.669	0.731	0.682

4.1.2　モデル条件

本実験では，PBT ラベルとして Table 1 のうち下降調，上昇調 1，上昇下降調の 3 クラスを用いた．まず Fig. 2(b) に示すような F0 から PBT ラベルを作成する PBT ラベル生成器を学習した．これは先行研究 [10] と同様の CSJ [19] コーパスを用いたモデルである．この正解率はマクロ平均で 94.2%であり，先行研究での複数ラベラー間の一致率が 87.7%である [20] ことを踏まえると十分高い精度であると言える．TTS モデルとしては FastSpeech2 [1] の PyTorch での公開実装[4]を用い，テキストを pyopenjtalk[5]によって音素に変換して入力し，学習をした．損失関数は FastSpeech2 [1] と同様である．ボコーダには事前学習済み HiFi-GAN [21] を使用した．次に (c) に示すように PBT ラベル予測器を学習した．ここでは事前学習済み BERT[6]の出力にテキスト特徴量の埋め込みを加えた．テキスト特徴量としては音素・アクセント・品詞 (part-of-speech: POS) タグ・アクセント句中のモーラ数・疑問の有無・呼気段落中のアクセント句位置・事前学習済み BERT の出力を用いた．

4.2　実験結果

以上のようなモデル条件で実験を行った．比較したモデルを Table 2 に記述した．

まず，DA ラベルが PBT ラベルの予測精度の向上に寄与するかを検証する．DA ラベル予測器の精度を Table 3 に示す．ここで acc は正解率，rec は再現率，prec は適合率，F は F 値のそれぞれマクロ平均を表す．結果は正解率において微小な改善が見られたが，DA ラベルが PBT ラベルの予測精度を大きくは向上させないことを示した．

Table 4　イントネーションの再現性に関する XAB テストの結果．

Method	score	p-value
DA vs NoLabel	0.520 vs. 0.480	2.58×10^{-1}
PBT(text&DA) vs NoLabel	0.472 vs. 0.528	1.15×10^{-1}
PBT(text)&DA vs NoLabel	0.488 vs. 0.512	4.86×10^{-1}

Table 5　ターンテイキングの自然性に関する AB テストの結果．

Method	score	p-value
DA vs NoLabel	0.516 vs. 0.484	3.78×10^{-1}
PBT(text&DA) vs NoLabel	0.550 vs. 0.450	4.64×10^{-3}
PBT(text)&DA vs NoLabel	0.482 vs. 0.518	3.23×10^{-1}

まずイントネーションの再現性に関する XAB テストとターンテイキングの自然性に関する AB テストを，NoLabel，DA，PBT(text&DA)，PBT(text)&DA の 4 つのモデルに実施した．ターンテイキングの自然性に関する AB テストは，テストデータのうち 106 発話を使用して 1 秒以上の発話時間がある前後各 2 発話の原音と評価対象の合成音声を結合し，一連の会話音声を受聴して評価をした．各 AB テストごとに約 40 名が 10 問ずつ回答した．まずイントネーションの再現性についてはどのモデルもベースラインとの有意な差が見られなかった．一方で，Table 5 からターンテイキングについては PBT ラベルの予測に DA ラベルを使用するモデルである PBT(text&DA) が有意にベースラインと比較して選択された．この結果から対話行為情報を用いて予測された PBT ラベルの入力により合成音声のターンテイキングの自然性が向上することが示唆された．

次にこのターンテイキングの主観評価結果を詳しく検証するため，一対比較法により PBT の精度による主観評価への影響と PBT ラベルと DA ラベルを両方使用した場合の影響を調べ，それぞれ Table 6, Table 7 に結果を示した．ここではそれぞれ 107 名，105 名が 12 問ずつ回答して，選択率の z 値により評価をした．Table 6 の結果から，ベースラインと比較して PBT ラベルを入力する 3 つのモデルでターンテイキングの自然性が向上した．一方で，F0 から生成した正解の PBT を用いる PBT(F0) のモデルがテキストから PBT を予測する PBT(text) よりも悪い結

[4] https://github.com/Wataru-Nakata/FastSpeech2-JSUT/tree/master/config/JSUT
[5] https://github.com/r9y9/pyopenjtalk
[6] https://huggingface.co/cl-tohoku/bert-base-japanese

Table 6　ターンテイキングの自然性に関する PBT の精度による一対比較．数値は z 値を表す．

モデル名	No...	...F0)	...DA)	...text)	平均
NoLabel		−0.082	0.059	−0.117	−0.047
PBT(F0)	0.082		−0.094	0.094	0.027
PBT(text&DA)	−0.059	0.094		−0.106	−0.023
PBT(text)	0.117	−0.094	0.106		0.043

Table 7　ターンテイキングの自然性に関する DA と PBT の利用による一対比較．数値は z 値を表す．

モデル名	No...	DA	...DA)	...&DA	平均
NoLabel		0.024	0.036	0.024	0.028
DA	−0.024		−0.060	0.048	−0.012
PBT(text&DA)	−0.036	0.060		−0.060	−0.012
PBT(text&DA)&DA	−0.024	−0.048	0.060		−0.004

果を示しており，PBT の精度の向上によっては自然性が改善されないことが示唆された．また，Table 7 の結果からは大きな差が見られず，DA ラベルを直接 FastSpeech2 の encoder の出力に入力しているモデルはターンテイキングの自然性を改善しなかった．また，これらの一対比較法の結果のうち NoLabel と PBT(text&DA) を比較している項目は少し NoLabel の方が選択率が高く，この結果は Table 5 の結果と矛盾している．これは Table 5 の評価数の方が 2 倍あり，Table 6，Table 7 での 2 モデルの差の方が小さいことから Table 5 の結果の方が妥当性が高いと判断できるが，より詳細な評価が必要であると考えられる．

最後に AB テストを行った各モデルの客観評価結果を Table 8 に示す．客観評価の指標として F0 RMSE [cent]，gross pitch error (GPE) [22]，mean cepstral distortion (MCD) [23] の 3 つの手法を用いた．比較対象とした 3 つのモデル全てが，F0 RMSE と GPE でベースラインを上回り，PBT(text&DA) と PBT(text)&DA は MCD についてもベースラインを上回った．その中でも特に PBT(text&DA) は全ての指標が最も高かった．この結果は DA ラベルを用いることで音声品質を改善することができ，DA ラベルの入力方法として PBT ラベルの予測に DA ラベルを利用することが良いことを示している．これは PBT ラベルの使用が TTS モデルの品質改善の主要因であり，ターンテイキングや情報の要求という対話行為情報が PBT を通して表現されることを示唆している．

5　結論

本研究では，ターンの維持・譲渡と情報要求に関する対話行為情報を用いることで TTS モデルのイントネーションやターンテイキングの自然性を改善することを目標とした．複数の対話行為情報の入力方法を検討し，対話行為情報を用いて句末境界音調を予測して TTS モデルに入力する手法が主観的なターンテイキングの自然性と客観評価結果を改善した．また，句末境界音調の入力がターンテイキングの伝達に有効であることが示唆された．

謝辞　本研究の一部は，JSPS 科研費 21H04900 と 21H05054 の助成の委託を受け実施した．

Table 8　各モデルの客観評価結果．

Method	F0 RMSE	GPE	MCD
NoLabel	346.79	0.239	11.62
DA	343.18	0.233	11.66
PBT(text&DA)	340.15	0.228	11.43
PBT(text)&DA	344.58	0.236	11.57

参考文献

[1] Y. Ren et al., “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in *Proc. ICLR*, May. 2021.

[2] H. Guo et al., “Conversational end-to-end tts for voice agents,” in *Proc. SSW11*. IEEE, 2021, pp. 403–409.

[3] Y. Iseki et al., “Characteristics of everyday conversation derived from the analysis of dialog act annotation,” in *Proc. Oriental COCOSDA*, 2019, pp. 1–6.

[4] G. Skantze, “Turn-taking in conversational systems and human-robot interaction: a review,” *Computer Speech and Language*, vol. 67, p. 101178, 2021.

[5] M. Kikuo et al., “X-JToBI: An extended J-ToBI for spontaneous speech,” in *Proc. ICSLP*, Sep. 2002, pp. 1545–1548.

[6] I. C. Toshinori, “The functions of phrase final tones in Japanese: Focus on turn-taking,” *Journal of the Phonetic Society of Japan*, vol. 10, no. 3, pp. 18–28, 2006.

[7] 小磯花絵, “話者交替における統語的・韻律的特徴の役割: 日本語三者会話の定量的分析に基づく考察,” *Journal of the Phonetic Society*, vol. 14, no. 3, pp. 13–26, 2010.

[8] Y. Yamashita et al., “Investigating effective additional contextual factors in DNN-based spontaneous speech synthesis,” in *Proc. INTERSPEECH*, Oct. 2020, pp. 3201–3205.

[9] J. O’Mahony et al., “Synthesising turn-taking cues using natural conversational data,” in *Proc. SSW12*, 2023.

[10] 佐藤匡紀 et al., “日本語音声合成におけるアクセント句韻律特徴量の表現と予測,” **電子情報通信学会技術研究報告**, vol. 122, no. 389, pp. 197–202, 2023.

[11] K. Maekawa, “Corpus of spontaneous Japanese: its design and evaluation,” in *Proc. SSPR*, 2003, pp. 7–12.

[12] M. Schuster, K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[13] C. d. Santos et al., “Attentive pooling networks,” *arXiv*, vol. arXiv:1602.03609, 2016.

[14] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, Jun. 2019, pp. 4171–4186.

[15] H. Koiso et al., “Design and evaluation of the corpus of everyday Japanese conversation,” in *Proc. LREC*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 5587–5594.

[16] H. Hemati, D. Borth, “SpeechBrain: A general-purpose speech toolkit,” *arXiv preprint arXiv:2106.04624*, 2021.

[17] G. Mittag et al., “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Proc. INTERSPEECH*. ISCA, aug 2021.

[18] G. Kazai et al., “Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking,” in *Proc. SIGIR*, ser. SIGIR ’11. New York, NY, USA: Association for Computing Machinery, 2011, p. 205–214.

[19] K. Maekawa et al., “Spontaneous speech corpus of Japanese.” in *Proc. LREC*, vol. 6. Athens, Greece: Citeseer, May. 2000, pp. 1–5.

[20] H. Kikuchi, K. Maekawa, “Performance of segmental and prosodic labeling of spontaneous speech,” in *SSPR*, Apr. 2003.

[21] J. Kong et al., “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NIPS*, ser. NIPS’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[22] T. Nakatani et al., “A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments,” *Speech Communication*, vol. 50, no. 3, pp. 203–214, 2008.

[23] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proc. IEEE pacrim*, vol. 1. IEEE, May. 1993, pp. 125–128.