STUDIES：表現豊かな音声合成に向けた日本語共感的対話音声コーパス*

○齋藤 佑樹，西邑 勇人，高道 慎之介（東大），橘 健太郎 (LINE)，猿渡 洋（東大）

1 はじめに

深層学習技術の進展に伴って音声合成の品質は著しく改善し，多様な話者の自然な音声を合成可能になりつつある [1, 2]．より表現豊かな音声合成を実現するには，多様なスタイルで発話された音声を含むコーパスが重要な役割を持ち，これまでにオーディオブックの朗読音声 [3, 4]，対話音声 [5, 6, 7, 8, 9]，演技感情音声 [11, 12] を収録したコーパスが公開されている．本稿では，ユーザに寄り添った発話が可能な音声 AI エージェントの実現を目指し，我々が新たに構築した **STUDIES コーパス**を紹介する．

STUDIES コーパスは，個別指導塾の女性講師が生徒と雑談する状況を想定して発話された，男性声優 1 名，女性声優 2 名による模擬対話音声を含む．雑談は非タスク志向型対話に分類されるが，本コーパスは女性講師が共感的な話し方で生徒の学習意欲を引き出し，学生生活の楽しさを向上させることを目的として演技した音声を含むように設計されている．本稿では，本コーパスの構築方法として，(1) クラウドソーシングによる雑談対話の台本収集，(2) 相手の感情を考慮して発話された模擬対話音声の収録，(3) 機械学習と人手によるアノテーションを説明する．また，本コーパスを用いた音声合成実験も行い，雑談対話において相手に寄り添う発話を合成するために有効な補助情報を検討する．実験結果より，対話相手の感情ラベルを用いることで，雑談対話音声合成において発話者自身の感情ラベルを用いる場合と同程度の自然性を持つ音声を合成できることを示す．本コーパスのプロジェクトページは `http://sython.org/Corpus/STUDIES` である．

2 コーパスデザイン

2.1 コーパス概要

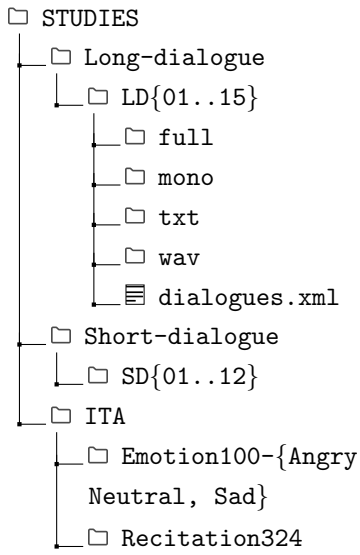
本コーパスは，**Long-dialogue**（LD：10–20 ターンで終了する対話），**Short-dialogue**（SD：4 ターンで終了する対話），**ITA**（ITA コーパス [12] の読み上げ）の 3 つのサブセットからなる．LD/SD サブセットは，個別指導塾の女性講師と { 男子，女子 } 生徒が一対一で雑談する状況を模擬して収録された音声を含む．ITA サブセットは，女性講師が ITA コーパスの Recitation324 文を平静感情で，Emotion100 文を { 平静，喜び，悲しみ，怒り } の 4 感情で読み上げた音声を含む．

2.2 構成要素

本コーパスのディレクトリ構造を以下に示す．なお，各サブセットの構成はディレクトリ名・ファイル名を除きほぼ同一であるため，以下では LD ディレクトリの内容のみを展開して示している．

Table 1　対話台本の例．対話状況と生徒の初期感情はそれぞれ「生徒がテストで良い点数を録って、嬉しそうにしているので、講師が生徒を褒める」と「喜び」である

話者	感情	台詞
男子生徒	喜び	先生、こんにちは！
女性講師	平静	あら、その顔。テスト良かったのかな？
男子生徒	喜び	当ったりー！苦手な数学と化学、点数アップしました！
女性講師	喜び	おめでとう！良かったね！



各サブセットの構成要素を以下に示す．

txt: 発話者，感情ラベル，かな漢字混じりテキスト，読み仮名を対話ごとに保存したもの

wav: 発話ごとの音声ファイル（48,000 Hz サンプリング，16 bit RIFF WAV フォーマット）

mono: 読み仮名から生成したモノフォンラベル

full: かな漢字混じりテキストから Open JTalk [13] で自動生成したフルコンテキストラベル

dialogues.xml: 生徒のペルソナ，対話状況，対話開始時の生徒の感情を含め，txt ディレクトリの内容を xml 形式でまとめたもの

Table 1 に対話台本の例を示す．この表に示す内容をディレクトリごとにまとめたファイルが dialogues.xml であるが，ITA サブセットは読み上げ音声であるため，dialogues.xml は存在しない．また，ITA サブセットの各 txt ディレクトリには，すべての発話の情報を含む 1 つのテキストファイルが存在する．また，モノフォン/フルコンテキストラベルは，HTS [14] を用いて付与された音素アラインメントを含む．

*STUDIES: A Japanese empathetic dialogue speech corpus towards expressive speech synthesis by SAITO, Yuki, NISHIMURA, Yuto, TAKAMICHI, Shinnosuke (UTokyo), TACHIBANA, Kentaro (LINE), and SARUWATARI, Hiroshi (UTokyo).

3 コーパス構築

3.1 雑談対話の台本収集

Lancers [15] でのクラウドソーシングにより，女性講師と生徒の雑談対話の台本を収集した．

3.1.1 LD 台本収集

Lancers で雑談対話台本作成のプロジェクトを 15 件開始し，各プロジェクトに対して 1 名ずつ作業者を割り当てた．このとき，女性講師のペルソナを「20 代前半の女性，東京方言話者，明るくて気さくな性格，語り口調は優しめ，多少は会話をリードしてくれそう」としてすべての作業者に同一のものを指定した．男子生徒と女子生徒のペルソナは指定しなかったが，各作業者が担当する 10 本（男子生徒・女子生徒との対話それぞれ 5 本ずつ）の中で同じペルソナを設定するように指示した．各作業者が設定した生徒のペルソナは，dialogue.xml の `<students_persona>` タグに記載されている．また，雑談の（対話状況，生徒の初期感情）を予め 25 セット用意し，各作業者に対し平静 ×2，喜び ×1，悲しみ ×1，怒り ×1 の 5 パターンの台本を男子生徒・女子生徒との対話で作成させた．各セットにおける台本作成は異なる 3 名の作業者に依頼した．最終的に，25（セット）×2（生徒の性別）×3（異なる作業者の数）=150 対話の台本を収集した．

3.1.2 SD 台本収集

LD 台本収集と同様に，雑談の（対話状況，生徒の初期感情）を予め 6 セット（喜び，悲しみ，怒りそれぞれ 2 つずつ）用意し，Lancers で雑談対話台本作成のタスクを各セット・生徒の性別ごとに 100 件ずつ依頼した．女性講師のペルソナは LD 台本収集と同一のものを指定したが，生徒のペルソナは指定しなかった．得られた台本のうち，指定した初期感情を用いなかったもの，誤字脱字を含むもの，公序良俗に反するものを除外し，最終的に 720 対話の台本を収集した．

3.2 音声収録

3.1 節で収集した台本に基づき，レコーディングスタジオで女性講師と生徒の模擬対話音声を収録した．収録は第一著者が立ち会い，Covid-19 対策として，話者ごとに日を分けて実施した．女性講師と生徒役の声優には事前に台本を配布し，自身の役だけでなく，対話相手の役の発話内容も確認するよう指示した．クラウドソーシングで収集した台本には文法上の誤りや解釈が一意に定まらない文章が含まれたため，声優との協議のもとで台本を部分的に修正しながら音声収録を実施した．助詞の欠落など，リテイクに至らない程度の台本の読み間違いは第一著者が記録し，音声収録終了後に読み間違えた内容で台本を修正した．

女性講師役の声優には，前述の LD/SD 台本に加え，ITA コーパスの読み上げを依頼した．読み上げはコーパス製作者の Github リポジトリ[1]で公開されている pdf 台本のページ単位で実施した．

Table 2　サブセットごとの発話数

話者	LD	SD	ITA	合計（時間数）
女性講師	1,201	1,440	724	3,365 (5.0)
男子生徒	609	720	N/A	1,329 (1.5)
女子生徒	621	720	N/A	1,341 (1.7)

Table 3　感情ラベルごとの発話数

話者	平静	喜び	悲しみ	怒り
女性講師	2,220	715	312	118
男子生徒	351	381	329	268
女子生徒	300	417	340	284

3.3 アノテーション

収録された音声をもとに，台本の読み仮名と音素アラインメントのアノテーションを実施した．読み仮名は pyopenjtalk[2]の g2p メソッドにより自動推定されたものを第一著者が手動で修正する形で付与した．音素アラインメントは (1) 前述の読み仮名から得られたモノフォンラベルと (2) 台本のテキストから Open JTalk で生成したフルコンテキストラベルの 2 つに対し，HTS [14] を用いて付与した．

4 コーパス分析

4.1 スペック

Table 2 に STUDIES コーパスのサブセットごとの発話数を示す．総データ量は女性講師だけで約 5 時間（ITA サブセットを除外すると約 4.3 時間）であり，コーパス全体では約 8.2 時間であった．

Table 3 に感情ラベルごとの発話数を示す．STUDIES コーパスで想定している女性講師と生徒の雑談において，女性講師の発話の過半数は「平静」の感情であり，生徒に同調して「怒り」の感情で発話するケースはごく僅かである．一方で，生徒の感情ラベルごとの発話数は概ね均衡が取れていることが確認できる．

4.2 既存のコーパスとの比較

音声合成システムの開発に用いられる代表的なコーパスを Table 4 に示す．我々の STUDIES コーパスは，対話相手の感情を考慮し，相手に寄り添った発話を生成する「共感的対話」という新たな音声合成のタスクを提唱する．これは，JNAS [16] や JSUT [17] におけるテキスト読み上げや，J-KAC [3] や J-MAC [4] におけるオーディオブックの音声合成とは大きく異なる．また，UUDB [7] や JTES [11] といった感情ラベル付きのコーパスと比較すると，本コーパスの話者あたりのデータ量は 1 時間を超えており，深層学習を用いた音声合成により適しているといえる．また，本コーパスの ITA サブセットは 4 感情のパラレル 100 文を含むため，感情音声変換にも利用できる．

4.3 音声分析の結果

STUDIES コーパスに含まれる対話音声の韻律的特徴を分析するために，対数 F_0 の平均と標準偏差 (std)，energy std，発話ごとのモーラ数を計算した．F_0 の分析には WORLD [18, 19] を使用した．

Figure 1 に韻律的特徴の統計量をバイオリンプロッ

[1] https://github.com/mmorise/ita-corpus

[2] https://github.com/r9y9/pyopenjtalk

Table 4　代表的な日本語コーパス

Corpus	Domain	Open-source	Dur [hour]	# of speakers	Emotion label
JNAS [16]	Reading	No	59	306	No
JSUT [17]	Reading	Yes	10	1	No
JVS [17]	Reading	Yes	30	100	No
JTES [11]	Reading	Yes	23	100	Yes
J-KAC [3]	Audiobook	Yes	9	1	No
J-MAC [4]	Audiobook	Yes	31	39	No
CSJ [5]	Spontaneous reading/dialogue	No	660	1,417	No
Chiba3Party [6]	Spontaneous dialogue	No	2	36	No
UUDB [7]	Spontaneous dialogue	Yes	2	14	Yes
OGVC [8]	Spontaneous/Acted dialogue	Yes	14	17	Yes
NICT-VADC [9]	Acted dialogue	Yes	7	1	No
STUDIES	Empathetic acted dialogue	Yes	8	3	Yes

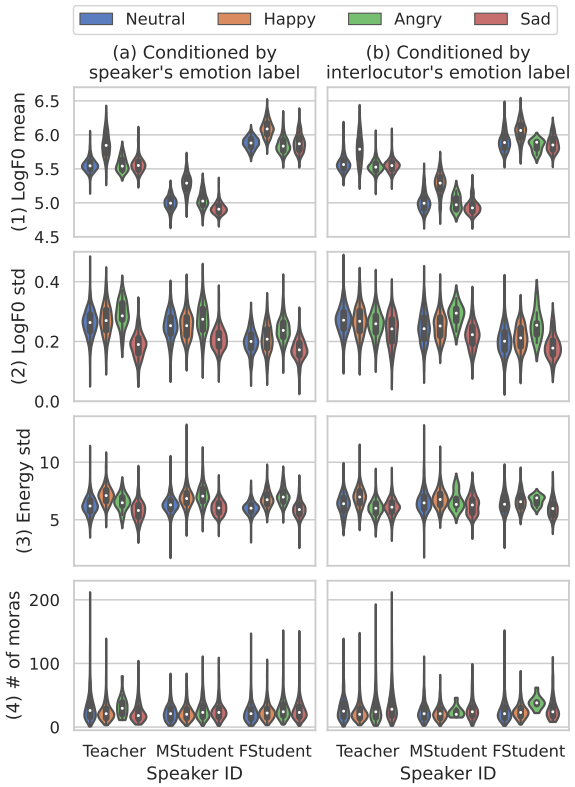


Fig. 1　韻律特徴の統計量

トで可視化した結果を示す．ここで，(a) と (b) はそれぞれ統計量の計算結果を当該話者の感情ラベルで集計した場合と対話相手の感情ラベルで集計した場合を意味する．まず，Fig. 1(1) より，対数 F_0 の平均については，(a) と (b) で顕著な差は見られない．一方で，Fig. 1(2) より，対話相手の感情ラベルで集計した場合，女性講師の対数 F_0 の標準偏差の感情間差異は小さくなることが確認できる．同様の傾向は，女性講師の energy std 分布の怒り・悲しみ間でも確認できる (Fig. 1(3))．最後に，Fig. 1(4) より，女性講師の発話ごとのモーラ数については，対話相手が怒り・悲しみというネガティブな感情なときに長くなる傾向にあることが確認できる．

5　音声合成実験

5.1　データ

STUDIES コーパスの LD/SD サブセットにおける女性講師のデータを Table 5 に示すように分割し，学習・検証・評価データとした．検証・評価データは，生徒の性別（男子，女子）と初期感情（喜び，悲しみ，怒り）がバランスされるように分割した．ITA サブセットの 724 発話はすべて学習データとした．音声データは 22,050 Hz にダウンサンプリングした．

Table 5　STUDIES コーパス女性講師音声の対話単位での学習・検証・評価データ分割．括弧内の数字は発話数を意味する

セット	LD		SD	
学習	126	(1,743)	600	(1,200)
検証	12	(91)	60	(120)
評価	12	(91)	60	(120)

5.2　実験条件

本稿では，音声合成の音響モデルとして FastSpeech2 [20] を使用し，Wataru-Nakata により公開されている日本語音声合成向けの FastSpeech2 pytorch 実装[3]をアクセント記号 [21] で条件付けしたものを用いて実験を行った．音素アラインメントとアクセントの情報はフルコンテキストラベルから抽出した．DNN アーキテクチャの設定（隠れ層数や隠れユニット数など）は前述の実装から変更せずに用いた．メルスペクトログラムの次元は 80 とし，STFT の分析パラメータは FFT 長 1,024，ホップサイズ 256，窓長を 1,024 サンプルとした．FastSpeech2 の variance adaptor で予測される音声特徴量は，基本周波数 (F_0) と energy とした．最適化アルゴリズムには Adam [22] を使用し，学習率の初期値は 0.0625，β_1 は 0.9，β_2 は 0.98 とした．

音声波形生成には HiFi-GAN [23] を用いた．HiFi-GAN は，JSUT [17] コーパスを用いて 100 エポックの事前学習を行った後に，FastSpeech2 と同じ学習データを用いて 200 エポック学習させた．最適化アルゴリズムには Adam を使用し，学習率の初期値は 0.0002，β_1 は 0.8，β_2 は 0.99 とした．

本稿では，以下の 6 つのモデルを比較した．

1. **FS2**：FastSpeech2 [20]
2. **FS2+T-EMO**：女性講師自身の感情ラベルで 1. を条件付け
3. **FS2+S-EMO**：生徒の直前の発話に対する感情ラベルで 1. を条件付け
4. **FS2+CCE**：1. に Conversational Context Encoder (CCE) [24] を導入
5. **FS2+CCE+T-EMO**：女性講師自身の感情ラベルで 4. を条件付け
6. **FS2+CCE+S-EMO**：生徒の直前の発話に対する感情ラベルで 4. を条件付け

[3] https://github.com/Wataru-Nakata/FastSpeech2-JSUT

4. 以降のモデルは，対話履歴のテキスト情報から日本語 pretrained BERT [25] を用いて抽出された sentence embeddding から対話の context embedding を抽出する CCE を導入し，対話の文脈情報を考慮して FastSpeech2 を学習する．CCE は，当該発話よりも 10 個前までの発話の sentence embedding から 256 次元の context embedding を抽出した．また，上記の 6 つのモデルに加え，HiFi-GAN に自然音声のメルスペクトログラムを入力して合成した “Reconst.” も比較した．

5.3 主観評価

前述した 6 つのモデルで合成された音声と “Reconst.” の音声の単一発話としての自然性を 5 段階の Mean Opinion Score (MOS) テストにより評価した．評価者はクラウドソーシングにより集められた 100 名であり，LD/SD サブセットの評価データ 221 個の音声サンプルの中からランダムに抽出された 35 サンプルの品質を評価した．

本稿ではさらに，合成音声の対話としての自然性に関する MOS テストも実施した．このテストでは，評価者に “高校生の男子生徒もしくは女子生徒が，個別指導塾の女性講師と雑談している状況を想定してください．対話の音声を聴いたあとに，講師の音声が，生徒の音声に対して自然に応答するものになっているか（生徒の気持ちに寄り添って発話できているか）を評価してください．” というインストラクションを提示し，女性講師の合成音声が対話として自然かどうかを 5 段階で評価させた．評価者はクラウドソーシングにより集められた 100 名であり，SD サブセットの評価データ 60 対話の中からランダムに抽出された 35 対話の品質を評価した．

Table 6 に評価結果を示す．まず，発話自然性に関する評価結果より，FastSpeech2 ベースの 6 モデルの MOS 間に有意差は確認できず，“Reconst.” が MOS 4 以上を達成している．一方で，対話自然性に関する評価結果では，“FS2+T-EMO,” “FS2+CCE,” “FS2+CCE+S-EMO” の 3 つが “FS2” を有意に上回る MOS を達成している．この結果は，対話において相手に寄り添うような音声合成は (1) 当該話者の感情ラベルが利用できる場合，(2) 対話履歴のテキスト情報が利用できる場合に実現できることを示唆する．さらに，(3) 当該話者の感情ラベルが利用不可能であっても，対話相手の感情が推定できれば，対話履歴のテキスト情報と組み合わせることで自然な対話音声合成を実現できることが確認された．主観評価で用いた音声サンプルは，`http://y-saito.sakura.ne.jp/sython/Corpus/STUDIES/demo_STUDIES.html` で公開した．

6 おわりに

本稿では，表現豊かな音声合成の実現を目指して我々が新たに構築した STUDIES コーパスを紹介した．また，本コーパスを用いた音声合成実験の結果より，対話音声合成において対話相手の感情ラベルを用いることの有効性を示した．今後は，本コーパスの台本のテキスト特徴の分析や，対話履歴を考慮した音声合成の音響モデリングを検討する．

Table 6 MOS テストの結果と 95%信頼区間（**太字**は “FS2” よりも $p < 0.05$ で有意に高いスコア）

手法	発話自然性	対話自然性
Reconst.	4.08±0.09	4.33±0.08
FS2	3.42±0.10	3.20±0.11
FS2+T-EMO	3.43±0.10	**3.38±0.11**
FS2+S-EMO	3.38±0.10	3.29±0.11
FS2+CCE	3.32±0.10	**3.37±0.10**
FS2+CCE+T-EMO	3.38±0.10	3.34±0.10
FS2+CCE+S-EMO	3.43±0.10	**3.41±0.11**

謝辞: 本研究は，LINE 株式会社と東京大学 猿渡・小山研究室の共同研究プロジェクトとして実施した．

参考文献

[1] J. Shen et al., Proc. ICASSP, pp. 4779—4783, Calgary, Canada, Apr. 2018.
[2] J. Kim et al., Proc. ICML, pp. 5530–5540, Virtual Conference, Jul. 2021.
[3] W. Nakata et al., Proc. SSW, pp. 211–215, Budapest, Hungary, Aug. 2021.
[4] S. Takamichi et al., `https://sites.google.com/site/shinnosuketakamichi/research-topics/j-mac_corpus`.
[5] K. Maekawa et al., Proc. LREC, pp. 947–952, Athens, Greece, May 2000.
[6] Y. Den et al., Conversational Informatics: An Engineering Approach, pp. 307–330, T. Nishida, Ed. Hoboken, NJ: John Wiley & Sons, Oct. 2007.
[7] H. Mori et al., Speech Commun., vol. 53, no. 1, pp. 36–50, Jan. 2011.
[8] Y. Arimoto et al., Acoust. Sci. & Tech., vol. 33, no. 6, pp. 359–369, Jun. 2012.
[9] K. Sugiura et al., Advanced Robotics, vol. 29, no. 7, pp. 449–456, Mar. 2015.
[10] R. Zandie et al., Proc. INTERSPEECH, pp. 2751–2755, Brno, Czech Republic, Aug. 2021.
[11] E. Takeishi et al., Proc. O-COCOSDA, pp. 16–21, Bali, Indonesia, Oct. 2016.
[12] 小口 他, 情報処理学会研究報告, vol. 2021-MUS-131, no. 31, pp. 1–6, Jun. 2021.
[13] Open JTalk, `http://open-jtalk.sourceforge.net/`.
[14] HMM-based Speech Synthesis System (HTS), `http://hts.sp.nitech.ac.jp/`.
[15] Lancers, `https://www.lancers.jp/`.
[16] K. Itou et al., Journal of ASJ (En), vol. 20, no. 3, pp. 199–206, May 1999.
[17] S. Takamichi et al., Acoust. Sci. & Tech., Vol. 41, No. 5, pp.761–768, Sep. 2020.
[18] M. Morise et al., IEICE Trans. on Info. & Sys., vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.
[19] M. Morise, Speech Commun., vol. 84, pp. 57–65, Nov. 2016.
[20] Y. Ren et al., Proc. ICLR, Virtual Conference, May 2021.
[21] 藤井 他, 情報処理学会研究報告, vol. 2021-SLP-138, no. 16, pp. 1–6, Oct. 2021.
[22] D. P. Kingma et al., Proc. ICLR, San Diego, U.S.A., May 2015.
[23] J. Kong et al., Proc. NeurIPS, Vancouver, Canada, Dec. 2020.
[24] H. Guo et al., Proc. SLT, pp. 403–409, Shenzhen, China, Jan. 2021.
[25] 柴田 他, 言語処理学会 年次大会 講演論文集, pp. 205–208, Mar. 2019.