Vol.2022-SLP-140 No.16
2022/3/1

対話履歴の韻律情報を考慮した共感的対話音声合成

西邑 勇人[1,a)]　齋藤 佑樹[1,b)]　高道 慎之介[1]　橘 健太郎[2]　猿渡 洋[1]

概要：
本稿では，対話履歴の言語・韻律情報を考慮し，対話相手に寄り添った発話を合成可能な共感的対話音声合成の手法を提案する．音声コミュニケーションにおいて人間は，対話の言語的・韻律的特徴から文脈を理解し，適切な韻律で対話相手に応答できる．しかし，この振る舞いをどのように計算機的に模擬し，音声合成に取り入れるかは詳細に検討されていない．提案法は，対話相手とエージェントの発話テキストと音声からクロスモーダル注意機構により推定される対話文脈埋め込みベクトルで音声合成の音響モデルを条件付けする．本研究ではさらに，対話履歴の文脈を考慮した学習を容易にするためのカリキュラム学習も検討する．実験的評価の結果より，提案法が従来法と比較して合成音声の発話自然性・対話自然性の両方を改善させることを示す．

Yuto Nishimura[1,a)]　Yuki Saito[1,b)]　Shinnosuke Takamichi[1]　Kentaro Tachibana[2]
Hiroshi Saruwatari[1]

1. はじめに

コミュニケーションの役割の一つは，互いの理解を深めて共通の認識を持ち，共有世界を構築することである [1]．また，それを実現する傾聴行為の一つが，本論文の扱う共感 (empathy) である．共感は，相手の世界をあたかも自分のように感じる行為 [1]，または，相手の内側に入り込もうとする能動的な試み [2] であり，相手と感情を同一化する同調 (sympathy) と異なる．近年は，この機能を非タスク指向型対話システム（コミュニケーションを目的とする対話システム） [3] に取り込むために，言語理解 [4] や応答言語生成 [5] が研究されている．他方，音声の感情と韻律も共感の主要素 [6] であり，これを合成音声に付与する技術は共感的対話音声合成 [7] と呼ばれる．この技術の実現には，当該対話中の対話相手（ユーザ）との対話履歴を考慮し，次の応答に寄与する適切な音声特徴量を推定しなければならない．

音声コミュニケーションにおいて人間は，対話の言語的・韻律的特徴から文脈を理解し，適切な韻律で対話相手に応答できる．この振る舞いを計算機的に模擬し，音声合成に導入するためのアプローチとしては，対話履歴の音声言語情報から文脈情報を推定し，音声合成の音響モデルをその文脈情報で条件付けして構築する手法が考えられる．Guo ら [8] は，事前学習済み大規模言語モデルから得られる文埋め込みベクトルを用いて言語的な文脈を考慮する対話音声合成を提案している．この手法では，対話相手の音声に現れる要素 [6] は考慮できない．他方，Yamazaki ら [9] は，対話相手による直前の発話から抽出された F0 系列から合成音声の F0 系列を予測する手法を提案している．しかし，音声の韻律的特徴のうち，単語レベルに量子化された F0 しか使用しておらず，深層学習モデルによる認識生成能力を十分に享受できていない．また，これら 2 つの手法ではそれぞれ言語もしくは音声という単一のモダリティのみが考慮されており，2 つのモダリティをどのように統合し，対話の文脈を推定して共感的対話音声合成を実現すべきかは明らかでない．

本研究は，音声対話における言語・音声の両モダリティを考慮した共感的対話音声合成の音響モデリング法を提案する．提案するアーキテクチャは，入力テキストから音声特徴量を生成する FastSpeech2 [10] ベースの音響モデルと，BERT [11] 由来の文埋め込みベクトル系列およびメルスペクトログラム由来の韻律埋め込みベクトル系列から対話文脈埋め込みベクトルを推定する Cross-Modal Conversational Context Encoder (CMCCE) で構成される．CMCCE は，文埋め込みベクトル系列と韻律埋め込みベクトル系列が持

[1] 東京大学
[2] LINE 株式会社
[a)] yutonishimurav20512@gmail.com
[b)] yuuki_saito@ipc.i.u-tokyo.ac.jp

つ対話履歴情報を統合するためのクロスモーダル注意機構を有しており，Guo らの先行研究 [8] における，言語モーダルのみを考慮した CCE の理論拡張といえる．本研究ではさらに，対話履歴の文脈を考慮した共感的対話音声合成の学習を容易にするために，(1) 対話エージェントのデータのみを用いた音響モデル・韻律埋め込みベクトルの学習と (2) 対話履歴からの韻律埋め込みベクトル予測の 2 段階学習を行うカリキュラム学習も検討する．実験的評価の結果より，(1) 提案するカリキュラム学習が合成音声の発話自然性改善に有効であること，(2) 言語・韻律のクロスモダリティを考慮した音響モデリング法が最も高い発話自然性・対話自然性を達成しうることを示す．

2. 関連研究

2.1 End-to-End 音声合成

音声合成は，テキストから音声を予測するタスクである．従来の統計的パラメトリック (Statistical Parametric Speech Synthesis: SPSS) 方式 [12] では，音声合成タスクを (1) テキストからの言語特徴量抽出，(2) 言語特徴量からの音響特徴量予測，そして (3) 音響特徴量生成からの音声波形生成の 3 つのサブタスクに分割して解いていた．近年では，深層学習技術の発展に伴い，End-to-End (E2E) 音声合成 [13] と呼ばれる，Deep Neural Network (DNN) を活用した手法が数多く提案されている．この方式は，SPSS における 3 つのサブタスクを 2 つに統合する手法 [13], [14]，1 つに統合する手法 [15], [16] など多岐に渡るが，本研究ではテキストから音声特徴量を予測するタスクを解く手法を採用する．また，学習速度の速さや推論の安定さの観点から，音響モデルとして FastSpeech2 [10] を用いる．

2.2 対話履歴を考慮した対話音声合成

E2E 音声合成の進展に伴い，人間の自然音声と同程度の品質の音声を合成できるようになりつつある．一方で，人間の音声コミュニケーションにおける共感・同調などの振る舞いを再現するまでには至っていない．特に，本研究の主題である共感的対話においては，対話の状況や相手の感情の推定など，人間は対話中の様々な情報を考慮して相手に発する言葉を考え，それに合った韻律を持つ音声を発する．しかし，通常の E2E 音声合成の音響モデルは対話履歴の文脈を考慮する機構を有さないため，音声対話の文脈に応じた合成音声の韻律制御を実現できない．

2.2.1 言語情報からの対話文脈埋め込みベクトル推定

Guo ら [8] は，対話履歴の言語情報を考慮した E2E 音声合成の枠組みを提案している．この手法では，対話履歴の各ターンにおけるテキストから BERT [11] によって文埋め込みベクトルを抽出し，それらを Bi-directional GRU (BGRU) によって固定次元に圧縮した対話文脈埋め込みベクトルで音響モデルを条件付けする．以降，このように対話履歴の言語情報から対話文脈ベクトルを予測する機構を Text-Modal Conversational Context Encoder (TMCCE) と称する．

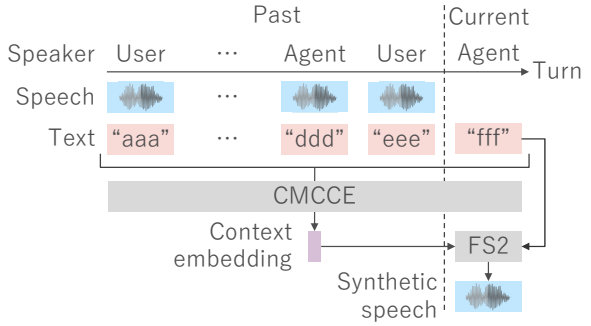

図 1 提案法のアーキテクチャ（“FS2” は FastSpeech2 を意味する）

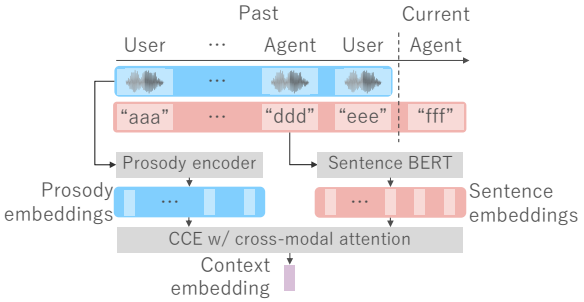

図 2 Cross-Modal CCE (CMCCE) のモデル構造

2.2.2 対話相手の F0 に基づく合成音声の韻律制御

Yamazaki ら [9] は，対話相手の直前の発話から F0 を抽出し，合成音声の F0 予測に用いる手法を提案している．この手法は，対話相手の F0 から推測可能な対話の短期的な文脈は考慮できるが，長期的な文脈と，F0 以外に表出する韻律的特徴を考慮できない．

これらの先行研究は，対話履歴を考慮することによって対話音声合成としてより自然な音声が合成できることを示したものである．しかし，利用可能な対話履歴の韻律情報の制限や，言語・音声のクロスモダリティを考慮できない．

3. 提案手法

本研究では，より高品質な共感的対話音声合成の実現を目指し，エージェントとユーザ間の対話履歴の言語情報と韻律情報の両方で条件付けされた FastSpeech2 ベースの音声合成手法を提案する．提案法のアーキテクチャを **Fig. 1** に示す．

3.1 Cross-Modal CCE (CMCCE)

本研究では，Guo ら [8] の研究における TMCCE を，韻律情報も同時に利用可能にした CMCCE で置換する．**Fig. 2** に CMCCE のモデル構造を示す．CMCCE は，(1) 対話履歴のメルスペクトログラムから各発話の韻律埋め込みベクトルを抽出する prosody encoder (**Fig. 3**)，(2) テキストから各発話の文埋め込みベクトルを抽出する事前学習済みの sentence BERT，そして (3) これらの埋め込みベクト

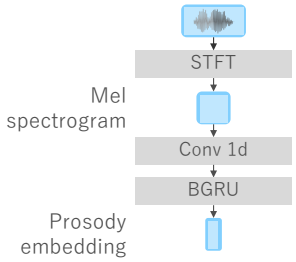

図 3 Prosody Encoder のモデル構造．この図では，ある対話ターンのみにおける韻律埋め込みベクトルの抽出を示している．

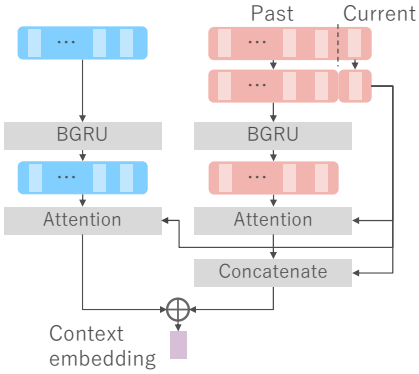

図 4 注意機構付き CMCCE のモデル構造．青色の図形は韻律埋め込みベクトルを表し，赤色の図形は文埋め込みベクトルを表す．

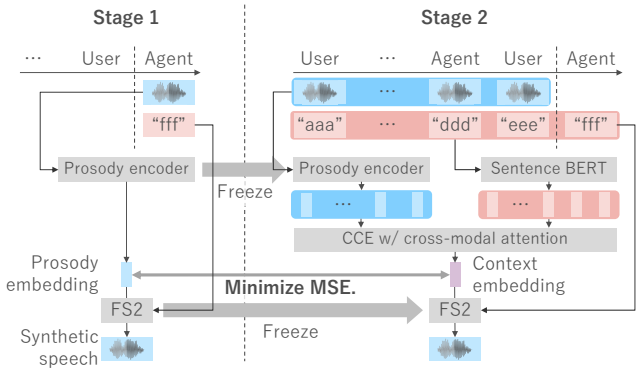

図 5 提案するカリキュラム学習の概念図

ルを統合して対話文脈埋め込みベクトルを抽出するクロスモーダル注意機構付き CCE により構成される．Prosody encoder は 入力された音声のメルスペクトログラムから 1 次元畳み込み層 (Conv 1d) と BGRU を用いて固定次元の韻律埋め込みベクトルを抽出する．クロスモーダル注意機構付き CCE の構造は TMCCE と類似しているが，**Fig. 4** に示すように，各モーダルから抽出された埋め込みベクトルを統合する際に，BGRU ではなく注意機構 (attention) を用いる．この際，attention の query として現在の対話ターンにおける文埋め込みベクトルを用いることで，そのターンにおけるエージェントの音声を合成する際に適した発話スタイルをクロスモーダルな対話履歴情報から選択的に学習可能にする．

3.2 共感的対話音声合成のためのカリキュラム学習

Guo らによる対話音声合成の音響モデリング法 [8] では，音響モデルと TMCCE を同時に学習し，一度に最適化していた．この手法は，文脈を考慮した対話音声合成を一気通貫的に学習できるが，文脈モデリング機構と音響モデルの同時学習の困難性により，必ずしも高品質な共感的対話音声合成を実現できるとは限らない．そこで本研究では，対話履歴の文脈を考慮した対話音声の学習を容易にするためのカリキュラム学習を検討する．**Fig. 5** にその概念図を示す．まず，stage 1 では，エージェントの学習データのみを用いて，エージェント音声から韻律埋め込みベクトルを抽出する prosody encoder と，その韻律埋め込みベクトルで条件付けした FastSpeech2 を学習する．次に，stage 2 ではこの FastSpeech2 と prosody encoder のモデルパラメータを固定し，現在の対話ターンのエージェント音声から抽出された韻律埋め込みベクトルを，対話履歴のテキストと音声を用いて予測するように CMCCE を学習する．これによって，同時に最適化するモデルパラメータの数が大幅に減少し，高品質な共感的対話音声合成が容易に実現可能となることを期待する．

4. 実験的評価

4.1 実験条件

データセットとして，共感的対話音声合成のためのコーパスとして我々が構築した STUDIES [7] を用いた．STUDIES は，個別指導塾の女性講師が男子生徒，女子生徒と雑談している共感的対話の状況を想定している．本研究では女性講師の音声合成モデルを構築したため，男子生徒と女子生徒はいずれも対話履歴においてのみ登場する．また，音響モデル (FastSpeech2) の事前学習用に JSUT コーパス [17] を用いた．音声のサンプリング周波数はすべて 22,050 Hz とした．使用する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 長を 1,024，ホップサイズを 256，窓長を 1,024 サンプルとした．学習データ，検証データ，テストデータの数はそれぞれ 2,209, 221, 211 発話とした．

FastSpeech2 は，Nakata-Wataru により公開されている日本語音声合成向けのオープンソース実装 [*1] を基に構築した．モデルパラメータはすべてこの実装と同じであり，Optimizer は学習率の初期値が 0.0625 の Adam [18] とした．メルスペクトログラムから音声波形を生成するニューラルボコーダには，HiFi-GAN [19] を使用した．HiFi-GAN のモデルは，GitHub 上で公開されている事前学習済みの UNIVERSAL_V1 モデル[*2]のモデルパラメータを初期値として，STUDIES の女性講師音声で fine-tuning したものを利用した．

[*1] https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[*2] https://github.com/jik876/hifi-gan

TMCCE [8] は，GitHub 上のオープンソース実装 [*3] を基に実装した． ただし，対話文脈ベクトルの統合方法が元論文とは異なっていたため，元論文と同じ構成に変更した．Guo らの先行研究では SPSS 由来の特徴量を考慮するための auxiliary encoder も用いられていたが，本研究では対話文脈埋め込みベクトルの予測部のみに着目するために除外した．

本論文では，対話文脈埋め込みベクトルを予測する以下の 3 手法を基本手法とする．

(1) **TMCCE**：BERT 由来の文埋め込みベクトル [8]
(2) **SMCCE**：メルスペクトログラム由来 (Speech-Modal: SM) の韻律埋め込みベクトル**（提案法）**
(3) **CMCCE**：文埋め込みベクトルと韻律埋め込みベクトル**（提案法）**

提案するクロスモーダル注意機構における attention は “TMCCE” でも利用可能であるため，attention を用いる手法を “(Attn)” と表記し，従来法 [8] と同様に BGRU を用いる場合を “(BGRU)” と表記する．また，各手法に対して提案するカリキュラム学習を用いる場合を “+CL” と表記する．さらに，より表現豊かな韻律埋め込みベクトルを学習するために，提案法の prosody encoder を多話者（JSUT, NICT 対話音声データベース [20]，STUDIES の男子生徒・女子生徒）の音声データで学習させる場合 (“+MS”) も検討した．

4.2 主観評価

各基本手法における追加要素の影響を詳細に調査するために，本研究では 2 段階に分けて主観評価を行った．

(1) 基本手法 3 つについて，それぞれで各要素の総組み合わせで主観評価を行う (**Section 4.2.1**).
(2) 第一段階の中で最も良い性能を示したものを選び，その 3 つに従来法 [8] などのベースラインモデルを加えたもので主観評価を行う (**Section 4.2.2**).

主観評価では，STUDIES での評価レギュレーション [7] を参考に，合成音声の単一発話としての自然性と，対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．この主観評価では，各基本手法の評価をクラウドソーシングにより集められた 50 名が実施したため，合計の評価者数は $50 \times 2 \times 3 = 300$ 名であった．

4.2.1 基本手法内での主観評価

まず，“TMCCE” に関する主観評価の結果を **Table 1**(a) に示す．評価結果から，言語情報から対話文脈ベクトルを推定する手法において，提案するカリキュラム学習を用いることで発話自然性が改善している．また，TMCCE を用いる従来法 [8] と比較すると，文埋め込みの統合方法を BGRU から attention に変更することで発話自然性は改善するが，対話自然性については一貫した傾向は見られない．手法間のスコアに有意差はなかったが，以降の評価では，“TMCCE” の最良手法として “TMCCE(BGRU)+CL” を使用する．

表 1 基本手法内で比較した MOS テストの結果と 95%信頼区間（**太字**は発話自然性と対話自然性の平均が最も高い手法）

(a) “TMCCE” 内での比較

手法	発話自然性	対話自然性
TMCCE(BGRU)	3.48±0.14	3.41±0.14
TMCCE(BGRU)+CL	3.66±0.13	3.36±0.14
TMCCE(Attn)	3.51±0.14	3.37±0.14
TMCCE(Attn)+CL	3.55±0.14	3.45±0.14

(b) “SMCCE” 内での比較

手法	発話自然性	対話自然性
SMCCE(BGRU)	3.48±0.14	3.30±0.14
SMCCE(BGRU)+CL	3.50±0.13	3.43±0.14
SMCCE(BGRU)+MS	3.49±0.13	3.34±0.14
SMCCE(BGRU)+CL+MS	3.51±0.13	3.36±0.14

(c) “CMCCE” 内での比較

手法	発話自然性	対話自然性
CMCCE(BGRU)	3.53±0.14	3.32±0.14
CMCCE(BGRU)+CL	3.54±0.14	3.49±0.14
CMCCE(Attn)	3.55±0.13	3.40±0.13
CMCCE(Attn)+CL	3.56±0.13	3.50±0.14
CMCCE(BGRU)+MS	3.40±0.15	3.35±0.15
CMCCE(BGRU)+CL+MS	3.54±0.13	3.48±0.14
CMCCE(Attn)+MS	3.56±0.13	3.43±0.13
CMCCE(Attn)+CL+MS	3.56±0.13	3.44±0.13

次に，“SMCCE” に関する主観評価の結果を **Table 1**(b) に示す．評価結果から，韻律情報から対話文脈ベクトルを推定する手法においても，提案するカリキュラム学習を用いることで発話自然性が改善している．また，prosody encoder を多話者データで事前学習させることの有効性は確認できない．“TMCCE” 内での比較と同様に，手法間のスコアに有意差はなかったが，以降の評価では，“SMCCE” の最良手法として “SMCCE(BGRU)+CL” を使用する．

最後に，”CMCCE” に関する主観評価の結果を **Table 1**(c) に示す．評価結果から，言語・韻律のクロスモダリティを考慮して対話文脈ベクトルを推定する手法の中では，提案するクロスモーダル注意機構とカリキュラム学習を用いた手法 (“CMCCE(Attn)+CL”) が最も高い発話自然性・対話自然性 MOS を達成している．そのため，手法間で有意差は付かなかったが，以降の評価では，“CMCCE” の最良手法として “CMCCE(Attn)+CL” を使用する．

4.2.2 基本手法間での主観評価

Section 4.2.1 で選択した最良の基本手法に以下のベースライン手法を加えて比較する主観評価を実施した．

(1) **FS2 only**：対話文脈を考慮しない FastSpeech2 [10]

[*3] https://github.com/keonlee9420/Expressive-FastSpeech2/tree/conversational

表 2 MOS テストの結果と 95%信頼区間

手法	発話自然性	対話自然性
FS2 only	3.47±0.10	3.46±0.10
TMCCE(BGRU)	3.41±0.10	3.57±0.10
TMCCE(BGRU)+CL	3.53±0.10	3.61±0.10
SMCCE(BGRU)+CL	3.56±0.10	3.58±0.09
CMCCE(Attn)+CL	3.63±0.10	3.68±0.09
CL (Stage1)	3.53±0.10	3.65±0.10

(2) **CL (Stage1)**：提案するカリキュラム学習における stage 1 のモデル
(3) **TMCCE(BGRU)**: 従来法 [8]

この主観評価では，クラウドソーシングにより集められた 100 名が参加したため，合計の評価者数は $100 \times 2 = 200$ 名であった[*4].

Table 2 に評価結果を示す．まず，発話自然性に関して，提案法である “SMCCE(BGRU)+CL” と “CMCCE(Attn)+CL” は，従来法である “TMCCE(BGRU)” よりも $p < 0.05$ で有意に高い MOS 値を達成している．また，“CMCCE(Attn)+CL” は発話自然性・対話自然性の両方で “FS2 only” よりも $p < 0.05$ で有意に高い MOS 値を達成しており，提案法で文脈を考慮した学習を行うことで，より高品質な共感的対話音声合成を実現できることを示唆している．そして，“SMCCE(BGRU)+CL” は “TMCCE(*)” と同程度以上の品質の音声を合成できており，共感的対話音声合成において，言語情報（対話相手の発話内容の書き起こし）を韻律情報で代替できる可能性を示唆している．

興味深い結果として，有意差はないものの，“CL (Stage1)” よりも “‘CMCCE(Attn)+CL” が高い MOS 値を達成している．後者の手法は “CL (Stage1)” で条件付けに用いている STUDIES 女性講師の現在ターン音声から抽出された韻律埋め込みベクトルを教師データとして学習を行うため．前者の手法が提案するカリキュラム学習における性能上界とみなせる．この結果は，文脈情報を考慮することで，共感的対話音声合成に適した対話文脈埋め込みベクトルが得られる可能性を示唆している．一方で，“CL (Stage1)” のモデル構造には更なる改善の余地があることも考えられる．

今回の評価において，文脈を考慮しない “FS2 only” と “TMCCE(BGRU)” の間に有意差は見られなかった．Guo らの先行研究 [8] では，よりリッチな言語特徴量を用いるために auxiliary encoder を導入していたが，本研究は除外していることが一因であると考えられる．故に，言語情報のみから対話の文脈を考慮して学習するためには SPSS 由来の言語特徴量を追加情報として用いることが重要である可能性が示唆された．

[*4] この主観評価で用いた共感的対話音声合成のサンプルは，`http://sython.org/Corpus/STUDIES/demo_empTTS.html` で確認できる．

5. おわりに

本研究は，音声対話における言語・音声の両モダリティを考慮した共感的対話音声合成の音響モデリング法を提案した．さらに，対話履歴の文脈を考慮した共感的対話音声合成の学習を容易にするために，(1) 対話エージェントのデータのみを用いた音響モデル・韻律埋め込みベクトルの学習と (2) 対話履歴からの韻律埋め込みベクトル予測の 2 段階学習を行うカリキュラム学習も検討した．実験的評価の結果より，(1) 提案するカリキュラム学習が合成音声の発話自然性改善に有効であること，(2) 言語・韻律のクロスモダリティを考慮した音響モデリング法が最も高い発話自然性・対話自然性を達成しうることを示した．今後は，発話単位よりも細かい単位で韻律埋め込みベクトルを抽出するための prosody encoder のアーキテクチャや，より微細な文構造を捉えるための PnG BERT [21] を導入した学習法について検討する．

謝辞 本研究は，LINE 株式会社と東京大学 猿渡・小山研究室の共同プロジェクトとして実施し，JSPS 科研費 21K21305 の助成を受けたものです．

参考文献

[1] 古宮 昇：プロカウンセラーが教える場面別傾聴術レッスン，ナツメ社 (2015).
[2] Davis, M. H.: *Empathy: A Social Psychological Approach*, Routledge (2018).
[3] 中野幹生，駒谷和範，船越孝太郎，中野有紀子，奥村学：対話システム，コロナ社 (2018).
[4] Lui, Q., Chen, H., Ren, Z., Ren, P., Tu, Z. and Chen, Z.: EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation, *Proc. COLING*, Barcelona, Spain, pp. 4454–4466 (2020).
[5] Rashkin, H., Smith, E. M., Li, M. and Boureau, Y.-L.: Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset, *Proc. ACL*, Florence, Italy, pp. 5370–5381 (2019).
[6] Regenbogen, C., Schneider, D. A., Finkelmeyer, A., Kohn, N., Derntl, B., Kellermann, T., Gur, R. E., Schneider, F. and Habel, U.: The Differential Contribution of Facial Expressions, Prosody, and Speech Content to Empathy, *Cognition & Emotion*, Vol. 26, No. 6, pp. 995–1014 (2012).
[7] 齋藤佑樹，西邑勇人，高道慎之介，橘健太郎，猿渡洋：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，音講論（春）(2022).
[8] Guo, H., Zhang, S., Soong, F. K., He, L. and Xie, L.: Conversational End-to-End TTS for Voice Agents, *Proc. SLT*, Shenzhen, China, pp. 403–409 (2021).
[9] Yamazaki, Y., Chiba, Y., Nose, T. and Ito, A.: Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems, *Proc. INTERSPEECH*, Brno, Czech Republic (2021).
[10] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, *Proc. ICLR*, Vienna, Austria (2021).

[11] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. NAACL-HLT*, Minneapolis, U.S.A., pp. 4171–4186 (2019).
[12] Zen, H., Tokuda, K. and Black, A.: Statistical Parametric Speech Synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009).
[13] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R. and Saurous, R. A.: Tacotron: Towards End-to-End Speech Synthesis, *Proc. INTERSPEECH*, Stockholm, Sweden, pp. 4006–4010 (2017).
[14] v. d. Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, Vol. abs/1609.03499 (online), available from ⟨http://arxiv.org/abs/1609.03499⟩ (2016).
[15] Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A. and Bengio, Y.: Char2Wav: End-to-End Speech Synthesis, *Proc. ICLR Workshop*, Toulon, France (2017).
[16] Ping, W., Peng, K. and Chen, J.: ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech, *Proc. ICLR*, New Orleans, U.S.A. (2019).
[17] Takamichi, S., Sonobe, R., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N. and Saruwatari, H.: JSUT and JVS: Free Japanese Voice Corpora for Accelerating Speech Synthesis Research, *Acoustical Science and Technology*, Vol. 41, No. 5, pp. 761–768 (2020).
[18] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proc. ICLR*, San Diego, California, U.S.A. (2015).
[19] Kong, J., Kim, J. and Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, *Proc. NeurIPS*, Vol. 33, Virtual Conference, pp. 17022–17033 (2020).
[20] Sugiura, K., Shiga, Y., Kawai, H., Misu, T. and Hori, C.: A Cloud Robotics Approach towards Dialogue-Oriented Robot Speech, *Advanced Robotics*, Vol. 29, No. 7, pp. 449–456 (2015).
[21] Jia, Y., Zen, H., Shen, J., Zhang, Y. and Wu, Y.: PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS, *Proc. INTERSPEECH*, Brno, Czech Republic, pp. 151–155 (2021).