ChatGPT-EDSS: ChatGPT由来のContext Word Embeddingから学習される共感的対話音声合成モデル

齋藤 佑樹[1,a)]　高道 慎之介[1]　飯森 栄治[1]　橘 健太郎[2]　猿渡 洋[1]

概要：本稿では，ChatGPT を活用して対話の文脈情報を自動的に抽出する共感的対話音声合成 (empathetic dialogue speech synthesis: EDSS) の手法である “ChatGPT-EDSS” を提案する．ChatGPT は，入力プロンプトの内容と意図を深く理解し，ユーザからの要求に対して適切に応答可能な最先端の AI チャットボットの 1 つである．我々は ChatGPT の文章読解力に着目し，対話相手の感情を考慮して共感的な音声を生成する EDSS タスクに ChatGPT を導入する．提案法である ChatGPT-EDSS では，まず ChatGPT に対話履歴のテキストをプロンプトとして与え，各話者の発話に対して意図，感情，発話スタイルを表現する 3 つの語（ChatGPT 文脈語）を回答させる．次に，得られた文脈語の word embedding で deep neural network (DNN) ベースの EDSS モデルを条件付けして学習し，ChatGPT 由来の文脈語で韻律を制御可能な音声合成を実現する．実験的評価の結果から，人手でアノテーションされた感情ラベルや，対話履歴から DNN で抽出された文脈情報で EDSS モデルを条件付けする従来法と同程度の合成音声品質を提案法が達成できることを示す．本研究で収集した ChatGPT 文脈語は，我々のプロジェクトページ `https://sarulab-speech.github.io/demo_ChatGPT_EDSS/` で公開している．

Yuki Saito[1,a)]　Shinnosuke Takamichi[1]　Eiji Iimori[1]　Kentaro Tachibana[2]
Hiroshi Saruwatari[1]

1. はじめに

対話音声合成 (dialogue speech synthesis: DSS) [1] は，音声対話システムのためのテキスト音声合成 (text-to-speech: TTS) 技術であり，人間–ロボット間での音声コミュニケーションの実現に必要不可欠な構成要素である．テキストに記述された内容を正確に伝達することが主目的である TTS に対し，DSS では，対話の状況（例えばレストランの予約 [2] や対話相手の説得 [3] など）に応じて合成音声の韻律を適切に制御する必要がある．発話意図 [4] や話者の感情 [5] などの対話文脈情報は，このような韻律制御のために DSS モデルを条件付けする特徴量として用いられる．

共感的 DSS (empathetic DSS: EDSS) [6] は，対話相手に共感するような韻律制御が可能な音声対話エージェントの実現を目的とした新興の技術である．テキスト対話における共感的対話生成 [7] と同様に，EDSS モデルは対話文脈情報を用いて対話相手に共感する音声を合成するように学習される．我々の先行研究 [6] は，話者の感情ラベルを文脈情報として用いた EDSS 手法が合成音声の品質を改善することを示している．しかし，この手法はアノテータ（即ち，人間の対話アドバイザー）に共感的対話における話者の発話単位での感情ラベルを付与させるため，アノテータが対話内容を深く理解しなければならない．Deep neural network (DNN) により対話履歴からデータ駆動で文脈情報の潜在特徴表現（文脈埋め込みベクトル）を獲得する学習法 [8] も表現力豊かな対話音声合成を実現できるが，学習された埋め込みベクトルは人間にとって解釈が困難である．

テキスト対話のパラダイムにおいて，最先端の AI チャットボットである ChatGPT (generative pre-trained Transformer)[*1] は，小説執筆や作詞などの多くの創作アプリケーションで目覚ましいブレークスルーをもたらした．Chat-

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

[*1] `https://chat.openai.com/chat`

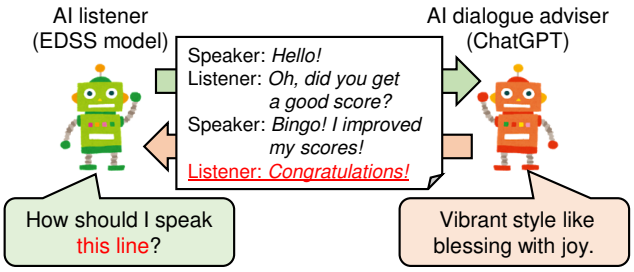

図 1 ChatGPT-EDSS のコンセプト図

GPT の基本構造は GPT-3 [9] に基づいており，人間の嗜好を反映させた応答を生成するための教師あり学習と強化学習 [10] に基づいて fine-tuning し続けられている．この学習機構により，ChatGPT は入力されたテキストプロンプトの意図と内容を深く理解し，ユーザからの要求に対して適切に応答することができる．しかしながら，この優れた文章読解を音声対話研究に活用できるかどうかは十分に調査されていない．

本研究では，ChatGPT を AI の対話アドバイザーとして活用し，合成音声の韻律を制御可能な “ChatGPT-EDSS”（図 1）を提案する．提案法は，ChatGPT を用いて共感的対話のテキスト履歴から各発話に対する 3 つの文脈語（発話意図，感情，発話スタイル）を抽出し，それらの word embedding で EDSS モデルを条件付けして表現力豊かな音声を合成する．本稿では，ChatGPT を用いて共感的対話の文脈語を収集する方法論，ChatGPT 文脈語の収集結果と，提案法による合成音声の品質に関する主観評価を実施した結果を報告する．本研究の貢献は以下のとおりである．

- 音声合成分野，特に，合成音声の適切な発話スタイル制御のために対話の文脈を深く理解する必要がある EDSS タスクにおいて，ChatGPT を導入する方法を世界に先駆けて検討した．
- ChatGPT を利用して有用な文脈語を得るためのプロンプトデザインを提示し，得られた文脈語を分析した．
- EDSS 実験の結果から，ChatGPT 由来の文脈語の word embedding を用いることで，話者の感情ラベルや深層学習由来の文脈埋め込みベクトル [8] で条件付けした EDSS モデルと同程度の合成音声の自然さとスタイル類似性を達成できることを示した．

2. 関連研究

2.1 共感的対話音声合成 (EDSS)

我々の先行研究 [6] では，発話単位の感情ラベルと対話履歴由来の文脈埋め込みベクトルをそれぞれ明示的な・データ駆動な文脈情報として活用する EDSS 手法を提案している．この手法は，conversational context encoder (CCE) [8] が対話履歴のテキスト列（文埋め込みベクトルの系列）から文脈埋め込みベクトルを抽出し，それに基づいて合成音声の韻律を制御することで，FastSpeech 2 [11] ベースの TTS モデルよりも表現力豊かな対話音声が合成できる．

2.2 ChatGPT の有用性

これまでに，多くの研究者が実世界での ChatGPT の能力（教育や評価 [12]）や心の理論 (Theory of Mind) [13] を調査している．また，ChatGPT をテキストによる人間の心的状態評価に導入した先行研究も存在する．例えば，パーソナリティ推定 [14] や感情推定 [15] である．これらの成果は，ChatGPT によって得られたテキスト対話の文脈情報を EDSS 技術に導入する動機となっている．

2.3 テキストプロンプトからのメディア生成

拡散確率モデル [16] などの深層生成モデリング技術の進展に伴い，テキストプロンプトからのメディア生成が広く研究されている．DaLL-E [17] は入力されたテキストプロンプトから写実的な画像を生成する初期のモデルの 1 つである．GPT-3 [9] は，プロンプトとして与えられた初期テキストから連続した文章を生成し続けることができる自己回帰型大規模言語モデル (large language model: LLM) である．AudioGen [18] と MusicLM [19] は，それぞれテキストプロンプトから環境音と音楽を生成するモデルである．これらの技術は，テキストプロンプト中の自然言語記述を変更することで，メディアの生成結果を直感的に制御する方法を提供する．

画像生成やテキスト生成に比べ，テキストプロンプトを用いた TTS 制御の研究は依然として発展途上である．その主な理由として，発話すべきテキストと音声，そして音声を説明するための自然言語記述の三つ組を多く含む，十分に大きなデータセットを構築することの難しさがあげられる．Guo らの研究 [20] では，音声の発話スタイルに関する説明文を専有のアノテータに記述させ，SimBERT [21] を用いてその内容を多様化することでこの難題を解決した．この先行研究により構築されたデータセットには，テキストプロンプトを考慮した TTS モデルの学習に使用できる 15 万件以上のデータが含まれているが，このようなデータセットの構築方法は非常にコストがかかり，スケーラビリティの面で課題がある．

3. ChatGPT-EDSS

図 2 に示すように，提案法である ChatGPT-EDSS は (1) ChatGPT を活用した文脈語の収集と (2) 文脈語で条件付けされた EDSS モデルの学習の 2 ステップで構成される．

3.1 文脈語の収集

ChatGPT に共感的対話の台本を与え，その各発話に対して文脈に関連する語を回答させる．図 2 の左側に示すように，ChatGPT へのテキストプロンプトは (1) 対話状況

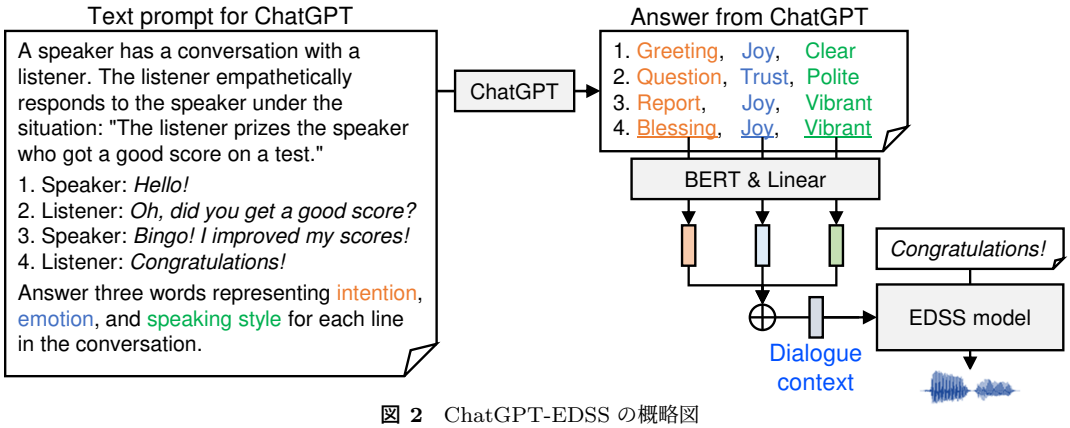

図 2 ChatGPT-EDSS の概略図

の説明，(2) 対話台本，そして (3) 文脈語を生成する指示から構成される．

(1) 対話状況の説明では，各話者の役割などを ChatGPT に伝える．特に，対話開始前の状況を追記することで，回答の文脈としての適切さが改善することを確認した．

(2) 対話台本では対話の内容を発話ごとに区切って記述する．フォーマットは “[ターン ID] [話者名] [発話内容]” の系列とした．ChatGPT による長すぎる対話への回答は途中で停止してしまう傾向にあったため，1 つのプロンプトに含まれる対話ターン数は 5 に制限した．ここで，全体の対話ターンが 6 以上であった場合，直前 2 発話の重複を含めて最大 5 ターンの複数プロンプトに分割した．例えば，10 ターンで終了する対話は，1–5, 3–7, 5–9, そして 7–10 の 4 つのプロンプトに分割される．

(3) 文脈語を生成する指示では，対話の各行に対して (1) 対話意図 [4]，(2) 感情 [5]，そして (3) 発話スタイルを意味する 3 つの単語を回答させるよう ChatGPT に指示した．感情は { 平静，喜び，期待，怒り，嫌悪，悲しみ，驚き，恐怖，信頼 }（即ち，Plutchik [22] により定義された 8 感情に平静を加えたもの）から，発話スタイルは { 可愛い，クール，落ち着いた，丁寧，知的，誠実，爽やか，穏やか，渋い，生き生きした } から回答させるよう指示した．

3.2 ChatGPT 文脈語を用いた EDSS モデル学習

ChatGPT を用いて収集した文脈語を BERT [23] により埋め込みベクトル化し，発話意図，感情，発話スタイルを表す 3 つの語の埋め込みを加算したもので EDSS モデルを条件付けして学習した．この学習法は，ChatGPT をインタラクションが可能な対話文脈ベクトル抽出器とみなし，従来の研究で用いられていた CCE を置換したものと解釈できる．

3.3 考察

ChatGPT-EDSS は，テキストから表現力豊かな発話スタイルを予測する TTS 手法である text-predicted global style tokens (TP-GSTs) [24] に関連する．この観点から，提案法は ChatGPT で発話スタイルに関する単語を対話台本から推測し，それを用いて韻律埋め込みを予測する手法であると解釈できる．TP-GSTs はベースラインである Tacotron [25] よりも合成音声の品質を改善するが，EDSS において共感的な発話スタイルの再現に必要不可欠な対話履歴 [6] を考慮できない．しかし，TP-GSTs において用いられている，GST の重み係数をテキストから予測する枠組みは ChatGPT-EDSS でも導入可能であると考えられる．

ChatGPT-EDSS は，発話単位での話者ごとの感情ラベルの代わりに ChatGPT 由来の文脈語を用いるという観点で，表現力豊かな TTS モデルの弱教師あり学習 [26] としても解釈できる．ChatGPT は与えられたテキストプロンプトに対して不正確な回答を生成することもあるため，本研究で収集した文脈語の信頼性については 4.2 節で議論する．

4. 実験的評価

4.1 実験条件

データセット：日本語の共感的対話音声コーパスである STUDIES コーパス [6] を用いた．STUDIES コーパスは，個別指導塾での講師と生徒による雑談を想定し，講師が生徒に共感して話すような模擬対話音声を含む．STUDIES コーパスの Long-dialogue（10–20 ターン）と Short-dialogue（4 ターン）に含まれる計 2,641 発話から 2,209/221/221 発話の学習/検証/評価セットを構築した．音声データは 22,050 Hz にダウンサンプリングした．

文脈語収集の条件：前述の STUDIES コーパスに含まれる対話データ 2,641 発話に対して ChatGPT で文脈語を収集した．31 名の作業者を雇用し，ChatGPT に対する文脈語回答の依頼と，得られた回答に対する信頼性スコアを 1（“非常に信頼できない”）から 5（“非常に信頼できる”）の整数値で回答させた．作業者は (1) 我々が作成した Google Sheets へのアクセス，(2) シートの第一列に記載されたテキストプロンプトをコピー，(3) ChatGPT の質問入力欄にペーストして送信，(4) 得られた回答をコピーしてシートの第二列にペースト，(5) 回答の信頼性スコアをシートの

表 1　平均された信頼性スコアと各感情カテゴリごとの再頻出文脈語

	信頼性スコア	対話意図	感情	発話スタイル
平静	3.95	問いかけ	期待	落ち着いた
喜び	4.04	祝福	喜び	穏やか
怒り	3.66	**共感**	信頼	丁寧
悲しみ	4.03	**共感**	悲しみ	丁寧

第三列に記入した[*2]．作業者には，(1) ChatGPT が文脈語の生成に失敗した場合，(2) 回答が文脈語以外の表現（例えば，話者名や元の発話など）を含む場合，(3) 日本語で回答されなかった場合にテキストプロンプトを再送信し，ChatGPT に再度回答させるよう指示した．

EDSS の音響モデル：テキストからメルスペクトログラムを生成する音響モデルは，FastSpeech 2 [11] の日本語 TTS 向け PyTorch 実装[*3]を用いた．DNN アーキテクチャや音声パラメータ抽出の設定は，この実装に基づいたものを用いた．F_0 の抽出には WORLD ボコーダ [27], [28] を用いた．Optimizer は Adam [29] であり，学習率 η は 0.0625，β_1 は 0.9，β_2 は 0.98 とした．FastSpeech 2 を JSUT コーパス [30] で 200K iteration 事前学習し，STUDIES コーパスで 100K iteration fine-tuning した．

ニューラルボコーダ：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は，HiFi-GAN [31] の公式 PyTorch 実装[*4]を用いた．HiFi-GAN を FastSpeech 2 の学習データで 350K iteration 学習した．Optimizer は Adam であり，η は 0.0003，β_1 は 0.8，β_2 は 0.99 とした．

4.2　ChatGPT 文脈語の分析

表 1 に STUDIES 講師の感情ラベルで集計した ChatGPT 文脈語の収集結果を示す．まず，信頼性スコアの平均はすべての感情カテゴリに対して 3.6 以上であった．次に，“怒り” と “悲しみ” の発話に対する再頻出の対話意図語は “**共感**” であった．そして，“平静” と “怒り” の発話を除き，再頻出の感情語は，各発話に付与された感情ラベルと一致した．最後に，再頻出の発話スタイル語は “落ち着いた，” “穏やか，” “丁寧” で構成された．これらの結果は，ChatGPT が (1) 確かに “共感的” な対話を理解し，(2) 表現力豊かな TTS のためのある程度信頼できる教師データを提供でき，(3) STUDIES 講師の発話スタイルを概して落ち着いたものであると推定したことを示した．

表 2 にユニークな ChatGPT 文脈語を STUDIES 講師の感情ラベルで集計した結果を示す．まず，対話意図語は非常に多様であり，“平静” 発話に対しては 100 以上のユニークな語が集められた．しかし，その中の 79%が 5 回以下しか出現していないことを確認した．同様の傾向は，事前に

[*2] 信頼性スコアの記入以外のすべての手順は，OpenAI の API により完全に自動化できる．
[*3] https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[*4] https://github.com/jik876/hifi-gan

表 2　感情ラベルごとのユニークな文脈語の数

	対話意図	感情	発話スタイル
平静	206	130	42
喜び	76	35	19
怒り	17	17	8
悲しみ	40	53	19

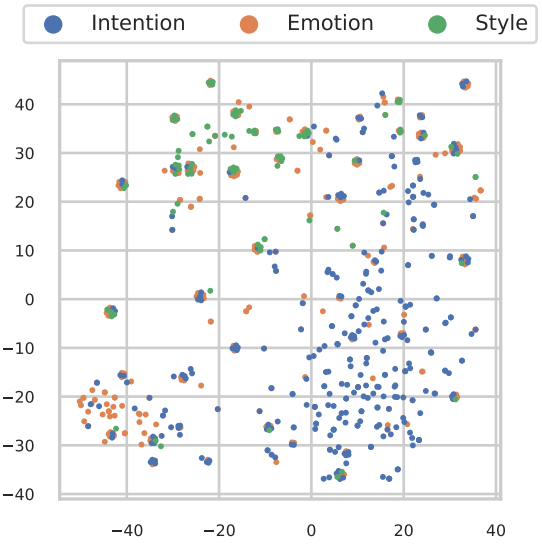

図 3　ユニークな ChatGPT 文脈語の BERT 埋め込みに対する t-SNE プロット

回答のカテゴリを指定したにもかかわらず，感情語と発話スタイル語の収集結果にも観測された．これらの多様性は，図 3 に示す文脈語の BERT embedding を t-SNE で可視化した結果にも確認できるが，カテゴリの違いが異なるクラスタを形成する傾向も示されている．以上より，ChatGPT は (1) 共感的対話の文脈を記述する際に多様な表現形式を考え，(2) 必ずしもテキストプロンプトで事前に定義した回答カテゴリに従わないことを示唆した．

4.3　主観評価

ChatGPT-EDSS が合成音声の自然性を劣化させずに共感的対話の発話スタイルをどの程度再現できるかを調査するための主観評価を実施した．

比較手法：FastSpeech 2 ベースの EDSS モデルを以下の特徴量で条件付けした．

- **Emo**: 人手で付与された感情ラベル
- **CCE**: データ駆動で学習される対話文脈埋め込み [8]
- **IES (ours)**: ChatGPT 由来の対話意図 (Intention),

表 3　MOS 評価結果と 95%信頼区間

Method			MOS	
Emo	CCE	IES **(ours)**	Naturalness	Similarity
✓			3.43±0.14	3.20±0.15
	✓		3.54±0.14	3.24±0.14
		✓	3.52±0.14	3.19±0.15
✓		✓	3.52±0.14	3.21±0.14
✓	✓		3.43±0.14	3.24±0.14
	✓	✓	3.49±0.14	3.20±0.14

感情 (Emotion)，発話スタイル (Style) を表す文脈語 **CCE** は話者 ID の one-hot ベクトルと，最大 4 発話までの 768 次元の sentence BERT[*5]による埋め込みベクトル（現在の発話と過去 3 つの発話）から 256 次元の対話文脈ベクトルを抽出した．BERT による文脈語の word embedding を 256 次元に射影するための全結合層を用意し，768 次元のベクトルを 256 次元の特徴空間に射影した．3.1 節で説明したように，1 つの発話に複数の文脈語が付与されることがあった．その際，複数の文脈語に対する word embedding を平均したもので EDSS モデルを条件付けした．

評価基準：合成音声の自然性と発話スタイル再現性に関する 5 段階 MOS 評価を実施した．自然性の評価では，評価者はランダムな順番で提示された 30 個の音声サンプルに対し，その自然性を 1 (“非常に不自然”) から 5 (“非常に自然”) の 5 段階で回答した．類似性の評価では，評価者はまず HiFi-GAN による分析再合成音声（参照音声）を聴き，その後に再生される合成音声の発話スタイルが参照音声とどの程度類似しているかを 1（“非常に似ていない”）から 5（“非常に似ている”）の 5 段階で回答した．評価者の数は 50 ずつ（合計 100 名）であり，クラウドソーシングによる評価システムで主観評価を実施した．

評価結果：表 3 に主観評価結果を示す．**IES** を EDSS モデルの条件付け特徴量として用いることで，**Emo** もしくは **CCE** を用いた場合に匹敵する合成音声品質が達成できている．即ち，ChatGPT を対話文脈情報の抽出器として利用でき，人間が付与した感情ラベルやデータ駆動で学習された文脈埋め込みベクトルを ChatGPT 文脈語で代替できる可能性を示唆した．また，**Emo** 単体もしくは **Emo** と **CCE** を組み合わせて条件付けした EDSS モデルはわずかに自然性 MOS を劣化させるが，**Emo** と **IES** の両方で条件付けすることで自然性と類似性がどちらも改善した．この理由として，表 2 に示すような多様な感情カテゴリが，事前に定義した 4 感情だけよりも高い表現力を持ち，合成音声の品質改善に結びついたことが考えられる．

ChatGPT 文脈語を導入することの影響をさらに詳細に調査するために，(1) **IES&Emo** と **Emo**，(2) **IES&CCE** と **CCE** の MOS 改善量を計算し，感情ラベルごとにその結果を集計した．図 4 に，横軸を信頼性スコアとした可視化結果を示す．結果から，信頼性スコアと MOS 改善量には相関が無く，信頼性スコアが 5 であっても改善量に大きなばらつきが観測される．即ち，ChatGPT 文脈語は合成音声品質に悪影響をもたらすことはないが，ChatGPT による回答のばらつきへの対策は検討する必要があることを示唆した．

[*5] https://huggingface.co/koheiduck/bert-japanese-finetuned-sentiment

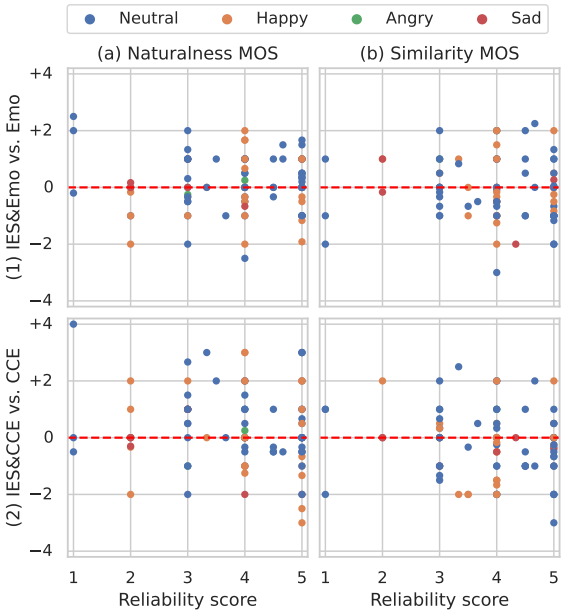

図 4　ChatGPT 回答の信頼性スコアに対する MOS 改善量

5. おわりに

本稿では ChatGPT を活用した対話文脈モデリング手法である ChatGPT-EDSS を提案し，人間が付与した感情ラベルや，DNN 由来の対話文脈埋め込みベクトルを用いる場合と同程度の合成音声品質を達成できることを示した．今後は，提案法における対話ドメインや，ChatGPT の hallucination が及ぼす影響を調査する．

謝辞　本研究は，LINE 株式会社と東京大学 猿渡・高道研究室の共同プロジェクトとして実施したものです．

参考文献

[1] Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D., Lastow, B. and Touati, P.: Speech synthesis in spoken dialogue research, *Proc. EUROSPEECH*, Madrid, Spain, pp. 1169–1172 (1995).
[2] Kim, S. and Banchs, R. E.: R-cube: a dialogue agent for restaurant recommendation and reservation, *Proc. APSIPA ASC*, Siem Reap, Cambodia (2014).
[3] Hiraoka, T., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Learning cooperative persuasive dialogue policies using framing, *Speech Communication*, Vol. 84, pp. 83–96 (2016).
[4] Hojo, N. and Miyazaki, N.: Evaluating intention communication by TTS using explicit definitions of illocutionary act performance, *Proc. INTERSPEECH*, Graz, Austria, pp. 1536–1540 (2019).

[5] Polzin, T. S. and Waibela, A.: Emotion-sensitive human-computer interfaces, *Proc. ITRW on Speech and Emotion*, Newcastle, Northern Ireland, U.K. (2000).
[6] Saito, Y., Nishimura, Y., Takamichi, S., Tachibana, K. and Saruwatari, H.: STUDIES: Corpus of Japanese empathetic dialogue speech towards friendly voice agent, *Proc. INTERSPEECH*, Incheon, South Korea, pp. 5155–5159 (2022).
[7] 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).
[8] Guo, H., Zhang, S., Soong, F. K., He, L. and Xie, L.: Conversational End-to-End TTS for Voice Agent, *Proc. SLT*, Shenzhen, China, pp. 403–409 (2021).
[9] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D.: Language models are few-shot learners, *Proc. NeurIPS*, Vancouver, Canada (2020).
[10] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. and Lowe, R.: Training language models to follow instructions with human feedback, *Proc. NeurIPS*, New Orleans, U.S.A. (2022).
[11] 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).
[12] Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V. and Eger, S.: ChatGPT: A Meta-Analysis after 2.5 Months, *arXiv*, Vol. arXiv:2302.1379 (2023).
[13] Kosinski, M.: Theory of Mind May Have Spontaneously Emerged in Large Language Models, *arXiv*, Vol. arXiv:2302.02083 (2023).
[14] Rao, H., Leung, C. and Miao, C.: Can ChatGPT Assess Human Personalities? A General Evaluation Framework, *arXiv*, Vol. arXiv:2303.01248 (2023).
[15] Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M. and Yang, D.: Is ChatGPT a General-Purpose Natural Language Processing Task Solver?, *arXiv*, Vol. arXiv:2302.06476 (2023).
[16] Ho, J., Jain, A. and Abbeel, P.: Denoising Diffusion Probabilistic Models, *Proc. NeurIPS*, Vancouver, Canada (2020).
[17] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I.: Zero-shot text-to-image generation, *Proc. ICML*, Virtual Conference, pp. 8821–8831 (2021).
[18] Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y. and Adi, Y.: AudioGen: Textually Guided Audio Generation, *Proc. ICLR*, Kigali, Rwanda (2023).
[19] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N. and Frank, C.: MusicLM: Generating Music From Text, *arXiv*, Vol. abs/2301.11325 (2023).
[20] Guo, Z., Leng, Y., Wu, Y., Zhao, S. and Tan, X.: PromptTTS: Controllable Text-to-Speech with Text Descriptions, *arXiv*, Vol. abs/2211.12171 (2022).
[21] Su, J.: SimBERT: Integrating Retrieval and Generation into BERT, Technical report (2020).
[22] Plutchik, R.: *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, New York: Academic (1980).
[23] 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).
[24] Stanton, D., Wang, Y. and Skerry-Ryan, R.: Predicting Expressive Speaking Style from Text in End-To-End Speech Synthesis, *Proc. SLT*, Athens, Greece, pp. 595–602 (2018).
[25] 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, *arXiv*, Vol. abs/1703.10135 (2017).
[26] Zhou, Z.-H.: A brief introduction to weakly supervised learning, *National Science Review*, Vol. 5, No. 1, pp. 44–53 (2018).
[27] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE Transactions on Information and Systems*, Vol. E99-D, No. 7, pp. 1877–1884 (2016).
[28] Morise, M.: D4C, a band-aperiodicity estimator for high-quality speech synthesis, *Speech Communication*, Vol. 84, pp. 57–65 (2016).
[29] Kingma, D. and Jimmy, B.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
[30] 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).
[31] Kong, J., Kim, J. and Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, *Proc. NeurIPS*, Vancouver, Canada (2020).