対話履歴の LLM 埋め込みを用いた音声合成のスタイル制御

小島淳嗣 藤田雄介 水本智也 吉川克正
SB Intuitions 株式会社
atsushi.kojima@sbintuitions.co.jp

概要

対話履歴に基づき、合成音声のスタイルを制御する手法を検討する。対話履歴を考慮するため、提案モデルは、大規模言語モデルによって得られた対話履歴の埋め込みに基づき音声を合成する。実験では、Emotional Speech Database を用いて擬似対話履歴付き音声データセットを作成し、モデルを学習した。実音声と合成音声の発話スタイルの類似度を 5 段階で主観評価し、提案モデルに対話履歴を入力せずに音声を合成する手法と性能を比較した。その結果、提案手法は 3.1 ± 0.18、比較手法は 2.3 ± 0.19 となり、対話履歴に基づく発話スタイル制御の有効性が示された。

1 はじめに

音声対話エージェントによる自然な応答の実現を目指し、表現力豊かな音声を合成する、対話音声合成の手法が探索されている [1, 2, 3]。これらの研究では、対話における感情、対話履歴、発話意図 (e.g., 質問、挨拶) に基づき、合成音声のスタイルを制御する。

本研究では、合成音声の発話スタイルを制御するため、これらの要素のうち対話履歴に着目する。対話中の発話スタイルは対話履歴に応じて変化する [4]。例えば、対話履歴の内容が明るい話題だった時の「そうなんだ」は喜びのスタイル、暗い話題だった時の「そうなんだ」は悲しいスタイルでそれぞれ発話されることが考えられる。このように、対話履歴は発話スタイルを決定づける重要な要素の 1 つであり、音声合成においても対話履歴を用いることで、自然な発話スタイルで音声対話エージェントに応答させることが期待できる。

対話履歴に基づいて発話スタイルを制御するため、提案モデルは、大規模言語モデルによって抽出された対話履歴を表す埋め込みを用いて音声を合成する。大規模言語モデルは膨大なテキストデータによって事前学習されており、高い精度で文脈を理解できるため [5]、効率的に、対話履歴のテキストから、文脈を表す隠れベクトルを得られる。この隠れベクトルを利用することで、対話履歴に基づいた、適切なスタイルでの音声合成が期待できる。

提案モデルは対話文脈エンコーダ、話者エンコーダ、離散ユニット抽出器、ユニット HiFi-GAN ボコーダ [6, 7] で構成される。話者エンコーダは、話者の音響情報を表す x-vector [8] を隠れベクトルに変換する。さらに、対話文脈エンコーダは、大規模言語モデルによって抽出された対話履歴を表す埋め込みを隠れベクトルに変換する。HiFi-GAN ボコーダは、これらの隠れベクトルと離散ユニット抽出器によって得られた hidden unit bidirectional encoder representations from Transformers (HuBERT) ユニット系列 [9] を用いて、音声を合成する。

実験では、提案モデルの有効性を評価するため、英語の演技音声のコーパスである Emotional Speech Database (ESD) [10] を用いて擬似対話履歴付き音声データセットを作成し提案モデルの学習を行う。合成音声の自然性を 5 段階評価する mean opinion score (MOS) と、実音声と合成音声の類似度 [1] をそれぞれ 5 段階で評価した結果について報告する。

2 関連研究

対話履歴を用いた対話音声合成モデルとして、gated recurrent unit (GRU) に基づく手法が提案されている [3]。この手法では、対話履歴のテキストを GRU に入力して得られた隠れベクトルを Tacotron2 [11] のエンコーダに補助特徴量として入力することで、合成音声のスタイルを制御する。評価では、自然性や文脈に応じた発話のスタイル制御といった観点で、オリジナルの Tacotron2 を上回る性能を達成している。一方、この対話履歴を表す隠れベクトルを抽出するエンコーダは、Tacotron2 の重みとともにフルスクラッチで学習されており、大規模言語モデルといった自己教師あり学習モデルの活用は検討さ

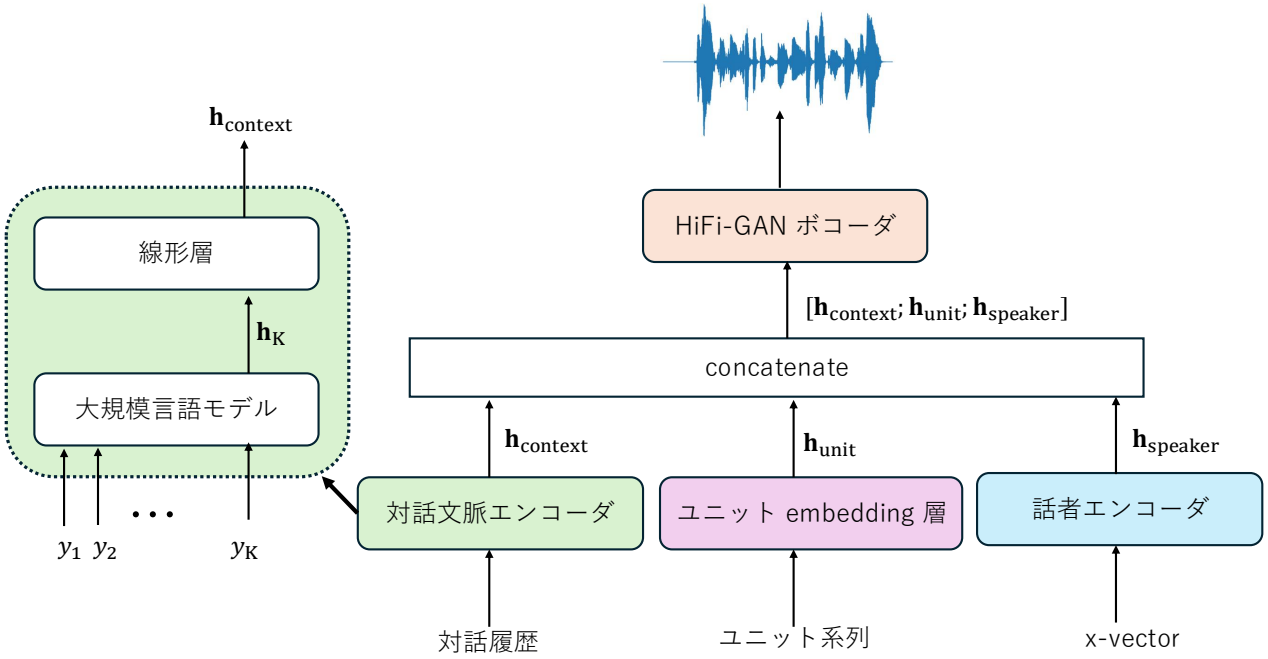

図 1　提案モデルの概要

れていない。

3　対話履歴の LLM 埋め込みを用いた音声合成のスタイル制御

提案モデルは、大規模言語モデルによって抽出された対話履歴を表す隠れベクトルを用いて、合成音声のスタイルを制御する。提案モデルの概要を図 1 に示す。提案モデルは、対話文脈エンコーダ、話者エンコーダ、離散ユニット抽出器、HiFi-GAN ボコーダから構成される。この図において、$\mathbf{h}_{\text{context}}$ と $\mathbf{h}_{\text{speaker}}$ は、それぞれ対話文脈と話者の音響情報を表す隠れベクトルを表す。対話文脈エンコーダは、対話履歴を表す隠れベクトル $\mathbf{h}_{\text{context}}$ を抽出する。y は、対話履歴を表すテキストのトークン、k はトークン数をそれぞれ示す。話者エンコーダは、事前学習された話者ベクトル抽出器によって抽出された x-vector を隠れベクトル $\mathbf{h}_{\text{speaker}}$ に変換する。HiFi-GAN ボコーダは、2 つのベクトル $\mathbf{h}_{\text{context}}$、$\mathbf{h}_{\text{speaker}}$ と、離散ユニット抽出器によって得られた HuBERT ユニット系列から音声を合成する。

対話文脈エンコーダは、事前学習された大規模言語モデルと線形層で構成される。大規模言語モデルは、直前 N ターン分の対話をテキストとして受け取り、ロジットを計算する。線形層は、最後のステップ k のロジットを対話履歴を表す隠れベクトルに変換する。対話文脈エンコーダへの入力の例を図 2 に示す。入力フォーマットは、1 発話につき、話者を示すアルファベットと発話内容のペアで構成される。この例では、話者 A と話者 B が対話を行う形式で、発話内容が対話履歴として入力される。話者に割り当てるアルファベットは対話セッションごとに設定する。なお、$N = 0$ の時は、対話文脈エンコーダへの入力は、ランダムに話者を表すアルファベットを 1 つ選択し、対話履歴の内容を空にする (e.g., C: 、 J:)。

```
### Current context:
===
A: Hey! Have you noticed how quiet it's been lately?
   It feels like we have some peace now.
B: Yeah, definitely quieter than usual.
   I guess all those construction sounds are gone for good.
A: True that! But hey, at least there aren't any loud noises
   bothering us anymore.
   Haha, seems oddly satisfying to me right about now.
===
```

図 2　対話文脈エンコーダの入力例

提案モデルの学習方法について述べる。対話文脈エンコーダにおいて、入力された対話のターン数が多いほど、文脈理解の難易度が上がり、合成音声のスタイル制御が困難になると考え、2 ステップのカリキュラム学習を導入する。1 つ目のステップでは、音声合成のための読み上げ発話を対話履歴が存在しないデータ (i.e., $N = 0$) とみなして、モデルを学習する。2 つ目のステップでは、対話音声データを用いて、対話文脈エンコーダに対話履歴を入力することで、1 つ目のステップで学習されたモデルを継続学習する。さらに、2 つ目のステップでは、対

話文脈エンコーダが発話スタイルに影響を与える主要な感情を抽出できるように、隠れベクトルを用いて感情を識別するタスクを追加し、generator の損失関数に組み込む。式 (1) に generator の損失関数を示す。λ_{mel} は、合成音声のメルスペクトログラムと実音声のメルスペクトログラムの L1 損失、λ_{feature} は feature matching 損失、λ_{adv} は、敵対的損失、λ_{emotion} は、感情識別タスクにおける交差エントロピー損失を示す。α、β、γ、ζ は、ハイパーパラメータを示す。なお、2 つ目のステップでは、大規模言語モデルの重みに加えて、ユニット embedding 層と話者エンコーダもフリーズする。

$$\lambda = \alpha(\beta\lambda_{\text{mel}} + \gamma\lambda_{\text{feature}} + \zeta\lambda_{\text{adv}}) + (1-\alpha)\lambda_{\text{emotion}}. \quad (1)$$

4 実験

4.1 実験条件

モデルの学習、評価に用いるデータセットについて述べる。1 つ目のステップでは、LibriTTS-R [12] から無作為に選択した 100 話者による読み上げ音声を対話履歴が存在しない (i.e., $N = 0$) 発話とみなしてモデルの学習に使用した。このデータセットの音声は、24 kHz で録音されており、学習時には 16 kHz にダウンサンプリングした。2 つ目のステップでは、英語演技感情音声コーパスである ESD を用いて作成した擬似対話履歴付き音声データセットをモデルの学習に使用した。このデータセットは、10 人の話者が 1750 発話を 5 通りの感情 (怒り、喜び、平常、悲しみ、驚き) で読み上げたデータで構成されており、16 kHz で録音されている。このコーパスから擬似音声対話データを作成するため、各発話に付与された感情ラベルと発話内容、対話の最大ターン数 N に基づき対話履歴を生成するように大規模言語モデルに指示することで、擬似的な対話履歴を生成した。具体的なプロンプトと推論パラメータ、生成に用いた大規模言語モデルについて Appendix A に示す。また、生成された対話履歴の例を Appendix B に示す。この例は、「Clear than clear water.」という発話内容を用いて生成された喜びと驚きに対応する対話履歴を示す。

提案モデルのアーキテクチャについて述べる。HiFI-GAN ボコーダにおける generator のユニット embedding 層の出力次元は 128 、upsampling 層には、転置畳み込みに基づく手法を用いた。upsampling factor は [10, 6, 4, 2]、upsampling kernel size は [16, 12, 14, 3] とした。discriminator の kernel と stride はそれぞれ 5 と 3 とした。ユニット系列抽出のために、LibriSpeech [13] を用いて学習された HuBERT の隠れベクトルを特徴量として、500 クラスの k-means クラスタリングモデルを学習した。対話文脈エンコーダの大規模言語モデルには、Phi-1.5[1] を利用し、2048 次元のロジットを線形層により 256 次元に変換する。対話文脈エンコーダに入力する対話の最大ターン数 N は 5 とした。話者エンコーダは、VoxCeleb [14] によって事前学習された話者ベクトル抽出器[2] を利用して 512 次元の x-vector を抽出し、線形層により 32 次元に変換する。

提案モデルの学習条件について述べる。提案モデルにおける損失関数のハイパーパラメータは、$\alpha = 0.9$、$\beta = 45$、$\gamma = 0.5$、$\zeta = 2$ とした。これらのパラメータは予備実験に基づき設定した。音声のログメルスペクトログラムは、FFT サイズを 1024 サンプル、フレームシフトを 160 サンプル、周波数のビンの数を 128 として計算した。感情識別の補助タスクでは、ESD に付与された 5 つの感情を識別するように学習した。また、ユニット系列は、HuBERT と k-means クラスタリングモデルを用いて実音声から抽出した。学習時のターゲットは、窓幅 16000 サンプル、フレームシフト 8000 サンプルとして、音声波形をチャンクに分割することで用意した。提案モデルの学習時の詳細なハイパーパラメータは、Appendix C に示す。

比較手法には、提案モデルにおいて対話文脈エンコーダに対話履歴を入力せず (i.e., $N = 0$) に音声を合成する手法を用いた。評価では、合成音声の自然性を 5 段階で評価する MOS 評価に加えて、実音声と合成音声の発話スタイルの類似度を 5 段階で主観評価した。MOS に関しては、1 を非常に不自然、5 を非常に自然とした。類似度に関しては、1 を非常に似ていない、5 を非常に似ているとして評価を行った。被験者は 3 名、サンプルは、擬似対話履歴付き音声データセットから選択された 10 発話とした。また、参考までに、提案モデルのアーキテクチャによって自然性に劣化が生じていないかを確認するため、LibriTTS-R を用いて学習された、対話文脈エンコーダなしの HiFi-GAN ボコーダとも性能を比較した。

1) microsoft/phi-1_5
2) speechbrain/spkrec-xvect-voxceleb

4.2 結果

表 1 に提案モデルにおける対話文脈エンコーダの MOS と類似度への影響を示す。類似度に関しては、対話文脈エンコーダに対話履歴を入力する時 (exp2) が 3.1 ± 0.18、入力しない時 (exp1) が 2.3 ± 0.19 となった。また、MOS に関しては、対話文脈エンコーダに対話履歴を入力する時 (exp2) が 3.4 ± 0.09、入力しない時 (exp1) が 3.5 ± 0.09 となった。この結果から、提案モデルの対話文脈エンコーダによって、対話履歴の内容に基づいて合成音声の発話スタイルを制御できていることがわかった。また、自然性に関しても音声品質の大幅な劣化は見られないことがわかった。

表 1 対話文脈エンコーダの MOS と類似度への影響 (95% 信頼区間付き)

ID	対話履歴の入力	MOS (↑)	類似度 (↑)
exp1		3.5 ± 0.09	2.3 ± 0.19
exp2	✓	3.4 ± 0.09	3.1 ± 0.18

表 2 に提案モデルと対話文脈エンコーダなしの HiFi-GAN ボコーダの MOS の比較結果を示す。提案モデル (exp2) が 3.4 ± 0.09、対話文脈エンコーダなしの HiFi-GAN ボコーダ (exp0) が 3.6 ± 0.07 となった。さらに、両側 t 検定を実施した結果、有意差は確認されなかった（$p > 0.05$)。この結果から、提案したアーキテクチャによる合成音声品質の有意な劣化は認められないと言える。

表 2 対話文脈エンコーダなしの HiFi-GAN ボコーダと提案モデルの比較 (95%信頼区間付き)

ID	Model	MOS (↑)
exp0	HiFi-GAN	3.6 ± 0.07
exp2	+ 対話文脈エンコーダ	3.4 ± 0.09

4.3 対話文脈エンコーダによって出力された隠れベクトルの分析

図 3 は、対話文脈エンコーダによって出力された隠れベクトルの分析結果の例を示す。この分析では、t-distributed stochastic neighbor embedding [15] によって対話文脈エンコーダの隠れベクトルを 2 次元に変換することで、感情ごとの分布を可視化した。感情ごとの隠れベクトルの数はそれぞれ 800 とし、テストデータから無作為に選択した。この図から、怒り、悲しみ、喜び、驚きに対応する隠れベクトルは、クラスタごとに分離される傾向があることがわかった。この結果から、対話文脈エンコーダが

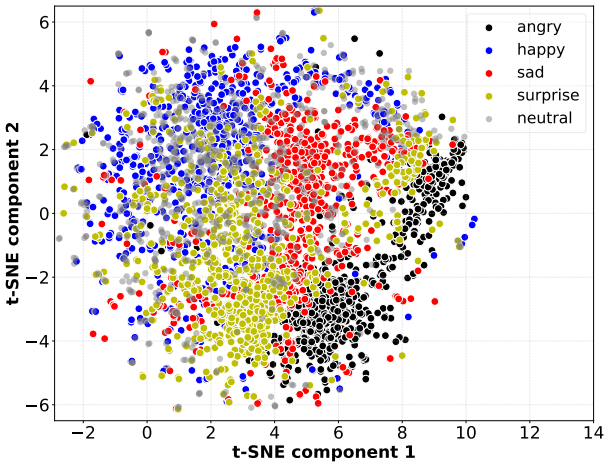

図 3 対話文脈エンコーダによって出力された隠れベクトルの可視化例

入力された対話履歴の内容に応じて、異なる感情を表す隠れベクトルに変換していると考えられる。例えば、怒りと喜びのクラスタは明確に分離される傾向にあり、これらの感情表現が提案モデルにおいて、識別的に学習されていることを示唆している。なお、平常に属する隠れベクトルは、比較的中心位置に分布しており、他の感情カテゴリーとの境界が緩やかであることわかった。これは、他の感情ラベルに属する対話履歴が、特定のキーワード、言い回し、文脈といった表現によって明示的に感情を表現できるのに対して、平常に属する対話履歴は識別的な要素が少ないためだと考えられる。

5 おわりに

大規模言語モデルによって抽出された対話履歴を表す隠れベクトルを用いて、合成音声のスタイルを制御する手法について検討した。提案する対話文脈エンコーダは、大規模言語モデルによって抽出された対話履歴を表す埋め込みを隠れベクトルに変換する。HiFi-GAN ボコーダは、この隠れベクトルに基づき音声を合成する。実験では、ESD を用いて擬似対話履歴付き音声データセットを作成し、提案モデルを学習した。実音声と合成音声との発話スタイルの類似度を 5 段階で主観評価し、対話文脈エンコーダに対話履歴を入力せずに音声を合成する手法と性能を比較した。その結果、提案手法は 3.1 ± 0.18、比較手法は 2.3 ± 0.19 となり、対話履歴に基づく発話スタイル制御の有効性が示された。今後は、提案したカリキュラム学習や感情識別の補助タスクの有効性について調査する。

参考文献

[1] Y. Saito, S. Takamichi, E. Iimori, K. Tachibana, H. Saruwatari, "ChatGPT-EDSS: Empathetic Dialogue Speech Synthesis Trained from ChatGPT-Derived Context Word Embeddings," in *Proc. INTERSPEECH*, 2023.

[2] G. Bruce, B. Granström, M. Filipsson, K. Gustafson, M. Horne, D. House, B. Lastow, P. Touati, "Speech Synthesis in Spoken Dialogue Research," in *Proc. EUROSPEECH*, 1995.

[3] H. Guo, S. Zhang, F. K. Soong, L. He, L. Xi, "Conversational End-to-End TTS for Voice Agent," in *Proc. SLT*, 2021.

[4] D. Wilson and D. Sperber, "Meaning and Relevance," Cambridge University Press, 2012.

[5] Y. Zhu, J. R. A. Moniz, S. Bhargava, J. Lu, D. Piraviperumal, S. Li, Y. Zhang, H. Yu, B.-H. Tseng, "Can Large Language Models Understand Context?," in *Proc. EACL*, 2024.

[6] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Proc. INTERSPEECH*, 2021.

[7] J. Kong, J. Kim, J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, 2020.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP*, 2018.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, Oct. 2021.

[10] K. Zhou, B. Sisman, R. Liu, H. Li, "Seen and Unseen Emotional Style Transfer for Voice Conversion with a New Emotional Speech Dataset," in *Proc. ICASSP*, 2021.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *Proc. ICASSP*, 2018.

[12] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, A. Bapna, "LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus," in *Proc. INTERSPEECH*, 2023.

[13] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *Proc. ICASSP*, 2015.

[14] A. Nagrani, J. S. Chung, A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. INTERSPEECH*, 2017.

[15] L. Maaten and G. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

A　対話履歴生成の条件

対話履歴の生成に用いたプロンプトを図 4 に示す。

```
Instruct: Given a final utterance and emotion,
create a natural short casual dialogue script before final utterance.
OUTPUT SHORT UTTERANCE.
The dialogue should be in A: and B: format with {max_turns} turns.
 Output only utterance content. DO NOT OUTPUT SITUATION EXPLAINING AND
THINKING AND SYMBOL.

Final utterance: {utterance}
Speaker emotion in Final utterance: {emotion}

Output the dialogue history:

Response:
```

図 4　対話文脈エンコーダの入力例

モデルは、meta-llama/Meta-Llama-3-8B-Instruct を利用し、推論アルゴリズムは、temperature sampling を採用し、temperature と max_sequence はそれぞれ 0.1、512 とした。さらに、冗長な繰り返しを避けるため、repetition penalty を 1.2 とした。また、生成する対話履歴の最大ターン N は 5 とした。

B　生成された対話履歴の例

生成された対話履歴の例を表 3 に示す。

表 3　生成された対話履歴の例 (対象発話は、Clear than clear water.)

感情	対話例
喜び	A: Have you seen this new spring? It's amazing how pure it is today! B: Yeah, I noticed that too! The clarity of the water really stands out compared to last weekend. A: Absolutely, nothing like seeing such crystal-clear streams; makes me feel so refreshed just looking at them.
驚き	A: Did you just see that? B: Yeah, it's like looking through crystal! A: I can hardly believe how transparent this pond is today. B: It must have rained recently; everything seems so fresh now. A: Clear as daylight indeed—it's quite surprising to witness such clarity here of all places.

C　提案モデルの学習時のハイパーパラメータ

提案モデルの学習パラメータを表 4 に示す。

表 4　提案手法の学習パラメータ

Parameter	Value
learning rate	10×10^{-4}
optimizer	Adam
warm-up steps	300
training steps	200,000
weight decay	10×10^{-1}
batch size	8