日本語コーパスを用いた混合感情音声合成の性能向上*

○坂田一成，小坂哲夫（山形大学）

1　はじめに

音声合成(Text-To-Speech: TTS)はテキストから音声を合成する技術である。近年の音声合成の研究ではテキストからメルスペクトログラムなどの中間表現を介さずに直接音声合成を行う End-To-End モデルを用いることにより、人間の声と同等の品質の合成音声を生成することが可能となった[1]。この発展により人間の感情表現を模倣する感情音声合成が研究されている。感情に関する理論として、アメリカの心理学者プルチックが提唱した感情の輪による混合感情がある。これは人間の感情は単一のカテゴリに収まらず、8 つの基本感情の混合状態にあるという理論である。そのため人間の感情表現を模倣するには混合感情を考慮した音声合成が重要となる。

我々はこれまで VITS ベースの End-To-End モデルを利用して、日本語を対象とした混合感情音声合成について検討してきた[2]。この研究では、日本語と中国語のゲーム音声で学習された事前学習済みモデルに対し、感情音声コーパスである声優統計コーパス[3]でファインチューニングしたモデルが使用されている。この研究の結果、感情ベクトルの線形結合による混合が、人間の知覚に近い感情制御を可能にすることが示された。しかし、2 ヶ国語で学習した事前学習済みモデルを使用したことでアライメントの精度が十分でなく、合成音声の品質が不十分であった。

そこで本研究では日本語多話者コーパスにより学習された事前学習済みモデルを用いることにより合成音声の品質向上について検討を行う。また、先行研究で検討されていなかった感情の組み合わせについて評価する。

2　混合感情音声合成モデル

2.1 wav2vec2.0 による感情特徴抽出モデル

本研究では感情特徴抽出モデルとして先行研究でも用いられた wav2vec2.0[4]をベースとした w2v2-L-robust-L-12[5]を使用した。wav2vec2.0 とは音声から特徴量を抽出する自己教師あり学習モデルである。w2v2-L-robust-L-12 は英語感情音声データである MSP-Podcast(v1.7)[6]で学習されており、音声から 1024 次元の感情特徴を抽出する。日本語感情音声コーパス JVNV[7]の感情特徴を t-SNE 分析により 2 次元に圧縮した散布図を Fig. 1 に示す。この図から、各感情がある程度分類されていることが確認でき、英語の感情音声データで学習されたモデルが、日本語の感情音声データを用いても感情特徴を適切に捉えていることがわかった。

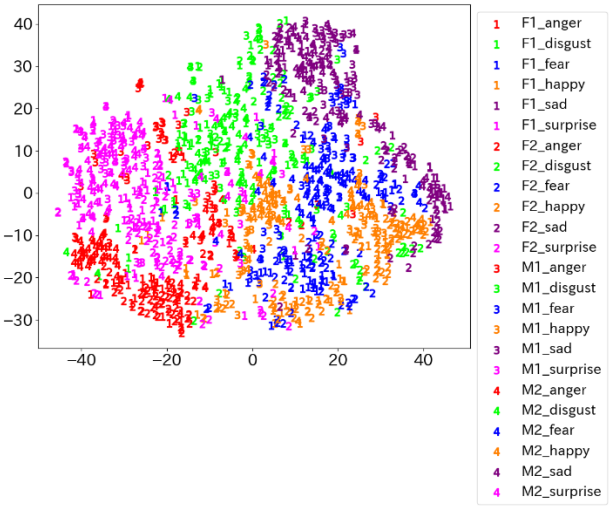

Fig. 1　t-SNE による感情ベクトル

2.2 Emotional-VITS

本研究では先行研究でも用いられた感情音声合成モデルである Emotional-VITS[8]を使用した。Emotional-VITS は End-To-End モデルの TTS である VITS と感情特徴抽出モデルの w2v2-L-robust-12 を統合したモデルである。構成図を Fig. 1 に示す。このモデルでは、w2v2-L-robust-12 から出力される 1024 次元の感情特徴を全結合層で 192 次元に圧縮し、192 次

* Performance Improvement of Mixed Emotion Speech Synthesis Using a Japanese Corpus, by SAKATA, Issei and KOSAKA, Tetsuo (Yamagata Univ.).

元のテキスト特徴とともに VITS のテキストエンコーダに入力する。学習にはテキストと感情特徴、及びその感情特徴に応じた線形スペクトログラムを用い、合成時にはテキストと感情特徴を入力することで、該当の感情特徴を持つ合成音声を出力することができる。

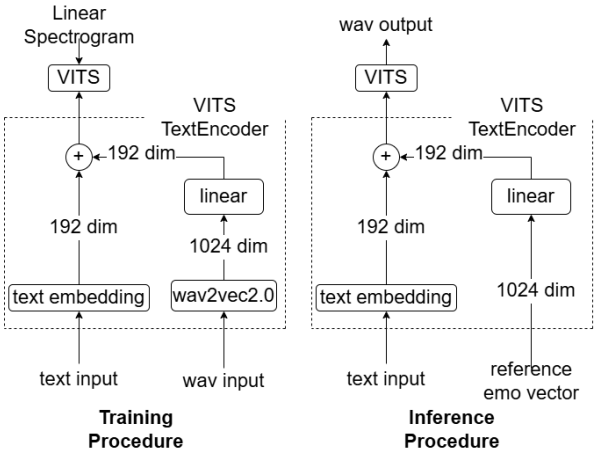

Fig. 2　Emotional-VITS の構成図

3　評価実験

3.1 実験条件

従来モデルの事前学習済みモデルは Hugging Face で公開されている日本語と中国語のゲーム音声 804 名分を使用して学習されたモデル[9]を使用した。このモデルはサンプリング周波数 22.05 kHz の音声で学習されているが、学習データの総量やモデルのステップ数は明らかにされていない。一方、提案モデルの事前学習には、日本語の多話者コーパスである JVS (Japanese Versatile Speech) [10] コーパスを使用した。JVS コーパスは、男女各 100 名の話者による読み上げ音声 130 発話、ささやき声 10 発話、裏声 10 発話で構成されている。本研究では、読み上げ音声 130 発話を学習データと検証データに 9:1 の割合で分割して使用した。サンプリング周波数を 22.05 kHz、バッチサイズを 64、学習率を2×10^{-4}として、40 万ステップの学習を行った。

両モデルの感情音声データによるファインチューニングには、JVNV (Japanese emotional speech corpus with Verbal content and Nonverbal Vocalizations)を使用した。JVNV コーパスは、日本語の感情音声データセットであり、言語音声と非言語音声から構成され、男女計 4 名が怒り、嫌悪、恐れ、喜び、悲しみ、驚きの 6 つの感情を表現する音声が収録されている。本研究では、非言語音声の発話を監督者が設計した「Regular」セッションの発話を、学習データ、検証データ、テストデータに 8:1:1 の割合で分割して使用した。サンプリング周波数を 22.05 kHz、バッチサイズを 64、学習率を2×10^{-4}として、4 万ステップの学習を行った。

3.2 合成音声の品質に関する評価実験

従来モデルと提案モデルの合成音声の品質について主観評価実験を行った。この実験では当該発話の感情ベクトル、本人平均および他人平均の感情ベクトルの 3 種類の感情ベクトルを使用した。ここで本人平均とは話者自身の全ての発話から算出した感情ベクトルの平均のベクトルであり、他人平均とは他の話者全員の発話から算出した感情ベクトルの平均のベクトルのことである。被験者 14 名に音声を聞いてもらい、自然性に関する 5 段階評価の MOS 評価を行った。結果を Table 1 に示す。結果から提案モデルはすべての種類の感情ベクトルにおいて従来モデルより自然性が高いという結果となった。また、5%水準で t 検定を行った結果、従来モデルと提案モデル間に有意差があることが確認された。提案モデルによる合成音声は、特に語尾の発音の明瞭性が向上し、イントネーションが改善された点で優れていた。これは日本語のみの音声データを学習に用いたことによるアライメント精度の向上が影響していると考えられる。以上の結果から、日本語多話者コーパスを学習した提案モデルは、日本語と中国語のゲーム音声で学習された従来モデルよりも自然性が向上することが明らかになった。

Table 1　自然性評価結果

モデル・感情ベクトル		MOS
Ground Truth		4.57
従来モデル	該当発話	1.68
	本人平均	1.74
	他人平均	1.57
提案モデル	該当発話	2.90
	本人平均	3.12
	他人平均	3.04

3.3 1 次感情に関する評価実験

混合感情の合成音声に対して感情分類実験を行う前に、予備実験として混合感情でない 1 次感情の合成音声に対して感情分類実験を行った。提案モデルを用いて JVNV に収録されている 6 つの 1 次感情のみを付与した合成音声がどの感情に聞こえるのかの主観評価を行った。被験者 6 名を対象に 6 つの感情のどれに聞こえるのかの 6 択での評価を行った。この実験では本人平均と他人平均の感情ベクトルを使用した。合成するテキストは「テキスト音声合成は、文字を解析して自然な音声に変換する技術です。」とした。このテキストはテキストから感情を予測できない平文として ChatGPT から生成されたものである。分類精度の結果を Table 2 に示す。この結果から、他人平均の感情ベクトルを使用することで、本人平均を使用した場合と比較して分類精度が向上することがわかった。特に、他人平均の感情ベクトルを用いた場合、喜び、悲しみ、驚きの分類精度が高くなる傾向が確認された。また、混同行列の結果を Fig. 2、3、4 に示す。この結果から怒りの音声を喜びと分類することが多いことがわかった。これは怒りと喜びの音声が音響的に似ていることが原因だと考えられる。また、恐れと悲しみの誤分類が多いことがわかった。

Table 2　感情分類精度

	本人平均	他人平均
怒り	0.333	0.292
嫌悪	0.250	0.375
恐れ	0.375	0.250
喜び	0.625	0.667
悲しみ	0.458	0.500
驚き	0.292	0.458
平均	0.389	0.424

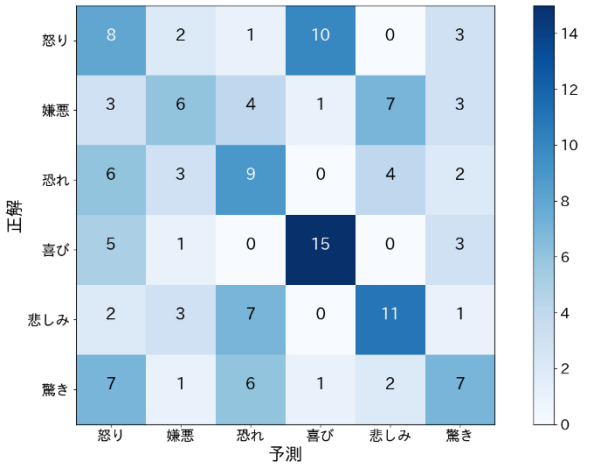

Fig. 3　混同行列（本人平均）

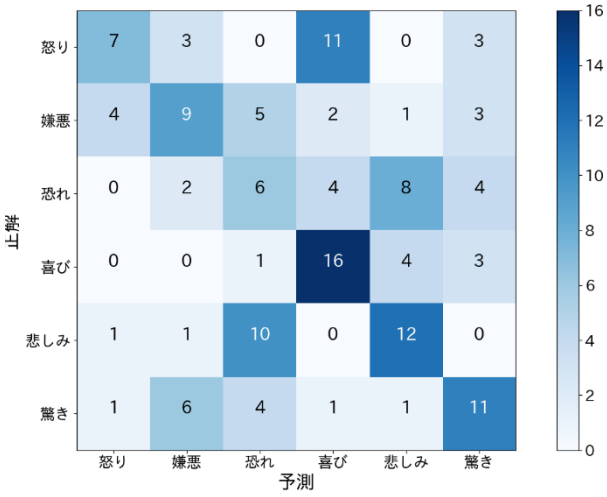

Fig. 4　混同行列（他人平均）

3.4 混合感情に関する評価実験

提案モデルを用いて、混合感情の合成音声がどの種類の感情に聞こえるかについて主観評価実験を行った。この実験では、前章 3.3 の一次感情に関する評価実験の結果を踏まえ、特に分類精度が高かった他人平均の感情ベクトルを使用し、被験者 14 名を対象に喜び、悲しみ、驚きの 3 つの感情を組み合わせた混合感情の合成音声に対して評価を行った。感情を混合する割合は、0:1、0:25:0.75、0.5:0.5、0.75:0.25、1:0 とした。結果を Fig. 5、6 に示す。この結果から、感情の割合を変えることで感情選択の割合も一貫して変化し、提案モデルが正しく感情制御を行えていることがわかった。喜びと驚きの混合では、割合が 0.25:0.75 で感情の選択が逆転しており、悲しみと驚きの混合では割合が 0.5:0.5 で感情の

選択が逆転する傾向が見られた。2 つの感情の割合が 0.5:0.5 のときにそれぞれの感情が同じ確率で選ばれることが理想的だが、喜びと驚きの混合では喜びに偏りがあることがわかった。また、喜びと驚きの混合の割合が 0:1 のときにその他・わからないと選ばれることが多いのは、Fig. 4 に示すように驚きの合成音声が嫌悪や悲しみなどの他の感情と認識していることが原因だと考えられる。

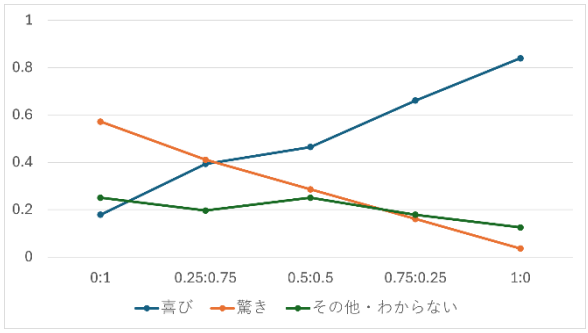

Fig. 5　混合感情に対する感情選択割合（喜びと驚き）

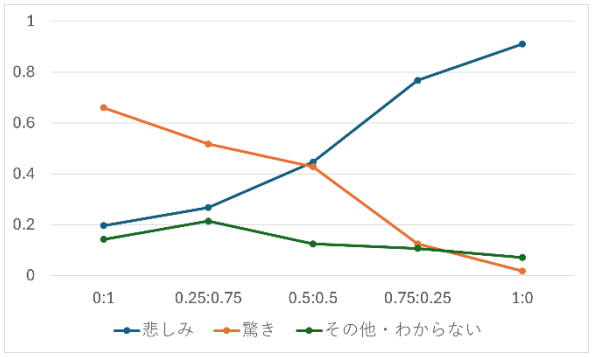

Fig. 6　混合感情に対する感情選択割合（悲しみと驚き）

4　まとめ

本稿では日本語多話者コーパスにより学習された事前学習済みモデルを用いることにより、従来モデルより自然性が向上したことが明らかになった。また、より多数の感情ベクトルを平均したベクトルを用いたほうが合成音声の感情認識精度が高くなり、感情を混合した場合でも正しく感情制御が可能であることがわかった。しかし、2 つの感情を混合した合成音声が混合感情を正しく表現できているかについては明確ではないため、今後は混合感情の合成音声に対する評価手法を検討していく。

参考文献

[1] J. Kim, et al., “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” Proc. ICML 2021, pp. 5530-5540, 2021.

[2] 李天毅, 他, "End-to-End モデルに基づく混合感情の音声合成に関する検討", 情報処理学会第 86 回全国大会, 2024

[3] 日本声優統計学会, https://voice-statistics.github.io/

[4] A. Baevski, et al., “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”, arXiv:2006.11477v3, 2020.

[5] J. Wagner, et al., “Dawn of the transformer era in speech emotion recognition: closing the valence gap”, arXiv: 2203.07378v4, 2023.

[6] MSP-Podcast corpus, https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html

[7] D. Xin, et al., "JVNV: A Corpus of Japanese Emotional Speech with Verbal Content and Nonverbal Expressions," arXiv preprint 2310.06072, Oct. 2023.

[8] Innnky, emotional-vits, https://github.com/innnky/emotional-vits

[9] zomehwh/vits-uma-genshin-honkai https://huggingface.co/spaces/zomehwh/vits-uma-genshin-honkai

[10] S. Takamichi, et al., "JVS corpus: free Japanese multi-speaker voice corpus", arXiv preprint, 1908.06248, Aug. 2019.