5R-06

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

李 天毅† 小坂 哲夫†
山形大学工学部†

1. はじめに

近年 End-to-end の音声合成(TTS:Text-to-speech)として VITS[1]が成功を収めている．VITS は変分オートエンコーダに基づいており感情などの情報による制御が容易と考えられる．この VITS をベースに感情音声合成が提案されている[2-4]．これらのシステムでは wav2vec2.0 を用いて感情特徴を抽出する．本研究では VITS を基に混合感情の音声合成について検討した．深層学習モデルによる混合感情の合成については拡散モデルに基づくもの[5],Encoder-decoder に基づくもの[6]が提案されているが，VITS ベースのシステムや日本語を対象としたものは十分に検討されていない．本研究では感情ベクトルの種類，感情の制御性などについて検討した結果について述べる．

2. 混合感情音声合成モデル

2.1 VITS

VITS(Conditional Variational Autoencoder with Adversarial Learning for End-to-End TTS)は，変分推論を正規化フロー[7]で強化し，敵対的なトレーニングプロセスを組み合わせた高性能なエンドツーエンドの TTS モデルの一つである.条件付き変分オートエンコーダ[8]を使用し，波形領域において敵対的学習を行い，合成音声の品質を向上させている．

2.2 wav2vec2.0 による感情音声ベクトル

w2v2-L-robust-12[9]は MSP-Podcast(v1.7)と呼ばれる英語音声コーパスを用い wav2vec2.0 に基づいてファインチューニングされた感情音声認識モデルであり，各発話から 1024 次元感情ベクトルを抽出することが可能である．日本語音声でも使えることを確認するため，声優統計コーパス[10]から抽出した感情ベクトルを主成分分析により 2 次元に圧縮し散布図を作成した(図1)．平静と喜びの分布は話者に関わらず重なっていることから，これらの感情ベクトルは話者性や言語に影響されないことが分かった．一方怒りの音声は 2 グループに分かれて観測される．音声を確認した結果，話者 tsuchiya は cold anger とよばれる冷たい怒りで発声しているのに対し，他の話者は hot anger で怒りを表現しており，怒りの表現方法が異なることが分かった．

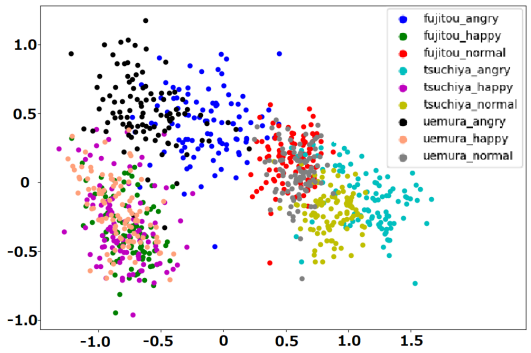

図 1 感情音声ベクトルの散布図

2.3 Emotional-VITS

Emotional-VITS[4]は，GitHub にてオープンソース化された VITS と w2v2-L-robust-12 を統合した TTS モデルである．学習時と合成時の構成図を図 2 に示す．このモデルでは VITS のテキストエンコーダ部分のみを修正し，全結合層を用いて 1024 次元の感情ベクトルを 192 次元まで圧縮する．この圧縮された感情ベクトルはテキストと共にテキストエンコーダに入力される．本研究ではこのモデルを用いて混合感情の検討を行う．

2.4 線形結合による混合感情の表現

アメリカの心理学者プルチックが提唱した感情の輪理論では，人間には 8 つの基本感情が存在し，他のすべての感情はこれらの基本感情の混合状態または派生状態として考えることができると述べている．

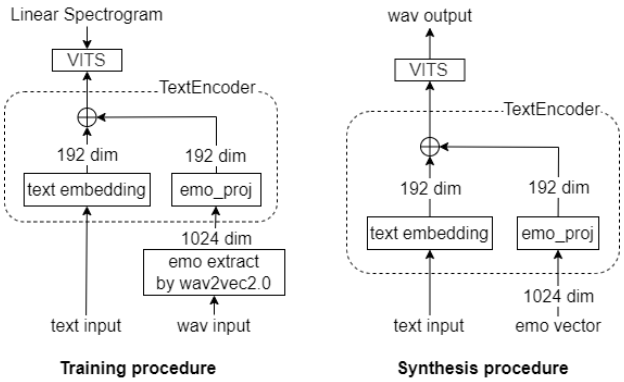

図 2 Emotional-VITS の構成図

Considerations on Mixed Emotion Speech Synthesis Based on E2E Model
†Tianyi LI and Tetsuo KOSAKA, Yamagata University

この考えに基づき本研究では感情ベクトルの線形結合により混合感情が表現できると仮定した．この方法では各感情を線形ベクトルとして扱い，単一または複数の感情を特定の割合で加え合わせることにより，新しい感情を生成する．

3. 評価実験

3.1 データセット

本研究では，声優統計コーパス[10]を使う．プロの女性声優 3 名によって，3 パターンの感情で読み上げられた音素バランス文から成る合計 900 文の音声データセットであり，そのうち学習と評価データはそれぞれ 855 文と 45 文とする．

3.2 感情ベクトルの種類による評価実験

線形結合で使う感情ベクトルの種類について主観評価実験を行った．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比較を行う．評価では 10 名の被験者に音声を聞いてもらい，自然性と音質に関する 5 段階評価の主観評価を実施した．その結果を表 1 に示す．結果から音質については 3 者で有意性は無かった．一方自然性では他人平均が最良で他の 2 方法と比べ 5%水準で有意性があった．以上より他人の感情ベクトルで制御可能なことが分かった．

3.3 感情の種類に関する評価実験

混合感情の合成音がどの種類の感情に聞こえるかの主観評価実験を行った．12 名の被験者を対象に 3 つの感情のどれに聞こえるか，または不明かの 4 択実験を行った．2 感情を混合する割合はそれぞれ 1:0，0.75:0.25，0.5:0.5，0.25:0.75，0:1 とした(図 3～5)．結果としては混合割合が増加すると，他の感情が選ばれる割合が一貫して上昇し，感情ベクトルの混合割合を変えることで人間の感覚に近い感情制御が可能であることがわかった．

4. まとめ

本稿では感情ベクトルの種類および感情の制御性について検討した．その結果，より多数の感情ベクトルを平均したものを使用して合成した音声の自然性が高いこと，線形結合により感情の割合を変えることで人間の感覚に近い感情制御が可能であることがわかった．

表 1　感情ベクトルの種類による MOS 評価

	音質	自然性
当該発話	2.76±0.82	2.93±0.76
本人平均	2.84±0.82	3.02±0.77
他人平均	2.84±0.79	3.13±0.79

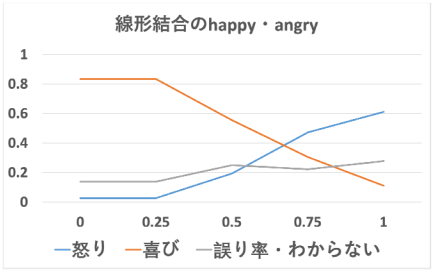

図 3　感情の選択割合（喜びと怒り）

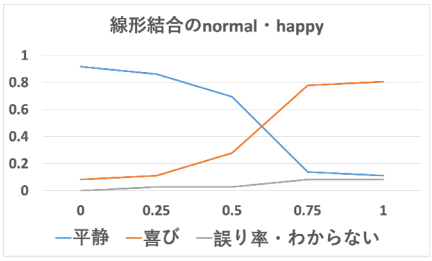

図 4　感情の選択割合（平静と喜び）

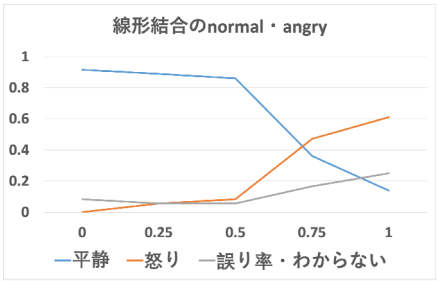

図 5　感情の選択割合（平静と怒り）

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] W. Li, *et al.*, "Emotion transfer with intensity control for text-to-speech synthesis based on self-supervised learning model," ASJ Autumn meeting, 2-Q-32, 2023.
[3] W. Zhao *et al.*, "An Emotion speech synthesis method based on VITS," Appl. Sci. 13(4), 2023.
[4] Innnky, emotional-vits,https://github.com/innnky/emotional-vits
[5] H. Tang, *et al.*, "EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis," Proc. Interspeech2023, pp.12-16, 2023.
[6] K. Zhou, *et al.*, "Speech synthesis with mixed emotions," arXiv:2208.05890, 2022.
[7] D. P. Kingma *et al.*, “Auto-encoding variational Bayes,” Proc. ICLR2014, 2014.
[8] D. J. Rezende *et al.*, “Variational inference with normalizing flows,” Proc. ICML2015, 2015
[9] J. Wagner, *et al.*, “Dawn of the transformer era in speech emotion recognition: closing the valence gap”, arXiv: 2203.07378v4, 2023.
[10] 日本声優統計学会, https://voice-statistics.github.io/