Vol.2023-MUS-137 No.50
Vol.2023-SLP-147 No.50
2023/6/24

DNN音声合成による嫌悪感情の表現と基礎評価

俣野 文義[1,a)]　松井 淑恵[2,b)]　森勢 将雅[1,c)]

概要：既存のテキスト音声合成システムでは，喜びや怒りなど様々な感情表現を制御する機能が備わっている．そこで本研究では，合成音声による新たな感情表現として「嫌悪」の感情音声を生成することを目的とし，既存のテキスト音声合成システムに筆者が発話した嫌悪感情の音声を学習させることで，嫌悪をどの程度表現できるのかを平静音声と比較する実験を実施した．その結果，同様の文脈の文章において7段階中，1段階程度の差が確認された．

1. はじめに

既存のテキスト音声合成システムには，様々な感情表現を制御する機能が備わっている．例として，株式会社エーアイによる商用ソフトウェア AiTalk では，喜び・怒り・悲しみの感情を強度指定して表現することができ [1]，オープンソースソフトウェアとして開発が行われている VOICEVOX では，様々な読み上げ方が「スタイル」として実装されている [2]．これらの機能は，ニコニコ動画[*1]といった動画投稿サイト等において，合成音声による表現の幅を広げるために活用されている．

そこで本研究では，合成音声による新たな感情表現として，ゲーム実況動画等で需要のある嫌悪の感情音声を生成することを目的とした．そのために，まず既存のテキスト音声合成システムに嫌悪感情音声を学習させることで，どの程度の強度で合成音声が嫌悪感情を表現できるかを，聴取による評価実験により調査した．実験では，第一著者が発話した平静感情音声と嫌悪感情音声からそれぞれの合成モデルを作成し，それらから生成された音声に，どの程度嫌悪感情の強度差が生じていたかを7段階で評価した．

2. 感情音声合成の現状と本研究の位置づけ

感情の分類について，Plutchik は「喜びと悲しみ，怒りと恐怖，受容と嫌悪，驚きと期待」を対極的な8つの基本感情とするモデルを提唱した [3]．これらの感情の内，喜び・悲しみ・怒りの表現については，既に様々な研究開発がされている．例えば，波形接続型音声合成において，DNN で感情強度に応じた差分スペクトルを予測し，平静感情の音声素片に対して畳み込む事で，これらの感情音声を生成した先行研究がある [4]．また VOICEVOX には，先に挙げた感情表現ではなく「ささやき」や「あまあま」といった表現が「スタイル」として実装されている [2]．

本研究では，テキスト音声合成における新たな感情表現として嫌悪の合成音声を生成することを目的とした．嫌悪の感情表現は，主に動画コンテンツの制作において需要があると考えている．例えば，料理動画において所謂ゲテモノ素材を使用する際や，ホラーゲーム実況動画において醜悪な外見の敵が出現した際に使用できる．

2023年現在，Tacotron 2 [5] を始めとした高品質な End-to-End（以下，E2E）音声合成が主流となっている．テキスト音声合成による感情表現についても，E2E 音声合成をベースに制御性を拡張するアイデアの元で研究が進められている [6]．一方で，本研究では音声の品質以上に感情の表出を重視するため，合成時のパラメータ操作により，E2E 音声合成と比較して出力波形を容易に制御できる [7] 利点のある，DNN 音声合成を合成手法として採用した．

3. 聴取実験による嫌悪モデルの評価

3.1 音声の収録

実験で用いる音声は，第一著者が表1に示す環境・機材で学習用音声と評価用音声について，それぞれ発話・収録をした．ここで学習用音声とは，合成モデルの学習に用いる音声である．また評価用音声とは，収録した音声が正確に嫌悪感情を表現できているか，及び被験者が嫌悪感情を認識できているかを評価するために，合成音声と同時に聴取実験で使用する音声である．

実験では平静感情の合成音声と嫌悪感情の合成音声を比較する．そのため平静と嫌悪のそれぞれの感情について，

[1] 明治大学
[2] 豊橋技術科学大学
[a)] matano.fumiyoshi.fx@tut.jp
[b)] tmatsui@cs.tut.ac.jp
[c)] mmorise@meiji.ac.jp
[*1] https://www.nicovideo.jp（最終検索日：2023年4月27日）

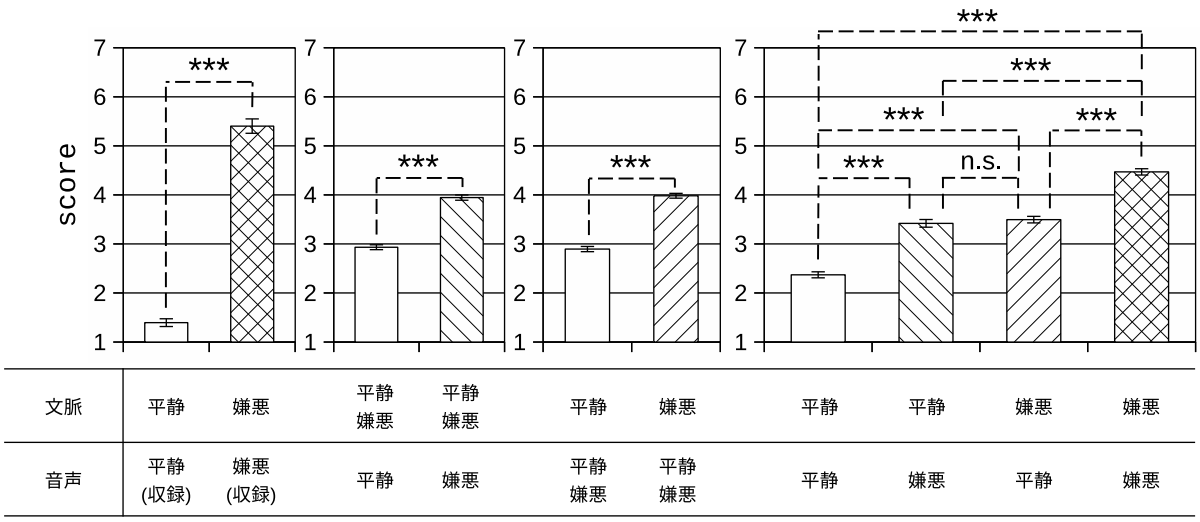

図 1 聴取実験の結果．縦軸は嫌悪強度スコアの平均値を示し，横軸は文脈と音声の組み合わせを示す．エラーバーは 95%信頼区間を示す．

表 1 音声の収録条件

収録環境	ミュージックキャビン SC-3
収録環境の騒音レベル	約 18 dB（A 特性）
収録用マイクロホン	Shure Microflex MX153
A/D 変換器	ZOOM UAC-2
収録ソフト	MathWorks MATLAB R2022a
A/D 変換（学習用音声）	44.1 kHz/24 bit
A/D 変換（評価用音声）	16 kHz/16 bit
発話者とマイクとの距離	約 10 cm

学習用音声を収録した．読み上げに用いた台本は，音素のバランス，読み上げの難易度，Julius [8], [9] によるラベリングの精度を考慮し，JSUT コーパス [10] の内，BASIC5000 の 1 番から 1000 番を使用した．評価用音声は，合成に使用する台本から，平静文 5 文を平静感情で読み上げた 5 発話と，嫌悪文 5 文を嫌悪感情で読み上げた 5 発話を収録した．

平静感情の音声は，いかなる感情も含まれないように発話した．嫌悪感情の音声は，嫌悪感情を表現する際の音響特徴量や表情の変化を調査した先行研究 [11] を参考に，嫌悪の感情が強く表現されるよう努めた．

3.2 合成モデルの学習と合成音声の生成

Python ライブラリである ttslearn [12] に含まれるモジュールの内，DNNTTS を使用して音声の学習と生成を行った．モデルの学習には ttslearn の公式リポジトリで配布されていたスクリプトを用いた．このスクリプトでは学習時間短縮のため，音声を 16 kHz・16 bit にダウンレートして，モデルの学習を行う．嫌悪感情を表現する主要な音響特徴量は 8 kHz を超えないことを示す先行研究 [11] が存在するため，今回のダウンレートによる感情表現の劣化は少ないと考えられる．

今回の実験では合成用の台本として，平静文と嫌悪文を 20 文ずつ，合計 40 文を用意した．平静文は「今日の晩御飯は○○です」の○○に一般的な 20 種類の料理名 [13] を当てはめたもので，嫌悪文は「私の嫌いな食べ物は△△です」の△△に嫌いな人の多い 20 種類の料理名・食材名 [14], [15], [16] を当てはめたものである．これを平静モデルと嫌悪モデルのそれぞれで音声を合成することにより，合成音声による読み上げ音声を合計 80 発話用意した．

表 2 音声評価実験の実験条件

実験環境	オンライン（Pavlovia）
評価者数	男女計 100 名
評価音声数	2 感情 × 40 文（合成音声）＋
	2 感情 × 5 文（評価用音声）= 90 発話
再生機材	被験者所有のイヤホン又はヘッドホン

3.3 実験

実験条件を表 2 に示す．本実験は，lab.js [17] を使用して作成された実験用アプリケーションを，オンライン実験プラットフォームの Pavlovia[*2]上で動作させることにより実施した．被験者はランサーズ[*3]上で募集した．募集対象は，「イヤホン，又はヘッドフォンを使用できる」及び「パソコンで作業できる」の各条件を満たす男女合計 100 名である．実験では被験者に対し，合成音声による読み上げ音声 80 発話と，肉声による読み上げ音声（評価用音声）10 発話の合計 90 文を聴取させ，各音声に対し嫌悪感情の強度を 7 段階で回答させることにより評価をした．

実験前にイヤホン又はヘッドホンの着用についてスクリーニング [18] を実施し，被験者が指示された環境で正しく実験を行っているかを確認した．実験後にはランダムに生成された回答者 ID を実験の募集サイトから回答させる

[*2] https://pavlovia.org（最終検索日：2023 年 4 月 27 日）
[*3] https://www.lancers.jp（最終検索日：2023 年 4 月 27 日）

ことで，同一人物による複数回答を防止した．

3.4 実験結果

スクリーニングの結果，有効な回答であると判断された74名分のデータの平均スコアと，95%信頼区間を図1に示す．

評価用音声に対する回答を対象として，感情を要因とする1要因分散分析（$\alpha = 0.05$）をANOVA君 [19] を用いて実施したところ，有意差が見られた（$F(1,73) = 2391.93$, $p < 0.01$，$\eta^2 = 0.75$）．次に，合成音声に対する回答のみを対象として，音声の感情と台本の文脈を要因とする2要因2水準分散分析を実施したところ，感情と文脈による効果がそれぞれ認められ（音声：$F(1,73) = 1296.77$, $p < 0.01$, $\eta^2 = 0.11$．文脈：$F(1,73) = 1049.11$, $p < 0.01$, $\eta^2 = 0.13$），その差を確認したところ，共に7段階中1段階程度だった．一方で，音声と文脈の交互作用については，有意傾向に留まった．（$F(1,73) = 3.03$，$p = 0.09$, $\eta^2 < 0.01$）．また，音声の感情2種類と台本の文脈2種類の全ての組み合わせについて，Shafferの方法による多重比較を用いて検定したところ（$\alpha = 0.05$），嫌悪音声・平静文脈と平静音声・嫌悪文脈の組み合わせ以外で，有意差が認められた．

4. 考察

評価用音声の実験結果より，今回収録した音声について，平静感情と嫌悪感情の表現が適切に出来ており，また被験者も感情表現の聞き分けができていたと考えられる．次に，作成した合成音声について，平静感情と嫌悪感情の間に有意差が認められたことから，合成音声を用いた嫌悪音声の生成が出来ていた考えられる．しかし，音声の感情による差と文脈による差が同程度であり，嫌悪音声・平静文脈と平静音声・嫌悪文脈の組み合わせの間に有意差が認められなかったことから，今回作成した合成音声による嫌悪感情の表現は，文脈による表現と同程度に留まっていると考えられる．

5. おわりに

本研究では，既存のTTSシステムに平静と嫌悪の各感情音声を学習させることで，それぞれを比較した時に，どの程度嫌悪感情が発露しているのかについて調査した．その結果，同じ文脈内において音声による感情の差は7段階中1段階程度であり，同じ感情の音声で発話した場合の，文脈による差も1段階程度であった．また，嫌悪感情・平静文脈の音声と平静感情・嫌悪文脈の音声の間に有意差が認められなかった．以上から，今回作成した嫌悪感情の合成音声による嫌悪表現は文脈による変化以上の表現は出来ていなかったと考えられる．

今後は，より嫌悪表現を強調するための方法を考える必要がある．そのために，今回生成した各感情の合成音声に対して音響特徴量の解析をすることで，どのような成分が音声に嫌悪感情を付与していたのかについて調査をする必要がある」また，モーフィングを用いることで平静感情と嫌悪感情の間がどのように補間されているかについての調査も実施する必要がある．

謝辞 本研究の一部は，JSPS科研費 JP21H04900, JP21K19794の支援を受けました．

参考文献

[1] 株式会社エーアイ：AITalk® 声の職人 S ｜製品｜音声合成ソフト、読み上げ、人工・電子音声の「株式会社エーアイ（AI）」，（オンライン），入手先 ⟨https://www.ai-j.jp/products/voice/⟩（参照 2023-04-17）．

[2] Kazuyuki, H.: VOICEVOX ｜無料のテキスト読み上げソフトウェア，（オンライン），入手先 ⟨https://voicevox.hiroshiba.jp/⟩（参照 2023-04-17）．

[3] Plutchik, R.: The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, *American Scientist*, Vol. 89, No. 4, pp. 344–350 (2001).

[4] 大谷大和，松永悟之，平井啓之：深層ニューラルネットワークを用いた波形接続型感情音声合成のための感情制御法，技術報告 39 (2019).

[5] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y. and Wu, Y.: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783 (online), DOI: 10.1109/ICASSP.2018.8461368 (2018).

[6] Kwon, O., Jang, I., Ahn, C. and Kang, H.-G.: Emotional Speech Synthesis Based on Style Embedded Tacotron2 Framework, *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pp. 1–4 (online), DOI: 10.1109/ITC-CSCC.2019.8793393 (2019).

[7] 山本龍一，高道慎之介：Pythonで学ぶ音声合成，株式会社インプレス (2021).

[8] Lee, A., Kawahara, T. and Shikano, K.: Julius — An Open Source Real-Time Large Vocabulary Recognition Engine, *In Proc. EUROSPEECH*, pp. 1691–1694 (2001).

[9] Lee, A. and Kawahara, T.: Recent Development of Open-Source Speech Recognition Engine Julius, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 131 – 137 (2009).

[10] 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 (online), DOI: 10.1250/ast.41.761 (2020).

[11] Chong, C. S., Kim, J. and Davis, C.: Disgust expressive speech: The acoustic consequences of the facial expression of emotion, *Speech Communication*, Vol. 98, pp. 68–72 (online), DOI: 10.1016/j.specom.2017.12.007 (2018).

[12] Yamamoto, R., Hayashibe, Y., zzxiang, Shidara, Y. and hrhr49: r9y9/ttslearn: v0.2.2 (2022-01-04) (2022).

[13] Gurunavi: 基本の100レシピ｜シェフごはん，（オンラ

イン），入手先 ⟨https://chefgohan.gnavi.co.jp/base100/⟩（参照 2023-05-10）.
[14] 藤原正光，番場梨彩：子どもの嫌いな食物と克服への支援：大学生の幼児期の回想による調査研究，教育学部紀要，Vol. 48, pp. 113–125（オンライン），入手先 ⟨https://cir.nii.ac.jp/crid/1050282676662737280⟩ (2014).
[15] 鈴木裕子，松野志穂，中野公美子：北海道医療大学生の食物嗜好，北海道医療大学人間基礎科学論集，Vol. 39, pp. A1–15（オンライン），入手先 ⟨https://cir.nii.ac.jp/crid/1050564288202726400⟩ (2013).
[16] 加藤　久, 小林　真：幼稚園児の食物の嗜好について：調理・加工による違いに注目して，富山大学人間発達科学部紀要，Vol. 15, No. 2, pp. 197–204（オンライン），DOI: 10.15099/00020731 (2021).
[17] Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J. and Hilbig, B. E.: lab.js: A free, open, online study builder, *Behavior Research Methods*, Vol. 54, No. 2, pp. 556–573 (2022).
[18] Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J. and Chait, M.: An online headphone screening test based on dichotic pitch, *Behavior Research Methods*, Vol. 53, No. 4, pp. 1551–1562 (2021).
[19] 井関龍太：anovakun version 4.8.7 (2022).