日本音響学会誌 81 巻 1 号（2025），pp. 64–72

論　　文

嫌悪感情を意図して発話された日本語演技音声の音響特徴量分析と話者間比較*

俣野文義*1　小口純矢*1　森勢将雅*2

［要旨］ 著者らは現在ボコーダで音響特徴量を操作することにより，平静音声から嫌悪感情音声を生成することを研究の目標に掲げている。この手法を用いることでテキスト音声合成や，声質変換後の音声への後処理として特徴量を付与することで，嫌悪音声を表現できるようになる。そこで本研究では複数名話者によりコーパス文を読み上げた平静音声と嫌悪演技音声について，音響特徴量に基づく解析・比較を行った。その結果一部話者について基本周波数の低下と，全話者に共通して日本語五母音の第 1，第 2 フォルマントにより表現される 5 角形の面積の減少が確認された。

キーワード　感情音声分析，嫌悪，基本周波数，フォルマント，スペクトル重心
Emotional speech analysis, Disgust, Fundamental frequency, Formant, Spectral centroid

1. はじめに

ニコニコ動画や YouTube といった動画投稿サイトにおいて，テキスト音声合成システム（以下，TTS）やボイスチェンジャー（以下，VC）を利用した動画コンテンツが数多く発信されている。その際に動画投稿者は，シーンに合った音声を生成するためにパラメータ調整機能を活用している。2020 年代に入り，TTS や VC による合成音声の品質は人間に匹敵するようになった。次のステップとして感情表現ができるようになれば表現の幅はさらに広がる。

感情を制御することについての研究は進められており，例えば現在主流となっている End-to-End 音声合成について，感情付与を目的としたモデルの改良が行われている [1, 2]。また，End-to-End ではない TTS に対しても感情表現をするための研究が行われており，感情音声の素片を接続することにより感情音声を合成する研究 [3] や深層学習（以下，DNN）を使用して平静音声の素片に感情パラメータを畳み込む研究 [4]，合成時の特徴量加工が比較的容易 [5] であるパラメトリック音声合成に対して感情成分を外挿する研究 [6] など，様々な手法が提案されている。

ここで著者らは，ボコーダによる DNN を利用しない簡便な特徴量変換により平静音声から嫌悪感情音声を生成することを研究の目標に掲げている。嫌悪感情を表現する TTS や VC を実現することにより，例えばゲーム実況動画において，醜悪な外見の敵が登場するシーンなどに使用する用途への需要を満足する。また，一般的な用途としても嫌悪は忌避の感覚であることから，TTS や VC によって拒絶の意思を示す際，相手に対して直感的に表現することができる。嫌悪は Ekman の六つの感情（驚き，恐怖，嫌悪，怒り，幸福，悲しみ）[7] や Plutchik の基本 8 感情（喜び，悲しみ，怒り，恐れ，受容，嫌悪，驚き，期待）[8] といった感情モデルで表現される基本的な感情である。しかしながら嫌悪感情を素片接続型音声合成で表現した先行研究は存在する [9] ものの，嫌悪感情と音響特徴量の結びつきが明らかにされていないため，ボコーダを用いた決定論的なアプローチによる感情変換はいまだ困難と言える。

嫌悪音声の学習データを用いずにボコーダを用いて TTS や VC に嫌悪感情を与えるためには，嫌悪感情を特徴付ける他の感情と競合しない固有の音響特徴量を明らかにする必要がある。そのため本研究では平静音声と嫌悪演技音声について，それぞれの音響特徴量を解析・比較する。また，その結果を元に，どの成分が強く嫌悪感情の表現に関連しているのかを考察する。

2. 関連研究と本研究の位置付け

感情音声分析の研究は様々な言語・条件の下で行われている。そこで本章では先行研究について述べ，本研究の位置づけを説明する。

---

* Acoustic features analysis of disgust emotion in Japanese play-acted speech,
by Fumiyoshi Matano, Junya Koguchi and Masanori Morise.

*1 明治大学大学院
*2 明治大学
（問合先：俣野文義　〒 164–8525 東京都中野区中野 4–21–1 明治大学大学院先端数理科学研究科 e-mail: matano.fumiyoshi.fx@tut.jp ）

（2024 年 1 月 29 日受付，2024 年 7 月 8 日採録決定）
[doi:10.20697/jasj.81.1_64]

2.1 外国語における嫌悪音声の研究

感情音声の音響的特徴を調査する研究は古くから行われており，例えば 1935 年に発表された Skinner による研究 [10] では，感情が誘導された状態で発話された “Ah” という音声のピッチ等を調査した。その後も時代と共に様々な研究が行われ，感情発露の方法も演技やロールプレイなど多様なものとなっている [11]。

Paeschke らによるドイツ語で実施された研究 [12] では，7 人の役者が六つの感情（Fear, Disgust, Happiness, Boredom, Sadness, Hot anger）で 20 文を読み上げたデータベースについて，基本周波数（F0）に着目した調査が行われた。この研究では平静音声と比較して嫌悪感情の発話は F0 が平均的に高くなり，不自然（Unnatural）な F0 の変動が発生することが認められている。一方 Plutchik の基本 8 感情に強度差を加えた感情の輪 [8] では，Disgust と同系統の，比較的弱い感情として表現されている Boredom（退屈）は，F0 が平均的に低くなり変動の幅も小さくなることが示されている。

2018 年に発表された，広東語で行われた研究 [13] では，嫌悪を表現する表情が，話し声の音響的特徴にどのような影響を与えるかが調査された。この研究では感情誘導された状態の表情と音声のペアデータから構成された the CAVES database [14] を使用している。結果は F0 の上昇とフォルマント周波数 F1, F2 の下降が認められた。考察では嫌悪演技音声においては F0 が下降することに触れつつ，この研究では感情誘導で発露した嫌悪感情であるため，F0 が上昇した可能性が示されていた。また，F1 と F2 の下降から口を閉じて，舌を引き込んだ可能性を示していた。

2.2 日本語における感情音声の研究

日本語の感情音声の研究も様々な感情で行われている。武田らは怒りの感情に着目して，音響特徴量の解析を行った [15]。この研究では怒りを 4 段階に分けて，音声パワー，時間構造，F0 を解析した。坂下らは発話者に対して詳細な設定の台本を提示し，八つの感情（平静，怒り（熱い），怒り（冷たい），嫌悪，悲しみ，喜び，驚き，恐怖）で「はい」と発話させた音声について F0 解析を行うことにより，台本が演技性に与える影響を調査した [16]。その結果，台本にかかわらず嫌悪は冷たい怒りと混同される傾向にあることが示された。また，嫌悪の平均 F0 は平静と比較して低くなることが示されている。

嫌悪感情音声の音響特徴を調査した研究として，重野による研究がある [17]。この研究では，日本人男性俳優 2 名が Ekman の六つの感情で 5 種類の文章を読み上げた 30 発話の音響特徴量を解析した。その結果，嫌悪音声の平均 F0 が最も低くなっていることが認められた。このように嫌悪感情については，研究により F0 の上昇・下降のどちらも観測されている。これは嫌悪表現には言語による差があることや，自然発話か演技発話かによる影響があることを示唆している。

2.3 本研究の位置付け

TTS や VC による表現では，発話者が何を意識して発話しているかが特に重要である。ここで感情音声の発話について，Jürgens らの研究 [18] や Chong らの研究 [13] により，自然な感情表現と演技による感情表現では変化する音響的特徴が異なることが示唆されている。TTS や VC への感情付与を目指す場合は，演技であっても発話者が意図した内容に意味がある。そのため本研究では，嫌悪感情を意図して演技発話をした際に，どのような音響特徴量の変化が現れるかを調査することとした。話者は男女 2 名ずつの合計 4 名とし，どの特徴量が共通して変化し，どの特徴量が個人性にどの程度依存するのかという観点から，嫌悪演技表現時における音響特徴量の変化を分析する。ここで坂下らの研究 [16] により，感情表現の指示の違いで演技性が変化することが示されている。しかし本研究では話者の意図する嫌悪表現を解析することが主な目的であるため，演技方法などの指示は最低限とし，詳細な表現は話者に委ねることとした。

評価対象とする音響特徴量は先に挙げた先行研究より，TTS や VC に応用することを念頭に音色に関連する特徴量として基本周波数（F0），フォルマント周波数（F1, F2）とする。また，嫌悪表現により声の明るさが変動することが予測されるため，スペクトル重心（Spectral centroid）も併せて調査の対象とする。

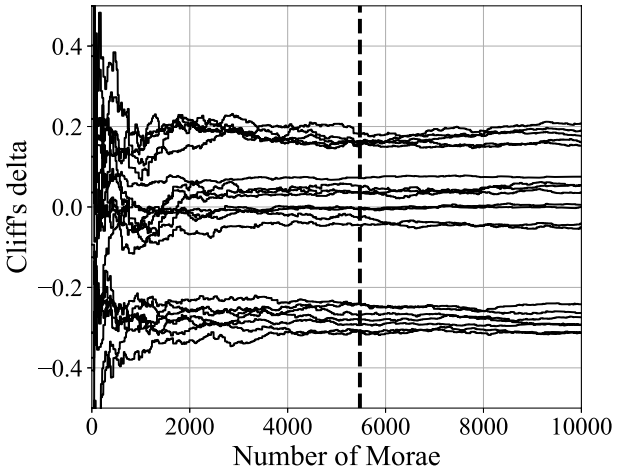
3. 音 声 収 録

既存の日本語感情音声コーパスとして，OGVC [19] や UUDB [20] などが挙げられる。しかし本研究の最終的な目標は嫌悪感情音声の合成であることから，音響特徴量解析ではモーラなどのバランスが保障されており，かつ発話文が全話者で統一されている必要がある。そのため本研究では，モーラのバランスが保証されているコーパス文を利用することとした。

3.1 収録に向けた予備検討

音響特徴量解析する上で，まず読み上げる文章数を決定する必要がある。そのために第一著者が JSUT コーパス [21] の BASIC5000 について，1 番から 1,000 番を平静音声と嫌悪演技音声のそれぞれで読み上げ [22]，モーラ数ごとに本研究で調査する音響特徴量の効果量（Cliff’s δ [23]）を求めた。

Cliff’s δ は二つの標本がどの程度異なるかを -1 か



図–1 収録モーラ数に対する平静音声と嫌悪演技音声の音響特徴量差を示す効果量（Cliff's δ）
各線は解析する音響特徴量であるが，本図ではモーラ数と各種音響特徴量差の関係を示すことが目的であるため，凡例は省略している。破線は本実験で使用する GUEST1000_5 のモーラ数（5,476 モーラ）。

ら 1 の範囲で表現する，ノンパラメトリック検定の効果量である。比較対象となる二つの標本について，効果量 Cliff's δ の絶対値が 0.147 未満の場合は無視できる規模（negligible），0.33 未満の場合は小規模（small），0.474 未満の場合は中規模（medium），それ以上の場合は大規模（large）の差であるとされている [24]。

この音声セットの収録には数日を要した影響から，安定した演技表現ができていない可能性がある。そのため収録順序による影響を減らすことを目的に，文章の解析順序をランダムに並び替えている。

その結果，**図–1** に示すとおり 5,000 モーラ程度で音響特徴量の効果量が概ね一定の値を示すことが認められた。そのため本研究ではモーラバランス型日本語コーパス ROHAN [25] より，1 文あたりのモーラ条件が課されていない GUEST1000_5 サブセット（5,476 モーラ）を使用して，平静音声と嫌悪演技音声をそれぞれ新規収録することとした。

GUEST1000_5 サブセットについて，再度第一著者が平静音声と嫌悪演技音声のそれぞれについて収録し，同様の条件でモーラ数ごとの音響特徴量解析を行った。その結果，図–1 と同様の傾向が認められた。一方で，収録開始直後と収録終了直前に演技性が不安定となる現象が予備収録データを解析した際に確認された。そのため本実験では発話者に対して，解析で使用する台本の前後にダミー文を各 25 文ずつ，合計 50 文を加えた 250 文を一つの文章セットとして提示することとした。解析時にはダミー文章を除外して解析する。

3.2 本実験に向けた音声収録

本研究では，プロとしての発話トレーニングを受けていない 20 代 4 名（男性 2 名，女性 2 名）について音声を収録した。収録条件を**表–1** に示す。音声の収録中に発話方法が変化することを防ぐために，平静音声と嫌悪演技音声をそれぞれ 1 日で収録し，読み間違い等が認められた場合は別日に再収録した。読み上げには予備実験で決定したダミー文を含む 250 文を使用した。このとき，話者に対して文章セットの中に解析対象外のダミー文が含まれていることは提示していない。口とマイクとの距離は約 15 cm を維持するよう指示し，ノイズを防止するためにポップガードを使用した。収録は 250 文を一つの音声ファイルとして連続して収録した後，収録時と同じビット深度・サンプリングレートの WAV ファイルとして発話ごとに分割した。読み間違い等により複数回発話された場合は，発話者自身が最も明瞭に発話されていたと考える発話を選択した。

表–1 本実験音声の収録条件

収録環境	防音室（ヤマハ AMDC08H）
吸音材	AURALEX LENRD Bass Traps
騒音レベル	約 16 dB
収録用マイク	audio-technica AT5040
マイクプリアンプ	GRACE DESIGN m201 mk2
入力ゲイン	+28 dB
オーディオ IF	RME ADI-2/4 PRO SE
A/D 変換	48 kHz/24 bit
収録ソフト	Cakewalk by BandLab

ここで感情音声を収録する際には，収録時の工夫により感情音声が現実離れしたものとならないよう注意すべきという指摘がある [26]。そのため本研究では話者が発話し易いよう，平静音声では感情を込めずに発話するように，嫌悪演技音声では強い不快感を持っている [27] 感情をロボットに対して表現する [28] ように，それぞれ指示した。また，発話者固有の嫌悪演技表現を失わないために，詳細な演技方法及びシチュエーションなどは定義していない。嫌悪演技音声については収録開始時と終了時で演技性の変化を防止するため，収録前にリファレンス音声を収録し，全体の 20%である 50 文を収録するごとに聴きなおす指示をした。リファレンス音声の台本は，ITA コーパス [29] より感情音声の読み上げに使用できる Emotion セットの 1 番から 5 番を使用した。また，発話者に対しては，リファレンス音声が解析の対象外となることを伝えた。

4. 収録音声の評価

4.1 主観評価実験

本研究で使用した音声は 3.2 節で述べたとおり，プロとしての発話トレーニングを受けていない発話者に

表–2 主観評価実験条件

発話数	400 発話（50 文 × 2 感情 × 4 話者）
評価者数	男女 20 名（19 歳–23 歳）
サンプリング	48 kHz/24 bit
聴取環境	防音室（表–1 と同じ）
オーディオ IF	RME ADI-2 DAC FS
ヘッドホン	SENNHEISER HD650

よるものである。そのため発話者が嫌悪感情を意識して発話したにもかかわらず，その感情が正確に伝わらない可能性がある。そこで収録音声について，主観評価実験を行うことで発話者が嫌悪感情を演技できているかを調査した。本実験は被験者に対して音声を聴取させ，話者がどの程度の強さで嫌悪感情を表現しているかについて回答させた。

音声は被験者ごとに，各話者各感情から 50 発話ずつランダムに非復元抽出した合計 400 発話を使用した。主観評価実験は被験者の疲労により回答精度が低下することを防ぐため，全 400 発話のうち半分にあたる 200 発話の時点で被験者に対して休憩を指示した。

本実験は音声の嫌悪感情強度を回答させるため，単極性の尺度となる。そのため選択肢は 0（平静）から 6（嫌悪）の 7 段階評価とした [30]。また，回答時は読み上げられたコーパス文の文脈は考慮せず，実験中は音量を変更をしないよう指示した。聴取する音声の順序はランダムとした。実験条件を**表–2** に示す。

4.2 主観評価結果

主観評価の結果について Kolmogorov–Smirnov 検定 [31] を用いて正規性を有意水準 5%で検定した結果，すべての話者について帰無仮説が棄却された。また，発話データを被験者ごとにランダムサンプリングをしていることから，評価に使用された平静音声と嫌悪音声の間には対応がない。以上より本評価では，対応がない 2 群間のノンパラメトリック検定である Mann–Whitney の U 検定 [32] を有意水準 5%で実施した。

結果を**表–3** に示す。これより，いずれの話者においても平静音声と嫌悪音声の間に嫌悪強度の有意差が認められた（$p < 10^{-126}$）。そのため本研究において収録された音声は，嫌悪感情を表現できていると考えられる。

5. 音響特徴量の解析

本研究ではボコーダで制御することを前提とした音声分析を実施するため，基本周波数（F0），第 1 フォルマント（F1），第 2 フォルマント（F2），スペクトル重心（Spectral centroid）について解析した。

ここでスペクトル重心は音の明るさに対応するパラメータであり，スペクトログラム $X[n]$ における計算方法を式 (1) に示す [33]。

$$S_c[n] = \frac{\sum_{k=0}^{N/2} f[k]|X[n,k]|}{\sum_{k=0}^{N/2} |X[n,k]|} \tag{1}$$

$S_c[n]$ は n フレームにおけるスペクトル重心であり，N は FFT (Fast Fourier Transform) 長に対応し，$f[k]$ は k 番目のビンに対応する周波数をそれぞれ示す [33]。

ダミー音声を除外した発話データの音声全体に対して音響パラメータの計算を行い，その結果から母音ごとにパラメータを抽出した。この際，平静と嫌悪演技がどちらも有声母音であった場合のみを解析の対象とし，いずれかが無声化した場合は解析の対象外としている。

表–3 主観評価実験のスコア平均値と 95%信頼区間（95% CI），及び 2 群間の効果量 δ（Cliff's δ）
今後の表において NEU は平静音声，DIS は嫌悪演技音声をそれぞれ示す。また，F は女性話者を，M は男性話者をそれぞれ示す。

		Score	95% CI	Cliff's δ
F #1	NEU	1.09	0.07	0.61
	DIS	2.86	0.10	
F #2	NEU	1.89	0.09	0.64
	DIS	3.94	0.10	
M #1	NEU	1.14	0.07	0.84
	DIS	4.04	0.09	
M #2	NEU	1.53	0.09	0.76
	DIS	4.13	0.09	

5.1 音響パラメータの推定

今回の解析では，F0 推定に Harvest [34] を，フォルマント推定に Praat [35] の Burg 法をそれぞれ用いた。スペクトル重心は CheapTrick [36] で予測されたスペクトル包絡から，高域の非周期的な成分による影響を抑制するために 8,000 Hz の帯域制限をかけて，フレームごとに計算した。いずれの計算も Python ラッパー（PyWORLD [37], Parselmouth [38]）を使用した。F0 とスペクトル包絡の推定はラッパーのデフォルトパラメータを使用した。フォルマント推定は Praat のマニュアルに従い，フォルマントの上限値を男性話者 5,000 Hz，女性話者 5,500 Hz に設定した。また，Harvest, CheapTrick と解析窓のシフト幅を揃えるために，フォルマント推定の Time step を Harvest, CheapTrick のデフォルトパラメータである 5 ms に設定している。

音素の時間情報は Julius [39, 40] による強制アライ

メントの結果を用いた。アライメントに用いる音素列は ROHAN の振り仮名情報から pyopenjtalk [41] を用いて生成した。pyopenjtalk が生成した音素の内，Julius が対応してない音素については対応する音素に置換した。取り出すパラメータについては，点ピッチパターン [42] の観点から母音区間の中央に対応するフレームを使用した。このとき，1 フレームのみで生じた推定誤差の影響を抑制するため，中央のフレームから前後 3 フレーム（約 30 ms）を候補として取り出し，パラメータの中央値を最終的な値として使用した。

5.2 集計及び統計量の計算

本研究では母音ごとの音響特徴量変化を観測対象としているため，解析対象となるサンプル数が非常に大きくなり，p 値の参照価値が低下すると考えられる。このことから本研究では検定により有意差が認められた特徴量について，効果量を用いた傾向の議論をする。

まず 5.1 節で述べた方法で母音ごとの各パラメータを計算した後，データを集計した。このときに音素アライメントの誤推定や読み間違い等から外れ値が発生する可能性があるため，集計後に四分位範囲法を用いてスクリーニングした。そのため特徴量ごとに標本数が異なる場合がある。このデータについて Kolmogorov–Smirnov 検定を用いて正規性を有意水準 5%で検定した結果，一部母音について帰無仮説が棄却されたことから，有意差検定は対応のない 2 群間のノンパラメトリック検定である Mann–Whitney の U 検定を有意水準 5%で実施した。その後，検定の結果有意差が認められたものについてのみ各種特徴量の効果量（Cliff's δ）を計算した。

また，先行研究 [13] より口の動かし方に差があるという可能性が示唆されているため，各母音のフォルマント周波数の平均値により表現される 5 角形の面積を MATLAB の polyarea 関数を用いて計算した。この面積の計算には平均値を用いることから Cliff's δ が求められないため，差分のみを求めた。

6. 客観評価の結果

6.1 基本周波数（F0）

F0 の解析結果を**表–4** に示す。女性話者 1 と男性話者 2 については，効果量に基づき小から中規模の下降が認められた。これらの話者について，母音の下降順序はそれぞれ異なった。一方で女性話者 2 と男性話者 1 は検定により有意差は確認されたものの，その変動幅は効果量より無視できる規模であった。

6.2 フォルマント周波数（F1, F2）

フォルマント周波数の解析結果を**表–5**, 6 及び**図–2** から**図–5** に示す。また，日本語五母音の第 1，第 2 フォルマント周波数により表現される 5 角形の面積を**表–7** に示す。

表–4 平静音声・嫌悪演技音声の母音別 F0 の平均値（単位は Hz）と検定結果及び効果量 δ（Cliff's δ）
今後の表において *** は $p < 0.001$ を，** は $p < 0.01$ を* は $p < 0.05$ を，n.s. は有意差なしを示す。有意差が認められなかった特徴量については — とし，効果量を表示していない。

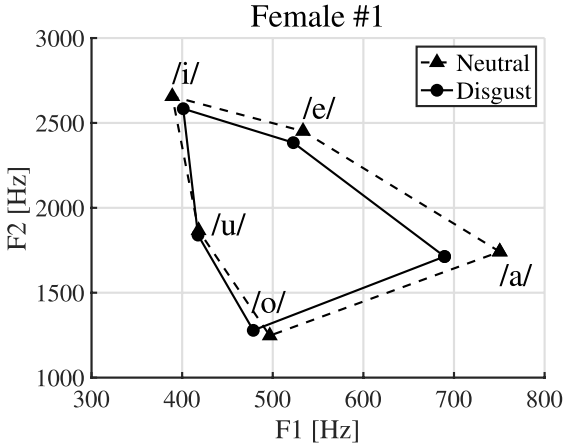
		n_{NEU}	n_{DIS}	NEU	DIS	Sig.	δ
F #1	/a/	1,252	1,146	231.3	215.8	***	−0.33
	/i/	859	812	235.4	219.1	***	−0.34
	/u/	760	729	240.1	221.0	***	−0.41
	/e/	662	610	232.6	216.6	***	−0.31
	/o/	1,143	1,080	234.1	215.7	***	−0.36
F #2	/a/	1,261	1,245	218.5	212.6	***	−0.10
	/i/	863	874	226.8	223.5	*	−0.06
	/u/	767	781	230.4	226.9	n.s.	—
	/e/	658	663	220.1	215.0	*	−0.07
	/o/	1,151	1,157	224.8	221.3	*	−0.06
M #1	/a/	1,303	1,287	147.7	147.1	**	0.07
	/i/	880	865	155.5	150.5	n.s.	—
	/u/	785	777	161.1	152.8	***	−0.10
	/e/	667	659	152.3	150.1	n.s.	—
	/o/	1,143	1,138	153.5	148.7	n.s.	—
M #2	/a/	1,312	1,309	172.6	150.3	***	−0.25
	/i/	877	879	185.5	161.9	***	−0.28
	/u/	800	800	191.5	169.0	***	−0.25
	/e/	676	676	180.1	157.8	***	−0.24
	/o/	1,160	1,158	185.4	160.8	***	−0.28

表–5 平静音声・嫌悪演技音声の母音別 F1 の平均値（単位は Hz）と検定結果及び効果量 δ（Cliff's δ）

		n_{NEU}	n_{DIS}	NEU	DIS	Sig.	δ
F #1	/a/	1,215	1,187	750.5	689.6	***	−0.27
	/i/	832	825	389.1	401.2	***	0.10
	/u/	743	734	418.1	417.5	n.s.	—
	/e/	668	667	533.4	522.7	n.s.	—
	/o/	1,106	1,121	496.7	478.5	***	−0.12
F #2	/a/	1,244	1,281	751.6	769.8	***	0.09
	/i/	824	846	410.7	425.7	**	0.08
	/u/	759	759	438.4	457.7	***	0.15
	/e/	670	671	533.3	546.1	*	0.08
	/o/	1,154	1,135	564.2	582.9	***	0.12
M #1	/a/	1,194	1,245	640.7	612.0	***	−0.22
	/i/	843	841	355.3	361.5	*	0.07
	/u/	743	759	370.1	380.2	***	0.12
	/e/	668	655	446.9	442.0	n.s.	—
	/o/	1,116	1,099	455.9	444.6	***	−0.13
M #2	/a/	1,291	1,306	615.7	621.8	n.s.	—
	/i/	846	827	397.6	400.9	n.s.	—
	/u/	754	761	413.3	423.3	**	0.09
	/e/	674	670	472.0	480.0	n.s.	—
	/o/	1,119	1,124	468.4	483.5	***	0.15

表–6 平静音声・嫌悪演技音声の母音別 F2 の平均値（単位は Hz）と検定結果及び効果量 δ（Cliff's δ）

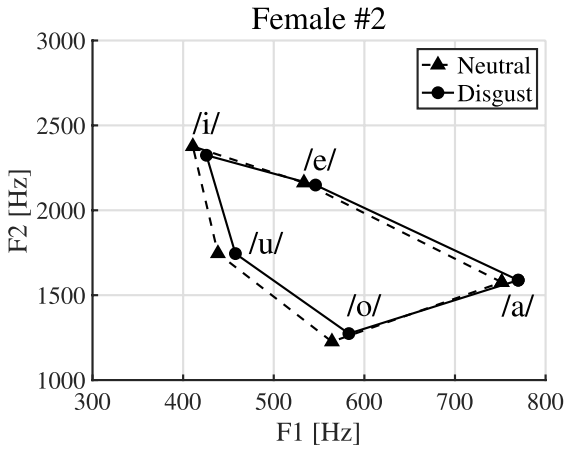
		n_{NEU}	n_{DIS}	NEU	DIS	Sig.	δ
F #1	/a/	1,248	1,257	1,741	1,713	**	−0.07
	/i/	786	787	2,657	2,584	***	−0.20
	/u/	745	748	1,867	1,840	n.s.	—
	/e/	650	654	2,451	2,384	***	−0.26
	/o/	1,124	1,123	1,247	1,278	***	0.09
F #2	/a/	1,290	1,283	1,576	1,588	n.s.	—
	/i/	852	866	2,376	2,324	***	−0.10
	/u/	756	777	1,746	1,745	n.s.	—
	/e/	653	656	2,162	2,149	n.s.	—
	/o/	1,143	1,148	1,226	1,274	***	0.11
M #1	/a/	1,292	1,296	1,457	1,516	***	0.20
	/i/	797	787	2,198	2,148	***	−0.15
	/u/	768	759	1,538	1,600	***	0.14
	/e/	665	653	2,004	1958	***	−0.15
	/o/	1,137	1,138	1,049	1,109	***	0.18
M #2	/a/	1,297	1,287	1,431	1,433	n.s.	—
	/i/	845	851	2,144	2,081	***	−0.15
	/u/	780	788	1,428	1,492	***	0.15
	/e/	667	664	1,957	1,922	***	−0.11
	/o/	1,125	1,116	1,033	1,085	***	0.20



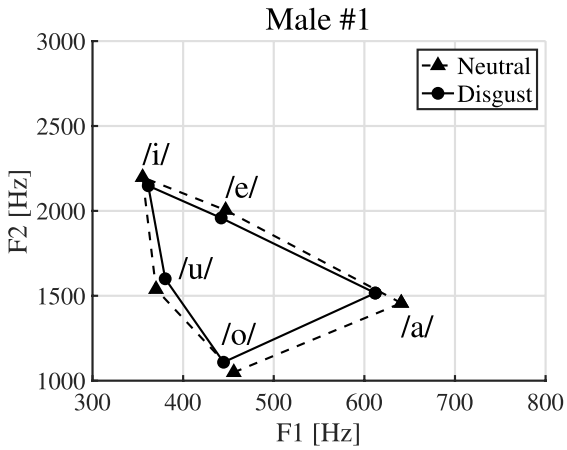
図–2 女性話者 1 の母音別フォルマント周波数
破線は平静感情音声を示し，実線は嫌悪感情音声を示す。

全話者に共通したフォルマント周波数の変化は認められなかった。また，話者単位での法則性も認められなかった。変動幅についても一部話者において，効果量に基づき小規模の変化が認められたが，約 68%は無視できる規模であった。

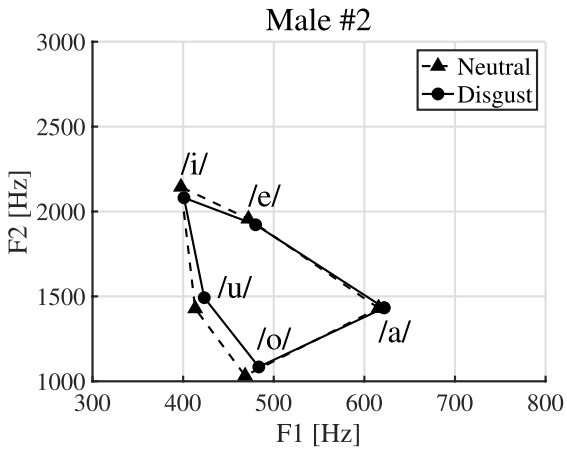
一方で，フォルマント周波数により表現される 5 角形の面積は全話者に共通して減少する傾向が観られた。また，実験条件や解析方法等が一部異なるため単純な比較はできないものの，本研究に先立って実施した第一著者の嫌悪演技に対しての客観評価 [43] と同様の傾向が確認されている。



図–3 女性話者 2 の母音別フォルマント周波数
破線は平静感情音声を示し，実線は嫌悪感情音声を示す。



図–4 男性話者 1 の母音別フォルマント周波数
破線は平静感情音声を示し，実線は嫌悪感情音声を示す。



図–5 男性話者 2 の母音別フォルマント周波数
破線は平静感情音声を示し，実線は嫌悪感情音声を示す。

6.3 スペクトル重心（Spectral centroid）

スペクトル重心の解析結果を表–8 に示す。全話者に共通した変化は認められなかった。また，話者ごとに結果を確認した場合においても，母音ごとに異なる傾向が観られた。女性話者 2 の/u, e, o/は効果量に基づき小規模上昇している。その一方で女性話者 1 及び男

表–7 フォルマント周波数により表現される 5 角形の面積と感情間の差分。単位は Hz^2。

	NEU	DIS	Difference
F #1	2.565×10^5	1.968×10^5	-5.969×10^4
F #2	1.797×10^5	1.657×10^5	-1.395×10^4
M #1	1.577×10^5	1.187×10^5	-3.900×10^4
M #2	1.187×10^5	1.045×10^5	-1.416×10^4

表–8 平静音声・嫌悪演技音声の母音別スペクトル重心の平均値（単位は Hz）と検定結果及び効果量 δ（Cliff's δ）

		n_{NEU}	n_{DIS}	NEU	DIS	Sig.	δ
F #1	/a/	1,243	1,216	1,100	1,054	***	−0.14
	/i/	826	827	754.8	802.4	***	0.10
	/u/	745	741	701.4	696.8	n.s.	—
	/e/	659	647	1,069	1,049	*	−0.08
	/o/	1,127	1,122	713.2	704.8	n.s.	—
F #2	/a/	1,282	1,287	1,141	1,192	***	0.11
	/i/	861	867	944.9	1,040	***	0.13
	/u/	773	773	862.1	945.6	***	0.17
	/e/	668	669	1,163	1,246	***	0.16
	/o/	1,145	1,148	841.6	908.2	***	0.19
M #1	/a/	1,303	1,302	1,010	1,032	*	0.06
	/i/	848	852	959.3	1,003	**	0.09
	/u/	771	767	769.8	820.2	***	0.11
	/e/	670	666	1,102	1,092	n.s.	—
	/o/	1,138	1,128	698.4	695.0	n.s.	—
M #2	/a/	1,306	1,308	1,004	986.3	n.s.	—
	/i/	853	863	1,013	1,029	n.s.	—
	/u/	784	792	877.3	880.3	n.s.	—
	/e/	674	675	1,061	1,068	n.s.	—
	/o/	1,139	1,155	777.5	791.1	*	0.05

性話者の発話については，いずれも無視できる規模の変動であった。

7. 考　　察

6 章にて，今回収録した話者 4 名の音響特徴量の変化がそれぞれ示された。本章では TTS や VC に向けて，どのような特徴量を与えることで話者が意図する嫌悪感情を表現できるのかを考察する。

7.1 基本周波数の低下

まず女性話者 1 と男性話者 2 で確認された特徴として，F0 が低下したことが挙げられる。2 章で述べたとおり，言語による差はあるものの，日本語の先行研究 [16, 17] では嫌悪感情を表現する際は F0 が低下する傾向にあることが示されている。しかし女性話者 2 と男性話者 1 において，嫌悪感情の発露が確認されているにもかかわらず，F0 が低下しないケースが認められた。嫌悪音声は冷たい怒りと混同される傾向にあり，F0 の低下は冷たい怒りでも発生する [16] ことが指摘されており，ドイツ語においては嫌悪の強度が強い場合は F0 が高く，不安定になることも示されている [12]。

このことから日本語で嫌悪表現をした場合 F0 は低下する傾向にあると考えられるものの，必須となる特徴量でなく，個人差に影響される特徴量と言える。また，冷たい怒りの感情等でも同様の傾向であることからも，F0 以外を含めた複数の特徴量により嫌悪固有の変化を策定することが必要である。

7.2 フォルマント周波数の変化

広東語における先行研究 [13] では，フォルマント周波数（F1, F2）がほとんどの母音において低下することが示された。一般的に F1 は開口度と対応し，F2 は舌の位置と対応するとされている [44] ことから，F1 と F2 の低下は口を閉じて舌を後退させることにより発生すると考えられる。そのため先行研究 [13] では嫌悪感情の表現は病原体などの異物回避に由来することに触れつつ，異物が体内に侵入することを防ぎ，口内に侵入した異物を排除する役割がある可能性を述べていた。しかし本研究の解析結果においてフォルマント周波数の個別変化は弱く，変化が確認された母音についても一部で上昇が確認された。

ここで本研究でみられた新たな傾向として，各母音のフォルマント周波数により表現される 5 角形の面積減少に着目する。面積の減少は母音ごとの F1 と F2 の変動幅，つまり口と舌の動きが小さくなることに由来すると考えられる。これについて先行研究 [13] と同様に異物回避の観点から考えると，口の動きを小さく押さえることで異物の侵入を防止する動きを表現したと考えられる。

この仮説を検証するためには，声道の動きを直接計測する必要がある。音声波形から声道断面積関数を推定することで近似的に解析することは可能であるが，推定誤差も生じるため新たに誤差を含めて追加の検討が必要となる。そのため本論文では声道形状については解析対象とせず，この検証は今後の検討項目とする。また，現時点ではフォルマント周波数により表現される 5 角形の面積の減少が知覚的に有意な変化をもたらしたとは言えないため，新たな実験により検討する必要がある。

7.3 今後の検討課題

以上の結果から，嫌悪感情を表現することでフォルマント周波数により表現される 5 角形の面積が減少する傾向にあると考えられる。しかし F0 などの特徴量については，話者により変化の傾向が異なることが認められた。このことからボコーダによる後処理で嫌悪感情を付与する場合，理想的には発話者ごとに変化パターンを明らかにすることが望まれる。一方で他者の

表現を用いても嫌悪を表現することができれば十分であることから，特定の話者の特徴付与により多くの話者に嫌悪を与えることが可能であるか検討する必要がある。特徴量を与えることで目的とする嫌悪感情を適切に付与できるかについては，次のステップとして検討する。

8. おわりに

本研究では嫌悪感情を意識して演技発話をした際にどのような音響特徴量の変化が現れるかを評価した。主観評価により嫌悪感情が確認された音声セットについて客観評価を実施したところ，一部の話者で F0 の低下傾向と，すべての話者でフォルマント周波数により表現される 5 角形の面積減少が確認された。

今後は TTS や VC に応用するため，本研究により確認できた嫌悪演技表現の音響特徴量変化を，第三者の平静音声に対して重畳することにより嫌悪感情を付与できるかを調査する。また，表現の幅を更に広げるため，合成時の嫌悪強度を変更する手法も同時に検討する必要がある。

謝　辞

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

文　献

[1] T. Li, S. Yang, L. Xue and L. Xie, “Controllable emotion transfer for end-to-end speech synthesis,” *Proc. 12th Int. Symp. Chinese Spoken Language Processing* (*ISCSLP 2021*), pp. 1–5 (2021).

[2] W. Zhao and Z. Yang, “An emotion speech synthesis method based on VITS,” *Appl. Sci.*, **13**(4), pp. 1–12 (2023).

[3] 飯田朱美, ニックキャンベル, 安村通晃, “感情表現が可能な合成音声の作成と評価,” 情報処理学会論文誌, **40**, 479–486 (1999).

[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,” 情処研報音声言語情報処理（SLP）, **2019-SLP-127**(39), pp. 1–6 (2019).

[5] 山本龍一, 高道慎之介, Python で学ぶ音声合成（インプレス, 東京, 2021）, p. 58.

[6] K. Inoue, S. Hara, M. Abe, N. Hojo and Y. Ijima, “Model architectures to extrapolate emotional expressions in DNN-based text-to-speech,” *arXiv:2102.10345* (2021).

[7] P. エクマン, W. V. フリーセン, 工藤　力, 表情分析入門：表情に隠された意味をさぐる（誠信書房, 東京, 1987）, p. 31.

[8] R. Plutchik, “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” *Am. Sci.*, **89**, 344–350 (2001).

[9] 小池和仁, 斎藤博昭, 中西正和, “感情音声の合成,” 情処研報音声言語情報処理（SLP）, **1998-SLP-024**, pp. 83–88 (1998).

[10] E. R. Skinner, “A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness; and a determination of the pitch and force of the subjective concepts of ordinary, soft, and loud tones,” *Speech Monogr.*, **2**, 81–137 (1935).

[11] K. R. Scherer, *Nonlinguistic Vocal Indicators of Emotion and Psychopathology* (Springer US, Boston, MA, 1979), pp. 493–529.

[12] A. Paeschke, M. Kienast and W. F. Sendlmeier, “F0-contours in emotional speech,” *Proc. 14th Int. Congr. Phonetic Sciences*, Vol. 2, pp. 929–932 (1999).

[13] C. S. Chong, J. Kim and C. Davis, “Disgust expressive speech: The acoustic consequences of the facial expression of emotion,” *Speech Commun.*, **98**, 68–72 (2018).

[14] C. S. Chong, C. Davis and J. Kim, “A cantonese audio-visual emotional speech (CAVES) dataset,” *Behav. Res. Methods*, **56**, 5264–5278 (2023).

[15] 武田昌一, 大山　玄, 朽谷綾香, 西澤良博, “日本語音声における「怒り」を表現する韻律的特徴の解析,” 音響学会誌, **58**, 561–568 (2002).

[16] 坂下尚史, 河原英紀, 松井淑恵, “演技未経験者の感情音声の演技における台本の影響：音響解析と主観評価による検討,” 音講論集, pp. 987–988 (2022.9).

[17] 重野　純, “感情を表現した音声の認知と音響的性質,” 心理学研究, **74**, 540–546 (2004).

[18] R. Jürgens, K. Hammerschmidt and J. Fischer, “Authentic and play-acted vocal emotion expressions reveal acoustic differences,” *Front. Psychol.*, **2**, Article 180 (2011).

[19] Y. Arimoto, H. Kawatsu, S. Ohno and H. Iida, “Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment,” *Acoust. Sci. & Tech.*, **33**, 359–369 (2012).

[20] 森　大毅, “宇都宮大学 パラ言語情報研究向け音声対話データベース（UUDB）,” https://doi.org/10.32130/src.UUDB（参照 2024-01-23）.

[21] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoust. Sci. & Tech.*, **41**, 761–768 (2020).

[22] 俣野文義, 松井淑恵, 森勢将雅, “DNN 音声合成による嫌悪感情の表現と基礎評価,” 情処研報音声言語情報処理（SLP）, **2023-SLP-147**(50), pp. 1–4 (2023).

[23] N. Cliff, “Dominance statistics: Ordinal analyses to answer ordinal questions,” *Psychol. Bull.*, **114**, 494–509 (1993).

[24] J. Romano and J. Kromrey, “Appropriate statistics for ordinal level data: Should we really be using t-test and cohen’s d for evaluating group differences on the nsse and other surveys?,” *Annu. Meet. Florida Association of Institutional Research*, pp. 1–32 (2006).

[25] 森勢将雅, “ROHAN：テキスト音声合成に向けたモーラバランス型日本語コーパス,” 音響学会誌, **79**, 9–17 (2023).

[26] 森　大毅, “感情音声の研究を始める人のための音声コーパス入門,” 音講論集, pp. 1345–1346 (2019.3).

[27] 小学館『大辞泉』編集部, デジタル大辞泉（小学館, 東京, 2012）.

[28] E. Takeishi, T. Nose, Y. Chiba and A. Ito, “Construction and analysis of phonetically and prosodically balanced emotional speech database,” *Proc. Conf. Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques* (*O-COCOSDA 2016*), pp. 16–21 (2016).

[29] 小口純矢, 金井郁也, 小田恭央, 齊藤剛史, 森勢将雅, “ITA コーパス：パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価,” 情処研報音声言語情報処理 (SLP), **2021-SLP-137**(31), pp. 1–4 (2021).
[30] 坂上貴之, 川原純一郎, 木村英司, 三浦佳代, 行場次朗, 石金浩史, 基礎心理学実験法ハンドブック (朝倉書店, 東京, 2018), pp. 151–152.
[31] F. J. Massey, Jr., “The Kolmogorov–Smirnov test for goodness of fit,” *J. Am. Stat. Assoc.*, **46**, 68–78 (1951).
[32] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *Ann. Math. Stati.*, **18**, 50–60 (1947).
[33] 森勢将雅, ひたすら楽して音響信号解析 MATLAB で学ぶ基礎理論と実装 (コロナ社, 東京, 2021), p. 83.
[34] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” *Proc. Interspeech 2017*, pp. 2321–2325 (2017).
[35] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot Int.*, **5**(9/10), 341–345 (2001).
[36] M. Morise, “Error evaluation of an f0-adaptive spectral envelope estimator in robustness against the additive noise and f0 error,” *IEICE Trans. Inf. & Syst.*, **E98-D**, 1405–1408 (2015).
[37] JeremyCCHsu, “JeremyCCHsu/Python-Wrapper-for-World-Vocoder,” https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder（参照 2024-01-15）.
[38] YannickJadoul, “YannickJadoul/Parselmouth,” https://github.com/YannickJadoul/Parselmouth（参照 2024-01-15）.
[39] A. Lee, T. Kawahara and K. Shikano, “Julius — an open source real-time large vocabulary recognition engine,” *Proc. EUROSPEECH*, pp. 1691–1694 (2001).
[40] A. Lee and T. Kawahara, “Recent development of open-source speech recognition engine julius,” *Proc. Asia-Pacific Signal and Information Processing Association Annu. Summit and Conf.*, pp. 131–137 (2009).
[41] R. Yamamoto, “r9y9/pyopenjtalk,” https://github.com/r9y9/pyopenjtalk（参照 2024-01-13）.
[42] 橋本新一郎, “日本語単語アクセントの諸性質,” 信学論 D, **J56-D**, 654–661 (1973).
[43] 俣野文義, 森勢将雅, “日本語嫌悪感情音声の音響特徴量解析,” 音講論集, pp. 1381–1382 (2023.9).
[44] 益子幸江, “日本語の母音の音色とフォルマントについての一研究,” 東京外国語大学論集, **82**, 105–121 (2011).

俣野　文義

2001 年生。2023 年豊橋技術科学大学卒業。同年 4 月より明治大学大学院博士前期課程にて音声合成・信号処理の研究に従事。日本音響学会，情報処理学会会員。

小口　純矢

1994 年生。2019 年明治大学卒業。2022 年明治大学大学院博士前期課程修了，修士（工学）取得。2022 年より，日本学術振興会特別研究員（DC1）として，信号処理や機械学習を用いた音声・楽器音の分析及び合成の研究に従事。日本音響学会，情報処理学会会員。

森勢　将雅

1981 年生。2008 年和歌山大学大学院博士課程修了，博士（工学）取得。関西学院大学，立命館大学，山梨大学を経て，2024 年より明治大学総合数理学部専任教授。人間の知覚情報を活用した，音声分析・合成・デザインの研究に従事。日本音響学会，電子情報通信学会，情報処理学会会員。