好感度自動推定モデルを利用した任意話者音声の好感度を制御可能な声質変換

須田 仁志[1,a)]　高道 慎之介[2,3,1,b)]　深山 覚[1,c)]

概要：音声の好感度は，人対人の直接的なコミュニケーションだけでなく，広告宣伝や政治活動など，音声が利用される様々な状況において，円滑あるいは効果的な情報伝達に重要である．主観的な音声の好感度を自動で評価できれば，好感度の評価結果を用いてより効果的な発声訓練や音声デザインに応用が可能である．そこで本稿では，主観的な音声好感度を収集したコーパスである CocoNut-Humoresque を用いて，音声の好感度を自動で推定する手法を提案する．実験的評価により，コーパス内の好感度評点と推定された評点の間に有意な相関が見られ，提案手法の有効性が確認された．さらに本稿では，音声–好感度のペアデータを活用した，入力音声の好感度を制御可能な声質変換手法を提案する．任意話者声質変換モデルの学習に用いられる高品質かつ多人数の音声合成用コーパスには，いまだ音声に対して好感度が付与されたものはない．そこで，本稿で提案する好感度自動推定手法により，音声合成用コーパス中の各音声に対して推定好感度評点を付与し，これを利用して好感度制御モデルを学習する．実験的評価により，入力となる目的好感度に応じて，音声の話者性および発話内容を維持した好感度制御が可能であることが示された．

1. はじめに

音声が伝達する情報は，1) シンボル列により表現可能な離散的な言語情報，2) 抑揚や発話スタイルなど話者が意図を持って制御可能なパラ言語情報，3) 声の個人性（話者性）など話者の意図によらない非言語情報，の 3 要素からなる [1]．与えられた言語情報にもとづき音声を合成するテキスト音声合成（text-to-speech; TTS）[2] においては，任意の話者での合成が可能な zero-shot text-to-speech（ZS-TTS）[3], [4]，感情をはじめとした発話スタイルを制御可能な手法 [5], [6], [7] など，パラ言語情報や非言語情報を制御する技術が広く研究されている．また，入力音声の言語情報を保持したまま，特定のパラ言語情報もしくは非言語情報を制御する音声変換技術についても，話者性を制御する話者変換，感情表現を制御する発話スタイル転写をはじめとして，数多く研究されている [6], [8], [9]．これらの技術においては，目標となる話者性や発話スタイルなどを持つ参照音声や，one-hot の感情ラベルなどを入力として，目的の情報を制御する．加えて，パラ言語情報あるいは非言語情報の制御に自然言語のプロンプトを用いる手法も提案されている [10], [11]．一方，こうした技術での制御は，話者性，感情，話速，発話スタイルなど，ある程度客観的で説明可能な情報にもとづく制御に限られる．すなわち，たとえば「事務手続きに追われ研究や教育に集中できない 30 代の大学教員にターゲティングした広告宣伝に適した声」や「若い男の子ウケする大学生バーチャル YouTuber（VTuber）[12] の声」など，極度に主観的な基準での音声制御は困難である．音声のパラ言語情報や非言語情報が，合成音声利用の主観的な目的に直接もとづいて制御できれば，より柔軟かつ適合性の高い音声デザインが可能になる．あるいは，たとえば，相手に応じてより効果的に意思伝達できる声を習得するために，ユーザの声を分析し手本となる合成音声を提示するような，発声訓練アプリケーションなどへの応用も可能である．

音声の基礎的かつ主観的なパラ言語情報および非言語情報として音声好感度が挙げられ，人間の社会的な活動における重要性が示されている [13]．たとえば，話者の性別に関係なく，異性からの音声好感度と外見の好感度は強い相関があることが示されており，音声の好感度はパートナーの選択に大きな影響を与えていることが示唆される [14], [15].

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] 慶應義塾大学
Keio University, 3-14-1, Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan
[3] 東京大学
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku 113-8656, Tokyo
[a)] suda.h@aist.go.jp
[b)] shinnosuke_takamichi@keio.jp
[c)] s.fukayama@aist.go.jp

また，政治活動における政治家の声質による得票数への影響や，経営者の声質による企業の売り上げへの影響も指摘されており，話者の声質は社会活動に——直接の因果関係とは到底思えないものの——大きく関与していると考えられる [16], [17], [18]．さらに，上述のような音声好感度に関する研究においては，話者と聴取者の性別の組み合わせによる影響が多く調査されている．YouTube 上の動画音声からなる音声好感度コーパス CocoNut-Humoresque に対する調査においても，同一の音声であってもその好感度が聴取者の性別や年代によって異なることが示されている [19]．したがって，音声好感度を扱うには，音声そのものだけでなくその聴き手を考慮する必要があり，聴き手の影響を受けるという観点においても主観的な特徴といえる．

音声の好感度は，基本周波数 f_o との関係について広く調査されている [20], [21], [22], [23]．一方で，音素継続長や発話速度（話速）など，他の音響特徴量も音声好感度に寄与している [24], [25]．CocoNut-Humoresque に対する調査においても，好感度に対する f_o の影響は大きいものの，他の要因の存在が強く示唆されている [19]．したがって，音声好感度は，パラ言語情報と非言語情報にわたって影響を受ける，単一の音響特徴量では十分に表現できない音声情報と位置付けられる．

本研究の目的は主観的かつ漠然とした音声情報の制御技術の構築であり，本稿ではその基礎として，主観的情報である音声好感度の制御手法を提案する．本稿ではとくに，任意話者による入力音声の話者性および言語情報を維持したまま，その好感度のみを選択的に制御する声質変換手法について議論する．主観的な音声情報である好感度を学習に用いるためには，本来すべての学習音声に対する複数の聴取者による主観評価が必要である．さらに，高品質な声質変換モデルの構築には多量の学習データを要するため，これらの音声すべてに主観評価を行うことは非現実的である．そこで本稿では，既存の音声好感度コーパスを利用して好感度自動推定器を構築し，これを用いて音声合成用コーパスに推定好感度評点を自動付与する手法を提案する．この手法により，話者性および言語情報を維持したまま，目標好感度に応じて好感度を制御する声質変換モデルを，学習データに対する人手での好感度評価なしに構築可能である．本稿では，提案する好感度自動推定手法の性能および，これを用いて構築した好感度制御モデルにおける好感度制御の有効性を評価する．

2. 音声好感度自動推定

本節では，主観的な音声好感度を収集したコーパスである CocoNut-Humoresque [19] を利用した，好感度自動推定手法を提案する．この手法により，単なる好感度の自動予測アプリケーションだけでなく，音声合成や音声変換での好感度制御などが可能になる．本稿で提案する好感度自動推定器のモデルアーキテクチャを図 1 に示す．提案手法では，入力を対数メルスペクトログラムとし，time-delay neural networks（TDNN）[26] を介して，聴取者 4 属性ごとによる好感度評点を推定する．話者認識に用いられる x-vector [27] のアーキテクチャと同様に，ニューラルネットワークの中間特徴量の時間平均および標準偏差を出力する stats pooling 層を介することで，可変長の入力信号から評点を出力する．

好感度自動推定は，合成音声の平均オピニオン評点（mean opinion scores; MOS）自動推定 [28], [29], [30] と類似した問題である．MOS 推定との相違点として，好感度自動推定の特徴にはとくに以下の 3 点が挙げられる．

(1) 同一音声であっても，聴取者によって評点が異なる
(2) 評点の分布が中央の評点に極度に集まりやすい
(3) コーパスが CocoNut-Humoresque に限られ，学習データ量が限られる

まず，(1) の問題に対応するため，聴取者を 30 代以下の男性（$n = 202$），40 代以上の男性（$n = 323$），30 代以下の女性（$n = 143$），40 代以上の女性（$n = 210$）の 4 属性に分割し，聴取者別に平均評点を学習および推定する．ここで，n は CocoNut-Humoresque 全体での聴取者数である．CocoNut-Humoresque では 10 歳ごとの年代のみが収集されているが，全性別–年代ごとに評点を集計するにはデータ量が不十分なため，これらの 4 属性に分割する．

次に (2) について，学習コーパスとして用いる CocoNut-Humoresque では好感度の評点の分布が一様でなく，中心の評点（−1 から 1 に変換した場合は 0）に集中する．このコーパスをそのまま学習に用いると，過学習により推定評点が平均値の 0 付近に集中する．そこで本手法では，学習データの検証セットから平均 μ および分散 σ^2 を得て，推定評点がこれらの統計量と一致するような次式に示す線形変換によるポストフィルタリングを適用する．

$$\hat{y}' = \frac{\sigma}{\hat{\sigma}} (\hat{y} - \hat{\mu}) + \mu \tag{1}$$

ここで，$\hat{y}$ は推定評点，$\hat{\mu}$ および $\hat{\sigma}^2$ はそれぞれ検証セットに対する推定評点の平均および分散，$\hat{y}'$ はポスタフィルタリング後の推定評点を表す．この変換は線形であるため，相関係数を利用した評価指標などには影響しない．

3. 話者性と言語情報を保持した音声好感度制御

本節では，任意話者声質変換を応用し，話者性および言語情報を保持したまま音声の好感度を制御できる任意話者声質変換手法を提案する．このアーキテクチャを図 2 に示す．このモデルは，自己教師あり学習（self-supervised learning; SSL）モデルにより得られた離散音声トークン列を用いた任意話者声質変換法 [35], [36] を拡張したものである．本手法では合成モデルとして FastSpeech 2 [32] を

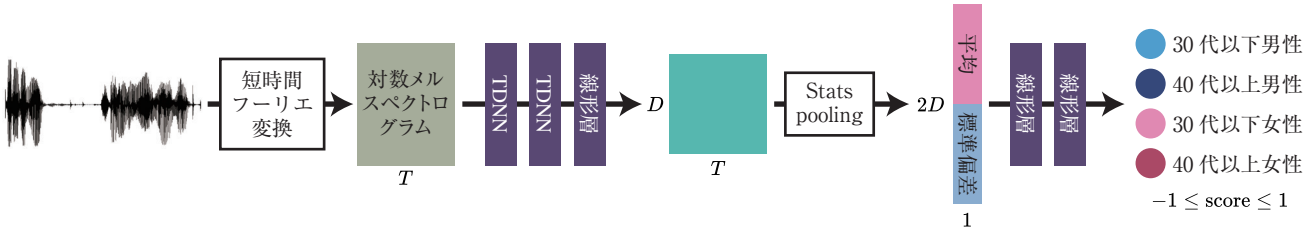


図 1 好感度自動推定器のアーキテクチャ．入力音声の対数メルスペクトログラムを利用し，ニューラルネットワークの中間特徴量に対する時間方向での stats pooling を介して，30 代以下の男性，40 代以上の男性，30 代以下の女性，40 代以上の女性からの好感度評点を −1 以上 1 以下でそれぞれ推定する．T は時間フレーム数を，D は中間特徴量の次元数を表す．

採用する．

提案手法では，SSL モデルとして hidden-unit BERT（HuBERT）[31] を利用する．事前に学習データから HuBERT 特徴量を抽出し，これに対して k-means 法によりクラスタリングを行い，k 個のクラスタ中心を得る．提案モデルは，各時刻フレームの HuBERT 特徴量の属するクラスタインデックスを入力離散音声トークン列として，ECAPA-TDNN による話者埋め込み表現 [33] および好感度評点を条件付けとして，音声信号にデコードするモデルにあたる．なお，SSL モデルによる離散トークン列を用いた音声合成では話者性を維持することが困難であることが知られており [37][*1]，本手法では離散トークン列に加えて話者埋め込み表現を個別に入力する．

任意話者好感度制御モデルの実現には，音声好感度評点が付与された多人数音声コーパスを利用する必要がある．しかし，好感度評点が付与されている CocoNut-Humoresque は，YouTube から収集され音源分離によって処理された音声からなる Coco-Nut [38] を利用しており，このコーパスのみでは高品質な音声合成を実現できない．そこで提案手法では，学習に用いる高品質な多人数の音声合成用コーパスに対し，第 2 節で述べた好感度自動推定手法により好感度評点を擬似的に付与する．具体的には，以下の手順により学習を行う．

(1) CocoNut-Humoresque にもとづき，第 2 節で述べた好感度自動推定モデルを学習する
(2) 多人数の音声合成用コーパス内の音声それぞれに対し，HuBERT 特徴量，ECAPA-TDNN による話者埋め込み表現，(1) で構築したモデルによる推定好感度評点を付与する
(3) 全 HuBERT 特徴量に対して k-means 法によるクラスタリングを行い，クラスタ中心を得る
(4) クラスタ中心にもとづき，各時刻フレームの属するクラスタインデックスの列を得て，離散音声トークン列とする
(5) 離散トークン列，話者埋め込み表現，推定好感度評点を用いて，図 2 に示す FastSpeech 2 を学習する

また，推論は以下の手順により行う．

(1) 入力音声から HuBERT 特徴量を抽出し，学習手順 (3) で得たクラスタ中心にもとづき離散トークン列を得る
(2) 入力音声から ECAPA-TDNN により話者埋め込み表現を得る
(3) 離散トークン列，話者埋め込み表現，目標の好感度評点をモデルに入力し，目的音声を得る

FastSpeech 2 はテキスト音声合成モデルとして提案されているため，トークン（シンボル）の継続長推定機構が含まれる．提案手法ではこの特徴を利用し，同じ離散トークン列が連続する場合は 1 つのトークンに圧縮する．たとえば，音声から [13, 7, 7, 21, 21, 5, 7] の離散トークンが得られた場合，モデルへの入力は [13, 7, 21, 5, 7] とする．この手法により，話者埋め込み表現や目標好感度評点に応じて継続長を変化させることが可能であり，目標好感度に応じた話速の変化が期待できる．

好感度制御は話者性の保持とのトレードオフであり，好感度制御の強度を高めるほど話者性の保持が困難になり，逆もまた同様である．提案モデルでは，好感度評点を埋め込む線形層の出力に対してスカラー値 s を乗算可能にすることで，このパラメータを利用し話者性とのトレードオフを制御する．学習時には $s = 1$ とし，推論時に s を増減して利用する．

4. 好感度自動推定手法の評価

4.1 実験条件

学習データとなる CocoNut-Humoresque は，1500 音声の train セット，300 音声の valid セット，300 音声の test セットからなる．このうち，train セットを学習データ，valid セットを検証データとして用い，test セットを評価に利用した．各音声は約 4 秒のステレオ信号であり，サンプリング周波数は 44 100 Hz である．これらの音声をチャンネル平均によりモノラル化し，22 050 Hz にダウンサンプ

[*1] 当該文献 [37] では話者性やピッチ情報の保持が困難であると指摘されている．著者の日本語音声に対する事前実験によれば，完全に話者性やピッチ情報が失われているのではなく，部分的に保持されているようである．

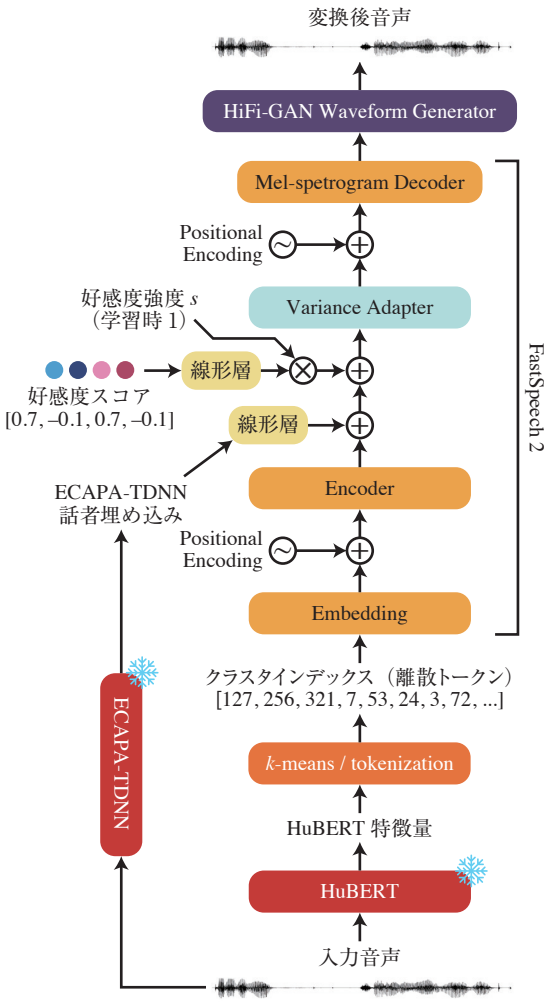


図 2 任意話者の音声好感度を制御可能な声質変換手法のネットワークアーキテクチャ．HuBERT [31] により得られた特徴量を k-means 法を利用してトークン化して利用する．得られた離散トークン列を用いて FastSpeech 2 [32] を駆動し，制御後のメルスペクトログラムを生成する．FastSpeech 2 のエンコーダ出力に，ECAPA-TDNN による話者埋め込み表現 [33] および好感度評点の埋め込み表現を加算する．波形生成には HiFi-GAN [34] を採用する．

リングして用いた．

アーキテクチャは図 1 に示すとおりで，各層の次元数は表 1 に示す．特徴量には 8000 Hz 以下の対数メルスペクトログラムを用いた．シフト長は 256 サンプル（0.12 s）とし，メル尺度方向のビン数は 80 とした．損失関数には評点の平均二乗誤差を用い，モデルは検証セットに対するスピアマンの順位相関係数にもとづき選択した．

学習時のデータ拡張として，学習データに対し，音声前後の無音の追加，音声の時間方向の反転，残響の追加，白色雑音の加算を無作為に適用した．残響のインパルス応答のコーパスには MIT Acoustical Reverberation Scene Statistics Survey [39] のデータセットを用いた．

表 1 好感度の自動推定に用いるニューラルネットワークの，各層の時間的コンテキストおよび次元数．Layer context はその層の考慮する時間的コンテキスト，Input×output は各層の入出力次元数を表す．t は時間フレームのインデックスを，T は総フレーム数を表す．

Layer	Layer context	Input×output
frame1	$\{t-2, t, t+2\}$	240×32
frame2	$\{t-6, t-3, t, t+3, t+6\}$	160×32
frame3	$[t]$	32×32
stats pooling	$[0, T)$	$32T \times 64$
segment4	[0]	64×32
segment5	[0]	32×4

表 2 CocoNut-Humoresque の test セットに対する好感度推定性能．MSE はコーパス内の評価との平均二乗誤差，Std はセット内の評点の標準偏差を表す．評価は –1 から 1 に線形変換されている．LCC，SRCC，KTAU は VoiceMOS Challenge [40] で用いられる評価指標で，それぞれ相関係数，スピアマンの順位相関係数，ケンドールの順位相関係数を表す．なお，これらの相関係数は，線形のポストフィルタリング前後では変化しない．

	フィルタ前		フィルタ後					正解
聴取者	MSE	Std	MSE	Std	LCC	SRCC	KTAU	Std
M –39	0.10	0.17	0.17	0.40	0.36	0.39	0.26	0.32
M 40–	0.07	0.16	0.12	0.34	0.41	0.43	0.28	0.28
F –39	0.11	0.17	0.19	0.44	0.40	0.41	0.28	0.35
F 40–	0.09	0.17	0.17	0.41	0.38	0.39	0.26	0.31
全体	0.05	0.16	0.08	0.29	0.46	0.49	0.33	0.26

4.2 結果

CocoNut-Humoresque の test セットに対する好感度予測の性能を表 2 に示す．提案手法と正解データ間で 0.46 の相関係数（$p < 3 \times 10^{-17}$）での相関が見られ，提案手法が好感度評点を有意に推定できることが確認された．また，線形変換によるポストフィルタリングを適用することで，推定評点–正解評点間の相関係数を保持したまま，推定結果の分散を正解データに近づけることが可能であると確認された．また，平均評点にもとづく音声に対する「好き」「嫌い」の 2 クラス分類を考えたとき，提案手法により 74%の正解率（$F = 0.70$）で推定され，提案手法は音声が好まれやすいか否かを推定可能であることが示された．

コーパス中の好感度評点および推定された好感度評点の関係を図 3 に示す．前述のように有意な相関が見られるものの，各音声に着目すれば正解評点と推定評点に大きな差が見られる場合があり，あらゆる音声に対する好感度自動推定についての精度は限定的であることが示唆された．

5. 音声好感度制御手法の評価

5.1 実験条件

学習および評価音声として，多人数音声コーパスである JVS コーパス [41] 中の 100 人の話者による 14997 発話

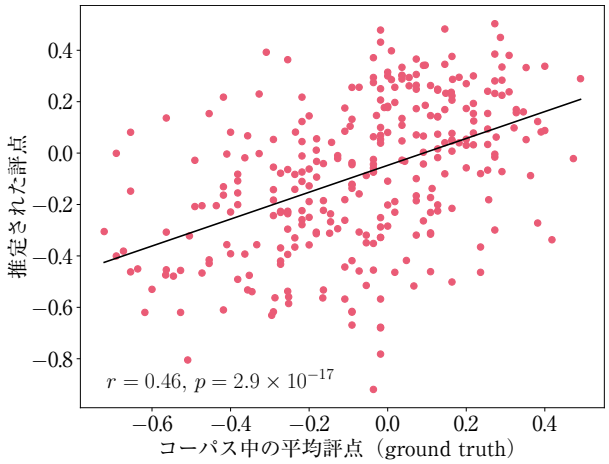


図 3 CocoNut-Humoresque の test セット中の各音声に対する平均好感度評点と推定好感度評点との比較．横軸はコーパス中に含まれる評点を，縦軸は推定評点を表す．r および p はそれぞれ相関係数および p 値を表す．

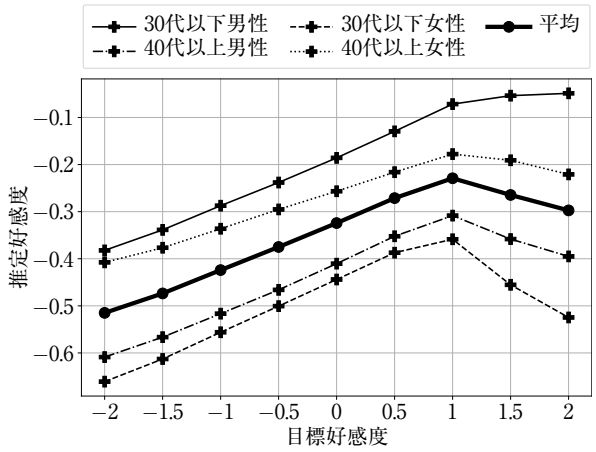


図 4 合成音声に対する好感度自動推定の結果．横軸は入力となる目標好感度，縦軸は合成音声の推定好感度を表す．細線の系列は対象の聴取者それぞれに対する系列を，太線の系列は 4 属性の結果を平均した系列を表す．

（ささやき声および裏声を含む）[*2]，および多人数感情音声コーパスである JTES コーパス [42] 中の 100 人の話者による 20000 発話を用いた．このうち，JTES コーパス中の話者 f49，f50，m49，m50 が発話した，平常文第 41 文から第 50 文の，合計 40 文を評価に用いた．また，JVS コーパスの全音声，および JTES コーパス中の話者 f01–f48 および m01–m48 の 96 話者の前述 10 文をのぞく 190 文を，学習および検証に用いた．すなわち，評価に用いた音声と同一の話者や言語的内容の音声は，学習データには含まれていない．話者埋め込み表現および好感度評点は各音声から個別に抽出して利用し，話者内での平均埋め込み表現および平均評点は用いなかった．

HuBERT 特徴量に対するクラスタリングは mini-batch k-means 法 [43] により行い，クラスタ数 k は 1000 とした[*3]．FastSpeech 2 にはオープンソース実装[*4]を用い，モデルアーキテクチャおよび学習スケジュールは LJSpeech データセットに対する標準設定をそのまま適用した．波形生成に用いる HiFi-GAN のモデルには学習済みの universal モデル[*5]を利用し，学習データに対して teacher-forcing により再生成したメルスペクトログラムを用いた fine-tuning [46] を行った．学習済み HuBERT モデルには，約 6 万時間の日本語テレビ放送音声を用いて学習された HuBERT Large モデル [47] を用いた．話者埋め込み表現抽出には，VoxCeleb1 および VoxCeleb2 で学習された ECAPA-TDNN を利用したモデルである `speechbrain/spkrec-ecapa-voxceleb`[*6] [48] を用いた．

各評価音声に対して，−2, −1.5, . . . 1.5, 2 の 9 通りの目標好感度で好感度制御を行った．学習データの好感度は −1 以上 1 以下であるため，目標好感度が −1.5 以下および 1.5 以上の場合は学習データには含まれておらず，これらの場合は好感度の外挿にあたる．制御の強度 s について，学習時には $s = 1$ とし，評価音声の合成の際には $s = 2.5$ とした．

5.2 客観的評価

まず，第 4 節で構築した好感度自動推定モデルを用いて，合成音声の好感度を推定した結果を図 4 に示す．いずれの話者においても，入力の目標好感度に応じて，合成音声の好感度が制御されていることが示された．ただし，目標好感度の範囲が −2 から 2 であるにもかかわらず，合成音声の平均評点は −0.51 から −0.23 であり，合成音声の評点の範囲は明らかに小さい．これは，本手法は話者性および言語情報を保持した好感度制御であり，話者性および言語的内容に制約されているためであると推測される．

さらに，言語情報に関する客観評価指標として，書き起こし文に対する合成音声の音声認識結果の文字誤り率（character error rate; CER）を調査した．音声認識器の特徴量には，離散トークン列の抽出に用いたモデルと同一の HuBERT Large モデルの出力特徴量を用いた．音声認識モデルは，LaboroTVSpeech [49] を用いて学習された Conformer [50] である．この結果を図 5 に示す．目標好感度が外挿となる場合をのぞき，目標好感度が言語情報に大きな影響を与えないことが示された．合成音声の CER は，男性話者の音声については自然音声との大きな CER の差

[*2] 話者 jvs030，jvs074，jvs089 について，それぞれ 1 音声収録されていない．
[*3] 離散トークンの種類数と合成品質に関しては，日本語の場合 $k = 1000$ のとき品質が高いとされている [44], [45]．一方，声質変換においての最適クラスタ数は十分に調査されておらず，文献によっては $k = 100$ とする場合も見られる [36]．本稿では前者の経験および事前実験の結果をもとに $k = 1000$ とした．
[*4] `https://github.com/ming024/FastSpeech2`
[*5] `https://github.com/jik876/hifi-gan`
[*6] `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`

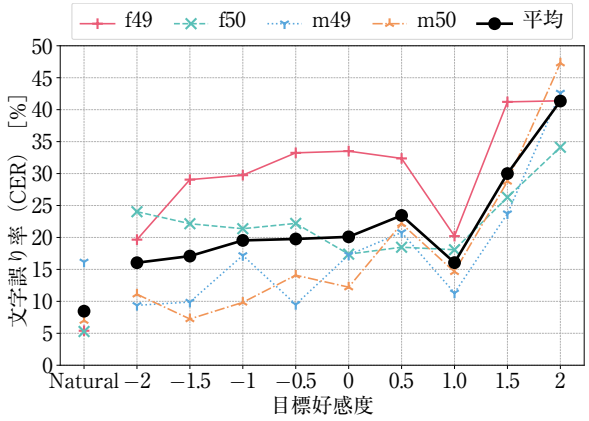


図 5 合成音声に対する音声認識の文字誤り率．横軸は入力となる目標好感度，縦軸は合成音声の文字誤り率を表す．最も左の列は，変換前の自然音声の文字誤り率を示す．線の系列は入力話者ごとの文字誤り率を，太線の系列は 4 話者の結果を平均した誤り率を示す．

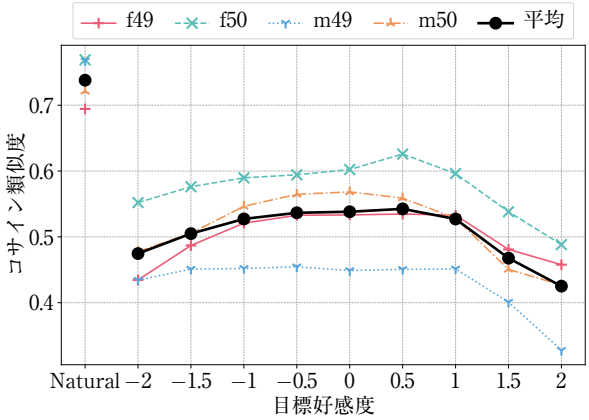
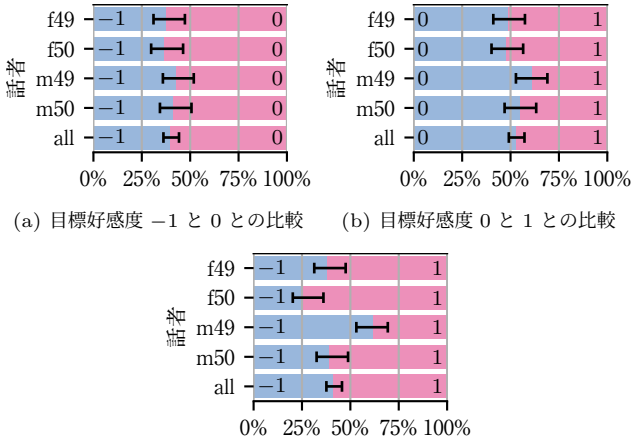


図 6 入力話者の音声および合成音声からそれぞれ抽出した話者埋め込み表現のコサイン類似度．コサイン類似度は −1 以上 1 以下で定義され，オープンソース実装[*8]によれば 0.25 を超えると同一の話者と推定される．横軸は入力となる目標好感度を表し，最も左の列は入力となる自然音声を表す．細線の系列は対象の聴取者それぞれに対する系列を，太線の系列は 4 属性の結果を平均した系列を表す．

は見られないものの，女性話者の音声では自然音声と比較して CER が大きく悪化した．これは，好感度制御の影響ではなく声質変換モデルによる制約と考えられ，さらに高品質な変換には，特徴量のクラスタリング手法やモデルアーキテクチャの改善が必要であると推察される．

また，話者性に関する客観評価指標として，参照音声と合成音声との話者埋め込み表現のコサイン類似度を調査した．ここで，参照音声は平常発話の第 1 文から第 40 文を用い，話者埋め込み表現には変換モデルと同一の `speechbrain/spkrec-ecapa-voxceleb` を用いた．この結果を図 6 に示す．目標好感度が −1 から 1 の範囲では話者表現の品質に大きな変化はなく，異なる目標好感度であっても話者性が維持されていることが示された．一方，目標好感度が 1 を超える場合には話者性が大きく損われ，変換品質が劣化することが確認された．

[*8] `https://github.com/speechbrain/speechbrain/blob/071b743520e5c2497d46dc6301b5d0957ab08121/speechbrain/inference/speaker.py#L61`



(a) 目標好感度 −1 と 0 との比較　(b) 目標好感度 0 と 1 との比較

(c) 目標好感度 −1 と 1 との比較

図 7 目標好感度の異なる音声組に対して，好感度が高いと評価された音声の割合．各行が話者を表し，all はすべての話者を合わせた結果を表す．エラーバーはカイ二乗検定にもとづく 95% 信頼区間を表す．

5.3 主観的評価

音声の好感度が制御されていることを主観的に評価するため，聴取者に合成音声を評価させた．評価に用いた 4 話者のそれぞれ 10 音声に対して，−1，0，1 の 3 通りの目標好感度で声質変換を行った．同一話者同一文で目標好感度が異なる 2 音声を聴取させ，どちらが好感度が高いかを強制選択させた．各聴取者は，すべての組み合わせ（−1 と 0，0 と 1，−1 と 1）で，無作為に選ばれた各話者 3 文を評価した．一人あたりの評価数は 36 組であり，音声の提示順序は無作為に並び替えた．評価人数は 100 人であり，各評価につき 120 円が支払われた．

結果を図 7 に示す．話者 m49 をのぞく 3 話者で，目標好感度に応じて好感度が制御されたことが示された．とくに，目標好感度が −1 と 0 の間で大きく好感度が変化し，総じて目標好感度 −1 と 1 の間において好感度の有意な差が見られた．一方，目標好感度が 0 の場合と 1 の場合では好感度の大きな変化は見られなかった．これは，図 5 が示唆するように，目標好感度を高めた場合，言語的な情報が保持されず自然性が劣化したためだと推察される．話者 m49 は他の話者と異なり，目標好感度が 0 から 1 にかけて好感度の減少が見られ，目標好感度 −1 と 1 の間においては目標好感度と反する有意な結果が表れた．話者 m49 については，図 6 が示唆するとおり話者性が維持されず不自然な音声が合成されており，加えて高い目標好感度の場合も自然性の劣化が示唆されることから，音声の自然性により主観的な好感度が低下したと推察される．

6. おわりに

本稿では，パラ言語情報と非言語情報にわたる主観的な音声情報として音声好感度に着目し，話者性および言語情報を保持して音声の好感度のみを選択的に制御する手法を提案した．提案手法では，話者埋め込み表現および目標好感度評点を補助特徴量として，HuBERT 特徴量にもとづく離散トークン列を用いた任意話者声質変換モデルにより好感度制御を実現した．任意話者声質変換モデルの学習には高品質な多人数音声コーパスを用いる必要があるが，音声に主観的な好感度の評点が付与されたコーパスはいまだない．そこで提案手法では，音声好感度コーパスを用いて事前に音声好感度自動推定モデルを学習し，これを用いて合成用音声コーパスに対し好感度評点を擬似的に付与した．客観的評価により，好感度推定モデルにより好感度評点が有意な相関係数をもって推定できること，また声質変換モデルにより目標好感度に応じて合成音声の好感度が制御できることが明らかになった．また，主観的評価においても，目標好感度にしたがって聴感上の好感度も制御可能であることが確認され，提案手法の有効性が示された．一方，目標好感度が極端な場合や話者によっては自然性が劣化する場合が見られ，さらなる頑健性の向上が必要であることが示唆された．

本稿で評価した好感度自動推定モデルは，全体を通じて好感度推定を達成したものの，個別の音声に着目すればその推定精度は限定的である．この原因として，音声好感度コーパスの規模が約 2 時間であり小さいこと，話速や言語情報などメルスペクトログラムでは表現困難な音響特徴量が好感度に影響していること，などが挙げられる．さらに高精度な推定のためには，大規模な音声好感度コーパスの構築や，SSL 特徴量の活用などを検討する必要がある．本稿で提案した好感度制御モデルにおいては，目標好感度によらず，モデル自体の制約による話者性および言語情報の劣化が示唆された．また，入力話者の特徴や極端な目標好感度評点の入力による，話者性や自然性の劣化が確認された．合成モデルの性能だけでなく，入力話者や目標評点に対する頑健性の向上が必要である．

本稿は，とくに主観的なパラ言語情報および非言語情報として音声好感度に着目した．一方，音声デザインなどの実アプリケーションを考慮すれば，好感度のような単一の評点で表せず，より曖昧かつ漠然とした表現文などを目標として制御できるべきである．本稿で議論した手法はその第一歩をなすものであり，提案手法をさらに拡張した，より柔軟で適合性の高い音声制御技術の構築が必要である．

謝辞 本研究は JSPS 科研費 21H04900，23K20017，23K24895，JST 創発的研究支援事業 JPMJFR226V の支援を受けて実施した．また本研究の一部は，産業技術総合研究所政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」の支援を受けて実施した．音声認識器での自動評価には，産業技術総合研究所の瀧澤大吾氏の協力を得た．

参考文献

[1] Fujisaki, H.: Prosody, models, and spontaneous speech, *Computing Prosody*, Springer US, pp. 27–42 (1997).

[2] Klatt, D. H.: Review of text-to-speech conversion for English, *The journal of the Acoustical Society of America*, Vol. 82, No. 3, pp. 737–793 (1987).

[3] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez Moreno, I. and Wu, Y.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis, *Advances in Neural Information Processing Systems*, Vol. 31 (2018).

[4] Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E. and Ponti, M. A.: YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone, *Proc. 39th International Conference on Machine Learning*, Vol. 162, pp. 2709–2720 (2022).

[5] Lee, Y., Rabiee, A. and Lee, S.-Y.: Emotional end-to-end neural speech synthesizer, *arXiv [cs.SD] 1711.05447* (2017).

[6] Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F. and Saurous, R. A.: Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis, *Proc. 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 80, pp. 5180–5189 (2018).

[7] Barakat, H., Turk, O. and Demiroglu, C.: Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources, *EURASIP journal on audio, speech, and music processing*, Vol. 2024, No. 1, pp. 1–34 (2024).

[8] Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H.: Voice conversion through vector quantization, *Journal of the Acoustical Society of Japan (E)*, Vol. 11, No. 2, pp. 71–76 (1990).

[9] Stylianou, Y.: Voice transformation: A survey, *Proc. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 3585–3588 (2009).

[10] Guo, Z., Leng, Y., Wu, Y., Zhao, S. and Tan, X.: PromptTTS: Controllable text-to-speech with text descriptions, *Proc. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023).

[11] Shimizu, R., Yamamoto, R., Kawamura, M., Shirahata, Y., Doi, H., Komatsu, T. and Tachibana, K.: PromptTTS++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions, *Proc. 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12672–12676 (2024).

[12] Lu, Z., Shen, C., Li, J., Shen, H. and Wigdor, D.: More Kawaii than a real-person live streamer: Understanding how the otaku community engages with and perceives virtual YouTubers, *Proc. 2021 CHI Conference on Human Factors in Computing Systems* (2021).

[13] Weiss, B., Trouvain, J., Barkat-Defradas, M. and Ohala, J. J.(eds.): *Voice Attractiveness: Studies on Sexy, Lik-*

able, and Charismatic Speakers, Springer Nature Singapore (2020).

[14] Collins, S. A. and Missing, C.: Vocal and visual attractiveness are related in women, *Animal behaviour*, Vol. 65, No. 5, pp. 997–1004 (2003).

[15] Saxton, T. K., Caryl, P. G. and Craig Roberts, S.: Vocal and facial attractiveness judgments of children, adolescents and adults: The ontogeny of mate choice, *Ethology: formerly Zeitschrift für Tierpsychologie*, Vol. 112, No. 12, pp. 1179–1185 (2006).

[16] Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C. and Feinberg, D. R.: Voice pitch influences voting behavior, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, Vol. 33, No. 3, pp. 210–216 (2012).

[17] Mayew, W. J., Parsons, C. A. and Venkatachalam, M.: Voice pitch and the labor market success of male chief executive officers, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, Vol. 34, No. 4, pp. 243–248 (2013).

[18] Pavela Banai, I., Banai, B. and Bovan, K.: Vocal characteristics of presidential candidates can predict the outcome of actual elections, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, Vol. 38, No. 3, pp. 309–314 (2017).

[19] Suda, H., Watanabe, A. and Takamichi, S.: Who finds this voice attractive? A large-scale experiment using in-the-wild data, *Proc. Interspeech 2024*, pp. 3165–3169 (2024).

[20] Skrinda, I., Krama, T., Kecko, S., Moore, F. R., Kaasik, A., Meija, L., Lietuvietis, V., Rantala, M. J. and Krams, I.: Body height, immunity, facial and vocal attractiveness in young men, *Die Naturwissenschaften*, Vol. 101, No. 12, pp. 1017–1025 (2014).

[21] Suire, A., Raymond, M. and Barkat-Defradas, M.: Human vocal behavior within competitive and courtship contexts and its relation to mating success, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, Vol. 39, No. 6, pp. 684–691 (2018).

[22] Borkowska, B. and Pawlowski, B.: Female voice frequency in the context of dominance and attractiveness perception, *Animal behaviour*, Vol. 82, No. 1, pp. 55–59 (2011).

[23] Puts, D. A., Barndt, J. L., Welling, L. L. M., Dawood, K. and Burriss, R. P.: Intrasexual competition among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness, *Personality and individual differences*, Vol. 50, No. 1, pp. 111–115 (2011).

[24] Burkhardt, F., Schuller, B., Weiss, B. and Weninger, F.: "Would you buy a car from me?" – On the likability of telephone voices, *Proc. Interspeech 2011* (2011).

[25] Ferdenzi, C., Patel, S., Mehu-Blantar, I., Khidasheli, M., Sander, D. and Delplanque, S.: Voice attractiveness: Influence of stimulus duration and type, *Behavior research methods*, Vol. 45, No. 2, pp. 405–413 (2013).

[26] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K. J.: Phoneme recognition using time-delay neural networks, *IEEE transactions on acoustics, speech, and signal processing*, Vol. 37, No. 3, pp. 328–339 (1989).

[27] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S.: X-vectors: Robust DNN embeddings for speaker recognition, *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333 (2018).

[28] Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y. and Wang, H.-M.: MOSNet: Deep learning-based objective assessment for voice conversion, *Proc. Interspeech 2019* (2019).

[29] Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K., Saurous, R. A. and Sculley, D.: AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech, *Proc. NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop* (2016).

[30] Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S. and Saruwatari, H.: UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022, *Proc. Interspeech 2022* (2022).

[31] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R. and Mohamed, A.: 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 (2021).

[32] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech 2: Fast and high-quality end-to-end text to speech, *Proc. International Conference on Learning Representations* (2021).

[33] Desplanques, B., Thienpondt, J. and Demuynck, K.: ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, *Proc. Interspeech 2020* (2020).

[34] Kong, J., Kim, J. and Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, *arXiv [cs.SD] 2010.05646* (2020).

[35] Huang, W.-C., Wu, Y.-C. and Hayashi, T.: Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations, *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5944–5948 (2021).

[36] van Niekerk, B., Carbonneau, M.-A. and Za A comparison of discrete and soft speech units for improved voice conversion.

[37] Nakata, W., Saeki, T., Saito, Y., Takamichi, S. and Saruwatari, H.: NecoBERT: Self-supervised learning model trained by masked language modeling on rich acoustic features derived from neural audio codec, *Proc. 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (2024).

[38] Watanabe, A., Takamichi, S., Saito, Y., Nakata, W., Xin, D. and Saruwatari, H.: Coco-Nut: Corpus of Japanese utterance and voice characteristics description for prompt-based control, *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2023).

[39] Traer, J. and McDermott, J. H.: Statistics of natural reverberation enable perceptual separation of sound and space, *Proc. National Academy of Sciences of the United States of America*, Vol. 113, No. 48, pp. E7856–E7865 (2016).

[40] Huang, W. C., Cooper, E., Tsao, Y., Wang, H.-M., Toda, T. and Yamagishi, J.: The VoiceMOS Challenge 2022, *Proc. Interspeech 2022*, pp. 4536–4540 (2022).

[41] Takamichi, S., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N. and Saruwatari, H.: JVS corpus: free Japanese multi-speaker voice corpus, *arXiv [cs.SD] 1908.06248* (2019).

[42] Takeishi, E., Nose, T., Chiba, Y. and Ito, A.: Construction and analysis of phonetically and prosodically balanced emotional speech database, *Proc. 2016*

Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 16–21 (2016).

[43] Sculley, D.: Web-scale k-means clustering, *Proc. 19th International Conference on World Wide Web* (2010).

[44] Onda, K., Park, J., Minematsu, N. and Saito, D.: A pilot study of GSLM-based simulation of foreign accentuation only using native speech corpora, *Proc. Interspeech 2024*, pp. 3600–3604 (2024).

[45] Park, J., Saito, D. and Minematsu, N.: Analytic study of text-free speech synthesis for raw audio using a self-supervised learning model, *Proc. 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (2024).

[46] 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, *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783 (2018).

[47] 瀧澤大吾，中村友彦，須田仁志，深山　覚：大規模自己教師あり学習モデルによる日本語方言の音声認識，日本音響学会 2025 年春季研究発表会講演論文集 (2025).

[48] Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D. and Bengio, Y.: SpeechBrain: A general-purpose speech toolkit, *arXiv [eess.AS] 2106.04624* (2021).

[49] Ando, S. and Fujihara, H.: Construction of a large-scale Japanese ASR corpus on TV recordings, *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6948–6952 (2021).

[50] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R.: Conformer: Convolution-augmented transformer for speech recognition, *Proc. Interspeech 2020*, pp. 5036–5040 (2020).