社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

人間の知覚評価フィードバックによる音声合成の話者適応

宇田川健太† 齋藤 佑樹† 猿渡 洋†

† 東京大学大学院 情報理工学系研究科 〒113–8656 東京都文京区本郷 7-3-1

あらまし 本稿では，人間の知覚評価をフィードバックに用いた多話者テキスト音声合成の話者適応を提案する．従来法では，話者識別によって事前学習した話者エンコーダを用いて目的話者の発話から話者埋め込みを抽出していた．しかし，従来法では参照音声を用意できない場合に目的話者の話者埋め込みを得ることができない．提案法では，探索パラメータ空間の線分上から人間に一点を選択させることを繰り返して探索する Sequential Line Search を利用して，目的話者の話者埋め込みを探索する．また，話者埋め込み空間から音声を選択するためのシステムとして，音素ごとに複数の話者の音声を切り替えるシステムを開発した．これらのシステムの実験的評価では，客観評価と主観評価により提案法の有効性を検証する．

キーワード DNN 音声合成，話者適応，人間参加型機械学習，ベイズ最適化，話者埋め込み

Kenta UDAGAWA†, Yuki SAITO†, and Hiroshi SARUWATARI†

† Graduate School of Information Science and Technology, The University of Tokyo, Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8656 Japan

1. はじめに

テキスト音声合成（Text to Speech; TTS）[1] とはコンピュータを用いてテキストから音声を合成する技術であり，近年 Deep Neural Network (DNN) に基づいたテキスト音声合成 [2], [3] が非常に高い自然性を持つ音声を合成できる手法として注目を集めている．テキスト音声合成において，複数の話者の音声を合成するための技術が多話者テキスト音声合成である．DNN に基づいた多話者テキスト音声合成 [4], [5] では，話者性を表す固定長の埋め込みベクトル（話者埋め込み）で音声合成の DNN 音響モデルを条件付けることによって合成音声の話者性を制御するが，話者埋め込みを DNN 音響モデルと共に学習する方式では学習データに含まれていない未知話者の音声を合成することができない．このような未知話者の音声を合成するための技術が話者適応 [6] である．

これまで，話者識別のタスクから転移学習を行うことで音声合成の話者適応を行う手法が提案されている [7]．この手法では，DNN 音響モデルとは別に話者識別のタスクで話者エンコーダを学習する．これにより，話者エンコーダの出力は話者性を表した表現になっていることが期待され，これを話者埋め込みとして DNN 音響モデルを条件付けする．音声合成の際は，まず目的話者の音声波形を話者エンコーダに入力し，目的話者の話者埋め込みベクトルを抽出する．その話者埋め込みベクトルでテキスト音声合成システムを条件付けして，目的話者の音声を合成する．この手法では，数秒程度の少量の音声データで話者適応が可能であるということや，適応の際にモデルのパラメータを更新する必要がないという利点がある．しかし，従来までの話者適応の手法では，目的話者の話者埋め込みを得るために目的話者の音声波形を用意しなければならないという問題点がある．これは参照音声を用意するのが不可能な状況で従来法を使うことができないということを意味する．

本稿では，目的話者の音声波形を用意することができない状況においても，ユーザの知覚評価をランダムな初期値の話者埋め込みにフィードバックすることによって話者適応を行う手法を提案する．提案法では，ユーザの知覚評価に基づく Sequential Line Search (SLS) [8] を用いて目的話者の話者埋め込みを探索する．SLS では探索パラメータ空間上の線分をユーザに選択肢として提示し，ユーザは一つのスライダーを操作することでその線分上から一点を選択する．この選択に基づいて，ベイズ最適化の手法により，ユーザに探索パラメータ空間上の新たな線分を選択肢として提示する．このプロセスを繰り返すことによって，ユーザが望むパラメータ点を探索する．ここで話者埋め込みを探索パラメータとして SLS を適用するとき，2 つの問題が発生する．1 つ目は探索パラメータが話者埋め込みという抽象的なパラメータであり，話者埋め込みから音声波形を直接変形することができないということである．2 つ目は音声が時系列データであり，SLS で操作するスライダー上での音声の変化を知覚できるユーザインターフェースの設計が容易ではないということである．提案法ではこれらの問題点を解決するために，あらかじめ話者埋め込み空間上の線分から複数の音声を合成し

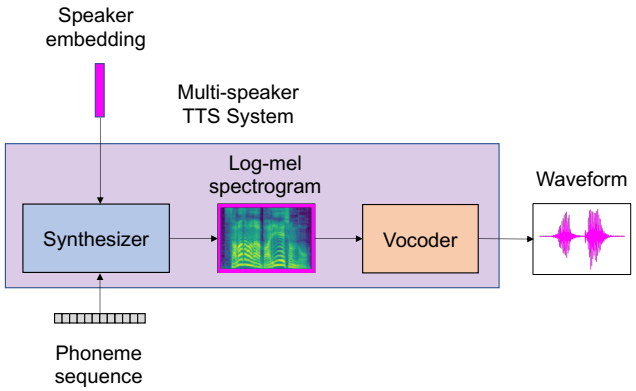

図 1　多話者テキスト音声合成システムの概略図

ておき，再生する音声を音素ごとに切り替えられるシステムを新たに開発している．本システムにより，ユーザは一つのスライダーを操作して，連続的に変化する音声から好みの音声を選択する．実験的評価により，提案法は主観的な自然性に改善の余地があるが，客観評価において従来法と匹敵しうる音声を合成できることを示す．

2.　従来の話者適応

2.1　多話者テキスト音声合成

2.1 節では，本稿で用いた多話者テキスト音声合成システムについて説明する．図 1 に多話者テキスト音声合成システムの概略図を示す．多話者テキスト音声合成システムは，音素列などのテキストと話者埋め込みベクトルから当該話者のメルスペクトログラムを生成するシンセサイザー，メルスペクトログラムから音声波形を合成するボコーダの 2 つの DNN で構成され，それぞれ個別に学習される．シンセサイザーはテキストと音声の対と話者埋め込みを用いて学習され，話者埋め込みが表す話者のメルスペクトログラムを合成する．ボコーダーはメルスペクトログラムと音声波形の対で学習され，メルスペクトログラムから音声波形を合成する．この多話者テキスト音声合成システムの構成は，従来法と提案法で共通した構成になっている．

2.2　話者識別モデルの転移学習

話者識別モデルの転移学習に基づく従来の話者適応 [7] では，図 2 のように音声波形から話者性を表す話者埋め込みベクトルを抽出する話者エンコーダを用いて，目的話者の話者埋め込みを得る．話者エンコーダは，Generalized end-to-end ロス [9] を最小化するように学習される．これにより，話者エンコーダ出力の埋め込みベクトルは同一話者で近く，異なる話者で遠くなるように学習される．話者エンコーダは背景雑音や残響を含んだ音声データで DNN 音響モデルとは別に学習され，学習データの話者数が多いほど話者適応の汎化性能が上がるという実験結果が報告されている [7]．シンセサイザーはテキストと音声の対に加えて，音声波形から事前学習済み話者エンコーダで抽出した話者埋め込みで学習される．従来法で話者適応を行う際は，目的話者の数秒ほどの参照音声から話者埋め込みを抽出し，抽出した話者埋め込みと任意のテキストを多話者テキスト音声

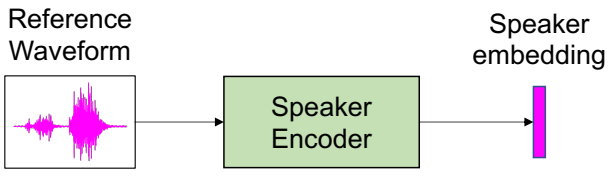

図 2　従来の話者適応の概略図

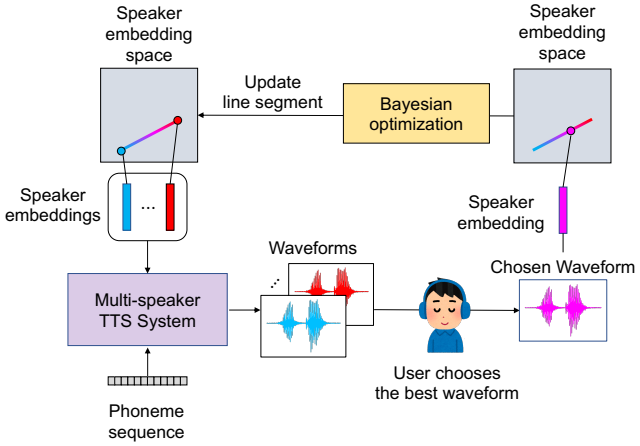

図 3　提案する話者適応の概略図

合成システムに入力することにより，目的話者の話者性を表した音声波形を合成する．

2.3　従来法の問題点

従来法では，目的話者の話者埋め込みを獲得するために目的話者の音声波形を用意する必要がある．これは，例えば目的話者が故人である場合や声を発することができない場合など，参照音声が用意できない状況で従来法を用いることができないということを意味する．また従来法では，合成された音声がユーザが望む話者性を表した音声になっていない場合にフィードバックを行う仕組みを有していない．

3.　提案する話者適応

従来の話者適応の問題点を受けて，本稿では人間の知覚評価をフィードバックに用いて話者埋め込みを探索する人間参加型の手法を提案する．提案法の概略図を図 3 に示す．提案法では，ユーザの知覚評価に基づく Sequential Line Search (SLS) [8] を用いて目的話者の話者埋め込みを探索する．

3.1　Sequential Line Search (SLS)

3.1 節では，提案法の探索アルゴリズムとして用いられている Sequential Line Search (SLS) [8] を説明する．D を探索パラメータの次元数，探索パラメータ空間を $\mathcal{X} = [0,1]^D$，ユーザの知覚評価を $g : \mathcal{X} \to \mathbb{R}$ とする．SLS ではパラメータ空間 $\mathcal{X}$ 上の線分からユーザにパラメータ点を選択させるプロセスを繰り返すことで，次式で定義される $\mathbf{x}^*$ を求める．

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) \tag{1}$$

SLS で t 回目に提案する線分を S_t，S_t の端点を $\mathbf{x}^+_{t-1}, \mathbf{x}^{\mathrm{EI}}_{t-1}$ とす

る．ただし，初期値の端点 $\mathbf{x}_0^+, \mathbf{x}_0^{\mathrm{EI}}$ は探索パラメータ空間上のランダムな 2 点とする，線分 S_t 上の選択肢から，ユーザがパラメータ点 $\mathbf{x}_t^{\mathrm{chosen}}$ を選択することを次式で表す．

$$\mathbf{x}_t^{\mathrm{chosen}} \succ \{\mathbf{x}_{t-1}^+, \mathbf{x}_{t-1}^{\mathrm{EI}}\} \tag{2}$$

また，t 回目までの選択の情報を次式で表す．

$$D_t = \{\mathbf{x}_{\mathrm{chosen}} \succ \{\mathbf{x}_{i-1}^+, \mathbf{x}_{i-1}^{\mathrm{EI}}\}\}_{i=1}^t \tag{3}$$

ここで，$t+1$ 回目に提案する線分 S_{t+1} の端点 $\mathbf{x}_t^+, \mathbf{x}_t^{\mathrm{EI}}$ は，次式から決定される[1]．

$$\mathbf{x}_t^+ = \mathbf{x}_t^{\mathrm{chosen}} \tag{4}$$

$$\mathbf{x}_t^{\mathrm{EI}} = \arg\max_{\mathbf{x}\in\mathcal{X}} a^{\mathrm{EI}}(\mathbf{x}; D_t) \tag{5}$$

$\mathbf{x}_t^+, \mathbf{x}_t^{\mathrm{EI}}$ はそれぞれ観測点のうち最良の評価値である点とパラメータ空間 $\mathcal{X}$ でより良い評価値を最も得やすい点を表す．ただし，$a^{\mathrm{EI}}(\cdot)$ は expected improvement (EI) [10] に基づく獲得関数であり，g^+ を観測点のうち最良の評価値として，次式で定義される．

$$a^{\mathrm{EI}}(\mathbf{x}; D_t) = \mathbb{E}[\max\{g(\mathbf{x}) - g^+, 0\}] \tag{6}$$

ここで獲得関数 $a^{\mathrm{EI}}(\mathbf{x}; D_t)$ を計算するために，$g(\cdot)$ にガウス過程を仮定する．これにより未観測点 $\mathbf{x}_*$ での $g(\mathbf{x}_*)$ の値は以下の正規分布に従う．

$$g(\mathbf{x}_*) \sim \mathcal{N}(\mu(\mathbf{x}_*), \sigma(\mathbf{x}_*)) \tag{7}$$

ただし，$\mu_t(\mathbf{x}_*), \sigma_t(\mathbf{x}_*)$ は，観測点 $\{\mathbf{x}_i\}_{i=1}^{N_t}$ での評価値 $\mathbf{g} = (g(\mathbf{x}_1), \cdots, g(\mathbf{x}_{N_t}))^\top$ とモデルパラメータ $\boldsymbol{\theta}$ に依存する．観測点での評価値 $\mathbf{g}$ とモデルパラメータ $\boldsymbol{\theta}$ は最大事後確率 (maximum a posteriori, MAP) 推定によって推定される．これにより，未観測点 $\mathbf{x}_*$ での評価値 $g(\mathbf{x}_*)$ の確率分布が閉形式で表され，獲得関数 $a^{\mathrm{EI}}(\mathbf{x}; D_t)$ を計算でき，$t+1$ 回目に提案する線分 S_{t+1} の端点 $\mathbf{x}_t^+, \mathbf{x}_t^{\mathrm{EI}}$ を (4), (5) 式から決定できる．

3.2 音声を対象とした SLS

提案法では，話者埋め込み空間の線分上からユーザが話者埋め込みを選択するために．話者埋め込みの線分上の複数の話者埋め込みから合成した音声を音素ごとに切り替えられるシステムを採用している．このシステムではスライダーの位置と合成した音声が対応しており，ユーザが音声を選択する際に，システムはユーザが操作しているスライダーの位置に対応する音声を再生する．再生している音声は音素から次の音素に移るタイミングでスライダーが指している音声に切り替えられる．このようにしてユーザは複数の音声を連続的に切り替えて，目的話者に近いと思う音声を選択する．合成音声の音素アライメントは用いたテキスト音声合成システムの継続長予測器の予測結果を用いている．また，音声を繋げる際はクロスフェード処理を施している．

(注1)：$\mathbf{x}_t^+$ の定義は観測点における後述の $\mu(\mathbf{x})$ の最大値とする設定もあるが，本稿では (4) 式の定義を用いている．

4. 実験的評価

4.1 実験条件

実験的評価において，話者エンコーダ学習用のデータは CSJ コーパス [11] を用いた．CSJ コーパスは日本人話者 1417 名（男性話者名 947 名，女性話者 470 名）の計 660 時間の発話データを含む．CSJ の音声データはサンプリング周波数 16 kHz にダウンサンプリングし，フレームシフトは 10 ms とした．シンセサイザーの学習・検証・評価用のデータは JVS コーパス [12] のパラレルデータを用いた．JVS コーパスのパラレルデータは日本人話者 100 名（男性話者 49 名，女性話者 51 名）の計 22 時間の発話データ（話者ごとに 100 文）を含む．学習データは話者 90 名分の計 20 時間の発話データを用い，学習データに含まれていない 10 名は男女 5 名ずつからランダムにサンプリングした．評価データは学習データに含まれていない 10 名から，話者適応の困難性におけるコーナーケースとして，学習データの話者との主観的類似度 [13] の平均が高い男性（“jvs078”），低い男性（“jvs005”），高い女性（“jvs060”），低い女性（“jvs010”）の 4 名分を選択した．残りの 6 名分は検証データとして使用した．JVS コーパスのパラレルデータには発話内容と発話ラベルが一致していないデータや，収録の失敗により音声が極端に短いデータが含まれているため，それらを除いたデータを用いた[2]．評価話者 4 名のうち “jvs060” の発話数は前処理により 99 であり，それ以外の話者では発話数は 100 であった．JVS の音声データは使用する学習済みニューラルボコーダーの設定に合わせるため，22.05 kHz の周波数にリサンプリングし，フレームシフトは 12 ms とした．

テキストからメルスペクトログラムを合成するためのシンセサイザーは，ming024 により公開されている FastSpeech 2 [14] のオープンソース実装[3]を用いた．FastSpeech 2 の学習時の最適化には，Warmup [15] を用いて学習率スケジューリングを行った Adam [16] を用い，Warmup step は 4000，学習率の初期値は 0.0625，バッチサイズは 8，学習ステップは 50000 とした．FastSpeech 2 を多話者テキスト音声合成に拡張するための話者埋め込みは以下の 2 種類を使用した．

- **EMB256dim**：話者エンコーダの LSTM の最終層の隠れ状態に続く 256 次元への全結合層の活性化関数に ReLU [17] を用い，その後 L1 正規化を用いて推論した 256 次元の話者埋め込み．これは従来法 [7] の話者エンコーダと同じ構造である．
- **EMB16dim**：話者エンコーダの LSTM の最終層の隠れ状態に続く全結合層の出力次元を 16 次元にし，活性化関数に sigmoid を使って推論した 16 次元の話者埋め込み．話者埋め込みを 16 次元に落とした理由は，ベイズ最適化において有効な探索パラメータの次元数は 10-20 程度が限界とされているからである [18]．また，活性化関数に sigmoid を用いている理由は，SLS が探索パラメータ空間として $[0,1]^D$ （D は探索パラメー

(注2)：JVS コーパスの前処理は以下のレポジトリを参考にした．https://github.com/Hiroshiba/jvs_hiho

(注3)：https://github.com/ming024/FastSpeech2

タの次元数）の超立方体を想定しているためである．
2 種類の話者埋め込みはどちらも全結合層＋ ReLU ＋全結合層を通して 256 次元に射影されてからテキストエンコーダ出力に加算される構成にした．話者エンコーダの学習時の最適化は，学習率 0.0001 とした Adam [16] を用い，バッチサイズは 8，学習ステップは 1000000 とした．本稿では，“EMB256dim” を用いた従来法，“EMB16dim” を用いた提案法に加えて “EMB16dim” を用いた従来法の性能も調査した．80 次元のメルスペクトログラムから時間領域の波形へ変換するためのボコーダーは ming024 により公開されている FastSpeech 2 のレポジトリにある HiFi-GAN [19] の generator_universal モデルを用いた．

SLS はユーザが最後に選んだパラメータ空間上の点を観測点の内で最良の点とする設定を用いた．SLS の線分上から選べる点は 20 点とした．また SLS は (4),(5) 式の端点を繋げた線分を 1.25 倍に拡大してからユーザに提示する設定を有しているが，今回はシミュレーションと人間が操作する場合で設定を同一にするために線分の拡大は使用していない．上記以外のハイパーパラメータはデフォルトのものを使用した．

話者適応による合成音声の客観評価指標はメルスペクトログラムの平均絶対誤差（Mean Absolute Error: MAE）を用いた．メルスペクトログラム MAE は，FastSpeech 2 に目的音声の正解音素継続長を与えて計算した．また提案法，従来法ともに FastSpeech 2 で合成した音声のポーズ部分にノイズが観測されたため，音声を合成する際はメルスペクトログラムのポーズ部分をマスクし，メルスペクトログラム MAE はポーズ部分を除く範囲を計算することで対処した．

4.2 シミュレーションによる提案法の客観評価

SLS が提案する線分上から参照音声とのメルスペクトログラム MAE を最小化するように選択して提案法をシミュレーションした．このシミュレーションは本来人間が音声を線分上から参照音声に近い音声を選択する部分を，機械的に参照音声とのメルスペクトログラム MAE が最小になる音声を選択するように置き換えた場合の提案法の性能を確かめたものになる．提案法で話者埋め込みを探索する際に合成する文は JVS コーパスのパラレルデータの 100 発話文のうち VOICEACTRESS100_001 を使用した．評価話者での発話 VOICEACTRESS100_001 の長さは約 8 秒であった．比較手法としては，以下の 2 つの話者埋め込みを従来法として，提案法と比較した．

（1） **TL256dim**：100 発話に対して従来法で抽出した “EMB256dim” の話者埋め込みの平均

（2） **TL16dim**：100 発話に対して従来法で抽出した “EMB16dim” の話者埋め込みの平均

提案法における SLS で線分上から音声を選択する過程を 1 ステップとし，図 4 に提案法を 10 回分シミュレーションした場合の発話 VOICEACTRESS100_001 のメルスペクトログラム MAE 遷移を示す．図 4 の赤線と緑線はそれぞれ 2 種類の従来法（“TL256dim”, “TL16dim”）の発話 VOICEACTRESS100_001 のメルスペクトログラム MAE を意味する．図 4 の濃い青線（提案法の 10 回分のシミュレーションの平均）に着目すると，男性話者 “jvs078”, “jvs005” においては，ランダムな初期値から 30 ステップで従来法と同程度のメルスペクトログラム MAE になっているが，女性話者 “jvs060”, “jvs010” においては 30 ステップで従来法より劣った性能になっている．これは，話者エンコーダの学習用データとして用いた CSJ コーパスに性別の偏りがあり，探索する話者埋め込み空間の性別の割合が偏っていたことが原因として考えられる．

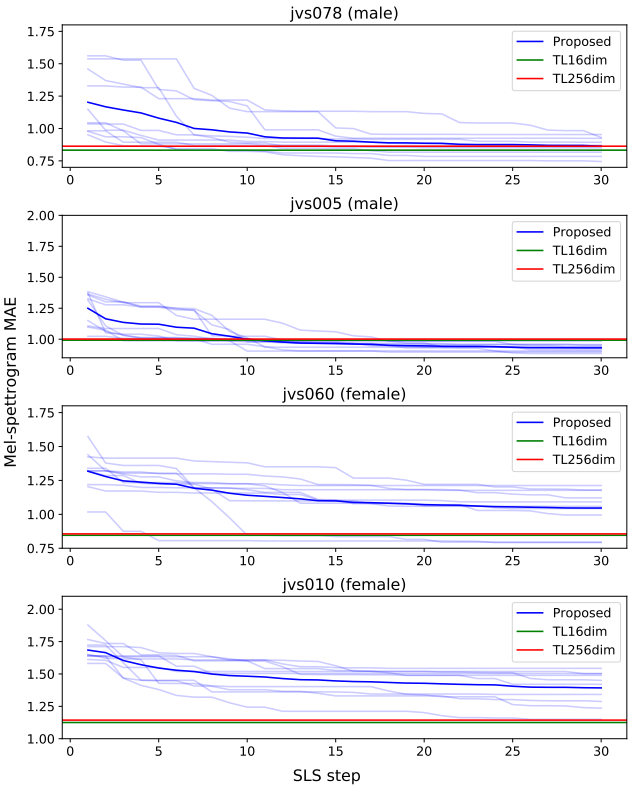

図 4　評価セットの 4 話者ごとに 10 回分のシミュレーションをしたときのメルスペクトログラム MAE 遷移．薄い青線がシミュレーション 1 回分を表し，濃い青線がシミュレーション 10 回分の平均を表す．赤線が “TL256dim”，緑線が “TL16dim” でのメルスペクトログラム MAE を表す．

4.3 人間が操作した場合の提案法の客観評価

提案法を操作者 8 名が操作して話者適応を行う実験を実施した．提案法は参照音声を使わずに人間の頭の中にある目的話者に対して話者適応が行えるが，本稿では操作者に目的話者の音声を十分に認知させるため，提案法の操作中に適宜参照音声を聴くことを許し，その参照音声に近い音声を探索するように指示した．操作者に初めに提示される SLS の線分は，図 4 におけるステップ 1 のメルスペクトログラム MAE が平均値に最も近い線分で話者ごとに固定した．図 4 におけるステップ 1 のメルスペクトログラム MAE は初めに提示される線分上のメルスペクトログラム MAE の最小値を意味する．線分上の選択肢から人間が選択するという条件と初期値の線分を固定する条件以外は，シミュレーション時の条件と同一とした．

図 5 に提案法を操作者 8 名が 30 ステップまで操作した場合の発話 VOICEACTRESS100_001 のメルスペクトログラム MAE 遷移を示す．図 4 と同様に，図 5 の赤線と緑線はそれぞれ 2 種類の従来法（“TL256dim”, “TL16dim”） の発話 VOICEAC-

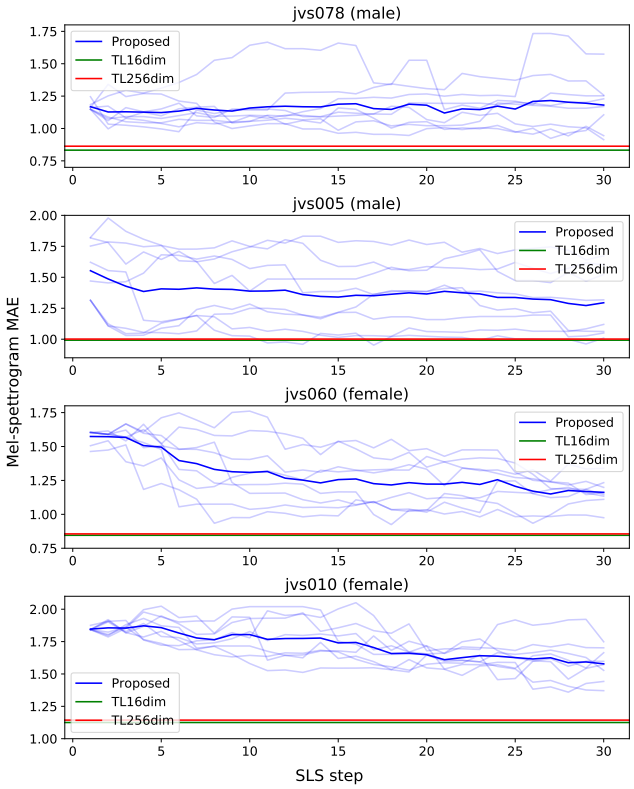

図 5　評価セットの 4 話者で提案法を人間 8 名が操作した時のメルスペクトログラム MAE 遷移，薄い青線が操作 1 回分を表し，濃い青線が操作 8 回分の平均を表す．赤線が “TL256dim”，緑線が “TL16dim” でのメルスペクトログラム MAE を表す．

TRESS100_001 のメルスペクトログラム MAE を意味する．図 5 の濃い青線（提案法の操作 8 回分の平均）に着目すると，“jvs078” 以外の話者で，SLS のステップが進むごとにメルスペクトログラム MAE が減少しており，“jvs060”（女性話者）ではステップ 30 でシミュレーションと同程度の性能になっている．“jvs078” でメルスペクトログラム MAE が減少傾向にない原因として，初期値に提案されている音声が既に操作者の知覚評価基準で目的音声に近く，人間が知覚できる範囲ではメルスペクトログラム MAE を改善できなかったことが挙げられる．

次に，評価話者の 100 発話を用いてメルスペクトログラム MAE を話者ごとに計算し，2 種類の従来法（“TL256dim”, “TL16dim”）に加えて，以下の 3 つの提案法を加えた 5 つの手法を比較した．

（1） **SLS-min**：図 5 の提案法の最終ステップにおける操作者 8 名分の話者埋め込みのうち探索に用いた発話とのメルスペクトログラム MAE が最小になる話者埋め込み

（2） **SLS-mean**：図 5 の提案法の最終ステップにおける操作者 8 名分の話者埋め込みのうち探索に用いた発話とのメルスペクトログラム MAE が平均に最も近い話者埋め込み

（3） **SLS-max**：図 5 の提案法の最終ステップにおける操作者 8 名分の話者埋め込みのうち探索に用いた発話とのメルスペクトログラム MAE が最大になる話者埋め込み

結果を表 1 に示す．従来法としては，“TL256dim” を用いた場合と “TL16dim” を用いた場合でそれほど客観評価指標は変わらないことが分かる．また提案法としては，男性話者 “jvs078”, “jvs005” において “SLS-min” が従来法に匹敵する性能を出している．一方で，女性話者 “jvs060”, “jvs010” においては “SLS-min” は従来法にやや劣る性能になっている．この原因としては，4.2 節で述べたシミュレーションの考察と同様に，話者エンコーダの学習用データとして用いた CSJ コーパスの話者の性別に偏りがあることが挙げられる．また，“SLS-mean”, “SLS-max” においては全話者で従来法より大きく劣る性能になっていることがわかる．この原因としては，

- SLS で探索する際の話者埋め込み空間に実データに現れないような話者埋め込みが含まれていることで探索効率が低くなっている．
- 提案法ではシステム上，音声の一部分のみで話者埋め込みを比較しており，広い範囲での比較がなされていない．
- 人間の知覚評価と客観評価指標の相関がそこまで大きくない．

といったものが考えられる．しかし，“SLS-min” の性能から，探索効率を高める工夫をすれば提案法が 30 程度のステップ数で従来法に匹敵する性能を出せるということを示唆している．

表 1　複数発話のメルスペクトログラム MAE の平均とその 95% 信頼区間

Method \ Speaker	“jvs078”	“jvs005”	“jvs060”	“jvs010”
“TL256dim”	0.82 ± 0.01	0.96 ± 0.01	0.84 ± 0.01	1.05 ± 0.02
“TL16dim”	0.81 ± 0.01	0.94 ± 0.01	0.84 ± 0.01	1.03 ± 0.01
“SLS-min”	0.88 ± 0.01	1.02 ± 0.01	1.02 ± 0.01	1.22 ± 0.01
“SLS-mean”	1.10 ± 0.02	1.23 ± 0.02	1.25 ± 0.02	1.43 ± 0.02
“SLS-max”	1.49 ± 0.03	1.68 ± 0.02	1.41 ± 0.03	1.55 ± 0.03

4.4　人間が操作した場合の提案法の主観評価

提案法と従来法の話者適応によって合成した評価話者 4 名の 100 発話に対して主観評価を実施した．本稿では，表 1 の実験で比較した 5 つの話者埋め込みから合成した音声の自然性と話者類似性を，5 段階の Mean Opinion Score (MOS) テストと Degradation MOS (DMOS) テストによりそれぞれ評価した．Ground-Truth の音声は自然音声とし，自然性の主観評価では比較する 5 種類に加えて Ground-Truth の音声を評価し，話者類似性の主観評価では参照音声として Ground-Truth の音声を用いた．主観評価は評価話者 4 名 × (MOS or DMOS) = 8 個分のタスクを個別に行った．タスクごとの評価者はクラウドソーシングによって集められた 50 名であり，100 発話からランダムに抽出された 5 発話 × 手法数（MOS テストでは 6 個, DMOS テストでは 5 個）の音声サンプルの品質を評価した．ただし，5 発話中 1 発話はダミー音声として用いた．合計の評価セット数は評価話者 4 名 × (MOS or DMOS) × 50（評価者数）= 400 であった．

自然性の主観評価結果を表 2 に示す．従来法としては，“TL256dim” と “TL16dim” に有意水準 5% で有意差はなかった．提案法としては従来法ほどの性能は出せておらず，“jvs060” 以外では提案法のうち最大の MOS 値が従来法の MOS 値から 1 ほど下がった値になっており，“jvs060” では提案法のうち最大の MOS 値が従来法の MOS 値から 0.5 ほど下がった値になって

表 2　合成音声の自然性に関する MOS 値とその 95% 信頼区間

Method \ Speaker	“jvs078”	“jvs005”	“jvs060”	“jvs010”
“TL256dim”	4.13 ± 0.10	4.08 ± 0.12	3.82 ± 0.12	4.07 ± 0.12
“TL16dim”	4.22 ± 0.10	4.10 ± 0.11	3.77 ± 0.13	4.13 ± 0.11
“SLS-min”	3.14 ± 0.11	3.11 ± 0.11	3.30 ± 0.13	2.42 ± 0.12
“SLS-mean”	2.73 ± 0.11	3.03 ± 0.12	2.86 ± 0.13	1.37 ± 0.09
“SLS-max”	1.79 ± 0.12	1.255 ± 0.10	2.88 ± 0.13	2.98 ± 0.14
Ground-Truth	4.56 ± 0.10	4.44 ± 0.11	4.36 ± 0.11	4.14 ± 0.14

表 3　合成音声の話者類似性に関する DMOS 値とその 95% 信頼区間

Method \ Speaker	“jvs078”	“jvs005”	“jvs060”	“jvs010”
“TL256dim”	3.15 ± 0.13	3.02 ± 0.14	3.68 ± 0.12	3.29 ± 0.13
“TL16dim”	3.29 ± 0.14	2.92 ± 0.13	3.49 ± 0.14	3.23 ± 0.13
“SLS-min”	2.54 ± 0.14	1.96 ± 0.12	2.23 ± 0.13	1.48 ± 0.09
“SLS-mean”	2.45 ± 0.13	2.04 ± 0.12	2.15 ± 0.13	1.57 ± 0.12
“SLS-max”	1.51 ± 0.11	1.16 ± 0.08	2.73 ± 0.14	2.35 ± 0.13

いる．しかし，提案法では目的話者の自然音声から当該話者の埋め込みを直接抽出していないのにも関わらず，全話者で MOS 値 3 程度の自然性を持つ音声を合成可能であることを示唆している．“jvs010” では “SLS-mean”，“SLS-min” が “SLS-max” より大きく劣る MOS 値となっているが，この原因として発話速度が平均と大きく離れていることが考えられる．客観評価では目的音声の音素継続長を与えているため，参照音声との発話速度の差はメルスペクトログラム MAE には表れていない．今後は，音素継続長の違いが合成音声の自然性に及ぼす影響を調査する．

話者類似性の主観評価結果を表 3 に示す．従来法としては，“TL256dim” と “TL16dim” に有意水準 5% で “jvs060” に有意差があり，他の話者では有意差はなかった．しかし，“jvs060” 以外の全ての話者で “TL256dim” が “TL16dim” より優れているわけではないため，“TL256dim” と “TL16dim” にそれほど性能差はないと考えられる．提案法としては，自然性評価の結果と同じく従来法ほどの性能は出せていない．自然性と話者類似性の相関を調べるため，全話者の従来法と提案法における自然性の MOS 値と話者類似性の DMOS 値の相関係数を計算したところ，0.90 という値になり，かなり強い相関が見られた．このことから，合成音声の自然性が話者類似性に大きく影響を与えていると考えられる．

5. おわりに

本稿では，参照音声を必要とせず，人間の知覚評価のみをフィードバックに用いる話者適応の手法を提案し，実験的評価によりその有効性を検証した．実験の結果，提案法は主観的な自然性に改善の余地があるが，客観評価において従来法と匹敵しうる音声を合成できることが分かった．今後は，探索する話者埋め込み空間を学習データの範囲に狭めるなど，提案法の探索効率と合成音声の品質を改善するための手法を検討する．

謝辞: 本研究は，JST，ムーンショット型研究開発事業，JPMJMS2011 の支援を受けたものです．

文　献

[1] Y. Sagisaka: “Speech synthesis by rule using an optimal selection of non-uniform synthesis units”, Proc. ICASSP, New York, U.S.A., pp. 679–682 (1988).

[2] H. Zen, A. Senior and M. Schuster: “Statistical parametric speech synthesis using deep neural networks”, Proc. ICASSP, Vancouver, Canada, pp. 7962–7966 (2013).

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis and Y. Wu: “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions”, Proc. ICASSP, Calgary, Canada, pp. 4779–4783 (2018).

[4] H.-T. Luong, S. Takaki, G. E. Henter and J. Yamagishi: “Adapting and controlling DNN-based speech synthesis using input codes”, Proc. ICASSP, New Orleans, U.S.A., pp. 1905–1909 (2017).

[5] N. Hojo, Y. Ijima and H. Mizuno: “DNN-based speech synthesis using speaker codes”, IEICE Transactions on Information and Systems, **E101-D**, 2, pp. 462–472 (2018).

[6] Z. Wu, P. Swietojanski, C. Veaux, S. Renals and S. King: “A study of speaker adaptation for DNN-based speech synthesis”, Proc. INTERSPEECH, Dresden, Germany, pp. 879–883 (2015).

[7] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno and Y. Wu: “Transfer learning from speaker verification to multispeaker text-to-speech synthesis”, Proc. NeurIPS, Montreal, Canada, pp. 4480–4490 (2018).

[8] Y. Koyama, I. Sato, D. Sakamoto and T. Igarashi: “Sequential line search for efficient visual design optimization by crowds”, ACM Trans. Graph., **36**, 4, pp. 1–11 (2017).

[9] L. Wan, Q. Wang, A. Papir and I. L. Moreno: “Generalized end-to-end loss for speaker verification”, Proc. ICASSP, Alberta, Canada, pp. 4879–4883 (2018).

[10] D. R. Jones, M. Schonlau and W. J. Welch: “Efficient global optimization of expensive black-box functions”, Journal of Global optimization, **13**, 4, pp. 455–492 (1998).

[11] K. Maekawa: “Corpus of spontaneous Japanese: Its design and evaluation”, Proc. SSPR, Tokyo, Japan, pp. 7–12 (2003).

[12] 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”, Acoustical Science and Technology, **41**, 5, pp. 761–768 (2020).

[13] Y. Saito, S. Takamichi and H. Saruwatari: “Perceptual-similarity-aware deep speaker representation learning for multi-speaker generative modeling”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **29**, pp. 1033–1048 (2021).

[14] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu: “Fastspeech 2: Fast and high-quality end-to-end text to speech”, Proc. ICLR, Virtual Conference (2021).

[15] P. Goyal, P. Dollár, R. B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia and K. He: “Accurate, large minibatch SGD: Training imagenet in 1 hour”, arXiv, **abs/1706.02677**, (2017).

[16] D. Kingma and B. Jimmy: “Adam: A method for stochastic optimization”, arXiv preprint arXiv:1412.6980 (2014).

[17] X. Glorot, A. Bordes and Y. Bengio: “Deep sparse rectifier neural networks”, Proc. AISTATS, Lauderdale, U.S.A., pp. 315–323 (2011).

[18] R. Moriconi, M. P. Deisenroth and K. S. Kumar: “High-dimensional bayesian optimization using low-dimensional feature spaces”, Machine Learning, **109**, 9, pp. 1925–1943 (2020).

[19] J. Kong, J. Kim and J. Bae: “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis”, Proc. NeurIPS, Virtual Conference, pp. 17022–17033 (2020).