人間の知覚評価をフィードバックに用いた音声合成の話者適応における探索手法の検討 *

☆宇田川 健太, 齋藤 佑樹, 猿渡 洋 (東大院・情報理工)

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 TTS [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．TTS 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 TTS である．DNN に基づいた多話者 TTS [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．

これまでに我々は人間の知覚評価のみをフィードバックに用いた Human-In-The-Loop (HITL) 型の話者適応手法を提案している [7]．この手法は，ユーザが思い浮かべる目的話者を再現するような話者適応が行えるという利点がある一方で，話者識別タスクで事前学習した話者エンコーダを用いる従来手法 [8] と比較して話者適応の品質が大きく劣るという問題点があった．この原因としては，HITL 型の話者適応手法で探索する話者埋め込み空間が，学習データ外の範囲を多く含んでいることが挙げられる．

本稿では，我々の HITL 型話者適応手法の問題点を解決するために，(1) 探索する話者埋め込みの初期値を学習データの平均話者にする方法と，(2) 探索する話者埋め込み空間を学習データの分位数に変換する方法の 2 つを検討し，シミュレーションによってその有効性を示す．また，探索効率を改善することで，HITL 型話者適応手法が従来法 [8] と同程度の性能を達成し得ることを示す．

2 従来の話者適応

2.1 話者エンコーダと参照音声を用いた話者適応

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 [8]．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．話者適応の際は，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 TTS 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，こ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する必要がある．これは参照音声を用意するのが不可能な状況でこの手法を使うことができないということを意味する．

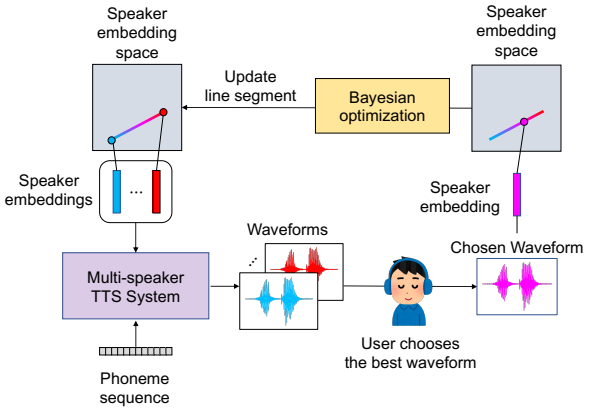

Fig. 1 HITL 型の話者適応の概略図

2.2 HITL 型の話者適応

我々はこれまでに，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している [7]．この手法では，ユーザの知覚評価に基づく Sequential Line Search (SLS) [9] を用いて目的話者の話者埋め込みを探索する．HITL 型話者適応手法の概要を Fig. 1 に示す．

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．我々は，話者埋め込み空間上の線分から音声を効率よく選択するために，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．HITL 型話者適応手法では参照音声を話者エンコーダの入力に使用せず，ユーザの頭の中の話者の音声を合成できるという利点がある．

しかし，この手法によって合成された音声は客観評価，主観評価において 2.1 節の従来法と比べて大きく劣った性能になっていることが分かっている [7]．この原因としては，探索する話者埋め込み空間に学習データの話者埋め込みの分布から離れた部分が多く

* Investigating on search method for speaker adaptation of speech synthesis using human perceptual evaluation as feedback, by UDAGAWA, Kenta, SAITO, Yuki, and SARUWATARI, Hiroshi (The University of Tokyo)

含まれており，探索結果の話者埋め込みから自然な音声が合成されにくいということが考えられる．

3 HITL 型話者適応の改善

2.2 節で述べた HITL 型話者適応手法の問題点を解決する手法を提案する．

3.1 探索の初期値を平均話者にする手法

SLS では探索する際に提示する線分の初期値は探索空間上のランダムな 2 点を端点とした線分としている．本稿では自然な音声から探索を開始するため，初期値の線分の端点は平均男性話者，平均女性話者の話者埋め込みとする．

3.2 探索空間を学習データの分位数に変換する手法

HITL 型話者適応手法では $[0,1]^{16}$ の超立方体に制限された話者埋め込み空間を直接探索していた [7]．探索する話者埋め込み空間を学習データにより近いものにするために，本稿では話者埋め込みの学習データの分位数を探索空間とする．SLS の探索パラメータ $\mathbf{q} \in [0,1]^{16}$ の第 i 成分を q_i，話者埋め込み $\mathbf{e} \in [0,1]^{16}$ の第 i 成分を e_i，N 個の学習データの話者埋め込みの第 i 成分のうち j 番目に小さいものを $e^{\text{train}}_{i,j}$（ただし，$e^{\text{train}}_{i,0} = -\infty, e^{\text{train}}_{i,N+1} = \infty$ とする）として，e_i から q_i への変換 Conv と q_i から e_i への変換 IConv を以下のように定義する（ただし $[\cdot]$ はガウス記号）．

$$\text{Conv}(e_i) = j/N, \ (e^{\text{train}}_{i,j} \leq e_i < e^{\text{train}}_{i,j+1}) \quad (1)$$

$$\text{IConv}(q_i) = e^{\text{train}}_{i,[q_i(N-1)+1]} \quad (2)$$

4 実験的評価

4.1 実験条件

本稿の実験条件は，我々の先行研究 [7] とほぼ同一である．話者エンコーダ学習には CSJ コーパス [10] を，TTS モデルの学習・検証・評価用には JVS コーパス [11] のパラレルデータを使用した．TTS モデルの学習データは話者 90 名分の計 20 時間であり，学習データに含まない 10 名は男女 5 名ずつランダムサンプリングした．評価データはこの 10 名のうち，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 [12] 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 名は検証データとして使用した．

テキストからメルスペクトログラムを合成するための DNN 音響モデルは，ming024 により公開されている FastSpeech 2 [13] のオープンソース実装[1]を用いた．FastSpeech 2 の学習時の最適化には，学習率スケジューリングを行った Adam [14] を用い，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 TTS 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 [15] 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 [8] 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．これは HITL 型話者適応手法で探索する話者埋め込みである．

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 [14] を用い，バッチサイズは 8，学習ステップ数は 1000000 とした．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 [16] 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 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 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

人間が複数の音声から目的話者に近いと思う音声を選択する確率を BTL モデル [17, 18] で定式化し，提案法をシミュレーションした．BTL モデルでは探索パラメータ空間を $\mathcal{X}$，ユーザの知覚評価を $g : \mathcal{X} \to \mathbb{R}$ として，人間が n 個の選択肢 $\mathbf{x}_1, \cdots, \mathbf{x}_n$ から $\mathbf{x}_i$ を選択する確率を以下のように定式化する．

$$p(\mathbf{x}_i; s) = \frac{\exp(g(\mathbf{x}_i)/s)}{\sum_j \exp(g(\mathbf{x}_j)/s)} \quad (3)$$

ただし s はスケール変数を表し，s が小さいとき $g(\mathbf{x})$ の値が高い選択肢がより選ばれやすくなり，s が大きいとき選択はランダムに近づく．本稿では，探索する話者埋め込みを $\mathbf{e}$，メルスペクトログラム MAE を M として，ユーザの知覚評価 g を

$$g(\mathbf{e}) = -\log M \quad (4)$$

のように定義し，SLS が提案する線分上から様々なスケール変数 $(s = \{1.0, 0.1, 0.01\})$ の BTL モデルによって音声を選択して，提案法をシミュレーションした．HITL 型話者適応で話者埋め込みを探索する際に合成する文（探索用発話）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．評価話者での探索用発話の長さは約 8 秒であった．

4.2.1 平均話者を初期値とする手法の有効性の検証

HITL 型話者適応において，平均男性話者，平均女性話者の話者埋め込みを端点とした線分を SLS の初

[1] https://github.com/ming024/FastSpeech2

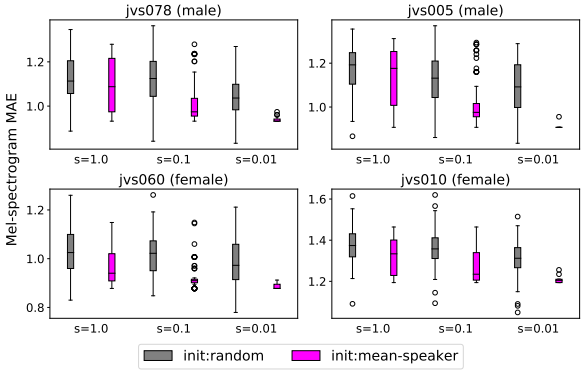

Fig. 2　HITL 型話者適応における SLS の初期値の線分から BTL モデルによって音声を 100 回分選択したときのメルスペクトログラム MAE の箱ひげ図.

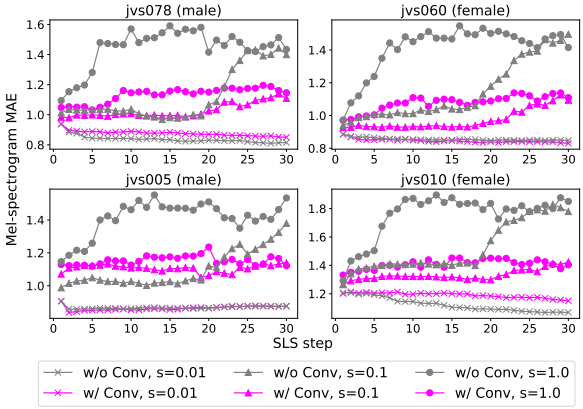

Fig. 3　HITL 型話者適応を 6 つの条件でそれぞれ 20 回分シミュレーションしたときのメルスペクトログラム MAE の遷移（20 回分で平均している）.

期値とする手法の有効性を検証した．探索空間は分位数に変換する条件にした．

SLS の初期値の線分から BTL モデルによって音声を選択したときの探索用発話のメルスペクトログラム MAE の箱ひげ図を Fig. 2 に示す．全ての話者，スケール変数において，初期値を平均話者にすることで，ランダムな初期値から探索開始する場合よりもメルスペクトログラム MAE が低い音声が選択されやすいことがわかる．これは HITL 型話者適応において初期値を平均話者にすることで，より自然な音声から探索を開始できる可能性を示唆している．

4.2.2　分位数に変換した探索空間の有効性の検証

HITL 型話者適応において探索する空間を JVS の学習データの分位数に変換する手法の有効性を検証した．初期値の線分は平均話者にする条件を用いた．

SLS で線分上から音声を選択する過程を 1 ステップとし，Fig. 3 に HITL 型話者適応をシミュレーションしたときの探索用発話のメルスペクトログラム MAE 遷移を示す．スケール変数が 0.01（バツマーク）のとき，分位数への変換によって性能にさほど差が出ていないが，スケール変数が 1.0, 0.1（丸，三角マーク）のとき，分位数への変換ありの方が変換なしの条件より，メルスペクトログラム MAE が低くなっていることがわかる．これは HITL 型話者適応において探索空間を分位数へ変換することで，自然な音声が探索されやすくなる可能性を示唆している．

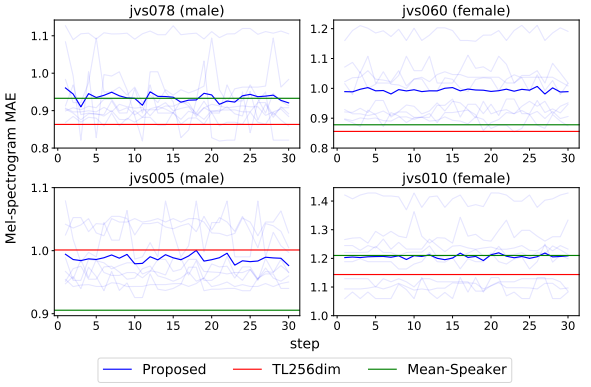

Fig. 4　提案法を人間 8 名が操作した時のメルスペクトログラム MAE 遷移，薄い青線が操作 1 回分を表し，濃い青線が操作 8 回分の平均を表す.

4.3　人間が操作した場合の提案法の客観評価

2 つの改善手法を導入した HITL 型話者適応を提案法として，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．比較手法としては，以下の二つを提案法と比較した．

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. **Mean-Speaker**：JVS の学習データから抽出した “EMB16dim” の話者埋め込みの平均．評価話者が男性のときは平均男性話者，評価話者が女性のときは平均女性話者としている.

“TL256dim” は従来法 [8]，“Mean-Speaker” は提案法の初期値にあたる.

Fig. 4 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探索用発話のメルスペクトログラム MAE 遷移を示す．Fig. 4 の赤線と緑線はそれぞれ “TL256dim”，“Mean-Speaker” の探索用発話のメルスペクトログラム MAE を意味する．Fig. 4 の薄い青線（操作一回分）のうち最もメルスペクトログラム MAE が低い青線に着目すると，提案法は客観評価において従来法に匹敵する音声を合成できることがわかる．しかし，Fig. 4 の濃い青線（提案法の操作 8 回分の平均）に着目すると，全ての評価話者において，人間の操作によってメルスペクトログラム MAE が改善傾向にないことがわかる．この原因としては，操作しているパラメータが抽象的であったために操作タスクが困難であったことや，提案法において初期値に用いている Mean-Speaker のメルスペクトログラム MAE が十分に低かったことが考えられる.

4.4　人間が操作した場合の提案法の主観評価

本稿では従来法，提案法によって合成した評価話者 4 名の 100 発話の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．比較手法としては TL256dim（従来法），Mean-Speaker（提案法の初期値）に加え，Fig. 4 の提案法の最終

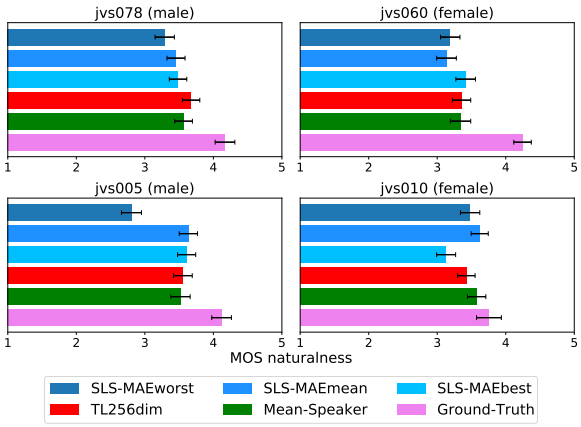

Fig. 5　合成音声の自然性に関する MOS 値とその 95%信頼区間

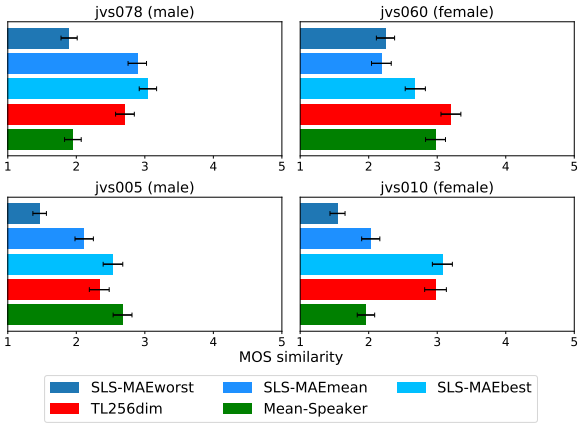

Fig. 6　合成音声の話者類似性に関する DMOS 値とその 95%信頼区間

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, 平均, 最大 } になる話者埋め込みを **SLS-MAE{best, mean, worst}** と定義して，この 3 つを提案法として比較した．

自然性の主観評価では比較する 5 種類の合成音声に加えて自然音声 (Ground-Truth) を評価し，話者類似性の主観評価では参照音声として自然音声を用いた．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．

自然性の主観評価結果を Fig. 5 に示す．まず，“Mean-Speaker” の音声は “TL256dim” と同程度の自然性になっている．これは提案法において従来法と同程度の自然性の音声を含む線分から SLS の探索を開始していることを示している．提案法としては，“jvs005” の “SLS-MAEworst” では従来法に大きく劣ってしまっているが，それ以外では従来法と同程度の自然性になっていることがわかる．

話者類似性の主観評価結果を Fig. 6 に示す．提案法では，“jvs060” 以外の話者で “SLS-MAEbest” が従来法を上回る性能になっており，また “jvs078”, “jvs005” の男性話者で “SLS-MAEmean” が従来法と同程度の性能になっていることがわかる．しかし，提案法では “Mean-Speaker” を下回る性能になっているものがいくつかあることがわかる．これは，初期値の線分に含まれる音声より，話者類似性が人間の探索によって下がってしまっているということを意味する．この原因としては，4.3 節での考察と同様に，操作タスクが困難であったことが考えられる．

5　おわりに

本稿では，これまでに我々が提案した HITL 型話者適応を改善する手法を検討し，実験的評価の結果からその有効性を示した．今後は，ユーザの操作するパラメータをより直感的なものにするために，話者埋め込みに解釈性を与える手法を検討する．

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．

参考文献

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

[2] H. Zen et al., “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 7962–7966.

[3] J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.

[4] H.-T. Luong et al., “Adapting and controlling DNN-based speech synthesis using input codes,” in *Proc. ICASSP*, New Orleans, U.S.A., May 2017, pp. 1905–1909.

[5] N. Hojo et al., “DNN-based speech synthesis using speaker codes,” *IEICE Transactions on Information and Systems*, vol. E101-D, no. 2, pp. 462–472, Feb. 2018.

[6] Z. Wu et al., “A study of speaker adaptation for DNN-based speech synthesis,” in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 879–883.

[7] 宇田川健太 et al., “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,” **信学技報**, vol. SP2021-33, pp. 46–51, Oct. 2021.

[8] Y. Jia et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Proc. NeurIPS*, Montreal, Canada, Dec. 2018, pp. 4480–4490.

[9] Y. Koyama et al., “Sequential line search for efficient visual design optimization by crowds,” *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–11, Jul. 2017.

[10] K. Maekawa, “Corpus of spontaneous Japanese: Its design and evaluation,” in *Proc. SSPR*, Tokyo, Japan, Apr. 2003, pp. 7–12.

[11] S. Takamichi et al., “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, Sep. 2020.

[12] Y. Saito et al., “Perceptual-similarity-aware deep speaker representation learning for multi-speaker generative modeling,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1033–1048, Feb. 2021.

[13] Y. Ren et al., “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *Proc. ICLR*, Virtual Conference, Sep. 2021.

[14] D. Kingma, B. Jimmy, “Adam: A method for stochastic optimization,” in *arXiv preprint arXiv:1412.6980*, 2014.

[15] X. Glorot et al., “Deep sparse rectifier neural networks,” in *Proc. AISTATS*, Lauderdale, U.S.A., Apr. 2011, pp. 315–323.

[16] J. Kong et al., “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NeurIPS*, Virtual Conference, Dec. 2020, pp. 17 022–17 033.

[17] R. A. Bradley, M. E. Terry, “Rank analysis of incomplete block designs: I. the method of paired comparisons,” *Biometrika*, vol. 39, p. 324, 1952.

[18] R. D. Luce, *Individual Choice Behavior: A Theoretical analysis*. New York, NY, USA: Wiley, 1959.