最適輸送による GMM 補間を用いた中間属性の非実在話者生成 *

☆渡邊 亞椰，高道 慎之介，齋藤 佑樹，辛 徳泰，猿渡 洋 (東大院・情報理工)

1 はじめに

ディープニューラルネットワーク (deep neural network: DNN) に基づくテキスト音声合成 (text-to-speech: TTS) [1,2] の拡張として，多話者コーパスで学習する多話者 TTS が提案されている．しかし，提案されたモデルの多くは生成対象の話者の発話データが実在しなければ学習できない．他方，非実在話者の音声を生成するタスクを，Stanton ら [3] は speaker generation （話者生成）と呼称している．

話者生成を達成する手法として，特定のカテゴリカルな話者属性（性別や母語など）を有する話者埋め込みの確率分布（以降，話者埋め込み分布と定義）を混合正規分布モデル (Gaussian mixture model: GMM) で表現する TacoSpawn がある [3]．この手法は，所望のカテゴリカルな話者属性を有する非実在話者を生成できるため，カテゴリカルな話者属性に限らない属性の非実在話者の生成に拡張できる可能性がある．

そこで本研究では，最適輸送に基づいて，複数のカテゴリカル話者属性に対応する話者埋め込み分布同士を補間することで，中間的な話者属性をもつ非実在話者を生成する手法を提案する．この手法では，学習済みの話者埋め込み分布の重み付き重心を計算する．この重み付き重心は元となる分布と対応する話者属性の中間的な話者属性に対応する GMM として推定される．実験的評価では，本手法により生成した非実在話者が知覚的に中間属性を有することを確認する．

2 関連研究

2.1 TacoSpawn

TacoSpawn [3] は多話者 Tacotron [1] の拡張であり，カテゴリカルな各話者属性に対応する話者埋め込み分布を GMM で表現する．

D 次元の話者埋め込み分布を K 混合 GMM で学習する．話者属性エンコーダは，カテゴリカル話者属性 c から，D 次元 K 混合 GMM を示すパラメータである，平均ベクトル系列 $\boldsymbol{\mu}(c) \in \mathbb{R}^{K\times D}$，対角分散系列 $\boldsymbol{\sigma}^2(c) \in \mathbb{R}^{K\times D}$，混合重み $\boldsymbol{\alpha}(c) \in \mathbb{R}^{K}$ を推定する．$\alpha_k(c)$，$\boldsymbol{\mu}_k(c)$，$\boldsymbol{\sigma}_k^2(c)$ をそれぞれ $\boldsymbol{\alpha}(c)$，$\boldsymbol{\mu}(c)$，$\boldsymbol{\sigma}^2(c)$ の k 番目の要素，$\mathcal{N}(\cdot)$ を正規分布とすると，話者属性 c に対応する話者埋め込み分布 $p(\boldsymbol{s}|c)$ は，

$$p(\boldsymbol{s}|c) = \sum_{k=1}^{K} \alpha_k(c)\mathcal{N}\left(\boldsymbol{s}; \boldsymbol{\mu}_k(c), \boldsymbol{I}\boldsymbol{\sigma}_k^2(c)\right) \quad (1)$$

と表せる．$\boldsymbol{I} \in \mathbb{R}^{D\times D}$ は単位行列である．

話者属性エンコーダは，全結合層と，各 GMM パラメータが定義条件を満たすための活性化関数からなる．このモデルは，損失関数に負の対数尤度を採り，TTS モデルと同時に学習される．TacoSpawn の概要を Fig. 1 に示す．

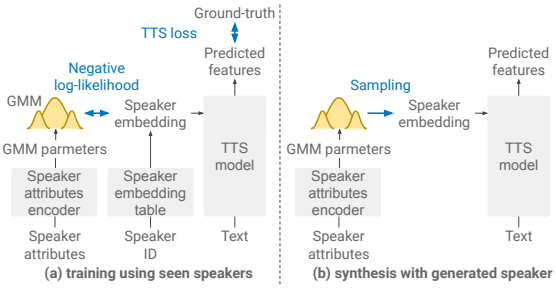

Fig. 1 TacoSpawn の学習及び推論

2.2 最適輸送

最適輸送は，確率分布間における移動を，点間の移動を評価するコスト関数の和に基づいて最適化する問題である．

総質量の等しい分布（例えば，確率分布）p_a から p_b への移動を写像 $T\colon \mathbb{R}^n \to \mathbb{R}^n\colon \boldsymbol{x} \mapsto T(\boldsymbol{x})$ で表現し，この最適化を考える．コスト関数を $C(\boldsymbol{x}, T(\boldsymbol{x}))$ としたとき，最適な T は，

$$\int_{\{\boldsymbol{x}\in\mathbb{R}^n; T(\boldsymbol{x})=\boldsymbol{y}\}} p_{\mathrm{a}}(\boldsymbol{x})d\boldsymbol{x} = p_{\mathrm{b}}(\boldsymbol{y}) \quad (2)$$

の条件下で

$$\min_{T} \int_{\mathbb{R}^n} C(\boldsymbol{x}, T(\boldsymbol{x}))\, p_{\mathrm{a}}(\boldsymbol{x})d\boldsymbol{x} \quad (3)$$

として求められる．ただし，これは不良設定問題であり，実際には Kantorovich 問題 [4] での定義を導入し，1 対 1 の写像ではなく 1 対多の移動について問題を解くことが多い．

ここで得られた最適な移動 T により，分布間の最適な移動経路を得られる．この経路を用いて，分布間の中間と成る確率分布（重みつき重心 [5]）を計算できる．この際，各分布の寄与する割合を任意に決定できるため，この割合を滑らかに変化させることで，分布間を滑らかに補間できる．このコスト関数の設計によって重みつき重心の形状が変化し，例えば，GMM 同士の重みつき重心が GMM となるコスト関数もある [6].

---

* Generation of Mid-Attribute Non-existent Speakers by Gaussian Mixture Model Interpolation based on Optimal-Transport, by Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Detai Xin, Hiroshi Saruwatari (The University of Tokyo).

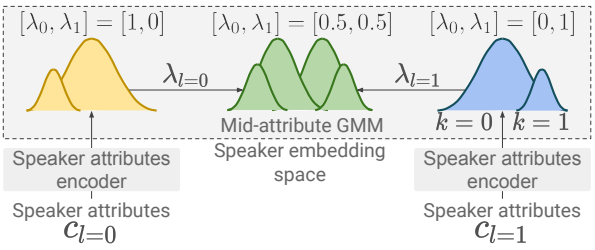

Fig. 2　中間話者属性を持つ GMM（話者属性数 $L = 2$，GMM 混合数 $K = 2$ と仮定）

3　提案手法

3.1　話者属性エンコーダ付き TTS モデル

話者生成のモデルと学習は TacoSpawn [3] と同様である．多話者 TTS モデルは，学習可能な話者埋め込みテーブルを有する系列変換モデルであり，入力テキストと話者埋め込みベクトル $\boldsymbol{s}$ から合成音声の特徴量を予測する．このモデルに付属する話者属性エンコーダは，話者属性 c から，それに対応する話者埋め込み分布（すなわち GMM）のパラメータを推定する．学習過程は Fig. 1 に示す TacoSpawn の学習過程と同様である．ここで，L 種類ある話者属性を $\{c_l | l = 1, 2, \ldots, L\}$，$l$ 番目の話者属性 c_l に対応する話者埋め込み分布である GMM を

$$p_l(\boldsymbol{s}) \equiv p(\boldsymbol{s}|c_l) = \sum_{k=1}^{K} \alpha_{l,k} \mathcal{N}\left(\boldsymbol{s}; \boldsymbol{\mu}_{l,k}, \boldsymbol{I}\boldsymbol{\sigma}_{l,k}^2\right) \quad (4)$$

と定義する．なお，$\alpha_{l,k}, \boldsymbol{\mu}_{l,k}, \boldsymbol{\sigma}_{l,k}$ をそれぞれ $p_l(\boldsymbol{s})$ の k 番目の混合要素の混合重み，平均ベクトル，対角標準偏差ベクトルとする．$p_l(\boldsymbol{s})$ から $\boldsymbol{s}$ をサンプリングすることで，話者属性 c_l を持つ非実在話者を生成できる．

3.2　話者属性補間

L 種類の学習済みの話者埋め込み分布 $\mathrm{GMM}\{p_l(\boldsymbol{s}) | l = 1, 2, \ldots, L\}$ と補間重み $\{\lambda_l | l = 1, 2, \ldots, L\}$ $(\sum_l \lambda_l = 1)$ から計算される重み付き重心を，中間的な話者属性の話者埋め込み分布と見做し，当該属性を持つ非実在話者を生成する方法を述べる．この計算の概略を Fig. 2 に示す．

3.2.1　正規分布同士の重み付き重心

まず，最も単純な場合として，各話者埋め込み分布に正規分布を採る場合の重み付き重心を求める．コスト関数はワッシャーシュタイン距離 W_2 である．2 つの正規分布 $\mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{I}\boldsymbol{\sigma}_0^2\right)$ と $\mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{I}\boldsymbol{\sigma}_1^2\right)$ のワッシャーシュタイン距離は，$\ell_{2,2}$ ノルム $\|\cdot\|_2^2$ を用いて，以下のように示される [7].

$$\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2^2 + \|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_1\|_2^2 \quad (5)$$

L 個の正規分布 $\left\{\mathcal{N}\left(\boldsymbol{\mu}_l, \boldsymbol{I}\boldsymbol{\sigma}_l^2\right) | l = 1, 2, \ldots, L\right\}$ と補間重み $\{\lambda_1, \lambda_2, \ldots, \lambda_L\}$ から算出する重み付き重心は，以下によって導かれる正規分布 $\mathcal{N}\left(\boldsymbol{\mu}', \boldsymbol{I}\boldsymbol{\sigma}'^2\right)$ として得られる [5].

$$\min_{\boldsymbol{\mu}', \boldsymbol{\sigma}'} \sum_{l=1}^{L} \lambda_l W_2\left(\mathcal{N}\left(\boldsymbol{\mu}', \boldsymbol{I}\boldsymbol{\sigma}'^2\right), \mathcal{N}\left(\boldsymbol{\mu}_l, \boldsymbol{I}\boldsymbol{\sigma}_l^2\right)\right)^2 \quad (6)$$

式 (6) を解くことで，求める重み付き重心の各パラメータは

$$\boldsymbol{\mu}' = \sum_{l=1}^{L} \lambda_l \boldsymbol{\mu}_l, \quad \boldsymbol{\sigma}' = \sum_{l=1}^{L} \lambda_l \boldsymbol{\sigma}_l \quad (7)$$

として得られる．

3.2.2　GMM 同士の重み付き重心

補間重み $\{\lambda_l\}$ で重み付けされた L 個の $\mathrm{GMM}\{p_l(\boldsymbol{s})\}$ の重み付き重心は，M 個の輸送先正規分布 $\left\{\mathcal{N}\left(\boldsymbol{\mu}'_m, \boldsymbol{I}\boldsymbol{\sigma}'^2_m\right) | m = 1, 2, \ldots, M\right\}$ とその混合重み α'_m から成る GMM として表現される．m 番目の輸送先正規分布 $\mathcal{N}\left(\boldsymbol{\mu}'_m, \boldsymbol{I}\boldsymbol{\sigma}'^2_m\right)$ のパラメータ，l 番目の GMM から $k_{l,m}$ 番目の正規分布 $p_{l,k_{l,m}}(\boldsymbol{s}) = \mathcal{N}\left(\boldsymbol{s}; \boldsymbol{\mu}_{l,k_{l,m}}, \boldsymbol{I}\boldsymbol{\sigma}^2_{l,k_{l,m}}\right)$ を選んだ場合の各組み合わせに対応して，

$$\min_{\boldsymbol{\mu}'_m, \boldsymbol{\sigma}'_m} \sum_{l=1}^{L} \lambda_l W_2\left(\mathcal{N}\left(\boldsymbol{\mu}'_m, \boldsymbol{I}\boldsymbol{\sigma}'^2_m\right), p_{l,k_{l,m}}(\boldsymbol{s})\right)^2 \quad (8)$$

によって得られる．つまり，輸送先正規分布は，l 番目の GMM から選択される混合要素のインデックスの系列 $[k_l, |l = 1, 2, \ldots, L] \in \{1, 2, \ldots, K\}^L$ のすべての可能な組み合わせに対して決定され，その総数は $M = K^L$ になる．各 $\boldsymbol{\mu}'_m$，$\boldsymbol{\sigma}'^2_m$ は式 (7) を解くことによって得られ，例えば，$\boldsymbol{\mu}'_m = \sum_{l=1}^{L} \lambda_l \boldsymbol{\mu}_{l,k_{l,m}}$ である．また，α'_m は

$$\sum_{m=1}^{M} \pi_{l,k,m} = \alpha_{l,k}, \quad \sum_{k=1}^{K} \pi_{1,k,m} = \cdots = \sum_{k=1}^{K} \pi_{L,k,m} \quad (9)$$

の条件下で

$$\min_{\pi_{l,k,m}} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{m=1}^{M} \lambda_l \pi_{l,k,m} W_2\left(\mathcal{N}\left(\boldsymbol{\mu}'_m, \boldsymbol{I}\boldsymbol{\sigma}'^2_m\right), p_{l,k}(\boldsymbol{s})\right)^2 \quad (10)$$

を解くことで

$$\alpha'_m = \sum_{k=1}^{K} \pi_{1,k,m} \quad (11)$$

として得られる．なお，式 (10) は離散分布における最適輸送の形で解くことができる．

なお，本研究では式 (9) の条件を無視した簡易的な実装で $\pi_{l,k,m}$ を求める．最適輸送を用いた分布全体での最適化ではなく，輸送先正規分布のうち最も移動コストの小さい正規分布への射影として個別の最適化に簡略化する．結果として，α'_m は，M 個の輸送先正規分布のうち m 番目の正規分布への輸送コストが最も小さい $p_{l,k_{l,m}}$ の混合重み $\alpha_{l,k_{l,m}}$ と補間重み λ_l の積の総和になる．輸送先に選択されない輸送先正規分布の混合重みが 0 となる．

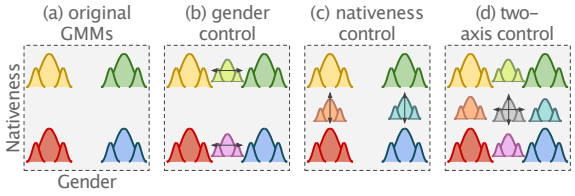

Fig. 3　主観評価実験で用いる分布

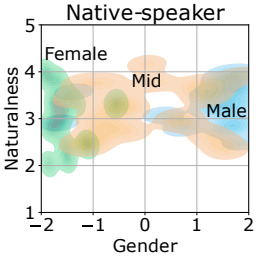

Fig. 4　母語話者の性別補間

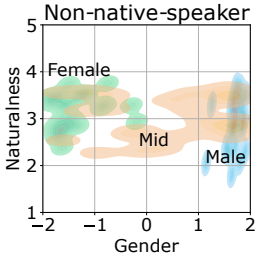

Fig. 5　非母語話者の性別補間

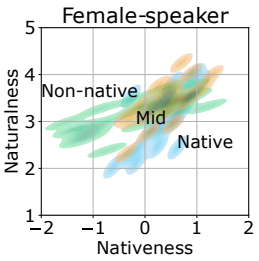

Fig. 6　女性話者の母語補間

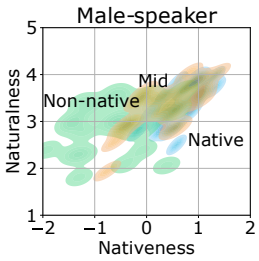

Fig. 7　男性話者の母語補間

3.2.3　中間的な話者属性の話者生成

中間的な話者属性の話者を生成するには，**3.2.2** 節の方法で求める重みつき重心から話者埋め込みをサンプリングし，**3.1** 節の TTS モデルに入力する．この重みつき重心は GMM で与えられるため，TacoSpawn と同様の方法でサンプリングできる．例えば，日本語母語男性の話者埋め込み分布と日本語母語女性の同分布を補間重み 0.5 でそれぞれ重みづけた重みつき重心からは，性別において中間的な日本語話者を生成できる．

4　実験的評価

4.1　実験条件

多言語多話者 TTS モデルを，性別と母語の 2 つの属性を持つ話者によって学習し，その日本語音声を評価した．具体的には，JVS コーパス [8] の日本語男性話者 49 人と日本語女性話者 51 人と VCTK コーパス [9] の英語男性話者 47 人と英語女性話者 61 人の音声データを用いて学習した．本実験のカテゴリカル話者属性を，性別，及び，母語が日本語か否かとする．例えば，日本語男性話者は [男性，母語] の属性を持ち，本実験の話者属性数は $L = 4$ である．TTS モデルは，JSUT コーパス [8] で事前学習した．使用した音声は事前に 22.05kHz にリサンプリングした．

外国語（本実験では英語）音声を用いて非母語（すなわち，外国語訛り．本実験では英語訛り日本語）音声を生成するために，TTS モデル学習に別の損失関数を追加した．具体的には，多言語 TTS モデルの出力から話者の母語を識別する識別器を設け，その cross entropy を損失関数に加える．識別器の構造は Xin ら [10] の論文に倣う．予備実験によってこの工夫により非母語話者の発話が実際の英語話者日本語発話に類似した特性を持つようになることを確認できた．

TTS モデルは FastSpeech2 [2][1]であり，ボコーダに学習済み HiFi-GAN [11][2]を使用した．入力テキストは，日本語は pyopenjtalk [3]で，英語は eSpeak NG[4] を用いて音素に変換した．

話者埋め込みの次元数 D は 256，GMM の混合数 K は 3 とした．Fig. 3 に示すとおり，用いる話者埋め込み分布の属性は，a) 4 種類のカテゴリカル話者属性，b) c) 母語あるいは性別のみについて，補間重みを 0.5 とした属性，d) 4 種類全てについて補間重みを 0.25 とした属性の合計 9 分布である．合成音声のサンプルはプロジェクトページ[5]にて聴取できる．実際の実験では，補間重みをより細かく設定した属性についても調査したが，紙幅の都合上割愛する．また，実在話者と生成話者の比較については，TacoSpawn の論文で言及されているため本論文では割愛する．

主観評価実験は，合成音声を聴取して知覚される話者性別，話者の母語，そして音声の自然性である．各話者埋め込み分布から 25 話者ずつ生成し，ITA コーパス[6]の RECITATION324 から話者毎にランダムに 5 文を割り当てて合成音声を生成した．ランサーズ[7]を通じて各評価項目ごとに 750 人の聴取者を雇用した．各聴取者は，各評価項目について 5 段階評価を行った．性別評価では −2（女性に聞こえる）〜+2（男性に聞こえる），母語評価では −2（非ネイティブらしく聞こえる）〜+2（ネイティブらしく聞こえる），自然性評価では 1（とても不自然に聞こえる）〜5（とても自然に聞こえる）の数値を割り当てて評価を行った．つまり，自然性評価は平均オピニオン評点 (mean opinion score: MOS) と同じ形式である．

4.2　実験結果と考察

主観評価結果をカーネル密度推定を用いて図示する．Fig. 4 と Fig. 5 は，Fig. 3b) に示す，性別補間における自然性と性別の評価結果である．“Mid” は補間した属性である．評価結果より，Mid 分布は Male と Female ではカバーされていない領域の話者を生成できることを確認できる．Fig. 6 と Fig. 7 は母語補

[1] https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[2] https://github.com/jik876/hifi-gan
[3] https://github.com/r9y9/pyopenjtalk
[4] https://github.com/espeak-ng/espeak-ng
[5] https://sarulab-speech.github.io/demo_mid-attribute-speaker-generation
[6] https://github.com/mmorise/ita-corpus
[7] https://www.lancers.jp

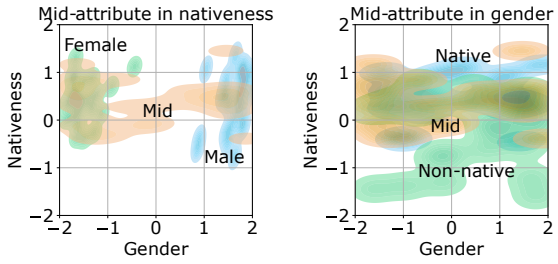

Fig. 8　母語における中間話者の性別補間

Fig. 9　性別における中間話者の母語補間

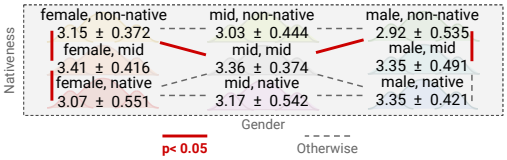

Fig. 10　各属性話者の MOS と標準偏差（有意差のある対は赤線表示）

間における自然性と自然性と母語の評価結果であり，前述した性別補間と同様の傾向が観測される．この図において非母語話者の音声の自然性が低いことを確認できるが，これは当該音声の流暢性の低さを鑑みると妥当である．また，性別補間の結果と比較して，母語補間では属性間でのスコアの重なりが大きい．これは，評価に用いた ITA コーパスがレアモーラを含むよう設計されており，母語話者の音声でも母語性が低く評価されたためだと考えられる．

Fig. 8 と Fig. 9 は，Fig. 3 の d) に示す，4 属性の混合において，中間的な母語話者における性別補間と，中間的な性別話者における母語補間の評価結果である．これらの結果から，母語および性別の両方について中間的な属性の話者を生成できることが分かる．以上の結果を総括し，提案法は，カテゴリカル話者属性を補間することで，その中間的な属性と知覚される話者を生成できること，また，複数のカテゴリカル属性を同時に補間できることが明らかになった．

Fig. 10 は，各話者埋め込み分布の MOS 値および標準偏差である．$p = 0.05$ で検定を行った結果，重みつき重心の MOS は，少なくともカテゴリカル分布の値と比較して有意に低い例はないことが確認できる．これにより，提案手法は自然性を減じることなく話者生成を拡張したと言える．

5　まとめ

本研究では，カテゴリカル話者属性の中間的な属性を持つ非実在話者を最適輸送に基づく確率分布補間で生成する方法を提案し，主観評価によってその話者が知覚的に中間的な属性を持つことを確認した．今後の展望として，更に多くの言語を利用した母語補間を検討する．

謝辞: 本研究は科研費 21H04900（実証実験）とムーンショット JPMJPS2011（アルゴリズム開発）の助成を受けたものです．

参考文献

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[2] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.

[3] Daisy Stanton, Matt Shannon, Soroosh Mariooryad, RJ Skerry-Ryan, Eric Battenberg, Tom Bagby, and David Kao, "Speaker generation," in *Proc. ICASSP*, 2022, pp. 7897–7901.

[4] Leonid V Kantorovich, "On the translocation of masses (in Russian)," *USSR Academy of Sciences*, vol. 37, pp. 199–201, 1942.

[5] Martial Agueh and Guillaume Carlier, "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[6] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum, "Optimal transport for Gaussian mixture models," *IEEE Access*, vol. 7, pp. 6269–6278, 2018.

[7] Asuka Takatsu, "Wasserstein geometry of Gaussian measures," *Osaka Journal of Mathematics*, vol. 48, no. 4, pp. 1005–1026, 2011.

[8] Shinnosuke Takamichi, Ryosuke Sonobe, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[9] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[10] Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, and Hiroshi Saruwatari, "Cross-lingual speaker adaptation using domain adaptation and speaker consistency loss for text-to-speech synthesis," in *Proc. Interspeech 2021*, 2021, pp. 1614–1618.

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.