Vol.2025-1 No.2-1
2025/2/22

ニューラルフォルマント合成における データ拡張手法の検討

小林　清流 †1　　山ノ内　裕翔 †1
小坂　哲夫 †1　　能勢　隆 †2

独立した音響パラメータから End-to-End で音声波形を合成する E2E-NF+は，ピッチや音韻性など発声の物理的特徴と 1：1 に対応した音響パラメータを制御した音声を高品質に合成可能である．一方で音響パラメータと音声波形の対応がデータ駆動型であることから，音響パラメータの制御性は学習データに含まれる音声の多様性に制約されるという課題があった．そこで本稿では，学習データの多様性を高めるデータ拡張手法を提案し，音響パラメータの制御性を向上させることを目指した．

Study of data extension methods for neural formant synthesis

Sumiharu Kobayashi,†1 Yuto Yamanouchi,†1
Tetsuo Kosaka†1 and Takashi Nose †2

E2E-NF+ is an end-to-end speech synthesis model capable of generating high-quality speech with fine-grained control over acoustic features like pitch and phoneme, directly from independent acoustic parameters. On the other hand, since the correspondence between acoustic parameters and speech waveforms is data-driven, there has been a problem that the controllability of acoustic parameters is limited by the diversity of speech included in the training data. In this paper, we propose a data augmentation method to enhance the diversity of training data and aim to improve the controllability of acoustic parameters.

†1 山形大学
Yamagata University
†2 東北大学
Tohoku University

1. はじめに

音響分析により得た特徴量から音声波形を生成するボコーダは，テキスト音声合成や声質変換の中核技術として広く使われている．深層学習を応用したニューラルボコーダ [1–3] は，メルスペクトログラムやメルケプストラムなどの音響特徴量に条件付けた学習により，元の音声波形を忠実に再現した高品質な音声を合成可能である．音響特徴量は，音声の物理的特性を考慮しかつ音声波形より単純で低レベルな表現であり，各時間フレームにおいて位相に不変であることから二乗誤差などの損失関数を用いて学習しやすいため，多くの研究 [4–6] で使われてきた． 一方で，音響特徴量は (D × T) 行列で表現される特徴であり，人間の発声の物理的意味を複数保持するため音高や音韻性などを個別に制御することは困難であった．

音響特徴量とは別に，基本周波数（F0）を条件付けるニューラルボコーダの提案 [7–9] によってピッチ制御性が向上したが，フォルマントやスペクトル傾斜に起因する声質の制御が困難であるという課題を抱えていた．フォルマントは音韻性と深く関係するパラメータであり，地域によって母音の区別に影響するフォルマント周波数が異なるなど方言のような特徴のある音声の理解において特に重要である．

この課題に対し neural formant synthesis (NF) [10] では，F0 やフォルマントなど独立した 9 つの音響パラメータから音響特徴量を予測し，予測した音響特徴量からニューラルボコーダによって音声波形を合成するボコーダシステムが提案された．独立した音響パラメータを個別に操作することで，声質を制御した音声波形の合成が可能である．またこれを発展させた各独立した音響パラメータから音声波形を直接合成する End-to-End Neural Formant Synthesis [11] が考案された．End-to-End の学習によって，合成品質の向上とピッチ制御性が NF と比較して向上した．一方で，フォルマント制御性について NF と同等であり依然として課題であった．これは，学習に用いたデータセットは読み上げ文章であるためと考えられた．フォルマントと音声はデータ駆動型の学習によって対応付けられることから，音韻における多様性に乏しいデータセットでは，十分なフォルマント制御性が得られなかったと考えられる．

そこでデータ拡張によって声質のバリエーションを増やすことで，音響パラメータの制御性の向上とロバスト性を向上させることを試みた．本研究では，音響パラメータのうち F0 とフォルマントについてそれぞれデータ拡張を行い評価した．実験の結果，F0 制御性と F1–F3 操作性の向上が確認された．

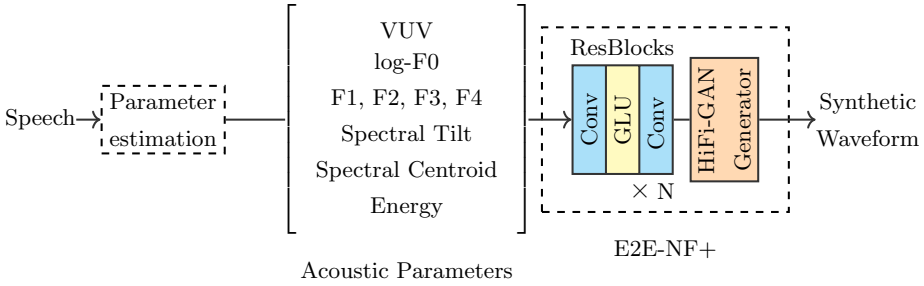

図 1 E2E-NF+による音声合成
Fig. 1 Synthesis of speech waveform by E2E-NF+

2. E2E-NF+

本研究では，[11] で提案された独立した 9 つの音響パラメータから End-to-End で音声波形合成するモデルのうち E2E-NF+ を用いた．E2E-NF+ の構造を**図 1** に示す．E2E-NF+ は，残差ブロックと HiFi-GAN Generator で構成されるモデルである．

先行研究において HiFi-GAN Generator のみで構成される E2E-NF は，音響パラメータと音声波形の対応を十分に獲得できなかった．そこでモデルパラメータを補完し長期の時間構造を捉えるために，Conv1D + Gated Linear Unit (GLU) + Conv1D で構成される残差ブロックを HiFi-GAN Generator の前にスタックした E2E-NF+ を提案した．各残差ブロックは，モデルの層が深くなることで適切に誤差伝播がされないことを考慮して，スキップ接続により接続された．

残差ブロックとスキップ接続の導入により，NF と同等の音響パラメータ制御性を持ちながら合成品質が改善された．

3. データ拡張手法

E2E-NF+は音響パラメータと音声波形の対応がデータ駆動型であることから，音響パラメータの制御性と制御時の合成音声のロバスト性は学習データの多様性に依存する．そこで本稿では 9 つの音響パラメータのうち，音高に対応するピッチと音韻に対応するフォルマントについて，音声加工により学習データの多様性を広げることを試みた．次に，学習に用いた音声コーパスと各データ拡張手法の詳細について述べる．

3.1 ピッチシフト

データセットにおける音高の多様性を広げるために，F0 伸縮によるデータの拡張をした．F0 の伸縮は，WSOLA (Waveform Similarity based OverLap-Add) [12] によるピッチシフトを用いた．

3.2 全域通過フィルタの α 係数シフト

データセットにおける音韻の多様性を広げるために，周波数方向の伸縮によるデータの拡張をした．周波数伸縮は，一次の全域通過フィルタの位相特性を制御する α 係数を操作することで行われた．

音声波形からメルケプストラムを算出した後，SPTK [13] の mgc2mgc コマンドを用いて全域通過フィルタの α 係数を操作したメルケプストラムへ変換した．この時，入力前後でメルケプストラムの次数は不変とした．シフトされたメルケプストラムは WORLD ボコーダ [14] によって音声波形に復元され，復元された音声を学習データに加えることでデータ拡張が達成された．

4. 実　　験

4.1 実 験 条 件

音声コーパスは，日本人男女 100 名で構成される JVS コーパス [15] のうち Parallel 100 を用いた．データは各話者の 100 発話について，学習：テスト：評価=90：5：5 の割合で分割した．ベースモデルとして，データ拡張していない JVS コーパスで学習された E2E-NF+ を用意した．評価は，男女 2 名，各 5 発話の計 20 発話に対して行った．音響パラメータの抽出パラメータは先行研究と同様のものを使用した．ホップサイズ 300，窓幅 1200，F0 抽出時の上下限をそれぞれ 75 Hz ，600 Hz とした．フォルマント抽出は，フォルマント数 5，窓幅 600，プリエンファシス 50 Hz とした．

ピッチシフトによるデータ拡張では，シフト幅を -600 cent から 600 cent までステップ幅 200 cent とし，各ステップにおいて学習データからランダムに抽出した 5 分の 1 を対象とした．

全域通過フィルタの α 係数シフトによるデータ拡張では，変換後の α 係数を，0.21, 0.28, 0.32, 0.52, 0.63 とし，各条件において学習データからランダムに抽出した 5 分の 1 を対象とした．また抽出時の周波数上限は男性 5,000 Hz，女性 5,500 Hz を基本とし，シフト後のフォルマント周波数を考慮して，α=0.21 では男女共に 6,000 Hz，α=0.28 では男性 5,500 Hz，女性 6,000 Hz，α=0.52, 0.63 では男性 4,500 Hz，女性 5,000 Hz とした．

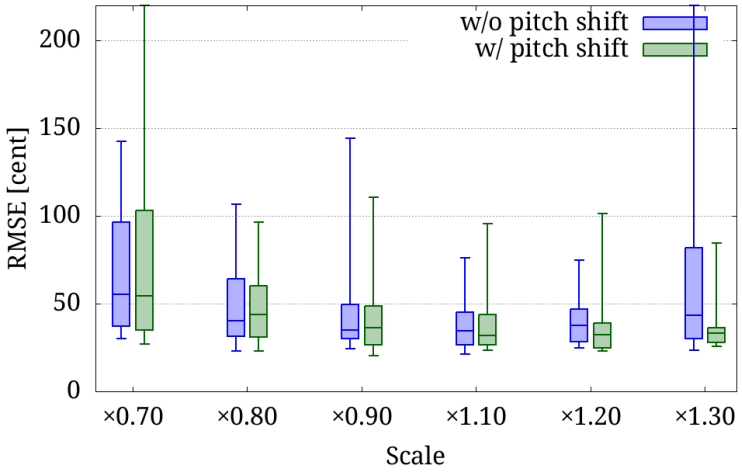

図 2 F0 シフトした合成音声における RMSE
Fig. 2 Plot of RMSE for manipulated F0

4.2 ピッチシフトの効果

F0 を × 0.7, × 0.8, × 0.9, × 1.1, × 1.2, × 1.3 のシフトスケールで合成した音声について客観評価実験により評価した．合成音声から音響分析により得た対数 F0 と，入力のシフトされた対数 F0 との平均平方二乗根誤差 (RMSE) を計算した結果を**図 2** に示す．

結果から，特にピッチを高くする場合において改善が見られた．また，多くの場合で分散が小さくなった．

ピッチスケール × 0.7 においてもっとも RMSE が高かった，拡張したデータで学習したモデルの合成音声（話者：JVS001，発話：VOICEACTRESS001）と，同じ音声をベースモデルで合成した音声の対数 F0 軌跡を**図 3** に示す．オレンジで示される軌跡は入力に用いたシフトされた F0 であり，青で示される合成音声の F0 の軌跡が一致しているほど，ピッチ制御が忠実であることを示す．拡張データを学習データとしたモデルで合成した音声は，左と同様にほとんどの場合でシフトされた F0 に近い F0 軌跡を持っていることが分かる．しかし，2.7 秒や 5.3 秒付近で F0 軌跡が大きく一致していない箇所が見られたことが，× 0.7 において RMSE の分散が高くなった原因と考えられる．この問題は他の評価音声でも同様に見られた．一方で該当箇所においてスペクトログラムや音声波形に差が見られず，

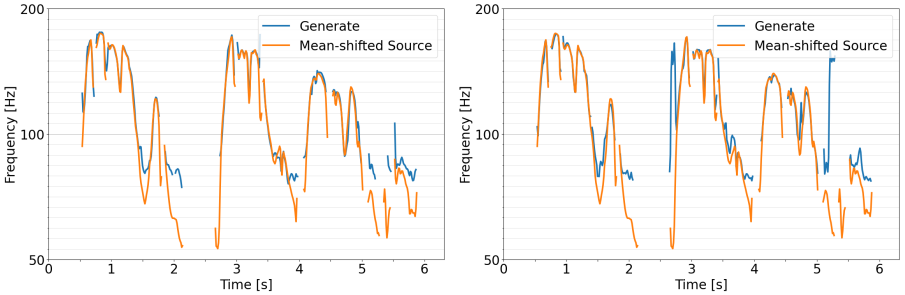

図 3 × 0.7 における最も RMSE が高い音声の対数 F0 の比較．
（左）：データ拡張無し，（右）：データ拡張有り
Fig. 3 Comparison of Log-F0 for the worst audio at × 0.7.
(Left) : Without data augmentation, (Right) : With data augmentation

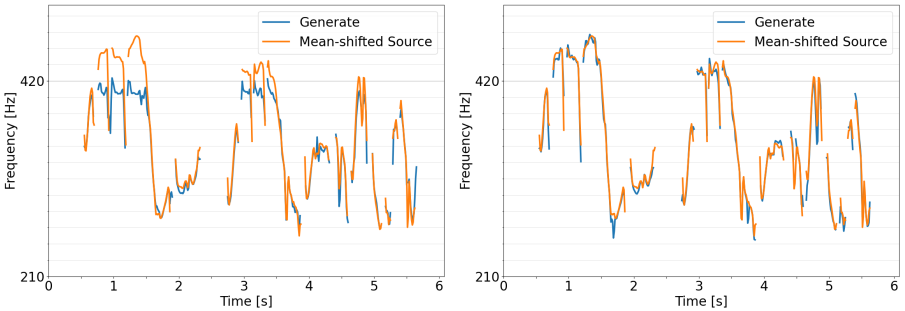

図 4 × 1.3 における最も RMSE が高い音声の対数 F0 の比較．
（左）：データ拡張無し，（右）：データ拡張有り
Fig. 4 Comparison of Log-F0 for the worst audio at × 1.3.
(Left) : Without data augmentation, (Right) : With data augmentation

少人数での主観評価においても差が見られなかった．そのため拡張データを学習データとしたモデルで 合成された音声は何らかの原因で F0 抽出の不備が起きやすく，これが結果に大きく影響している可能性が考えられる．

ピッチスケール × 1.3 においてもっとも RMSE が高かった，ベースモデルの合成音声（話者：JVS004，発話：VOICEACTRESS003）と，同じ音声について拡張データを学習データとしたモデルで合成した音声の対数 F0 軌跡を**図 4** に示す．拡張データを学習デー

タとしたモデルで合成した音声は，左のベースモデルよりも高い F0 追随性を持っていることが分かる．また少人数での主観評価においても，拡張データを学習としたモデルが良いとされた．

以上の結果から，ピッチシフトによるデータ拡張は一定の有効性を持つと示された．

4.3 周波数伸縮の効果

先行研究よりも大きいフォルマントシフトを実現するために，F1–F4 全てについて，× 0.5, × 0.7, × 0.8, × 0.9, × 1.1, × 1.2, × 1.3, × 1.5 のスケールシフトで操作し音声を合成した．合成音声から音響分析により得た F1–F4 と，入力のシフトされた F1–F4 との RMSE による客観評価実験で評価した．フォルマント抽出時の周波数上限は男性 5,000 Hz, 女性 5,500 Hz を基本とし，シフト後のフォルマント周波数を考慮して，× 1.2，× 1.3 では+ 500 Hz，× 1.5 では+ 1,000 Hz とした．RMSE の結果を**図 5** に示す．F1–F3 において，スケールシフトが大きいほど RMSE が改善する傾向が見られたものの有意な差は見られなかった．また F4 においては周波数を高くする場合は改善されたが，低くする場合は悪化した．

加工された音声を観察したところ各フォルマントはシフトしているもののその幅が小さく，データ分布を広げるには不十分であった可能性がある．また， α 係数のシフト幅を大きくすると加工音声の品質が劣化も大きくなり音響分析が難しくなることから，パラメータ抽出が不適当であった可能性が考えられる．

各フォルマントを× 1.5 して合成した音声を**図 6** に示す．ベースモデルにおいてシフトスケールを大きくして合成した音声は，高周波成分に歪みに起因するノイズが生じた．一方で拡張データで学習された音声のノイズが低減されたことから，フォルマント制御時におけるロバスト性の改善に寄与していることが示された．

5. ま と め

ピッチシフトによるデータ拡張によって，ピッチ制御性の改善を試みた．客観評価実験より，特にピッチを高くする場合における RMSE が改善されたそのため，SoX を用いたピッチシフトによるデータ拡張は一定の有効性を持つと示された．

また，全域通過フィルタの α 係数シフトによるデータ拡張によるフォルマント制御の改善を試みた．F1–F3 において，スケールシフトが大きいほど RMSE が改善する傾向が見られたものの有意差が見られなかった．フォルマントシフトを大きくした場合に生じる合成音声の歪みが低減されたことから，フォルマント操作時におけるロバスト性の向上には一定

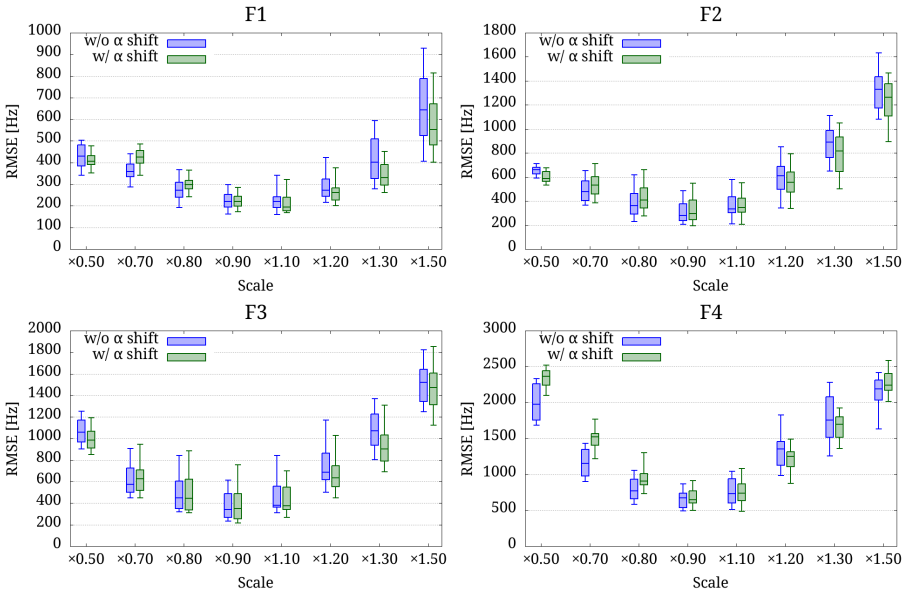

図 5 F1–F4 をシフトした合成音声から抽出された各フォルマントにおける RMSE
Fig. 5 Plot of RMSE for manipulated formants

の効果を示した．一方で，声質を大きく変化させる音声加工はデータ品質の劣化が生じやすいため，音響分析によるパラメータ抽出が難しくなる傾向にある．また， α 係数は全域通過フィルタの位相特性を変換させるパラメータであり，フォルマントシフトとの因果関係の理解が難しい．これらのことから，今回検証したデータ拡張手法は有効なものであると言い難い．

フォルマント制御性の改善における今後の検討として，マルチリンガルなデータセットでの学習とフォルマントに依存しないモデル構造の設計が上げられる．E2E-NF+はユニバーサルボコーダであり，データセットの言語性に依存しない．フォルマントは音韻性に関係することから，マルチリンガルな音声データを用意すればフォルマント制御性が向上する可能性がある．またフォルマント抽出は線形予測を伴う難しいタスクであることから，より簡単に抽出可能な線スペクトル対やメルケプストラムを用いる従来型のボコーダを深層化したニューラルボコーダの設計を検討していく．

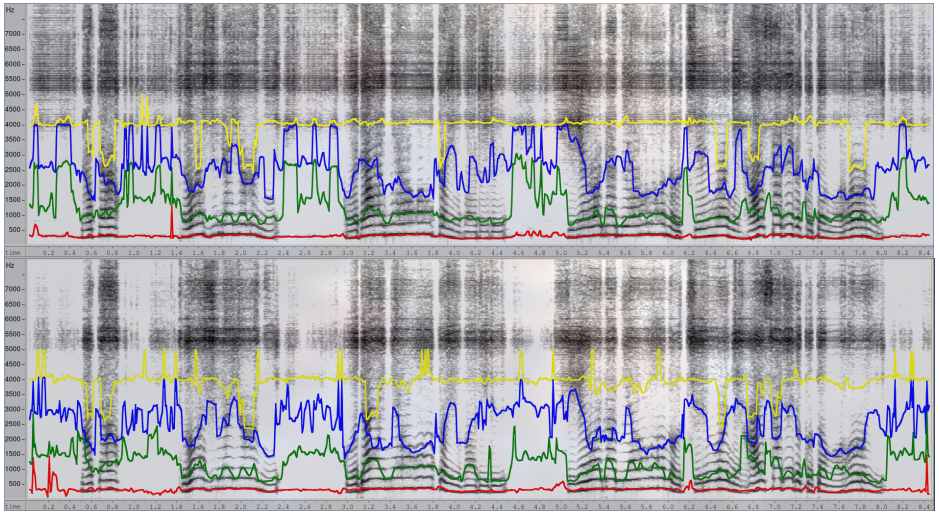

図 6 F1–F4 を× 1.5 して合成した音声
（上）：データ拡張無し，（下）：データ拡張有り
Fig. 6 Plot of RMSE for manipulated formants
(Top) : Without data augmentation, (Bottom) : With data augmentation

参 考 文 献

1) Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S. and Kavukcuoglu, K.: Efficient Neural Audio Synthesis, *PMLR 2018*, pp.2410–2419 (2018).
2) Kong, J., Kim, J. and Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, *Advances in Neural Information Processing Systems*, Vol.33, pp.17022–17033 (2020).
3) vanden Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, *arXiv preprint arXiv:1609.03499* (2016).
4) Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K. and Toda, T.: Speaker-Dependent WaveNet Vocoder, *Interspeech 2017*, pp.1118–1122 (2017).
5) 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, *ICASSP 2018*, pp.4779–4783 (2018).
6) 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, *ICLR 2021* (2021).
7) Wang, X., Takaki, S. and Yamagishi, J.: Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.28, pp.402–415 (2020).
8) Yamamoto, R., Song, E. and Kim, J.-M.: Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram, *ICASSP 2020*, pp.6199–6203 (2020).
9) Yoneyama, R., Wu, Y.-C. and Toda, T.: Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder, *ICASSP 2023*, pp.1–5 (2023).
10) PérezZarazaga, P., Malisz, Z., Henter, G.E. and Juvela, L.: Speaker-Independent Neural Formant Synthesis, *Interspeech 2023*, pp.5556–5560 (2023).
11) Kobayashi, S., Kosaka, T. and Nose, T.: End-to-End Neural Formant Synthesis Using Low-Dimensional Acoustic Parameters, *GCCE 2024*, pp.820–823 (2024).
12) Verhelst, W. and Roelands, M.: An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech, *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.2, pp.554–557 vol.2 (1993).
13) Yoshimura, T., Fujimoto, T., Oura, K. and Tokuda, K.: SPTK4: An open-source software toolkit for speech signal processing, *SSW 2023*, pp.211–217 (2023).
14) Morise, M., Yokomori, F. and Ozawa, K.: WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications, *IEICE TRANSACTIONS on Information and Systems*, Vol.E99-D, No.7, pp.1877–1884 (2016).
15) Takamichi, S., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N. and Saruwatari, H.: JVS Corpus: Free Japanese Multi-Speaker Voice Corpus, *arXiv preprint arXiv:1609.03499* (2019).