補助情報を加えた話者ベクトルの操作によりシームレスに話者性を制御できる End-to-End 音声合成方式の検討

Study of End-to-End Text-to-Speech that can seamlessly control speaker’s individuality by Manipulating Speaker features with Auxiliary Information

青谷 直樹
Naoki Aotani

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 本研究では，話者固有の声質や平均的な声の高さなどの情報を表現した話者ベクトルに補助情報を加え，操作を行うことにより，シームレスに話者性を制御できる End-to-End 音声合成の方式を検討する．話者性の制御は，2 名の話者の音声から話者ベクトルを抽出し，それぞれの話者ベクトルに重みづけ加算し，2 名の話者特徴量を内挿した特徴ベクトルを生成することで行う．本報告では，重みの調整によって話者性を徐々に変化させることを話者性のシームレスな制御と呼ぶ．本報告では，平均基本周波数の情報を分離した特徴として制御することで，話者性の変化をシームレスに行えるようにする方式を検討する．評価実験では，合成音声の声の高さの制御性能，声質を客観評価実験で評価し，話者性の制御性能を主観評価実験で評価した．

1 はじめに

音声を人工的に生成する技術として，音声合成技術がある．音声合成技術の中で，テキストから音声を合成する技術は Text to Speech(TTS) と呼ばれ，スマートスピーカ，音声対話アプリなどのサービスに利用されている．今後，更に多くの新たなサービスが生み出されることが予想され，これらのサービスには，利用者のニーズに合わせた多様な話者性を持つ合成音声が必要であると考えられる．

End-to-End 音声合成では，学習に大量なデータが必要なことから，様々な話者性を持つシステムを構築することは困難である．そこで，本研究では，End-to-End 音声合成の方式において，多様な話者性を持つ音声を合成することを目指す．また，話者性を積極的に制御し，データ収集した話者性のみならず，それ以外の話者性の実現も試みる．本報告では，第一ステップとして，話者ベクトルに対して重みづけの操作を行ってから，音声合成に使用することで話者性を徐々に変化させることが可能な End-to-End 音声合成システムについて検討する．このとき，重みを少しずつ変えることで，ある話者の音声が少しずつもう一方の話者の音声のように聞こえることを目指す．本報告では，話者性を徐々に変化させるため，従来方式の x-vector のみを話者特徴量として用いた方式に対して，平均的な声の高さの情報を欠落させて抽出した non-f0 x-vector と音の平均的な声の高さ f0 を特徴量として用いる方式を提案する．

2 提案方式

2.1 話者性の制御

提案方式は，図 1 に示す話者ベクトルを用いた End-to-End 音声合成方式に基づいている．この方式は，End-to-End 音声合成時に使用する話者ベクトルを操作することで話者性の制御を目指す．

話者性の操作について説明する．話者毎に求められた x-vector の操作の概要を図 1 の話者ベクトル操作部に示す．話者ベクトルの操作として，まず，2 名の異なる話者から x-vector を抽出する．これらの足し合わせ操作を行うことで，話者性ベクトル空間上において 2 名の話者から得た話者ベクトルを線形補間した話者ベクトルを得る．この操作において，重みづけの重み変更の度合いに応じて，話者性の制御度合いが，同程度に変化することが望ましい．この時の重みづけおよび足し合わせの操作を話者 A，話者 B で行った場合，この操作は式 1 で表せる．

$$V_{(A,B)w} = (1-w)V_{(A)} + wV_{(B)} \tag{1}$$

$V_{(A,B)w}$ は話者 A の話者ベクトルと話者 B の話者ベクトルを重み w で足し合わせることで得られる操作を行った話者ベクトル．$V_{(A)}$ は話者 A の話者ベクトル，$V_{(B)}$ は話者 B の話者ベクトルを表す．w は重みを表し，0 から 1 の間の値を取る．

2.2 話者性の制御向上を目指した話者ベクトルの作成と音声合成器の学習

話者性の制御能力を向上させることを目的として，話者ベクトルとして，音声の平均基本周波数を除いた non-f0 x-vector と音声の平均基本周波数 f0 を用いる方式を提案する．x-vector のみを話者ベクトルとして用いた場合，声質のみならず，声の高さも制御されるが，重みに対する両特徴の変化は安定しておらず，シームレスな制御が困難である．そこで，提案方式では，x-vector による F0 制御能力を除くことで声質の制御のみに注力させ，F0 の制御には別途与えた F0 情報を利用するように，End-to-End 音声合成器を学習する．

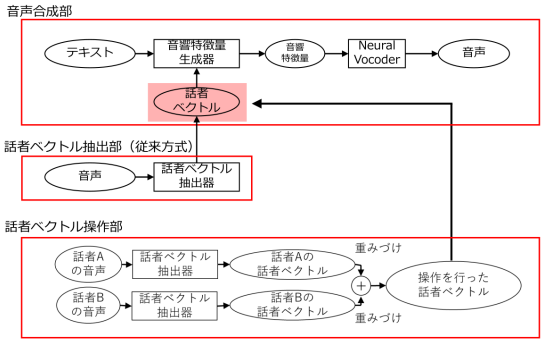

図 1: End-to-End 音声合成の概要図．提案方式では，話者ベクトル抽出器に図 2 の方式のものを使用する

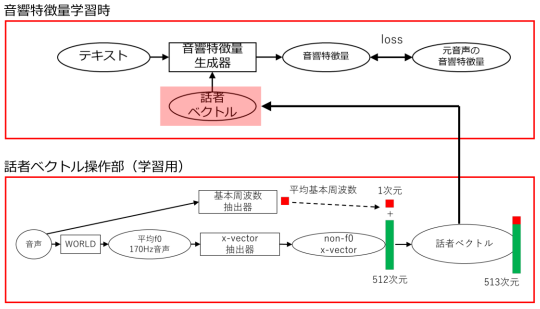

図 2: 平均 f0 情報を含む話者ベクトルを使用した音響特徴量生成器学習の概要図

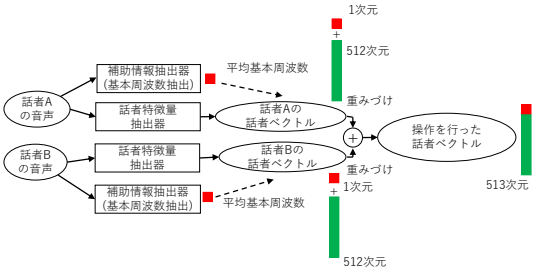

図 3: 平均 f0 情報を含む話者ベクトル生成の概要図

学習時の話者ベクトルの操作の概要図を図 2 に示す．学習時に使用する話者ベクトルの抽出元となる音声を WORLD[9](D4C edition[10]) を用いて平均基本周波数が 170Hz になるように分析合成を行ってから，話者ベクトルの抽出を行う．この話者ベクトルと分析合成前の音声から得た平均基本周波数（$\bar{f}_0$）の情報の結合を行う．このようにして得た補助情報を持つ話者ベクトルを音響特徴量生成器の学習時に使用する．

音声合成時の話者ベクトルの操作を図 3 に示す．操作後に得られる補助情報を持つ話者ベクトルを用いて，図 1 の合成方式により End-to-End 音声合成を行う．

3 評価実験

3.1 実験条件

音響特徴量生成器は Transformer TTS[7] のモデルを使用し，学習データは新聞読み上げコーパス JNAS を使用し，男性 126 名，女性 125 名の計 251 名の話者からなる 24857 文である．Vocoder は Parallel WaveGAN[8] のモデルであり，学習済みモデルを使用した．学習データは新聞読み上げコーパス JNAS であり，男性 131 名，女性 130 名の計 261 名の話者の音声である．

話者ベクトルの抽出元とする音声は次の 4 種類の音声を用いた．これらの音声は，これらの音声は，音響特徴量生成器，ボコーダーの学習に使用されている．

- M_{110}：平均基本周波数が 110Hz の男性の音声
- M_{150}：平均基本周波数が 150Hz の男性の音声
- F_{190}：平均基本周波数が 190Hz の女性の音声
- F_{230}：平均基本周波数が 230Hz の女性の音声

これらの音声は，これらの音声は，音響特徴量生成器，ボコーダーの学習に使用されている．

上記の音声から次の 3 組の組み合わせを作り音声合成を行った．すべての組み合わせで平均基本周波数の差が 40Hz である．

- $F_{190} - F_{230}$ ペア
- $M_{110} - M_{150}$ ペア
- $M_{150} - F_{190}$ ペア

評価を行うために合成した音声のテキストは，ATR 音素バランス文 [11] の I セットから選んだ 10 文を使用した．これらの文章を 2.1 節で述べた話者ベクトルの操作における重み w を 0 から 1 の間を 0, 1 刻みごとに作成した．各話者の組合せで 110 文を評価に使用した．

3.2 客観評価実験

客観評価実験では音声の平均基本周波数とメルケプストラム歪みを用いて評価する．音声の平均基本周波数は WORLD 分析，メルケプストラムは，pysptk [*1] の mcep 関数を用いて，24 次のメルケプストラムとして抽出した．

$$\bar{f}_0 = \frac{1}{T}\sum_{T}^{t=0} f_0(t) \tag{2}$$

ここで，$f_0(t)$ は WORLD 分析により得られた元音声の各時刻 t における $F_0[Hz]$ であり，T は発話の長さ，$\bar{f}_0$ は発話内の平均 F_0 である．1 発話の平均基本周波数は，式 (2) により計算した．この時，無声区間は計算に含めていない．

各重みごとの平均基本周波数を図 4 に示す．図 4 より，補助情報の追加によって，$F_{190} - F_{230}$ ペアの結果の重み 0.7 から重み 1 にかけて平均基本周波数の値の制御が変わっていることがわかる．また，$M_{150} - F_{190}$ ペア（補助情報なし），$M_{150} - F_{190}$ ペア（補助情報あ

[*1] https://github.com/r9y9/pysptk

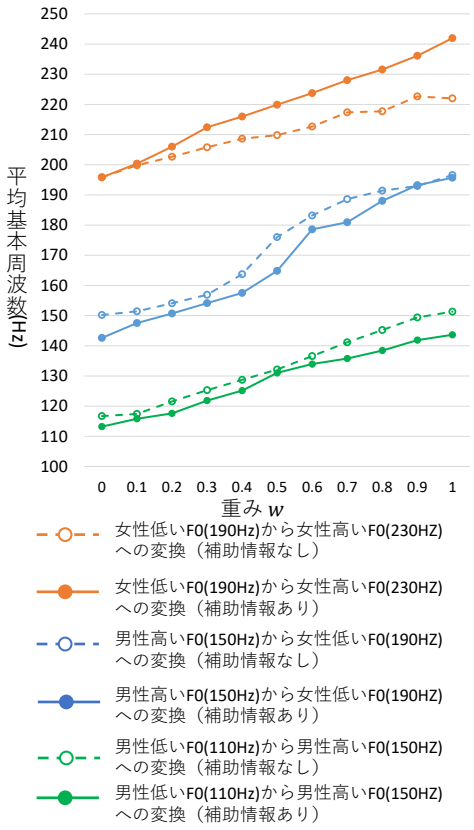

図 4: 各重みごとの合成音声の平均基本周波数

り） のいずれにおいても重み 0.4 から 0.6 にかけてグラフの傾きが大きくなっており，大きく平均基本周波数が変わってしまっていることがわかる．

各重みごとのメルケプストラム歪みを図 5 に示す．図 5 より，片方のメルケプストラム歪みが大きくなるにつれて，もう一方のメルケプストラム歪みが小さくなっていることがわかる．

3.3 主観評価実験

主観評価実験は，3 つの音声を A,B,X の順番で聞いてもらい，X の音声の話者が A の音声の話者，B の音声の話者のどちらに近いかを被験者に回答してもらうことで評価を行った．このとき A，B の音声はそれぞれ重み 0 または 1 の音声を使用し，X の音声は重み 0.1 から 0.9 のいづれかの音声を使用した．被験者は 6 名であり，評価文は 2 種類の内容のテキストをそれぞれ重み 0.1 から 0.9 まで 0.1 刻みで合成したものを使用した．評価分数は 42 文用いた．回答は以下の 7 つより選択してもらった．

- 1：A に非常に近い
- 2：A に近い
- 3：どちらかというと A に近い
- 4：どちらともいえない
- 5：どちらかというと B に近い

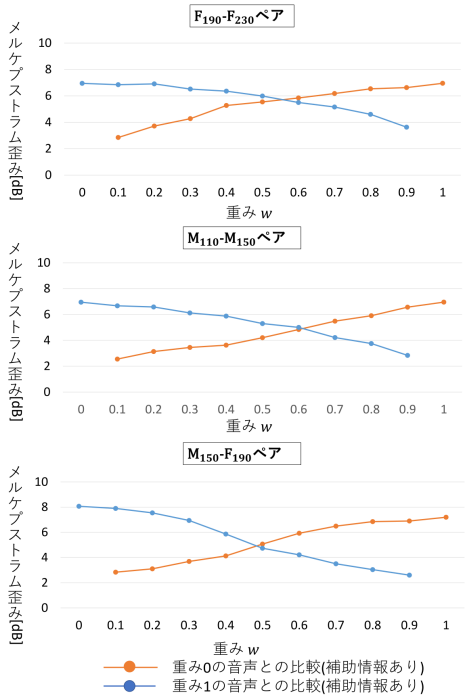

図 5: 各重みごとの合成音声のメルケプストラム歪み

- 6：B に近い
- 7：B に非常に近い

主観評価実験の結果を図 6 に示す．

図 6 の箱ひげ図より，$F_{190} - F_{230}$ ペアは重み 0.5 から 0.6 の間で，平均値が大きく変わっていること，$M_{110} - M_{150}$ ペアは重み 0.1 から 0.4 までの平均値の変化量が小さいこと，がわかる．また，$M_{150} - F_{190}$ ペア は箱ひげ図と平均値の値より，重み 0.7 から 0.9 までの間でほとんど X の音声が B の音声だと判定されていることがわかる．$F_{190} - F_{230}$ ペアの重み 0.5 から 0.6 の間で，平均値が大きく変わっていることは音声にどちらかの話者らしさが残っており，4：どちらともいえない があまり選ばれなかったことが原因であると考えられる．

結果より，男性から女性への話者性の変換において，重み 0.1 から 0.3,0.7 から 0.9 のような広い範囲でどちらかの話者と判定されてしまい，十分に話者性が制御できていない部分があることが分かった．

今後の課題として，男性から女性への話者性の制御の際に，大きく平均基本周波数が変わってしまう箇所を補助情報や制約などを加えることで線形に変化するようにする必要がある．このように変化するようにできれば，重み 0.1 から 0.3,0.7 から 0.9 のような広い範囲でどちらかの話者と判定される問題が改善すると考えられる．

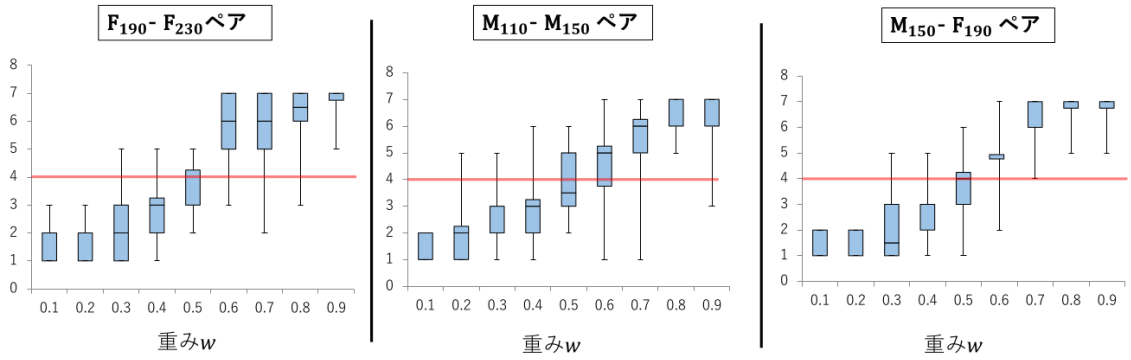

図 6: 主観評価実験の結果

4 まとめ

本報告では，話者特徴量の操作により，話者性のシームレスな制御を実現する End-to-End 音声合成方式の検討を行った．話者性の制御を行うために，2 名の話者から話者ベクトルを抽出し，それぞれの話者ベクトルに重みづけを行い，足し合わせを行う方式を検討した．このとき，話者性の制御性能向上のために，non-f0 x-vector と音声の平均基本周波数 f0 を結合した話者ベクトルを利用した．

今後の課題として，男性から女性への話者性の制御の際に，大きく平均基本周波数が変わってしまう箇所に注目し，新たな制約を話者特徴量か学習方式に加えることで，重みの変更による制御を精度良くすることが考えられる．

参考文献

[1] H. Zen, A.Senior, and M.Schuster, “Statistical Parametric Speech Synthesis Using Deep Neural Networks,“ In Proc. ICASSP, pp.7962–7966, 2013.

[2] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based Speech Synthesis System Applied to English,“ in Proc. IEEE Workshop Speech Synth., pp.227-230, Sep. 2002.

[3] J.Shen, R.Pang, R.J.Weiss, M.Schuster, N.Jaitly, Z.Yang, Z.Chen, Y.Zhang, Y.Wang, R.Skerrv-Ryan, et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,“ arXiv preprint arXiv:1712.05884, 2017.

[4] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio,“ arXiv:1609.03499, 2016.

[5] N. Hojo, Y. Ijima, and H. Mizuno, “DNN-Based Speech Synthesis Using Speaker Codes,“ IEICE Transactions on Information and Systems, 2018.

[6] D.Raj, D.Snyder, D.Povey, and S.Khudanpur, “Probing the information encoded in x-vectors,“ 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp.726–733, 2019.

[7] N. Li and Shujie Liu and Yanqing Liu and Sheng Zhao and Ming Liu and M. Zhou, “Close to Human Quality TTS with Transformer,“ ArXiv, 2018.

[8] R.Yamamoto, E.Song, and JM.Kim, “Parallel WaveGAN: Afast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,“ In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6199–6203, 2020.

[9] M. Morise, F. Yokomori, and K. Ozawa, “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.

[10] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, vol.84, pp.57-65, Nov. 2016.

[11] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis,“ Speech Communication, 9, pp.357–363, Aug. 2005.