話者タグによるTacotron2の話者性制御*

○太刀岡 勇気 (デンソーアイティーラボラトリ)

1 はじめに

複数の話者やスタイルを制御できる音声合成方式が望まれている。end-to-end 音声合成システムでは、中間表現に埋め込み表現を入力する方式 [1] が一般的であるが、より簡単な方式として、タグづけによりスタイルを制御する方法 [2] が提案されている。前者の方式では埋め込み表現を工夫することで、話者・スタイルの制御を行うことができるが、後者の方式ではそれらの切り替えはできるものの細かな制御 (内挿等) が難しいという問題がある。ここでは、タグを入力した際の中間表現を分析することで、話者・スタイルをタグにより制御する方式を提案する。

2 Tacotron2と話者・スタイルタグ

Tacotron2 [3] では、文字列を 5 文字ずつまとめてエンコーダーに入力し、512 次元の中間表現を得て、そこから自己回帰型のデコーダーにより、メルスペクトログラムを推定する。Tacotron2 に入力するかな文字列の前後にスタイル記号を付与することで、複数のスタイルの制御を行う方法が提案されている [2]。ここではスタイル制御に加えて話者の制御も行いたいため、感情のタグに加えて話者のタグを付与するように拡張する。すなわち「話者タグ」「スタイルタグ」「かな文字列」の順に入力する。例を図 1 に示す。

文献 [2] の方法では、文頭と文末に記号を配置するが、中間表現を分析しようとすると、文末まで確定できないという問題がある。文頭のみに記号を置いた場合と文末にも置いた場合を比較し、双方の性能差があまりないことを確認したため、中間表現の表現を見やすくするため、ここでは文頭のみに記号を置く。

3 実験

3.1 実験条件

Tacotron2 の学習に用いた音声は、プロのナレーター (男女各 7 名) が上述の 4 スタイルで決められた (2,456 文 (うち 4 名)、600 文 (残り 10 名) からなる) 原稿を読み上げた音声である。評価には各スタイル 30 文章からなる学習データとは別の文を用いた。「話者タグ」「スタイルタグ」に制御性を持たせるために、タグを以下のルールにより作成した。「話者タグ」は、4 文字のアルファベットからなり、はじめの 1 文字目が性別 (女性 (f)、男性 (m)) で残りの 3 文字が任意のアルファベットである。任意のアルファベット部分をランダムに生成することで、学習データにない新しい話者が創造できることを期待している。「スタイルタグ」は、アルファベット 1 文字とした。これは読み方の 4 つスタイル (平静 (h)、悲しげ (k)、楽しげ (t)、ぞんざい (z)) を表す。Tacotron2 により推定されたメルスペクトログラムから、waveglow [4] の配布されているモデルにより音声波形を生成した。サンプリング周波数は 22.5kHz とした。

[fhar] [h] ニ’ チベー ノ$ シュノー カ’ イダン ガ$@
ヒラカ レ’ ル$ コト’ ニ$ ナリ マ’ シ% タ

Fig. 1 Example of an input text.

話者タグを 20 人分ランダムに生成して、話者・スタイルタグのみを入力 (例えば [fhar] [k] まで入力) した際の中間表現を観察し、学習話者から離れているものを実際に合成した。各スタイル 30 文ずつ音声を合成し、合成音声を話者ごとの x-vector [5] に変換し、話者性を評価した。x-vector は 512 次元で、日本語話し言葉コーパス (CSJ) の学会講演音声から構築した DNN model を用いて抽出した。可視化のために、中間表現・x-vector ともに主成分分析により 2 次元空間にプロットして分離性を検討した。

3.2 中間表現での評価

ftak,fhar の 2 人の女性話者について、4 スタイルでの中間表現を主成分分析した結果を図 2 に示す。話者によって分離できていることと、k,z,t,h の順に上から並んでいることから、感情が同じものは近くに配置されていることが推測される。話者タグのみ変化 (スタイルタグは [h] で固定) させた場合の中間表現を主成分分析した結果を図 3 に示す。学習話者を ‘o’ でプロットしたのに加え、上述の通り 20 人分ランダムに生成した話者タグのうち、学習話者と比較的離れていた話者を男女 3 名ずつ ‘x’ で表示している。このように男女は明確に分かれていることから、はじめの 1 文字が性別であることは理解されていると考えられる。ただし、それ以外の話者はほぼ学習話者と同じところにマッピングされたことから、tacotron2 は新たな話者を創造する能力はあまり高くないことが推察される。

*Speaker control of Tacotron2 by using speaker tags. by TACHIOKA, Yuuki (Denso IT Laboratory)

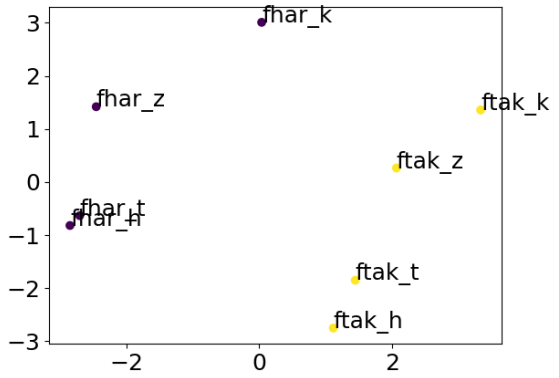

Fig. 2　Principal component analysis of embedded vectors in terms of a style.

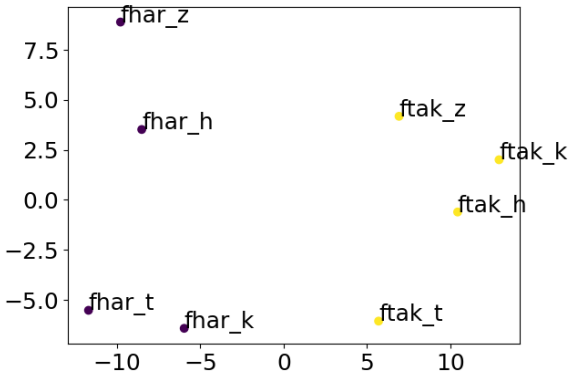

Fig. 4　Principal component analysis of x-vectors of synthesized speech in terms of a style.

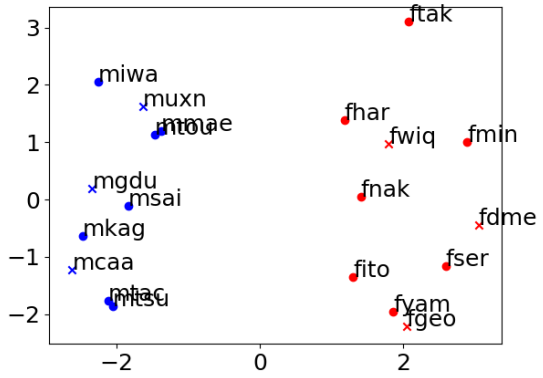

Fig. 3　Principal component analysis of embedded vectors in terms of a speaker, where ‘o’ indicates the speaker in training set and ‘x’ indicates the speaker who does not exist in the training set.

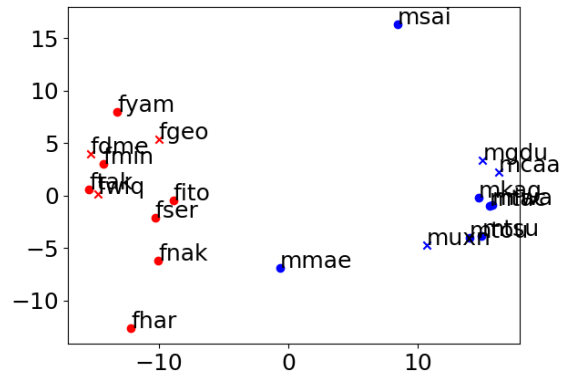

Fig. 5　Principal component analysis of x-vectors of synthesized speech in terms of a speaker.

3.3　合成音声の x-vector での評価

上と同様、2 人の女性話者について、4 スタイルの合成音声の x-vector の主成分分析結果を図 4 に示す。話者は分離されていることと ftak_k を除けばおおむね感情ごとに分類されていることがわかる。話者での結果を図 5 に示す。このように新たに加えた話者は学習話者とは異なる点にマッピングされている。ただオリジナルほどのばらつきは見られないことがわかる。

4　まとめ

合成音声の話者性と感情性を制御することを目的として、Tacotron2 に話者タグとスタイルタグを加えた。複数話者から学習したモデルにおいて、話者タグをランダムにすることで学習話者とは異なる音声を得られる可能性を示した。ただ学習話者のばらつきに比べると新たに生成された話者のばらつきは小さく、新たに話者を創造する能力は Tacotron2 では高くないことが推察される。

参考文献

[1] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, “End-to-end emotional speech synthesis using style tokens and semi-supervised training,” Proc. APSIPA, pp.623–627 (2019).

[2] 栗原清，清山信正，熊野正，今井篤，“End-to-end 音声合成における発話スタイル制御に関する音質評価，” 信学技報 SP2018-58，**118**，pp.29–34 (2019).

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

[4] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” Proc. ICASSP, pp.3617–3621 (2019).

[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” Proc. ICASSP, pp.5329–5333 (2018).