声道のMRIデータに基づくDNN音声合成の予備的な検討*

☆小澤凜夏，竹本浩典（千葉工大），平井啓之

1　はじめに

近年, DNN を用いた音声合成では自然音声と同等の品質の音声が出力できるようになった[1]。また，任意の話者性を再現する音声合成も検討されている[2]。そのモデルには x-vector[3]など話者認証で用いられる埋め込みベクトルが広く用いられている[4]。しかし，x-vector はそのベクトルを出力するのに音声が必要であり，病気などで声を失った話者の音声は再現できないという問題がある。それに対し声道の MRI データは声を失った話者でも撮像が可能である。また声道の MRI データは，声道断面積関数を計測し伝達関数のピークを計算することで音声の個人性に寄与する母音のフォルマントを取り出せる[5]。これにより最終的に MRI データからフォルマントなどの音声の個人性を表す音響的特徴を抽出することで，声を失った話者に対しても元の音声の個人性を再現した音声を合成できるのではないかと考えた。また，音響的特徴を話者埋め込みベクトルにすることで値が操作しやすくなるという利点がある。x-vector はどの値がどの程度話者性に影響するのか定かではなく，操作しづらい。しかし，音響的特徴ならどの程度値を変えればよいか明確であり話者性が制御しやすくなると考えられる。

本稿では MRI データから話者性を再現した音声合成をする前段階の実験として，MRI から抽出できる音響的特徴である母音のフォルマントや，音声の特徴に寄与する非周期性指標・基本周波数（F0）などの音響的特徴を埋め込みベクトルとして音声合成し，話者性が再現できるか検討した。

2　提案手法

話者の特徴を表すベクトルとして，音声から分析できる物理的な特徴である音響的特徴を使用した話者埋め込みベクトルを提案する。埋め込みベクトルに用いる音響的特徴は 5 母音の定常発話時の F1，F2，F3・帯域毎に平均した非周期性指標 3 点･F0 の自然対数（logF0）の平均と分散の合計 20 次元である。

今回使用する音声データである JVS コーパス[6]には定常母音が収録されていないため，フォルマントの抽出には x-vector で学習したモデルを用いて合成した定常母音を使用する。帯域毎の非周期性指標と F0 は JVS コーパス内の話者共通の 100 文を用いて WORLD[7] (D4C edition [8])で分析したものを用いた。

ここで，埋め込みベクトルの複雑さに対する学習データの不足による過学習を避けるため，非周期性指標・logF0 の平均と分散をそれぞれベクトル量子化し，代表ベクトルを用いて学習した。

3　評価実験

3.1 実験条件

提案手法の有効性を評価するため，男女混合 100 名の読み上げ音声が収録された JVS コーパスの話者間で共通の 100 文と，話者毎に異なる 30 文を用いて評価実験を行った。学習データは男性 3 名女性 1 名の 4 名の話者を除外した 96 名とした。サンプリング周波数は 24 kHz とした。学習時には埋め込みベクトルに用いた音響的特徴，および予測する WORLD の F0 を除く全てのパラメータは, 全データの平均が 0，分散が 1 になるように正規化した。F0 は話者毎に平均が 0，分散が 1 になるように正規化した。音声合成のモデルは, Tacotron[9]をベースに decoder でメルスペクトログラムを予測した後にポストネットで 60 次元のメルケプストラム，3 次元の非周期性指標，および F0 を予測するモデルを用いた。音声の生成には WORLD を用いた。今回の実験ではフレームシフト 10 ms にした以外は Tacotron[9]に記載のネットワークのパラメータを用いた。また，x-vector は，KALDI[10] の pre-trained xvector model[11]を使用した。

3.2 結果

客観評価として，合成音声が話者の音響的特徴を再現しているのか確認するために，学習データから除外した 4 名の話者で定常母音と話者で共通の 100 文を合成・分析し，入力した値との誤差を確認した。誤差と，その誤

* Preliminary study of DNN speech synthesis based on vocal tract MRI data, by OZAWA, Rika, TAKEMOTO, Hironori (Chiba Institute of Technology), and HIRAI, Hiroyuki.

差の比較の参考とするために話者間の標準偏差を Table 1 に示す。フォルマントは定常母音から、非周期性指標(AP)，F0 は文章発話から計算した結果である。太字は誤差が標準偏差と比較し大きかった値である。合わせて参考のため合成音声とベクトルとして入力した値の/a/の F1-F2 平面図を Fig. 1 に示す。三角が合成音声，丸が入力した値である。Table 1 より，話者間の標準偏差と比較し，各誤差が概ね小さいことがわかった。また，Fig. 1 より合成音声と元話者で近い値に位置していることがわかる。これらにより合成音声は元の話者の音響的特徴を再現していると考えられる。

主観評価実験として x-vector と提案手法で DMOS 評価をした。被験者は 7 名，テストデータは前述の 4 名で，共通の 100 文から 10 文を抽出し，話者類似度を 1 を似ていない，5 を似ているとして 5 段階で評価した。結果を Fig. 2 に示す。jvs001 と jvs002 では x-vector よりも類似度が高く，jvs028 では同程度の類似度であった。jvs006 では x-vector の方が類似度が高かった。これは Table 1 に示すように，jvs006 は話者性に大きく寄与する F0 の平均の誤差が大きかったためだと考えられる。

上記結果より提案手法は x-vector と概ね同程度の話者類似度が得られたと考えられる。

4　まとめ

MRI データから音声合成するための予備的な実験として音響的特徴をベクトルとした音声合成を提案した。評価実験から提案手法を用いても x-vector と同程度の話者類似度が得られることがわかった。

今後はさらに話者類似度を上げるため，今回では考慮しなかった音声の個人性を表すパラメータである音源の形状や梨状窩のディップ周波数などの追加の検討，及び MRI データからフォルマントを取り出した場合同じ結果が得られるかを検討する。

参考文献

[1] Shen *et al.*, Proc. ICASSP, pp. 4779- 4783, 2018.
[2] Ping *et al.*, Proc. ICLR, pp.214-217 2018.
[3] Snyder *et al.*, Proc. ICASSP, pp. 5329- 5333, 2018.
[4] Fang, *et al.*, Proc. SSW, pp. 155-160, 2019.
[5] Takemoto *et al.*, J. Acoust. Soc. Amer., vol. 119, no. 2, pp. 1037-1049, 2006.
[6] Takamichi *et al.*, arXiv, arXiv:1908.06248, 2019.
[7] Morise *et al.*, IEICE, pp. 1877-1884, 2007
[8] Morise, Speech Communication, vol. 84, pp. 57-65, 2016

Table 1 合成した音声とベクトルとして入力した話者の特徴の誤差

	jvs001	jvs002	jvs006	jvs028	話者間の標準偏差
/a/ F1	48.81	2.51	99.40	44.67	194.91
/a/ F2	98.82	34.07	28.46	5.11	214.36
/a/ F3	14.39	**868.02**	**472.66**	118.72	429.71
/i/ F1	5.12	12.87	22.52	3.78	39.25
/i/ F2	135.05	179.61	117.31	96.93	424.04
/i/ F3	120.97	**553.92**	**328.09**	131.59	305.61
/u/ F1	6.98	10.74	19.95	7.87	36.84
/u/ F2	49.56	122.70	86.83	44.03	219.03
/u/ F3	239.38	88.52	326.17	31.63	400.76
/e/ F1	8.34	**58.43**	6.83	24.17	45.79
/e/ F2	153.60	280.11	17.58	28.09	498.56
/e/ F3	47.29	**570.15**	**324.85**	**318.75**	266.12
/o/ F1	**65.77**	28.61	15.38	10.99	45.72
/o/ F2	20.33	6.83	108.18	**831.82**	300.28
/o/ F3	49.18	66.08	98.83	35.43	375.51
AP1次元	1.5792	**2.4608**	1.6264	0.8560	1.8122
AP2次元	0.3703	**1.4609**	0.5992	0.3331	1.1371
AP3次元	0.0841	0.4671	0.0593	0.4784	1.0192
logF0の平均	0.1225	0.0286	**0.8566**	0.0615	0.3021
logF0の分散	**0.0388**	0.0064	0.0001	0.0083	0.0195

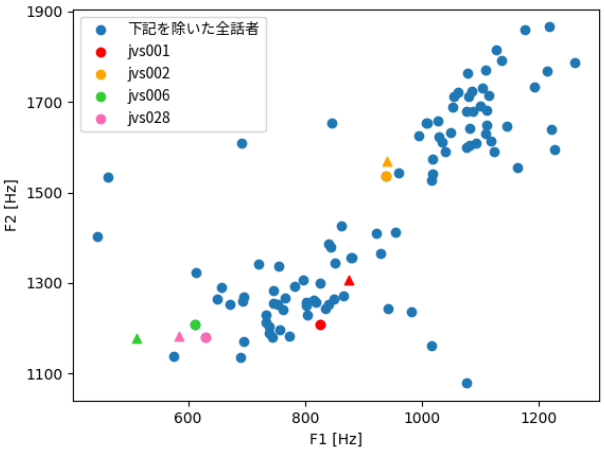

Fig. 1 提案手法で合成した音声のフォルマントとベクトルとして入力したフォルマントの F1-F2 平面図

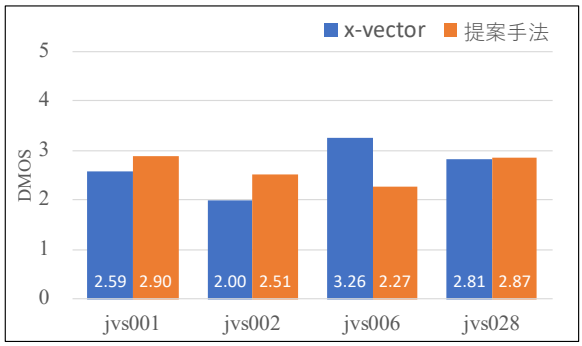

Fig. 2 x-vector と提案手法の合成音声の話者類似度の評価結果

[9] Wang *et al.*, arXiv, arXiv:1703.10135, 2017.
[10] Povey, *et al.*, Proc. ASRU, 2011.
[11] http://www.kaldi-asr.org/models/m3