明示的な音声特徴量に基づくDNN音声合成*

☆後藤仁，小澤凜夏（千葉工大），竹本浩典（千葉工大），平井啓之，
前川喜久雄（国語研）

1　はじめに

近年，音声合成は DNN によって自然音声と同等の評価が得られるようになり[1]，任意の話者性を再現する多話者音声合成が検討されている[2]。そのために音声から DNN でその話者の特徴を抽出した x-vector[3]を話者埋め込みベクトルとして用いる手法が広く用いられている。しかし，x-vector はどの値が話者のどのような音声特徴量を表しているのか不明瞭であり，個々の特徴量を制御することが困難であるという問題がある。

そこで，音声の個人性を表す明示的な音声特徴量を話者埋め込みベクトルとして用いて学習すれば，音響特徴量を操作しやすくなるのではないかと考えた。そこで，予備実験として 5 母音のフォルマントや基本周波数（F0）などの明示的な音響特徴量を話者埋め込みベクトルとして音声合成を行ったところ，従来法と同程度の話者類似度が得られた[4]。

フォルマントや F0 は声道形状や喉頭の構造や運動から推定できる可能性がある。そこで本稿では，まず発話中の話者の頭頸部を撮像した MRI データから，5 母音のフォルマント，F0，非周期性指標を抽出できるか検討する。そして，これらを話者埋め込みベクトルとして音声を合成し，話者性を評価したので報告する。

2　提案手法

2.1 フォルマントの抽出

日本語 5 母音（/a/，/i/，/u/，/e/，/o/）を発声中の声道形状を MRI で撮像した。撮像した MRI データに歯列を補填し，Takemoto ら[5]の手法を用いて，声門から口唇まで 2.5 mm 間隔でグリッドを設定して（Fig. 1 左）声道断面積関数を抽出した。得られた断面積関数から声道伝達関数を計算した。その第 1～4 ピークの周波数を第 1～4 フォルマント（F1～F4）として抽出した。

2.2 F0 の推定

F0 は声帯の長さとその伸長・喉頭の上下動などから推定できるとされている[6]。すでに先行研究で MRI 画像から CNN でメルケプストラムを推定する研究が検討されている[7]。

そこで今回はそのモデルを参考に発話中の MRI 動画のフレームから喉頭周辺を切り出した画像（Fig. 1 右）と同時に録音した音声から抽出した F0 を用いて CNN モデルを作成した。そして，このモデルを用いて MRI 動画のフレームから F0 を推定した。

2.3 非周期性指標

非周期性指標は声門で生じる乱流雑音に由来する。そこで MRI 画像から声門の開度などを検討したが，空間・時間解像度が低いため，画像から非周期性を推定することはできなかった。そこで，本研究では音声から抽出することとした。

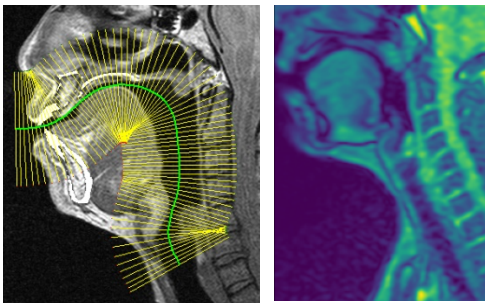

Fig. 1 左：声道断面積関数を抽出するグリッド，右：F0 推定に用いる画像の領域

3　音声特徴量と音声合成の評価

3.1 音声特徴量の再現結果

MRI データから計算により求めたフォルマントと，実音声から抽出したフォルマントの誤差を検討した。日本人話者 14 名（男性 12 名，女性 2 名）×5 母音×4 フォルマントの計 280 の周波数のうち，4 つが誤差 3%を超えたが，その他は誤差 1%以下であった。よって，画像からフォルマントは比較的正確に抽出できたといえる。

Table 1 は，3 名の話者の画像から推定した logF0（F0 の自然対数をとったもの）と，実

* DNN speech synthesis based on explicit speech features, by GOTO, Jin, OZAWA, Rika, TAKEMOTO, Hironori (Chiba Institute of Technology), HIRAI, Hiroyuki, and MAEKAWA, Kikuo (NINJAL).

際の音声から抽出した logF0 の平均と分散の誤差である。平均の誤差は少ないが，分散の誤差は大きい傾向が見られた。

Table 1 3 名の MRI データから推定した logF0 と音声から抽出した logF0 の誤差

平均	M1	F1	M2
正解	4.8763	5.3769	4.9752
予測	4.7971	5.3718	4.9357
誤差[Hz]	0.0791	0.0051	0.0395
誤差[%]	-1.6	-0.1	-0.8
分散			
正解	0.0424	0.0328	0.0311
予測	0.0475	0.0329	0.0391
誤差[Hz]	-0.0050	-0.0001	-0.0081
誤差[%]	**11.9**	0.3	**26.0**

3.2 音声合成の評価結果と考察

音声合成は小澤らの手法[4]を用いて行った。従来の x-vector を用いて合成した音声と，提案手法で合成した音声をそれぞれ実音声と比較し，DMOS 評価を行った。評価に用いた音声は，3 名が「北風と太陽」の 5 文を読み上げたもので，暗騒音を含んでいた。

実験参加者は聴力に問題のない 4 名で，1 を似ていない，5 を似ているとして音声の話者類似度を 5 段階で評価した。結果を Fig. 2 に示す。

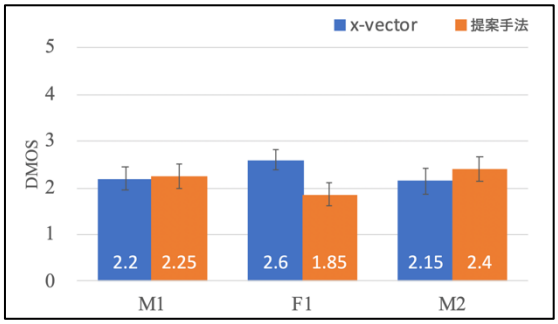

Fig. 2 x-vector と提案手法の話者類似度の評価結果

この結果より，提案手法は x-vector と概ね同程度の話者類似度が得られた。しかし，全体的に類似度が提案手法，x-vector ともに 3 を下回った。x-vector の類似度が低い要因の一つは，録音した音声に暗騒音が含まれていたため，DNN で出力した x-vector に話者性以外の情報が含まれていたことがあげられる。一方，提案手法の類似度が低い要因として，MRI では仰臥位で撮像するため，舌の位置が通常発話時と異なり，その影響を受けてフォルマントが変化したことがあげられる。

4 まとめ

本研究では，まず MRI 動画からフォルマント・F0・非周期性指標を抽出できるか検討した。その結果，フォルマントは声道断面積関数を抽出して伝達関数を計算することで求めることができた。F0 は喉頭周辺の画像と実音声の F0 を学習させた CNN により抽出することができた。しかし，声門の開度などに由来すると思われる非周期性指標は画像から抽出することができなかった。非周期性指標のみ実音声から抽出し，これと MRI 動画から抽出したフォルマントと F0 を埋め込みベクトルとして音声を合成した。その結果，従来の x-vector を用いて合成した音声と同程度に話者類似度を再現できたが，いずれも値は低かった。その要因として，録音した音声の暗騒音や MRI 撮像時の姿勢の影響などの影響が考えられる。

謝辞

本研究は JSPS 科研費 20H01265 の助成を受けて実施した。

参考文献

[1] J. Shen, et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions", ICASSP, pp. 4779-4783, 2018.

[2] Ping et al., Proc. ICLR, pp.214-217 2018.

[3] D. Snyder, et al., “X-vectors: Robust DNN embeddings for speaker recognition,” ICASSP, pp. 5329- 5333, 2018.

[4] 小澤凜夏ら，音講論(春), pp. 823-824, 2021.

[5] Takemoto et al., J. Acoust. Soc. Am., 119, pp. 1037‐1049, 2006.

[6] 本多清志，「実験音声科学-音声事象の成立過程を探る-」 コロナ社, 186p, 2018.

[7] 丹治涼ら，春季日本音響学会講演論文集, pp.749-750, 2021.