スタイル適応したDNN音声合成における話者性の検討*

☆蛭田宜樹（東工大），郡山知樹（東大），
太刀岡勇気（デンソーITラボラトリ），小林隆夫（東工大）

1 はじめに

スタイル付与とは読上げ音声のみ収録した話者（目標話者）に対して，他の話者（学習話者）から学習した感情表現や発話様式（スタイル）を付与することで目標話者のスタイル音声を合成する手法である．我々はこれまでにDNN音声合成におけるスタイル付与において，目標話者の読上げスタイル音声が少ない場合に話者を示す情報としてi-vectorを用いることで話者性やスタイル再現度の向上がみられることを報告した [1]．しかし合成音声の話者性と話者依存の変換層（話者層）の位置の関係について検討していなかった．

そこで本稿では，話者層の位置の違いと合成音声の話者性への影響からスタイル付与における適切な話者層の位置について検討した結果を報告する．なお本稿でのスタイル付与は，学習話者と目標話者の読上げ音声を用いて学習した複数話者読上げモデルを，学習話者のスタイル音声を用いてスタイル適応することにより行う．

2 DNN音声合成におけるスタイル付与

DNN音声合成におけるスタイル付与では話者やスタイルをone-hotの話者選択ベクトルとスタイル選択ベクトルで指定する手法が提案されている [2]．この手法では言語特徴量に話者選択ベクトルとスタイル選択ベクトルを結合するモデル（AIM），話者毎の最終隠れ層とスタイル毎の出力層を持つモデル（SM）と話者毎の出力層とスタイル毎の出力層を持つモデル（PM）を検討し，AIMとPMが有用とされた．

また我々は目標話者の学習用読上げスタイル音声が少ない場合に着目し，話者に関する情報としてi-vectorを用いることを提案した [1]．その結果i-vectorを用いることでAIMでの話者性やスタイル再現度が向上することを示した．また出力特徴量を話者毎に平均0分散1となるように正規化することで，全データで同様の正規化を行うより自然性や話者性が向上することを報告した [3]．

3 音響モデルとその学習法

一般的な順伝播型DNNの構造と話者層の位置の候補を図1に示す．順伝播型DNNは大きく分けて入力層，隠れ層と出力層からなる．したがって話者層の位置もこのうちの1つ以上が考えられる．本稿では位置の違いによる合成音声の話者性への影響を調べるため，入力層，隠れ層と出力層のいずれか1ヶ所を話者層にする．それぞれのモデルを便宜上IM，HMとOMと呼ぶ．井上らのSMを参考に，HMでは最終隠れ層のみを話者層とする．

話者層の構造を図2に示す．話者層はPMやSMと同様one-hot表現の話者選択ベクトルによりどの話者の線形変換を用いるかを決定する．事前実験の結果より，話者層の出力は話者依存の線形変換の結果に話者共通の線形変換の結果を足して活性化関数を適用したものとする．

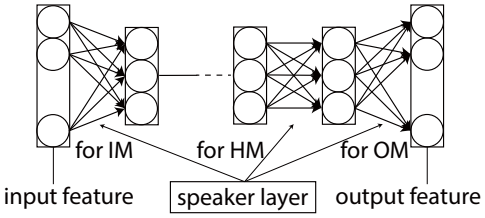

Fig. 1 一般的な順伝播型DNNの構造と話者層の位置の候補．

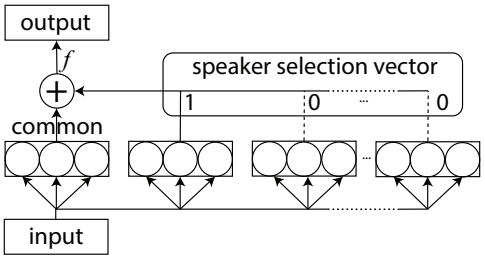

Fig. 2 話者層の構造．fは活性化関数を表す．

学習は2段階に分けて行う．まず，学習話者と目標話者の読上げ音声のみを用いて複数話者読上げモデルを学習する．続いて，学習話者の各スタイル音声を用いてスタイル適応を行う．複数話者読上げモデルを学習することで話者層が目標話者の話者性を獲得することができる．また，話者適応の際は目標話者の線形変換パラメータを除いた全てのパラメータを更新する．この結果学習話者の線形変換パラメータが話者性と話者特有のスタイル表現を，全話者共通のパラメータが学習話者に共通するスタイル表現を獲得することが期待できる．結果，読上げで学習した目標話者の話者性と学習話者共通のスタイル表現により目標話者の話者性を維持したままそのスタイルに聞こえる音声を合成できると考えられる．

4 実験

4.1 条件

実験には2種のデータセットを用いた．1つ目は男性4名，女性3名がATR音素バランス文503文章を悲しげ，楽しげ，読上げ，ぞんざいの4つのスタイルで発話したものである．2つ目は男女5名ずつがATR音素バランス文からスタイル毎に異なる100文と各スタイルによる発話を想定した30文 [4] を発話したものである．2つ目のデータセットの女性2名（F-1, F-2）と男性1名（M）を目標話者とし，その他を学習話者として用いた．

*A study on speaker reproducibility in style adapted DNN speech synthesis. by Hiruta, Yoshiki (Tokyo Institute of Technology), Koriyama, Tomoki (University of Tokyo), TACHIOKA, Yuuki (Denso IT Laboratory), KOBAYASHI, Takao (Tokyo Institute of Technology)

Table 1　各話者各スタイルのテスト文の自然音声全体の i-vector と読上げの i-vector の cos 類似度.

	sad	joyful	reading	rough
F-1	0.80	0.72	0.96	0.89
F-2	0.78	0.75	0.96	0.85
M	0.68	0.81	0.96	0.78

データセットは学習セット，学習の停止基準に用いる開発セットと評価セットに分割した．目標話者は読上げのみ 100 文を用い，これも学習セットと開発セットに分割した．複数話者読上げモデル学習時の学習セットと開発セットはそれぞれ 4050 文と 471 文，それぞれのスタイルへの適応時は 3780 文と 441 文となった．評価セットには各スタイルによる発話を想定した 30 文を用いた．

言語特徴量は 443 次元，音響特徴量は 139 次元とした．言語特徴量にはトライフォン，アクセント，文長とフレームの位置に関する情報が含まれる．音響特徴量は 40 次元のメルケプストラム，対数 F0，有声無声フラグと 5 次元の帯域平均非周期性指標とそれらの 1 次，2 次差分からなる．これらは周波数 16 kHz でサンプルした音声から STRAIGHT を用いて 5 ms 単位で抽出した特徴量から計算した．

音響モデルの隠れ層は 3 層とし，1 層の隠れ素子数は 512 とした．隠れ層の活性化関数は sigmoid，出力層の活性化関数は恒等関数とした．最適化手法，初期学習率とミニバッチサイズは複数話者読上げモデル学習時は Adam，0.001 と 1024 とし，スタイル適応時は MomentumSGD（moment=0.9），0.05 と 128 とした．

評価指標には音響特徴量歪と，話者性の評価のため各目標話者について各スタイルの合成音声と学習セットの読上げ音声間の i-vector の cos 類似度を用いた．i-vector は 50 次元とし，日本語話し言葉コーパス (CSJ) の学会講演音声を対象とした 2048 混合の universal background model を用いて抽出した．

Table 1 に各目標話者の各スタイルの自然音声と学習セットの読上げ音声間の i-vector の cos 類似度を示す．Table 1 より，目標話者の音声であればスタイル音声でも cos 類似度が概ね 0.7 より大きくなる．

4.2　客観評価結果

Table 2 に各モデルのスタイル毎の音響特徴量歪を示す．メルケプストラム距離では OM が，対数 F0 の RMS 誤差では HM が概ね最小であった．メルケプストラム距離では楽しげでの IM と OM 間以外では有意差が見られた．また対数 F0 の RMS 誤差では悲しげの全モデル間と読上げの HM と OM 間では有意差が見られなかった．

Table 3 に各スタイルの合成音声の i-vector について，各目標話者の読上げ音声の i-vector との cos 類似度を示す．いずれの話者，スタイルでも OM が最大であり，Table 1 に示した自然音声の値に近くなった．

4.3　考察

客観評価結果から，話者層は出力層に配置すべきと言える．Table 2 のメルケプストラム距離の結果と Table 3 は OM がスタイル付与に適していることを示してる．ただし，対数 F0 の RMS 誤差は HM が最小だったため，隠れ層での話者モデリングについても検討する必要がある．

Table 2　合成音声と自然音声間の音響特徴量歪による客観評価結果.

メルケプストラム距離 [dB]			
	IM	HM	OM
sad	6.14	6.17	**5.87**
joyful	5.98	6.09	**5.97**
reading	5.15	5.04	**4.95**
rough	5.85	5.69	**5.54**
対数 F0 の RMS 誤差 [cent]			
sad	418.2	420.7	**417.5**
joyful	414.3	**347.5**	356.5
reading	197.9	**191.3**	193.5
rough	278.0	**258.3**	262.4

Table 3　各話者各スタイルのテスト文の合成音声全体の i-vector と読上げの i-vector の cos 類似度.

	sad			joyful		
	IM	HM	OM	IM	HM	OM
F-1	0.63	0.52	**0.82**	0.56	0.65	**0.72**
F-2	0.57	0.46	**0.75**	0.51	0.55	**0.60**
M	0.72	0.76	**0.89**	0.62	0.65	**0.79**
	reading			rough		
F-1	0.85	0.88	**0.90**	0.69	0.73	**0.86**
F-2	0.81	0.85	**0.87**	0.60	0.65	**0.79**
M	0.88	0.88	**0.91**	0.72	0.85	**0.89**

男性 M の悲しげの合成音声の i-vector の cos 類似度が自然音声のものを上回っているが，これは自然音声のスタイル表現の度合いが合成音声に付与されたスタイル表現の度合いより大きかったことが原因だと考えられる．このような話者特有のスタイル表現は本稿の手法では再現することが難しく，検討の余地がある．

5　おわりに

本稿ではスタイル適応した DNN 音声合成における話者性と話者層の位置の関係を検討した．i-vector の cos 類似度による評価の結果，話者層が出力層にある場合に最も話者性が高くなることが示唆された．今後の課題として主観評価実験による合成音声の話者性の評価があげられる．

参考文献

[1] 蛭田 他，“DNN 音声合成における少量の学習データを用いたスタイル付与の検討,” 音講論（春），1-P-33, pp.1119-1120, 2019.

[2] K. Inoue *et al.*, “An investigation to transplant emotional expressions in DNN-based TTS synthesis” in Proc. APSIPA ASC, pp. 1253–1258, 2017.

[3] 蛭田 他, “DNN 音声合成におけるスタイル付与モデル学習法の検討,” 電子情報通信学会技術研究報告, vol.119, no.80, SP2019-1, pp.1-6, 2019.

[4] 橘 他，“平均声と重回帰 HSMM を用いた合成音声の多様なスタイル・声質制御の検討，” 日本音響学会 2009 年春季研究発表会講演論文集，1–6–3, pp.293–296，2009.