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*

☆蛭田宜樹，郡山知樹（東工大），
太刀岡勇気（デンソーアイティーラボラトリー），小林隆夫（東工大）

1 はじめに

音声合成の課題として多様な話者・スタイルの取り扱いがある．DNN 音声合成では一般に多量の学習データが必要とされるが，目標話者の多様なスタイル音声を長時間音声収録をすることは難しい．これに対し，目標話者音声の話者性を読上げスタイル音声から学習し，他の話者から学習されたスタイルを付与することにより，多様なスタイルでの音声合成を可能とする借用感情音声合成が提案されている [1]．本稿では目標話者の学習音声が少ない状況において，同様に多様なスタイルでの音声合成を可能とする手法について検討を行う．

2 DNN 借用感情音声合成

DNN 借用感情音声合成 [1] では，目標話者の音声は読上げスタイルしかないが，既存話者の音声は読上げスタイル以外の音声も存在する場合を想定し，既存話者と目標話者を合わせた複数話者・複数感情モデルを学習する．このモデル学習の際，話者と感情表現に関する特徴量を明示的に与えて話者・感情と出力特徴量の対応が明らかなモデルを構築し，合成時にはそれらの特徴量を操作することで目標話者の所望の感情音声を合成する．話者および感情を制御するための特徴量としては one-hot 表現の話者選択ベクトルと感情選択ベクトルを使用する．借用感情合成音声の評価結果 [1] より，言語特徴量に話者・感情選択ベクトルを結合して入力特徴量とする Auxiliary Input Model （AIM）と，それぞれの話者とそれぞれのスタイルに特有の出力層を持つ Parallel Model （PM）の 2 つが有効だと報告されている．

3 提案法

本稿では目標話者の読上げ音声が少量しか得られない場合を想定する．この場合，目標話者に対応するモデルパラメータが十分に学習されないことが考えられる．そこで，話者選択ベクトルを話者間の類似度を考慮できるようなものに変更し，「声が似ている話者」という情報を利用することを考える．ここでは文献 [2] において音声合成での有用性が認められている i-vector を採用する．

DNN のモデル構造として，[1] で提案されている AIM と PM を採用する．本稿では PM のモデル構造として全ての話者とスタイルで共通の出力層も持つ PM+を用いる．AIM においては入力特徴量の話者選択ベクトルを該当する話者の i-vector に置き換える．このモデルを AIMIV とする．一方，PM においては話者ごとに出力層が存在するため，単純に置き換えることはできない．ここでは文献 [3] を参考に，次のように話者 i の i-vector の $\boldsymbol{i}_i$ と直前の隠れ層 $\boldsymbol{h}$ から話者層の出力 $\boldsymbol{O}_{sp}$ を計算する．

$$\left[\boldsymbol{a}^{\mathrm{T}}, \boldsymbol{b}^{\mathrm{T}}\right]^{\mathrm{T}} = \boldsymbol{W}_{spout} f\left(\boldsymbol{W}_{iv}\boldsymbol{i}_i + \boldsymbol{b}_{iv}\right) + \boldsymbol{b}_{spout} \quad (1)$$

$$\boldsymbol{O}_{sp} = \boldsymbol{W}_{out}\left(\boldsymbol{a} \odot \boldsymbol{h}\right) + \boldsymbol{b}_{out} + \boldsymbol{b} \quad (2)$$

ただし，$\boldsymbol{W}_{iv}$ は $N_h \times N_{iv}$，$\boldsymbol{W}_{spout}$ は $(N_h + N_{out}) \times N_h$，そして $\boldsymbol{W}_{out}$ は $N_h \times N_{out}$ の行列を表す．また，

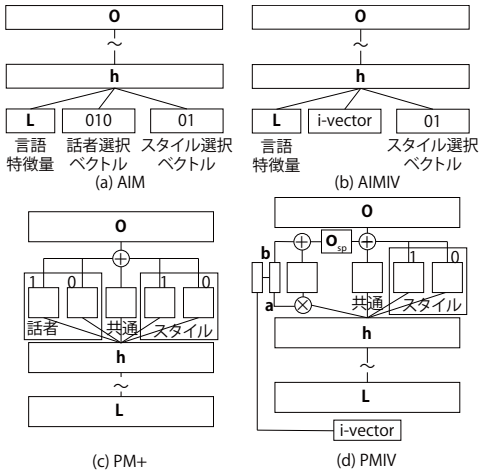

Fig. 1 各モデルの構造の比較．

$\boldsymbol{b}_{iv}$ は N_h，$\boldsymbol{b}_{spout}$ は $N_h + N_{out}$，$\boldsymbol{a}$ は N_h，そして $\boldsymbol{b}_{out}$ は N_{out} 次元のベクトルを表す．ただし，N_h は隠れ層の出力次元数，N_{iv} は i-vector の次元数，N_{out} は出力特徴量の次元数である．また，$f(\cdot)$ は活性化関数を表す．$\boldsymbol{a}$ は文献 [2] での Learning Hidden Unit Contribution（LHUC）と，$\boldsymbol{b}$ は話者特有のバイアスとみなせる．このモデルを PMIV とする．それぞれのモデル構造を Fig. 1 に示す．

4 評価実験

4.1 実験条件

本実験では 3 種のデータセットを用いた．1 つ目は ATR 日本語音声データベースセット B で，全話者（男性 6 名，女性 4 名）の音声を読上げ（reading）スタイルとして用いた．2 つ目は ATR 音素バランス文セットを男性 2 名と女性 1 名のプロのナレーターが悲しげ（sad），楽しげ（joyful）と読上げ （reading）各スタイルで発話したデータセットである．3 つ目は ATR 音素バランス文セットのうち 100 文を男性 10 名，女性 1 名の学生が 2 つ目と同じ 3 つのスタイルで発話したものである．

目標話者は 2 つ目のデータセットのうちの男性 1 名と女性 1 名とし，それぞれモデルを構築した．テストセットは 50 文とした．目標話者はテストセットと異なる読上げスタイル 10 文（約 42 秒）のみ学習セットに用いた．その他の話者はテストセットを除き学習セットと学習停止基準に用いる開発セットに分けた．学習セットのうち，悲しげと楽しげスタイルの音声は 1285 文ずつ，読上げスタイル音声は目標話者の音声を含め 5245 文となった．

i-vector の抽出には Kaldi を用いた．日本語話し言葉コーパス（CSJ）の学会講演音声を対象として 2048 混合の universal background model を構築し，50 次元の男女共通の i-vector を抽出した．

Fig. 1 における言語特徴量 $\mathbf{L}$ には，音素継続長では 412 次元，音響特徴量モデルでは位置情報の 4 次元を加えた 416 次元ベクトルを用いた．スタイル選択ベクトルは悲しげ，楽しげ，読上げの値をそれぞれ $(1,0),(0,1),(0,0)$ とした．話者選択ベクトルには 24

*Style transplant with small amount of training data for DNN-based speech synthesis. by HIRUTA, Yoshiki, KORIYAMA, Tomoki (Tokyo Institute of Technology), TACHIOKA, Yuuki (Denso IT Laboratory), KOBAYASHI, Takao (Tokyo Institute of Technology)

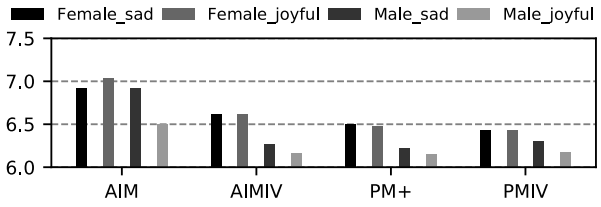

Fig. 2　メルケプストラム距離 [dB]

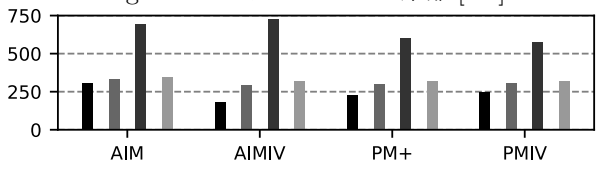

Fig. 3　対数 F0 の RMS 誤差 [cent]

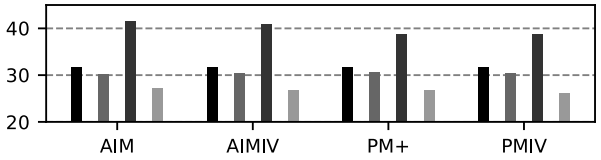

Fig. 4　音素継続長誤差 [ms]

次元の one-hot ベクトルを用いた．音素継続長モデルの出力特徴量はフレーム単位の音素継続長とした．音響モデルの出力特徴量は 40 次元のメルケプストラム，対数 F0，有声無声フラグ，5 次元の帯域平均非周期性指標とそれらの 1 次，2 次差分の計 139 次元とした．これらは周波数 16 kHz でサンプルされた音声から STRAIGHT [4] を用いて 5 ms 単位で抽出したスペクトル包絡，非周期性指標と F0 から計算された．出力特徴量は平均 0 分散 1 に正規化した．

各モデル構造について，音素継続長モデルは隠れ素子数 64 の隠れ層を 2 層，音響モデルは隠れ素子数 512 の隠れ層を 3 層持つフィードフォワードニューラルネットワークとした．隠れ層の活性化関数として sigmoid 関数を用いた．最適化手法は MomentumSGD（momentum=0.9）とし，初期学習率とミニバッチサイズはそれぞれ音素継続長モデルは 0.001 と 64，音響モデルでは 0.05 と 128 とした．

4.2　客観評価結果

客観評価として，合成音声と原音声間のメルケプストラム距離，対数 F0 と音素継続長の RMS 誤差を計算した．Figs. 2〜4 に結果を示す．

まず，AIM と AIMIV を比較すると，AIMIV の方がスペクトル特徴量の予測精度が高いという結果となった．一方，PM+と PMIV の音響特徴量の予測精度に大きな差は見られなかった．また，音素継続長の RMS 誤差は各モデルでほぼ同じような値となった．なお，男性目標話者の悲しげにおける対数 F0 の RMS 誤差が他より大きくなっているが，これは学習に用いた話者の悲しげの表現と目標話者のそれが大きく異なることを表していると考えられる．

4.3　主観評価結果

テストセットを用いて話者性の再現度とスタイル表出度を評価する主観評価実験を行なった．音素継続長モデルと音響モデルは同一のモデル構造とした．比較のため分析合成音声（NAT）も用いた．聴取者は 5 名で，防音室内でのヘッドホン聴取により実験を行なった．聴取者に目標話者の学習セットに用いた音声と合成音声を 1 つずつ続けて聞かせ，合成音声のスタイルと 5 段階で話者性の再現度を回答させた．合成音声は 1 話者 1 モデル 1 スタイルで 10 文とし，事前にランダムに文と順番を決定した．

Fig. 5 に各モデルでの聴取者の回答の F 値を示す．

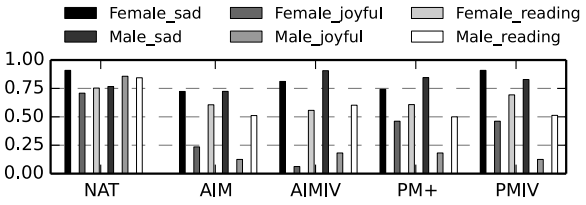

Fig. 5　各モデルでの聴取者の回答の F 値

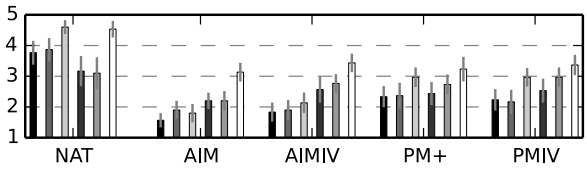

Fig. 6　各モデルの話者再現度の評価

女性話者では提案法は従来法より悲しげの表出度合いが大きいことがわかった．男性話者では AIM と AIMIV において同様の結果となった．楽しげの F 値はどのモデルでもその他と比べ小さくなった．詳細に調べると，楽しげとして合成した音声の多くが読上げとして回答されていた．

Fig. 6 に各モデルの話者性の再現度の評価結果を示す．AIM と AIMIV を比較すると概ね AIMIV の方が評価が高く，特に男性の楽しげスタイルでは有意水準 0.01 で有意差が見られた．一方，PM+と PMIV では PMIV の方が高い評価を得たが有意差があるとは言えない結果となった．

5　考察

i-vector を用いる AIMIV のメルケプストラムの予測精度が AIM を上回った理由として，MFCC から計算される i-vector がスペクトル包絡の情報を保持していることから，学習データが少量でもより良いモデルが構築できたためと考えられる．しかし，i-vector は韻律を考慮していないため，音素継続長については改善が見られなかったと考えられる．

また，PM+と PMIV で音響特徴量の予測精度に差が出なかったのは，PM+では話者特有のパラメータが出力層 1 層しかなく，10 文章程度でも十分学習できたためと考えられる．そのため，PM における i-vector の利用にはより詳細な検討が必要である．

6　おわりに

本稿では借用感情音声合成において，目標話者の読上げ音声が少量である状況を想定し，話者類似度を考慮できる i-vector を話者選択ベクトルの代わりに用いることを検討した．その結果，AIM において音響特徴量の予測精度が向上した．主観評価の結果においても AIM での性能向上が確認できた．今後の課題として，自然性に関する主観評価実験が挙げられる．

参考文献

[1] 井上 他, 音講論（秋）, 1-4-9, pp. 1105–1108, 2018.

[2] Z. Wu *et al.*, *Proc. INTERSPEECH*, pp. 879–883, 2015.

[3] H. T. Luong, J. Yamagishi, *Proc. IEEE SLT*, 2018.

[4] H. Kawahara *et al.*, Speech Communication, 27, pp.187–207 (1999).