主観的話者間類似度のグラフ埋め込みに基づく DNN 話者埋め込み *

◎齋藤 佑樹，高道 慎之介，猿渡 洋（東大院・情報理工）

1 はじめに

人間の主観的印象と対応した音声潜在表現の学習は，多様な話者性の音声を高品質に合成可能な技術の実現に重要である．これまでに我々は，クラウドソーシングで収集した主観的話者間類似度スコアに基づく deep neural network (DNN) 話者埋め込みの学習法として，(1) 類似度スコアベクトル埋め込みと，(2) 類似度スコア行列埋め込みの 2 つを提案し，多話者音声生成における品質改善を確認している [1]．本稿では，類似度スコアの新たな表現形を用いた学習法として，主観的話者間類似度のグラフ埋め込みに基づく DNN 話者埋め込みを提案する．提案法では，入力音声特徴量から，話者間の関係性（主観的に類似度しているかどうか）を表すグラフの構造を予測するように DNN を学習する．実験的評価の結果より，この提案法で学習された話者埋め込みを用いた多話者音声生成の品質が改善することを示す．

2 主観的話者間類似度を考慮した従来の DNN 話者埋め込み

我々の従来法 [1] では，受聴者によって知覚される主観的話者間類似度を定義した行列を用いて DNN を学習する．N_{s} を知覚評価に用いる話者数，$\mathbf{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_i, \cdots, \boldsymbol{s}_{N_{\mathrm{s}}}]$ を $N_{\mathrm{s}} \times N_{\mathrm{s}}$ の類似度スコア行列，$\boldsymbol{s}_i = [s_{i,1}, \cdots, s_{i,j}, \cdots, s_{i,N_{\mathrm{s}}}]^\top$ を i 番目の話者の N_{s} 次元の類似度スコアベクトルとする．行列の各要素 $s_{i,j}$ は，i 番目と j 番目の主観的話者間類似度を表す．本稿では，“i 番目と j 番目の話者の声はどれだけ類似しているか？” を評価基準とした 7 段階評価（-3：全く似ていない，$+3$：非常に似ている）の主観評価スコアの平均値として類似度スコア $s_{i,j}$ を定義する．また，行列 $\mathbf{S}$ は対称行列とし，同一話者内の類似度を示す対角成分の値は 3（スコアの最大値）とする．図 1(a) に類似度スコア行列の例を示す．

我々は，この類似度スコアを用いた DNN 話者埋め込みの学習法として，入力音声特徴量から当該話者の類似度スコアベクトルを予測する類似度スコアベクトル埋め込みと，話者埋め込みの Gram 行列（埋め込み由来の話者間類似度）を類似度スコア行列に近づける類似度スコア行列埋め込みを提案している．

3 主観的話者間類似度のグラフ埋め込み

3.1 グラフ埋め込み

グラフ埋め込みとは，データ間の関係性がグラフとして与えられたもとで，各データの潜在表現（埋め込みベクトル）を学習する技術である [2]．本稿では特に，グラフの各節点に補助特徴ベクトルが割り当てられているもとで，節点間の関係性（即ち，各節点の接続情報）を保持しながら各節点の埋め込みベクトルを学習するタスクを考える．

3.2 類似度グラフ埋め込みの学習法

話者を節点とし，類似話者対に辺が張られる類似度グラフ（図 1(b)）を構築する．提案法では，音声

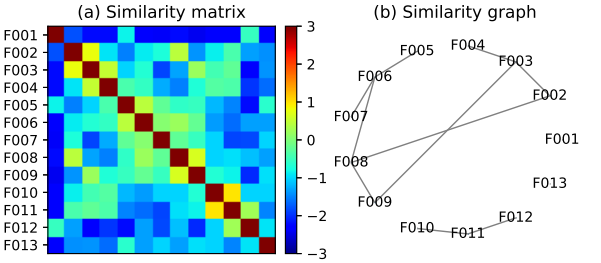

Fig. 1 (a)13 名の日本人女性話者の類似度スコア行列および (b) 対応する類似度グラフ

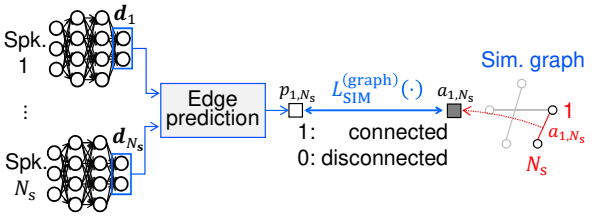

Fig. 2 類似度グラフ埋め込みの学習

特徴量から抽出される話者埋め込みを用いて，類似度グラフの辺を予測するように話者埋め込み DNN を学習する．学習時の損失関数は，次式で与えられる．

$$L_{\mathrm{SIM}}^{(\mathrm{graph})}(\boldsymbol{d}_i, \boldsymbol{d}_j) = -a_{i,j} \log p_{i,j} - (1 - a_{i,j}) \log (1 - p_{i,j}) \quad (1)$$

ここで，d_i と d_j はそれぞれ i 番目と j 番目の話者の音声特徴量から抽出される話者埋め込みである．$a_{i,j}$ は類似度グラフの隣接行列の要素であり，類似度スコア $s_{i,j}$ の値に基づいてグラフの辺の有無を決定する．$p_{i,j}$ は話者埋め込みから計算される辺の生起確率であり，本稿では文献 [5] を参考に $p_{i,j} = \exp(-||\boldsymbol{d}_i - \boldsymbol{d}_j||_2^2)$ とする．図 2 に提案法の損失関数の計算手順を示す．

3.3 考察

従来法の類似度スコア行列埋め込み [1] は，埋め込み空間内での話者間の距離を類似度スコアの値に基づいて直接的に決定する．一方で，提案法の類似度グラフ埋め込みは，スコアの値から間接的に決まる辺の有無を予測する．また，提案法では，グラフ信号処理 [3] や graph NN [4] を導入した学習も実現できる．

4 実験的評価

4.1 実験条件

本稿では，我々の従来法 [1] と同様に，JNAS コーパス [6] の 153 名の日本人女性話者の類似度スコア行列 $\mathbf{S}$ を用いた．音声データのサンプリング周波数は 16 kHz，フレームシフトは 5 ms とした．スペクトル特徴量として STRAIGHT 分析 [7] により得られた 39 次のメルケプストラム係数を，音源特徴量として対数 F0，5 帯域の非周期性指標 [8] を用いた．話者埋め込みと多数話者音声合成の DNN 学習時には，図 1(b) に示す “F001” から “F013” の 13 名以外の 140 名のデータのうち，話者間類似度の主観スコアリングに

* DNN-based speaker embedding using graph embedding of subjective inter-speaker similarity by SAITO, Yuki, TAKAMICHI, Shinnosuke, and SARUWATARI, Hiroshi (The University of Tokyo).

Table 1　合成音声の自然性に関する評価結果

	d-vec.	vs.	Prop.	p-value
F001	0.388	-	**0.612**	$< 10^{-6}$
F002	0.444	-	**0.556**	0.012
F003	0.444	-	**0.556**	0.012
F004	0.448	-	**0.552**	0.020
F005	0.432	-	**0.568**	0.002
F006	0.448	-	**0.552**	0.020
F007	0.424	-	**0.576**	$< 10^{-3}$
F008	0.392	-	**0.608**	$< 10^{-3}$
F009	0.364	-	**0.636**	$< 10^{-9}$
F010	0.448	-	**0.552**	0.020
F011	0.312	-	**0.688**	$< 10^{-9}$
F012	0.440	-	**0.560**	0.007
F013	0.372	-	**0.628**	$< 10^{-6}$

用いた各話者の 5 発話を除く全発話の 9 割を用いた．

4.1.1　DNN 話者埋め込みの条件

本稿では，話者認識に基づく学習 (d-vec.) [9] と提案法 (Prop.) を比較した．話者埋め込み DNN のアーキテクチャは，隠れ層数 4，隠れ層の活性化関数に tanh 関数を用いた Feed-Forward 型とした．1 層から 3 層までの隠れ層のユニット数は 256，話者埋め込みの抽出に用いる 4 層目の隠れ層のユニット数は 8 とした．DNN の入力は，1 次から 39 次のメルケプストラム係数とその動的特徴量の結合ベクトルとした．DNN 学習時には，入力特徴量を平均 0，分散 1 となるように正規化した．DNN 学習時の最適化アルゴリズムには，学習率を 0.01 とした AdaGrad を用いた．学習の反復回数は 100 とした．提案法の学習では，類似度スコア行列の各成分を [0, 1] の範囲に収まるように正規化して類似度グラフの隣接行列を定義した．

4.1.2　多話者音声生成の条件

本稿では，多話者音声生成モデルとして，音素事後確率 [10] と話者埋め込みで条件付けた Variational AutoEncoder (VAE) [11, 12] を用いた．音素事後確率を予測する DNN のアーキテクチャは，隠れ層数 4，隠れ層の活性化関数に tanh 関数を用いた Feed-Forward 型ネットワークとして構築した．隠れ層のユニット数は，全ての層で 1024 とした．DNN の入力は話者埋め込みのものと同じであり，話者埋め込み DNN 学習時に用いた各話者のおよそ 50 文ずつの音声を用いて，43 次元の日本語音素事後確率を推定するように学習した．学習の反復回数は 100 とした．VAE の DNN アーキテクチャは，encoder と decoder から構成される Feed-Forward 型ネットワークとした．Encoder は，活性化関数に ReLU を用いた 2 層の隠れ層を持ち，メルケプストラム係数とその動的特徴量と 43 次元の音素事後確率の結合ベクトルから 64 次元の潜在変数を抽出するように構成した．第 1 層と第 2 層の隠れ層のユニット数はそれぞれ 128，64 とした．Decoder は，encoder と対称の隠れ層を持ち，潜在変数，音素事後確率，話者埋め込みの結合ベクトルから，メルケプストラム係数とその動的特徴量を復元するように構成した．学習の反復回数 25 とした．合成音声波形の生成には，生成された 1 次から 39 次のメルケプストラム係数，自然音声の 0 次メルケプストラム係数，F0，非周期性指標を用いた．

4.2　合成音声品質の評価結果

VAE を用いた多数話者音声生成における話者埋め込みの影響を調査するために，学習に用いなかった 13 名の話者の合成音声の自然性と話者類似性に関する主観評価を実施した．各話者の 50 発話を，当該話者の話者埋め込みの推定および主観評価に用いた．クラウドソーシングによる主観評価システムを用いて，自然性の評価にプリファレンス AB テストを実施し，話者類似性の評価に当該話者の自然音声を参照音声としたプリファレンス XAB テストを実施した．主観評価の各被験者は，50 発話からランダムに抽出された 10 発話を評価した．被験者の総数は 650 人であった．

Table 2　合成音声の話者類似性に関する評価結果

	d-vec.	vs.	Prop.	p-value
F001	0.412	-	**0.588**	$< 10^{-3}$
F002	0.384	-	**0.616**	$< 10^{-6}$
F003	0.440	-	**0.560**	0.007
F004	0.448	-	**0.552**	0.020
F005	0.432	-	**0.568**	0.002
F006	0.376	-	**0.624**	$< 10^{-6}$
F007	0.452	-	**0.548**	0.032
F008	0.400	-	**0.600**	$< 10^{-3}$
F009	0.428	-	**0.572**	0.001
F010	0.420	-	**0.580**	$< 10^{-3}$
F011	0.356	-	**0.644**	$< 10^{-9}$
F012	0.484	-	0.516	0.475
F013	0.436	-	**0.564**	0.004

表 1 と 2 にそれぞれ自然性に関する評価結果と話者類似性に関する評価結果を示す．ここで，太字で示された結果は 2 手法のスコアに $p < 0.05$ で有意差があることを示す．評価結果より，“Prop.” は自然性，話者類似性の両方で常に “d-vec.” よりも高いスコアを獲得していることが確認でき，提案法による合成音声の品質改善効果が示された．

5　結論

本稿では，主観的話者間類似度のグラフ埋め込みに基づく DNN 話者埋め込みを提案し，多話者音声生成における音質改善効果を確認した．類似度スコアの半教師あり学習の枠組みを検討する．

謝辞: 本研究の一部は，JSPS 科研費 18J22090 の助成，及び総務省 SCOPE(受付番号 182103104) の委託を受け実施した．

参考文献

[1] Y. Saito *et al.*, *Proc. SSW,* pp. 51–56, Vienna, Austria, Sep. 2019.
[2] P. Goyal *et al.*, *Knowledge-Based Systems,* vol. 151, pp. 78–94, Jul. 2018.
[3] D. I. Shuman *et al.*, *IEEE Signal Processing Magazine,* vol. 30, no. 3, pp. 83-98, May. 2013.
[4] Z. Wu *et al.*, *IEEE Trans. on Neural Networks and Learning Systems,* Mar. 2020. (Early Access)
[5] J. Li *et al.*, *Proc. IJCAI,* pp. 1554–1560, Stockholm, Sweden, Jul. 2018.
[6] K. Itou *et al.*, *Journal of the ASJ (E),* vol. 20, no. 3, pp. 199—206, May 1999.
[7] H. Kawahara *et al.*, *Speech Communication,* vol. 27, no. 3–4, pp. 187—207, Apr. 1999.
[8] Y. Ohtani *et al.*, *Proc. INTERSPEECH,* pp. 2266-–2269 Pittsburgh, U.S.A., Sep. 2006.
[9] E. Variani *et al.*, *Proc. ICASSP,* pp. 4080—4084, Florence, Italy, May 2014.
[10] L. Sun *et al.*, *Proc. ICME,* Seattle, U.S.A., Jul. 2016.
[11] D. P. Kingma *et al.*, *arXiv:1312.6114,* 2013.
[12] Y. Saito *et al.*, *Proc. ICASSP*, pp. 5274–5278, Apr. 2018.