主観的話者間類似度に基づく DNN 話者埋め込みを用いた多数話者 DNN 音声合成の実験的評価 *

◎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の潜在表現の学習は，ユーザが解釈しやすく，直感的に制御しやすい音声合成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行列に基づく 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 つを提案している [1]．前者は類似度スコア行列のベクトルを予測する学習法であり，後者は話者埋め込み間のカーネル関数の値を格納したグラム行列を類似度スコア行列に近づける学習法である．この手法を用いることで，話者認識に基づく DNN 話者埋め込み [2] と比較して主観的話者間類似度と強い相関を持つ話者埋め込みが得られる．

本稿では，この手法 [1] の有効性を，variational autoencoder (VAE) [3] に基づく多数話者音声生成 [4] で実験的に評価する．主観的話者間類似度を考慮して学習された話者埋め込み空間により，学習データに含まれない未知話者も高精度に再現可能となることが期待できる．本稿ではさらに，類似度スコア行列埋め込みにおけるカーネル関数の影響を調査するために，従来の sigmoid カーネルと Gauss カーネルを比較する．評価結果より，(1) 話者認識に基づく埋め込み [2] と比較して，類似度スコアベクトル埋め込みが合成音声の品質改善に有効であること，(2) Gauss カーネルを用いた類似度スコア行列埋め込みが合成音声の品質を劣化させる傾向にあることを示す．

2 主観的話者間類似度に基づく DNN 話者埋め込み

2.1 主観スコアリングと類似度スコア行列

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 $-v$（全く似ていない）から v（非常に似ている）の間の値を取り，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は主観評価スコアの最大値 v とする．図 1(a) と (b) にそれぞれ 153 名の日本人女性話者の類似度スコア行列とその部分行列を示す．

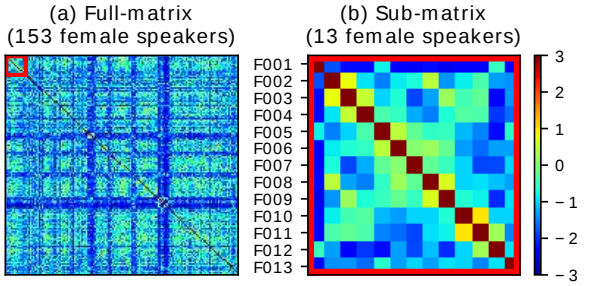

Fig. 1 (a) 大規模な主観スコアリングにより得られた 153 名の日本人女性話者の類似度スコア行列および (b) その部分行列

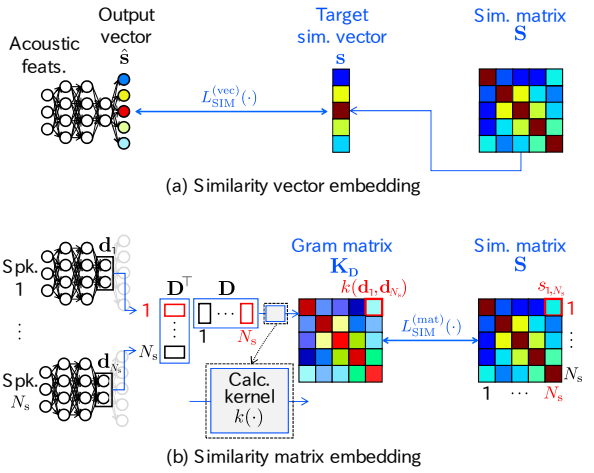

Fig. 2 (a) 類似度スコアベクトル埋め込みに基づく学習および (b) 類似度スコア行列埋め込みに基づく学習における損失関数の計算手順

2.2 類似度スコアベクトル埋め込み

類似度スコアベクトル埋め込みでは，話者埋め込みの DNN は，音声特徴量を入力とし，類似度スコア行列のベクトルを予測するように学習される．学習時の損失関数は，次式で与えられる．

$$L_\mathrm{SIM}^{(\mathrm{vec})}(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \frac{1}{N_\mathrm{s}} (\hat{\boldsymbol{s}} - \boldsymbol{s})^\top (\hat{\boldsymbol{s}} - \boldsymbol{s}) \qquad (1)$$

ここで，$\boldsymbol{s} \in \mathbf{S}$ と $\hat{\boldsymbol{s}}$ はそれぞれターゲットの類似度スコアベクトルと DNN の予測結果である．図 2(a) に損失関数 $L_\mathrm{SIM}^{(\mathrm{vec})}(\cdot)$ の計算手順を示す．

2.3 類似度スコア行列埋め込み

類似度スコア行列埋め込みでは，行列 $\mathbf{S}$ によって話者埋め込み空間の配置に制約を与えて DNN を学習す

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

る．$\boldsymbol{d}_i = [d_i(1), \cdots, d_i(N_\mathrm{d})]^\top$ を i 番目の話者の N_d 次元話者埋め込み，$\mathbf{D} = [\boldsymbol{d}_1, \cdots, \boldsymbol{d}_{N_\mathrm{s}}]$ を学習データに含まれる全話者の話者埋め込みを含む $N_\mathrm{d} \times N_\mathrm{s}$ の行列とする．学習時の損失関数は，次式で与えられる．

$$L_{\mathrm{SIM}}^{(\mathrm{mat})}(\mathbf{D}, \mathbf{S}) = \frac{2}{\|\mathbf{1}_{N_\mathrm{s}} - \mathbf{I}_{N_\mathrm{s}}\|_F^2} \left\| \widetilde{\mathbf{K}}_{\mathbf{D}} - \widetilde{\mathbf{S}} \right\|_F^2 \quad (2)$$

$$\widetilde{\mathbf{K}}_{\mathbf{D}} = \mathbf{K}_{\mathbf{D}} - (\mathbf{K}_{\mathbf{D}} \odot \mathbf{I}_{N_\mathrm{s}}) \quad (3)$$

$$\widetilde{\mathbf{S}} = \mathbf{S} - v\mathbf{I}_{N_\mathrm{s}} \quad (4)$$

ここで，$\|\cdot\|_F^2$，$\odot$，$\mathbf{1}_{N_\mathrm{s}}$，そして $\mathbf{I}_{N_\mathrm{s}}$ はそれぞれ与えられた行列のフロベニウスノルム，Hadamard 積，全ての要素が 1 である $N_\mathrm{s} \times N_\mathrm{s}$ の行列，そして $N_\mathrm{s} \times N_\mathrm{s}$ の単位行列である．$2/\|\mathbf{1}_{N_\mathrm{s}} - \mathbf{I}_{N_\mathrm{s}}\|_F^2$ は行列 $\widetilde{\mathbf{K}}_{\mathbf{D}} - \widetilde{\mathbf{S}}$ の自由度に対応し，損失関数 $L_{\mathrm{SIM}}^{(\mathrm{mat})}(\cdot)$ のスケールを正規化する役割を持つ．$\mathbf{K}_{\mathbf{D}}$ は話者埋め込みから計算される Gram 行列であり，次式で与えられる．

$$\mathbf{K}_{\mathbf{D}} = \begin{bmatrix} k(\boldsymbol{d}_1, \boldsymbol{d}_1) & \cdots & k(\boldsymbol{d}_1, \boldsymbol{d}_{N_\mathrm{s}}) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{d}_{N_\mathrm{s}}, \boldsymbol{d}_1) & \cdots & k(\boldsymbol{d}_{N_\mathrm{s}} \boldsymbol{d}_{N_\mathrm{s}}) \end{bmatrix} \quad (5)$$

ここで，$k(\boldsymbol{d}_i, \boldsymbol{d}_j)$ は $\boldsymbol{d}_i$ と $\boldsymbol{d}_j$ から計算されるカーネル関数であり，話者埋め込みに由来する話者間類似度に対応する．先行研究 [1] では話者認識とのマルチタスク学習に基づく定式化を行っていたが，この定式化が話者類似度スコアとカーネル関数の値の相関を弱めることが明らかになったため，本稿では式 (2) のみを考慮する学習を行う．図 2(b) に損失関数 $L_{\mathrm{SIM}}^{(\mathrm{mat})}(\cdot)$ の計算手順を示す．

先行研究 [1] では，主観的に類似した話者対のみを考慮する手法も提案している．この手法では，類似度スコアが 0 以下の話者対をフィルタリングする行列 $\mathbf{W}$ を導入し，式 (2) を次式に示すように書き換える．

$$L_{\mathrm{SIM}}^{(\mathrm{mat-re})}(\mathbf{D}, \mathbf{S}) = \frac{2}{\|\mathbf{W} - \mathbf{I}_\mathrm{s}\|_F^2} \left\| \mathbf{W} \odot \left(\widetilde{\mathbf{K}}_{\mathbf{D}} - \widetilde{\mathbf{S}} \right) \right\|_F^2 \quad (6)$$

ここで，$\mathbf{W}$ の i 行 j 列目の要素 $w_{i,j}$ は，$s_{i,j} > 0$ なら 1，それ以外で 0 をとる．$2/\|\mathbf{W} - \mathbf{I}_\mathrm{s}\|_F^2$ は行列 $\mathbf{W}$ の自由度（即ち，類似話者対の数）に対応し，損失関数 $L_{\mathrm{SIM}}^{(\mathrm{mat-re})}(\cdot)$ のスケールを正規化する役割を持つ．

2.4 Gauss カーネルの利用

類似度スコア行列埋め込みで用いるカーネル関数は，最終的に学習される埋め込み空間に大きく影響すると考えられる．本稿では，先行研究 [1] で用いていた sigmoid カーネル $k(\boldsymbol{d}_i, \boldsymbol{d}_j) = \tanh(\boldsymbol{d}_i^\top \boldsymbol{d}_j)$ と，Gauss カーネル $k(\boldsymbol{d}_i, \boldsymbol{d}_j) = \exp(-\gamma||\boldsymbol{d}_i - \boldsymbol{d}_j||^2)$ を比較する．ここで，γ は Gauss 分布の分散の逆数に対応するハイパーパラメータである．前者を用いた場合は，異なる話者埋め込み間がなす角に依存する形で話者間の類似度が定義される．一方で，後者を用いた場合は，各話者を原点としたときの距離に依存する形で話者間の類似度が定義される．

3 実験的評価

3.1 実験条件

本稿では，我々の従来法 [1] と同様に，JNAS コーパス [5] 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 [6] 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 [7] 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用いた各話者の 5 発話を除く全発話の 9 割を用いた．

3.1.1 DNN 話者埋め込みの条件

本稿では，以下の 4 つの学習法を用いた．

Conv.：話者認識を用いた学習 [2]

Prop. (vec)：類似度スコアベクトル埋め込み

Prop. (mat)：類似度スコア行列埋め込み

Prop. (mat-re)：同上（類似話者対のみを考慮）

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 型ネットワークとして構築した．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 [8] を用いた．学習の反復回数は 100 とした．

類似度スコアベクトル埋め込みに基づく学習では，類似度スコア行列 $\mathbf{S}$ の各成分を $[-1, +1]$ の範囲に収まるように正規化し，ユニット数を 140，活性化関数を tanh 関数とする出力層を追加した．sigmoid カーネルを用いる類似度スコア行列埋め込みでも同様の正規化を行ったが，Gauss カーネルを用いる場合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た．各話者の埋め込みは，当該話者の全発話の有声区間における音声特徴量から得られる埋め込みの平均として推定した．

3.1.2 多数話者 DNN 音声変換の条件

本稿では，多数話者の音声特徴量の生成モデルとして，音素事後確率 [9] と話者埋め込みで条件付けた VAE [4] 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 [10] 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．VAE は，話者埋め込み DNN 学習と同様のデータを用いて，音声特徴量の対数尤度の変分下限 [3] を最大化するような 25 反復の学習により構築した．音声パラメータの生成には，最尤パラメータ生成 [11] を用いた．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．

3.2 合成音声品質における類似度スコアベクトル・行列埋め込みの効果

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．

ここでは，話者認識に基づく学習法 (“Conv.”) と主観的話者間類似度に基づく学習法 (“Prop. (*)”) の比較を実施した．この主観評価に参加した被験者の総数は，2 (AB もしくは XAB) × 3 (“Prop. (*)” の数) × 13 (話者数) × 25 (1 評価あたりの被験者) = 1,950 人であった．

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，表中の各項目の左側が “Conv.” のスコアを，右側が “Prop. (*)” のスコアを表し，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．これらの結果より，“Prop. (vec)” は自然性，話者類似性の両方で常に “Conv.” よりも高いスコアを獲得していることが確認でき，話者間類似度を考慮した話者埋め込みは，多数話者音声生成における学習に用いられない話者の合成音声の品質改善に有効であることが示唆された．同様に，“Prop. (mat)” も合成音声の自然性を改善しているが，いくつかの話者（例えば “F005” や “F012”）では話者類似性を有意に劣化させていることが確認できる．同様の傾向は，“Prop. (mat-re)” の結果にも観測された．この理由を調査するために，各話者の類似している話者の数を計算したところ，“F005” と “F012” の類似話者数はそれぞれ 7 名と 1 名だけであった．故に，類似話者数が少ない場合では，類似度スコア行列埋め込みの性能が劣化することが推測される．

Table 1　合成音声の自然性に関する主観評価結果（学習法の比較）

	Prop. (vec)	Prop. (mat)	Prop. (mat-re)
F001	0.408 - **0.592**	0.456 - **0.544**	0.448 - **0.552**
F002	0.456 - **0.544**	0.456 - **0.544**	0.504 - 0.496
F003	0.416 - **0.584**	0.444 - **0.556**	0.448 - **0.552**
F004	0.452 - **0.548**	0.460 - 0.540	0.432 - **0.568**
F005	0.380 - **0.620**	0.484 - 0.516	0.484 - 0.516
F006	0.400 - **0.600**	0.452 - **0.548**	0.424 - **0.576**
F007	0.424 - **0.576**	0.484 - 0.516	0.492 - 0.508
F008	0.436 - **0.564**	0.384 - **0.616**	0.436 - **0.564**
F009	0.428 - **0.572**	0.492 - 0.508	0.460 - 0.540
F010	0.436 - **0.564**	0.464 - 0.536	0.452 - **0.548**
F011	0.460 - 0.540	0.428 - **0.572**	0.452 - **0.548**
F012	0.436 - **0.564**	0.460 - 0.540	0.524 - 0.476
F013	0.428 - **0.572**	0.436 - **0.564**	0.412 - **0.588**

Table 2　合成音声の話者類似性に関する主観評価結果（学習法の比較）

	Prop. (vec)	Prop. (mat)	Prop. (mat-re)
F001	0.436 - **0.564**	0.488 - 0.512	0.528 - 0.472
F002	0.468 - 0.532	0.496 - 0.504	0.488 - 0.512
F003	0.432 - **0.568**	0.504 - 0.496	**0.604** - 0.396
F004	0.380 - **0.620**	0.404 - **0.596**	0.488 - 0.512
F005	0.428 - **0.572**	**0.616** - 0.384	**0.596** - 0.404
F006	0.428 - **0.572**	0.444 - **0.556**	0.464 - 0.536
F007	0.492 - 0.508	**0.568** - 0.432	**0.548** - 0.452
F008	0.424 - **0.576**	0.500 - 0.500	0.504 - 0.496
F009	0.400 - **0.600**	0.500 - 0.500	0.448 - **0.552**
F010	0.432 - **0.568**	0.404 - **0.596**	0.496 - 0.504
F011	0.348 - **0.652**	0.444 - **0.556**	0.536 - 0.464
F012	0.492 - 0.508	**0.544** - 0.456	**0.564** - 0.436
F013	0.372 - **0.628**	**0.564** - 0.436	0.452 - **0.548**

3.3 Gauss カーネル導入の効果

ここでは，Gauss カーネルを類似度スコア行列埋め込みに導入した影響を調査する．

3.3.1 Gauss カーネルのハイパーパラメータの調査

まず，Gauss カーネルにおけるハイパーパラメータ γ の影響を調査する．ここでは，$\gamma = \{0.125, 0.25, 0.5, 1.0\}$ の 4 通りを用いて話者埋め込み DNN を学習させ，類似度スコア $s_{i,j}$ とカーネル関数 $k(\boldsymbol{d}_i, \boldsymbol{d}_j)$ の値の Pearson の相関係数を計算した．

図 3 に全ての話者対を用いた類似度スコア行列埋め込みにおける相関係数 r の値および類似度スコアとカーネル関数の散布図を示す．ここで，“Closed” と “Open” はそれぞれ学習データに含まれる 140 名の話者とそれ以外の 13 名の話者を意味する．この図より，ハイパーパラメータ γ が大きくなるにつれて，より類似度スコアと強い相関を持つ話者埋め込みが学習されていることが確認できる．同様の傾向は，類似話者対のみを用いた類似度スコア行列埋め込み (図 4) においても確認できる．

3.3.2 合成音声品質に対する効果

次に，3.2 節と同様の合成音声品質評価で，sigmoid カーネルと Gauss カーネルを比較した．3.2 節の調査結果に基づき，Gauss カーネルのハイパーパラメータ

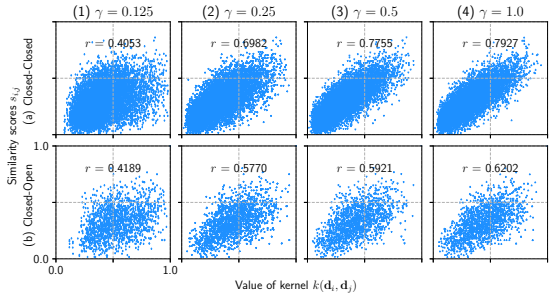

Fig. 3　全ての話者対を用いた類似度スコア行列埋め込みにおける，話者間類似度スコア $s_{i,j}$ とカーネル関数 $k(\boldsymbol{d}_i, \boldsymbol{d}_j)$ の値の散布図と，これらの相関係数 r の値．この図に含まれる全ての散布図は，全ての話者対から作成された．

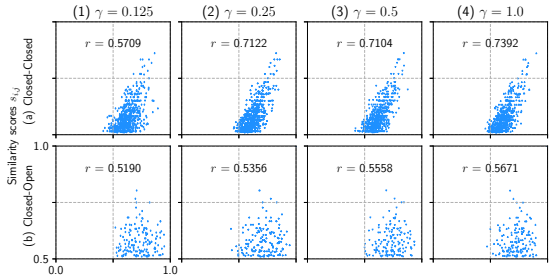

Fig. 4　類似話者対のみを用いた類似度スコア行列埋め込みにおける，話者間類似度スコア $s_{i,j}$ とカーネル関数 $k(\boldsymbol{d}_i, \boldsymbol{d}_j)$ の値の散布図と，これらの相関係数 r の値．この図に含まれる全ての散布図は，類似度スコアが 0 より大きい話者対から作成された．

γ は 1.0 とした．この主観評価に参加した被験者の総数は，2 (AB もしくは XAB) × 2 (“Prop. (mat)” もしくは “Prop. (mat-re)”) × 13 (話者数) × 25 (1 評価あたりの被験者) = 1,300 人であった．

表 3 と 4 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，表中の各項目の左側が sigmoid カーネルを用いた場合のスコアを，右側が Gauss カーネルを用いた場合のスコアを表し，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．これらの結果より，“Prop. (mat)” で Gauss カーネルを用いた場合，ほぼ全ての話者で合成音声の自然性，話者類似性が有意に劣化していることが確認できる．“Prop. (mat-re)” でも同様の傾向が確認できるが，(1) 話者 “F013” では自然性，話者類似性の両方を有意に改善させているという点，(2) いくつかの話者では話者類似性を有意に改善させているという点で異なる．以上の結果から，Gauss カーネルを用いた学習により，主観的話者間類似度とカーネル関数の値が強い相関を持つような埋め込み空間は学習できるが，話者の配置に自由度が高すぎるような空間は高品質な合成音声生成には不適切であることが示唆される．

4　結論

本稿では，主観的話者間類似度に基づく DNN 話者埋め込みを多数話者音声生成に導入し，(1) 類似度スコアベクトル埋め込みによる合成音声の品質改善，および (2) Gauss カーネルを用いた類似度スコア行列埋め込みによる合成音声の品質劣化を確認した．今後は，話者内の変動を考慮した学習法を検討する．

Table 3　合成音声の自然性に関する主観評価結果（カーネル関数の比較）

	Prop. (mat)	Prop. (mat-re)
F001	0.512 - 0.488	0.518 - 0.472
F002	**0.604** - 0.396	0.536 - 0.464
F003	**0.636** - 0.364	**0.564** - 0.436
F004	**0.744** - 0.256	**0.564** - 0.436
F005	**0.716** - 0.284	**0.564** - 0.436
F006	**0.656** - 0.344	0.540 - 0.460
F007	**0.644** - 0.356	0.548 - 0.452
F008	**0.596** - 0.404	0.496 - 0.504
F009	**0.720** - 0.280	**0.572** - 0.428
F010	**0.676** - 0.324	**0.560** - 0.440
F011	**0.640** - 0.360	0.532 - 0.468
F012	**0.576** - 0.424	**0.556** - 0.444
F013	0.488 - 0.512	0.388 - **0.612**

Table 4　合成音声の話者類似性に関する主観評価結果（カーネル関数の比較）

	Prop. (mat)	Prop. (mat-re)
F001	**0.544** - 0.456	0.508 - 0.492
F002	**0.580** - 0.420	0.532 - 0.468
F003	**0.612** - 0.388	0.432 - **0.568**
F004	**0.696** - 0.304	0.488 - 0.512
F005	**0.664** - 0.336	0.500 - 0.500
F006	**0.612** - 0.388	0.512 - 0.488
F007	**0.616** - 0.384	0.452 - **0.548**
F008	**0.592** - 0.408	0.432 - **0.568**
F009	**0.696** - 0.304	**0.572** - 0.428
F010	**0.592** - 0.408	0.528 - 0.472
F011	**0.636** - 0.364	0.504 - 0.496
F012	**0.584** - 0.416	0.540 - 0.460
F013	**0.548** - 0.452	0.436 - **0.564**

謝辞: 本研究の一部は，総務省の委託「知覚モデルに基づくストレスフリーなリアルタイム広帯域音声変換の研究」，セコム科学技術支援財団および JSPS 科研費 18J22090，17H06101 の助成を受け実施した．

参考文献

[1] 齋藤 他, 音講論（春）, 3-10-7, 2019 年 3 月.
[2] E. Variani *et al.*, *Proc. ICASSP,* pp. 4080-4084, Florence, Italy, May 2014,
[3] D. P. Kingma et al., *arXiv:1312.6114,* 2013.
[4] Y. Saito et al., *Proc. ICASSP*, pp. 5274–5278, 2018.
[5] K. Itou *et al.*, *Journal of the ASJ (E),* vol. 20, no. 3, pp. 199-206, May 1999.
[6] H. Kawahara *et al.*, *Speech Communication,* vol. 27, no. 3–4, pp. 187-207, Apr. 1999.
[7] Y. Ohtani *et al.*, *Proc. INTERSPEECH,* pp. 2266-2269 Pittsburgh, U.S.A., Sep. 2006.
[8] J. Duchi *et al.*, *Journal of Machine Learning Research,* vol. 12, 2121-2159, Jul. 2011.
[9] L. Sun *et al.*, *Proc. ICME,* Seattle, U.S.A., Jul. 2016.
[10] X. Glorot *et al.*, *Proc. AISTATS,* pp.315-32, Lauderdale, U.S.A., Apr. 2011.
[11] K. Tokuda *et al.*, *Proc. ICASSP,* pp. 1315–1318, Istanbul, Turkey, June 2000.