DNN 音声合成に向けた主観的話者間類似度を考慮した 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

Deep Neural Network (DNN) を用いた統計的音声合成 [1] の発展により，自然音声とほぼ同程度の品質の合成音声が生成可能となった [2]．一方で，音声合成技術のさらなる発展および普及には，合成音声の高品質化だけでなく，解釈が容易で制御しやすい音声の潜在表現（すなわち，人間の主観的印象と対応付けが取れた表現）を学習するための技術が重要となる．本稿では，多様な声色・音高の音声を自在に生成可能な DNN 音声合成の実現を目的として，DNN 音声合成に適した話者表現の学習法を提案する．

DNN 音声合成における最も単純な話者表現は，話者の情報を one-hot ベクトルで与える話者コード [3] である．学習データに含まれる話者の声質を高精度に再現できるが，(1) 学習データに含まれていない未知話者は表現できず，また，(2) 話者数が増えるにつれてユーザが所望する声質の話者を選択することが困難となる．話者コードの埋め込みおよび適応 [3, 4] によりこれらの問題は緩和できるが，未知話者の声質を完全に再現することは困難である．より効率的な話者表現は，入力音声の話者を特定するように学習された DNN の中間表現（すなわち，連続的な特徴空間への話者埋め込み）を用いることである．d-vector [5] はこの DNN 話者埋め込みの代表例であり，多人数話者の声質変換における話者表現として利用されている [6]．しかし，d-vector は話者を特定するための中間表現であり，その特徴空間内での配置と主観的話者間類似度には相関がないため，ユーザが所望する声質の再現および適応に適した表現であるとは限らない．

本稿では，ユーザが解釈しやすく，また，制御しやすい話者表現を学習するための手法として，主観的話者間類似度を考慮した DNN 話者埋め込みの学習アルゴリズムを提案する．まず，話者間類似度を表す類似度スコア行列を定義するための大規模な主観スコアリングを実施する．提案アルゴリズムでは，話者埋め込みの DNN は類似度スコア行列により定義される損失関数を最小化する．類似度スコアを用いる学習法として，本稿では (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 つのアプローチを取る．(1) 類似度スコアベクトル埋め込みに基づく学習では，話者埋め込みの DNN は，従来の話者コードの代わりに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．(2) 類似度スコア行列埋め込みに基づく手法では，話者埋め込みの DNN は，話者認識の softmax cross-entropy と，類似度スコア行列と d-vector のグラム行列（すなわち，d-vector から定義される話者間類似度）の差のフロベニウスノルムの重み付き和を最小化するように学習される．実験的評価では，提案アルゴリズムが従来の d-vector よりも主観的話者間類似度と強い相関を持つ話者埋め込みを学習できることを示す．

2 従来の DNN 話者埋め込み

2.1 One-hot 話者コード

話者コード [3] $\boldsymbol{c} = [c(1), \cdots, c(n), \cdots, c(N_\mathrm{s})]^\top$ は学習データに含まれる N_s 人の話者のうちの 1 人を特定するための離散的な話者表現である．i 番目の話者に対する話者コード $\boldsymbol{c}_i$ は，次式で定義される．

$$c_i(n) = \begin{cases} 1 & \text{if } n = i \\ 0 & \text{otherwise} \end{cases} \quad (1 \leq n \leq N_\mathrm{s}) \tag{1}$$

学習データに含まれる話者の声質は再現できるが，未知話者の話者コードは定義不可能であり，コードの次元が話者数に比例して大きくなる．さらに，主観的話者間類似度を完全に無視した表現であるため，ユーザが所望する声質を持つ話者の選択は困難である．

2.2 d-vector

d-vector [5] は学習済みの話者認識 DNN のボトルネック特徴量として得られる連続的な話者表現である．話者認識 DNN は，入力された音声特徴量からその話者を特定するように学習される．学習時の損失関数は，次式の softmax cross-entropy で与えられる．

$$L_\mathrm{SCE}\left(\boldsymbol{c}, \hat{\boldsymbol{c}}\right) = -\sum_{n=1}^{N_\mathrm{s}} c\left(n\right) \log \hat{c}\left(n\right) \tag{2}$$

ここで，$\hat{\boldsymbol{c}} = [\hat{c}(1), \cdots, \hat{c}(n), \cdots, \hat{c}(N_\mathrm{s})]^\top$ は DNN の予測結果である．N_d 次元の d-vector $\boldsymbol{d} = [d(1), \cdots, d(N_\mathrm{d})]^\top$ は，DNN のボトルネック層から抽出される．d-vector の次元数 N_d を話者数 N_s よりも小さくすることで，低次元特徴空間に連続的に分布する話者表現が得られる．i 番目の話者に対する話者表現 $\boldsymbol{d}_i$ は，当該話者の音声特徴量から抽出された全ての d-vector の平均によって定義される．連続的な特徴空間への話者埋め込みによって，未知話者の話者表現も定義可能 [6] だが，特徴空間での配置と主観的話者間類似度に相関がないため，ユーザにとって解釈しやすく，制御しやすい話者表現であるとは限らない．

3 主観的話者間類似度を考慮した DNN 話者埋め込み

本稿では，主観的話者間類似度と強い相関を持つ話者埋め込みを学習するための DNN の学習アルゴリズムを 2 つ提案する．

3.1 類似度スコア行列

提案アルゴリズムでは，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．$\mathbf{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_i, \cdots, \boldsymbol{s}_{N_\mathrm{s}}]$ 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 $-v$ から v 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 $s_{i,j}$ 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．

---

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

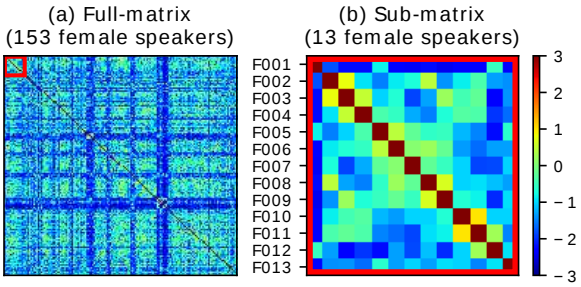

Fig. 1　(a) 大規模な主観スコアリングにより得られた 153 名の日本人女性話者の類似度スコア行列および (b) その部分行列

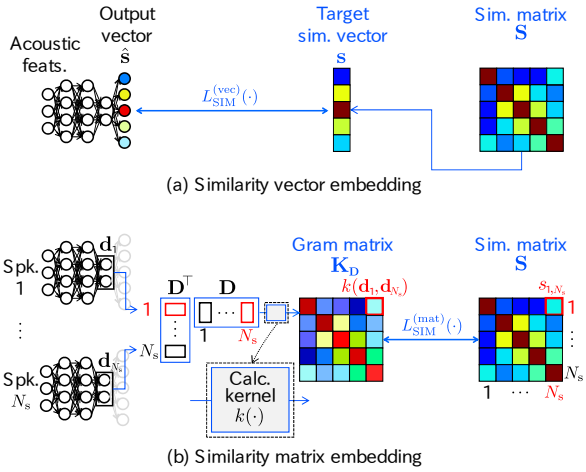

Fig. 2　(a) 類似度スコアベクトル埋め込みに基づく学習および (b) 類似度スコア行列埋め込みに基づく学習における損失関数の計算手順

3.2　類似度スコアベクトル埋め込みに基づく学習

類似度スコアベクトル埋め込みに基づく学習では，話者埋め込みの DNN は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．

$$L_{\mathrm{SIM}}^{(\mathrm{vec})}\left(\boldsymbol{s},\hat{\boldsymbol{s}}\right)=\frac{1}{N_{\mathrm{s}}}\left(\hat{\boldsymbol{s}}-\boldsymbol{s}\right)^{\top}\left(\hat{\boldsymbol{s}}-\boldsymbol{s}\right) \tag{3}$$

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．類似度スコア行列の対角成分は主観評価スコアの最大値を取るように定義されているため，予測スコアの最大値を取ることで，提案アルゴリズムは話者認識 DNN の学習にも適用できる．図 2(a) に提案アルゴリズムにおける損失関数 $L_{\mathrm{SIM}}^{(\mathrm{vec})}(\cdot)$ の計算手順を示す．

3.3　類似度スコア行列埋め込みに基づく学習

類似度スコア行列埋め込みに基づく学習では，行列 $\mathbf{S}$ によって d-vector 空間の配置に制約を与えて DNN を学習する．$\mathbf{D}=[\boldsymbol{d}_1,\cdots,\boldsymbol{d}_i,\cdots,\boldsymbol{d}_{N_{\mathrm{s}}}]$ を学習データに含まれる全話者の d-vector 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．

$$L\left(\boldsymbol{c},\hat{\boldsymbol{c}},\mathbf{D},\mathbf{S}\right)=L_{\mathrm{SCE}}\left(\boldsymbol{c},\hat{\boldsymbol{c}}\right)+\omega_{\mathrm{s}}L_{\mathrm{SIM}}^{(\mathrm{mat})}\left(\mathbf{D},\mathbf{S}\right) \tag{4}$$

$$L_{\mathrm{SIM}}^{(\mathrm{mat})}\left(\mathbf{D},\mathbf{S}\right)=\frac{2}{\|\mathbf{1}_{N_{\mathrm{s}}}-\mathbf{I}_{N_{\mathrm{s}}}\|_F^2}\left\|\widetilde{\mathbf{K}}_{\mathbf{D}}-\widetilde{\mathbf{S}}\right\|_F^2 \tag{5}$$

$$\widetilde{\mathbf{K}}_{\mathbf{D}}=\mathbf{K}_{\mathbf{D}}-\left(\mathbf{K}_{\mathbf{D}}\odot\mathbf{I}_{N_{\mathrm{s}}}\right) \tag{6}$$

$$\widetilde{\mathbf{S}}=\mathbf{S}-v\mathbf{I}_{N_{\mathrm{s}}} \tag{7}$$

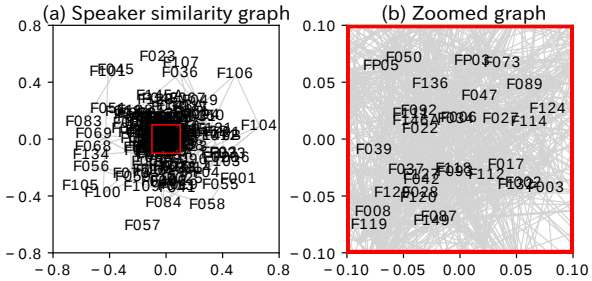

Fig. 3　(a) 図 1(a) に示す間類似度スコア行列に対応する話者グラフおよび (b) 赤枠で拡大された部分

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．ω_{s} は，式 (4) の第 2 項に対する重みを調整するハイパーパラメータである．$\mathbf{K}_{\mathbf{D}}$ は d-vector から計算される Gram 行列であり，次式で与えられる．

$$\mathbf{K}_{\mathbf{D}}=\begin{bmatrix}k\left(\boldsymbol{d}_1,\boldsymbol{d}_1\right) & \cdots & k\left(\boldsymbol{d}_1,\boldsymbol{d}_{N_{\mathrm{s}}}\right)\\ \vdots & \ddots & \vdots\\ k\left(\boldsymbol{d}_{N_{\mathrm{s}}},\boldsymbol{d}_1\right) & \cdots & k\left(\boldsymbol{d}_{N_{\mathrm{s}}}\boldsymbol{d}_{N_{\mathrm{s}}}\right)\end{bmatrix} \tag{8}$$

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，d-vector に由来する話者間類似度に対応する．故に，提案アルゴリズムは，主観的話者間類似度を考慮しつつ，話者を特定できるように話者埋め込みの DNN を学習させる．図 2(b) に提案アルゴリズムにおける損失関数 $L_{\mathrm{SIM}}^{(\mathrm{mat})}(\cdot)$ の計算手順を示す．

式 (5) の最小化は，(1) 主観的に類似した話者対は話者埋め込み空間内で近づき，かつ，(2) そうでない話者対は遠ざかるような制約を与えて DNN を学習させる．制御が容易な DNN 音声合成の実現には，少なくとも第一の条件は満たされることが望ましい．そのため，式 (5) を次式に示すように書き換える．

$$L_{\mathrm{SIM}}^{(\mathrm{mat-re})}\left(\mathbf{D},\mathbf{S}\right)=\frac{2}{\|\mathbf{W}-\mathbf{I}_{\mathrm{s}}\|_F^2}\left\|\mathbf{W}\odot\left(\widetilde{\mathbf{K}}_{\mathbf{D}}-\widetilde{\mathbf{S}}\right)\right\|_F^2 \tag{9}$$

ここで，$\mathbf{W}=[w_{i,j}]_{1\le i,j\le N_{\mathrm{s}}}$ は各成分に対応する類似度スコア行列の成分が 0 より大きいときに 1 となる行列である．$2/\|\mathbf{W}-\mathbf{I}_{\mathrm{s}}\|_F^2$ は行列 $\mathbf{W}$ の自由度に対応し，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．この定式化を用いた学習では，類似度スコアが 0 以下の話者対は学習時の制約として考慮されない．

3.4　考察

提案アルゴリズムで用いる類似度スコア行列から得られる話者間の関係性は，話者を頂点とし，類似話者間に辺を持つ無向グラフとして可視化できる．図 3(a) および (b) に図 1(a) の類似度スコア行列から定義される話者グラフとその一部を拡大したものを示す．話者グラフを定義する隣接行列は，式 (9) 中の $\mathbf{W}$ と同じである．図 3 における各話者の配置は，類似度スコア行列を用いた多次元尺度構成法により決定した．式 (9) の最小化は，音声特徴量から話者グラフを学習する枠組みと解釈できるため，DNN 音声合成における多人数話者モデリングにグラフ信号処理 [7] やグラフ埋め込み [8] の知見を取り入れることが可能となる．

これまでに，hidden Markov model (HMM) を用いたテキスト音声合成 [9] や Gaussian mixture model (GMM) を用いた声質変換 [10] における制御が容易な潜在変数のモデリング手法が提案されている．これらの手法では，ある特定の話者の声質に対する “warm – cold” や “clear – hoarse” といった印象語の対と，HMM や GMM の潜在変数が関連付けされるように学習される．提案アルゴリズムは，これらの手法を発展させたものであり，従来の話者ごとの印象ではなく，話者間の主観的類似度を考慮して話者埋め込み DNN を学習させる．さらに，話者の意図と受聴者の知覚の差（例えば，話者が表現した感情と，受聴者が知覚した感情の違い [11] など）もモデル化できる．

類似度スコア行列埋め込みに基づく提案アルゴリズムでは，d-vector の Gram 行列が類似度スコア行列に近づくように DNN を学習させる．埋め込み空間上での類似度を定義するカーネル関数には任意のものが選択でき，線形カーネルを用いた場合，式 (5) は deep clustering [12] と等価になる（ただし，対角成分を除外する点でわずかに異なる）．また，線形カーネルだけでなく，Gaussian カーネルなどのより複雑なカーネルも選択できる．

4　実験的評価

4.1　実験条件

4.1.1　大規模な主観スコアリングの条件

提案法で用いる類似度スコア行列 $\mathbf{S}$ を得るためのクラウドソーシングを利用した大規模な主観スコアリングを実施した．音声コーパスとして JNAS [13] を使用し，その中から 153 名の日本人女性話者の音声を用いた．コーパスに含まれる各話者は少なくとも 150 文の読み上げ発話を収録しており，本稿では，その中から選択した 5 発話をテキスト非依存な主観的話者間類似度の評価に用いた．各評価者は，全ての 11,628 話者対からランダムに選択された 34 話者対の主観的類似度を -3（完全に異なる）と $+3$（非常に似ている）の間の整数値で評価した．全ての話者対は，少なくとも 10 名の異なる評価者により評価された．最終的に，4,060 名の評価者がスコアリングに参加し，138,040 件の回答が得られた．

4.1.2　DNN 話者埋め込みの条件

本稿では，JNAS コーパスを DNN 話者埋め込みの学習にも用いた．各話者の全発話の 9 割を学習に，残りを評価に用いた．大規模な主観スコアリングに用いた 5 発話は，学習・評価データの両方から除外した．DNN 学習に用いる話者数 N_{s} は，35, 70, 140 とした．図 1(b) に示す 13 名の話者は学習データには含まれなかったが，話者埋め込みと類似度スコアの相関の計算に用いた．類似度スコア行列の各成分は $[-3, +3]$ の範囲に収まるように定義したが，DNN 学習時には各成分が $[-1, +1]$ の範囲に収まるように正規化した．したがって，3.3 節で述べた類似度スコア行列埋め込みに基づく提案アルゴリズムでは，カーネル関数に $k(\boldsymbol{d}_i, \boldsymbol{d}_j) = \tanh(\boldsymbol{d}_i^{\top} \boldsymbol{d}_j)$ として定義されるシグモイドカーネルを用いた．各話者の d-vector 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 d-vector の平均として推定した．ここでの有声・無声判定には，STRAIGHT 分析 [14] によって得られた F_0 を用いた．

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，d-vector 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．2.2 節および 3.3 節に示す学習アルゴリズムを用いる場合，出力層の活性化関数は softmax 関数とした．類似度スコア行列 $\mathbf{S}$ の各成分は $[-1, +1]$ の範囲に収まるように正規化したため，3.2 節に示す学習アルゴリズムを用いる場合，出力層の活性化関数は tanh 関数とした．d-vector を話者表現として用いた先行研究 [6] と同様に，$N_{\mathrm{s}} = \{35, 70, 140\}$ 名の話者に加え，無声区間を表す 1 次元スカラー値を話者コードに追加した．DNN の入力は，当該フレームと前後 2 フレームにおける 1 次から 39 次のメルケプストラム係数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．メルケプストラム係数の抽出には，STRAIGHT 分析 [14] を用いた．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 [15] を用いた．2.2 節，3.2 節，そして 3.3 節で述べた全ての学習アルゴリズムの反復回数は 100 とした．学習時のミニバッチサイズは 2,048 とし，式 (4) に示すハイパーパラメータ ω_{s} は 10.0 とした．

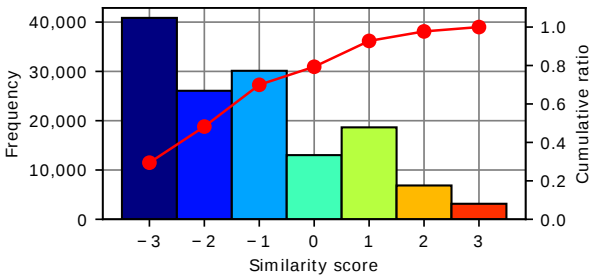

Fig. 4　主観的話者間類似度スコアのヒストグラム（赤線は累積度数）

4.2　大規模な主観スコアリング

図 1(a) および (b) にそれぞれ 153 名の女性話者の類似度スコア行列とその部分行列を示す．この図より，多くの話者対の類似度スコアは 0 以下の値となっている中で，いくつかの話者対（例えば，図 1(b) の “F010” と “F011”）の類似度スコアは確かに 0 より大きな値となっていることが確認できる．図 4 に得られた類似度スコアのヒストグラムを示す．

4.3　客観評価

従来のアルゴリズムと提案アルゴリズムを比較するために，話者認識の精度と，提案アルゴリズムにおける損失関数 $L_{\mathrm{SIM}}^{(\mathrm{vec})}(\cdot)$, $L_{\mathrm{SIM}}^{(\mathrm{mat})}(\cdot)$, そして $L_{\mathrm{SIM}}^{(\mathrm{mat-re})}(\cdot)$ の値を計算した．本稿では，以下の 4 つのアルゴリズムを比較した．

Conv. ：式 (2) の最小化

Prop. (vec) ：式 (3) の最小化

Prop. (mat) ：式 (4) の最小化

Prop. (mat-re) ：式 (2) + ω_{s}×(9) の最小化

客観評価結果を図 5 に示す．ここで，“Prop. (vec)” において話者認識の精度を計算する際には，DNN の出力の最大値を取る要素に対応する話者を認識結果とした．評価結果より，softmax cross-entropy 最小化に基づく従来アルゴリズムが最も高い認識精度を達成しているが，主観的話者間類似度に関する損失はいずれも減少していないことが確認できる．一方で，提案アルゴリズムでは，話者認識精度は低下しているものの，学習に用いた損失関数は確かに減少させ

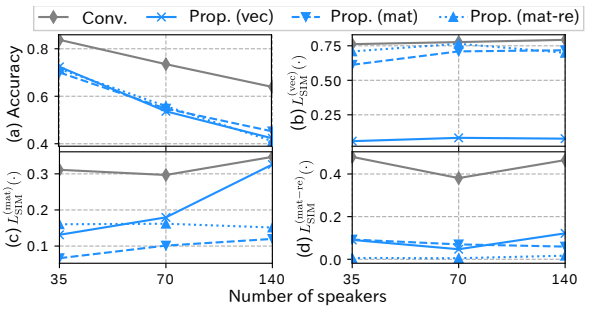

Fig. 5 客観評価結果: (a) 話者認識の精度, (b)$L_{\mathrm{SIM}}^{(\mathrm{vec})}(\cdot)$ の値, (c) $L_{\mathrm{SIM}}^{(\mathrm{mat})}(\cdot)$ の値, そして (d) $L_{\mathrm{SIM}}^{(\mathrm{mat-re})}(\cdot)$ の値

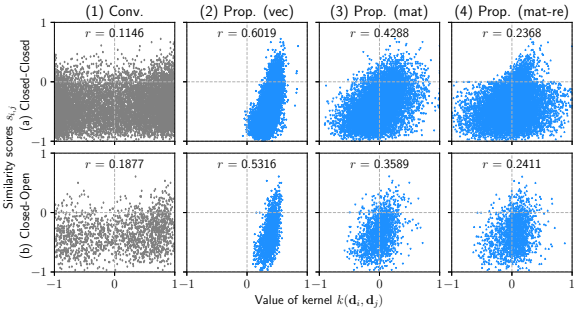

Fig. 6 話者間類似度スコア $s_{i,j}$ とカーネル関数 $k(\boldsymbol{d}_i, \boldsymbol{d}_j)$ の値の散布図と, これらの相関係数 r の値. この図に含まれる全ての散布図は, 全ての話者対から作成された.

るように話者埋め込みの DNN を学習させていることが確認できる. 故に, 提案アルゴリズムは主観的話者間類似度を考慮して DNN 話者埋め込みを学習可能であることが示された.

話者埋め込みと主観的話者間類似度の相関を調査するために, 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 6 に相関係数 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. 図 6(1a) および (1b) より, 話者認識の softmax cross-entropy 最小化のみで学習された DNN から抽出された従来の d-vector は類似度スコアと弱い相関を持つことが確認できる. 一方で, 図 6(2a) および (3a) に示す結果から, 提案アルゴリズムである “Prop. (vec)” and “Prop. (mat)” は, 類似度スコアと強い相関を持つ DNN 話者埋め込みを学習していることが確認できる. さらに, 図 6(2b) および (3b) に示す結果は, 提案アルゴリズムが学習データに含まれる話者とそうでない話者間の類似度スコアと強い相関を持つ DNN 話者埋め込みも学習可能であることを示唆している. これらの結果より, 提案アルゴリズムは従来の d-vector よりも主観的話者間類似度と強い相関を持つ DNN 話者埋め込みを学習することが示された. また, 図 7 に示す, 類似度スコアが 0 より大きい話者対のみに着目した結果より, “Prop. (mat-re)” は 4 つのアルゴリズムの中で学習データに含まれる類似話者対の類似度スコアと最も強い相関を持つ話者埋め込みを学習していることが確認でき, このアルゴリズムが主観的に類似した話者対を考慮した DNN 話者埋め込みに最適であることを示唆している.

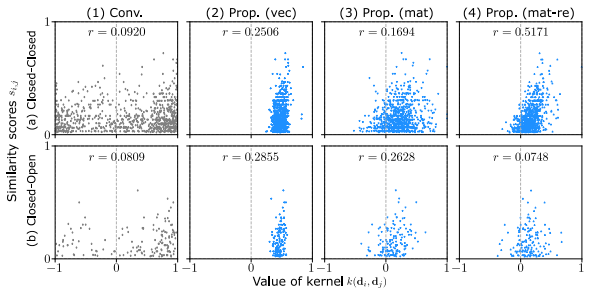

Fig. 7 話者間類似度スコア $s_{i,j}$ とカーネル関数 $k(\boldsymbol{d}_i, \boldsymbol{d}_j)$ の値の散布図と, これらの相関係数 r の値. この図に含まれる全ての散布図は, 類似度スコアが 0 より大きい話者対から作成された.

5 結論

本稿では, DNN 音声合成に適した話者表現を学習するための手法として, 主観的話者間類似度を考慮した話者埋め込み DNN の学習アルゴリズムを 2 つ提案した. 実験的評価より, 提案アルゴリズムは従来の DNN 話者埋め込みである d-vector 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学習可能であることを示した. 今後は, 提案アルゴリズムにおけるハイパーパラメータの影響を調査し, DNN 音声合成における多人数話者モデリングへ提案アルゴリズムを適用する.

謝辞: 本研究の一部は, 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] H. Zen *et al.*, *Proc. ICASSP,* pp. 3872–3876, Vancouver, Canada, May 2013.
[2] J. Shen *et al.*, *Proc. ICASSP,* pp. 4779-4783, Calgary, Canada, Apr. 2018.
[3] N. Hojo *et al.*, *IEICE Transactions on Information Systems,* vol. E101-D, no. 2, pp. 462-472, Feb. 2018.
[4] H.-T. Luong *et al.*, *Proc. ICASSP,* pp. 1905–1909, Calgary, Canada, Apr. 2018.
[5] E. Variani *et al.*, *Proc. ICASSP,* pp. 4080-4084, Florence, Italy, May 2014,
[6] Y. Saito *et al.*, *Proc. ICASSP,* pp. 5274–5278, Calgary, Canada, Apr. 2018.
[7] D. I. Shuman *et al.*, *IEEE Signal Processing Magazine,* vol. 30, no. 3, pp. 83-98, May. 2013.
[8] P. Goyal *et al.*, *Knowledge-Based Systems,* vol. 151, pp. 78-94, Jul. 2018.
[9] M. Tachibana *et al.*, *Proc. ICSLP,* pp. 2438–2411, Pittsburgh, U.S.A., Sep. 2006.
[10] K. Ohta *et al.*, *Proc. ICSLP,* pp. 101–106, Bonn, Germany, Aug. 2007.
[11] J. Lorenzo-Trueba *et al.*, *Speech Communication,* vol. 99, pp. 135-143, May 2018.
[12] J. R. Hershey *et al.*, *Proc. ICASSP,* pp. 31-35, Shanghai, Chaina, Mar. 2016.
[13] K. Itou *et al.*, *Journal of the ASJ (E),* vol. 20, no. 3, pp. 199-206, May 1999.
[14] H. Kawahara *et al.*, *Speech Communication,* vol. 27, no. 3–4, pp. 187-207, Apr. 1999.
[15] J. Duchi *et al.*, *Journal of Machine Learning Research,* vol. 12, 2121-2159, Jul. 2011.