主観的話者間類似度を考慮したDNN話者埋め込みのための Active Learning

齋藤 佑樹[1,a)]　高道 慎之介[1,b)]　猿渡 洋[1,c)]

概要：本稿では，主観的話者間類似度のスコアリングと，Deep Neural Network (DNN) を用いた話者埋め込みの学習を反復する active learning を提案する．我々はこれまでに，話者間類似度の主観スコアリング結果に基づく DNN 話者埋め込みの学習法を提案し，多話者音声生成における合成音声の品質と制御性の改善効果を確認している．しかしながら，この学習法は，主観スコアリングの作業コストと，話者埋め込み学習時の計算コストを要する．提案法では，DNN 話者埋め込みの学習に用いる話者間の主観的な類似度スコアが一部だけ観測されていると仮定し，(1) 観測されているスコアを用いた DNN 話者埋め込み学習と (2) 話者埋め込み由来の類似度に基づく優先度付きのスコアリングを反復する．実験的評価の結果より，提案法がコストを削減しつつ多話者音声生成モデリングに適した DNN 話者埋め込みを学習することを示す．

Yuki Saito[1,a)]　Shinnosuke Takamichi[1,b)]　Hiroshi Saruwatari[1,c)]

1. はじめに

話者埋め込みとは，音声からその話者を特定するための特徴量であり，話者性に関する分散表現である．特に，Deep Neural Network (DNN) に基づく speaker encoder により学習される DNN 話者埋め込み [1] は，話者認識 [2] や話者ダイアライゼーション [3] などの識別的タスクにおいて精度の改善に大きく貢献している．また，代表的な DNN 話者埋め込みである d-vector [2] は，テキスト音声合成 [4] や音声変換 [5] などの生成的タスクにおいて合成音声の話者性を制御するための話者表現としても利用されている．しかしながら，話者識別に基づく DNN 話者埋め込みは，話者間の主観的な類似度を考慮していないため，人間にとって解釈しにくい話者表現であり，話者性制御の困難性や，音声合成モデルの話者適応時の品質劣化 [4], [6] といった問題が生じる．

この問題に対し我々は，話者間の主観的な類似度を考慮し，人間にとって解釈しやすい DNN 話者埋め込みを学習する手法を提案している [7], [8]．この手法は，図 1 に示すように，話者間類似度の主観スコアリングと，類似度スコアを用いた DNN 話者埋め込みの学習から構成される．

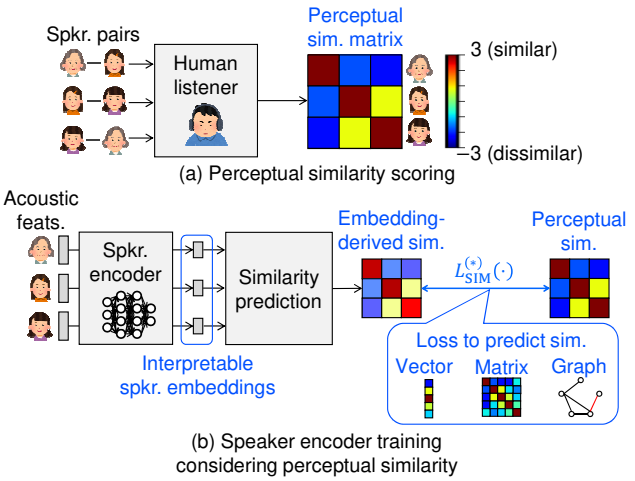

図 1　主観的話者間類似度を考慮した DNN 話者埋め込み

主観スコアリングでは，話者間の主観的な類似度を表す類似度スコア行列を定義し，話者埋め込みの学習では，話者埋め込みを用いて主観的な話者間の類似度を予測するように speaker encoder を学習する．この学習法として，類似度スコア行列のベクトル，行列全体，そして行列から導出されるグラフを用いる手法を提案している．この手法は，従来の d-vector よりも主観的話者間類似度と強い相関を持ち [7]，Variational AutoEncoder (VAE) [9] に基づく多対多音声変換 [5] における話者適応 [7] や，Tacotron2 [10] に基づく End-to-End クロスリンガル音声合成 [11] の品質改

[1] 東京大学, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[a)] yuuki_saito@ipc.i.u-tokyo.ac.jp
[b)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[c)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

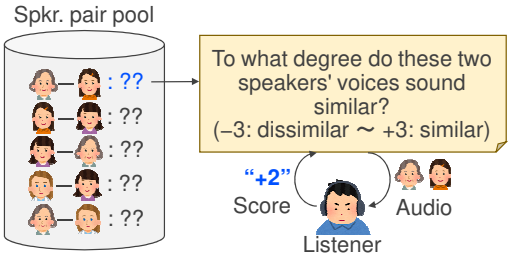

図 2 話者間類似度の主観スコアリング．各評価者は，提示された話者対の音声の主観的な類似度を $-v$ から v の間の整数で評価する．ここでは，$v = 3$ とした．

善に有効な DNN 話者埋め込みを学習できる．一方で，この手法では，主観スコアリングの作業コストや，DNN 話者埋め込み学習の計算コストを要するため，学習話者数に対するスケーラビリティの観点で課題が残っている．

本稿では，主観的話者間類似度を考慮した DNN 話者埋め込みをより効率的に学習するための active learning を提案する．提案法では，話者間の主観的な類似度スコアが一部だけ観測されていると仮定し，(1) 観測されているスコアを用いた DNN 話者埋め込みの学習と (2) 話者埋め込み由来の類似度に基づく優先度付きの主観スコアリングを反復することで，作業コストと計算コストの両方を削減しつつよりよい話者埋め込みを学習する．また，本稿では，主観スコアリングにおける優先度を決定するクエリ戦略の影響も実験的に調査する．実験的評価の結果から，提案法が主観スコアリングの作業コストと DNN 話者埋め込み学習時の計算コストの両方を削減しつつ，多話者音声生成モデリングに適した話者表現を学習することを示す．

2. 主観的話者間類似度を考慮した DNN 話者埋め込み [7], [8]

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 [7], [8] では，音声の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 speaker encoder を学習する．N_s を主観スコアリングに用いる話者数，$\mathbf{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_i, \cdots, \boldsymbol{s}_{N_\mathrm{s}}]$ 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，図 2 に示すように，“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 3(a) と (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．

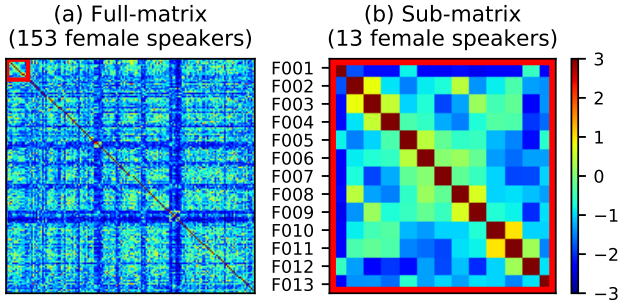

図 3 (a) 153 名の女性話者間の類似度スコア行列と (b) その部分行列

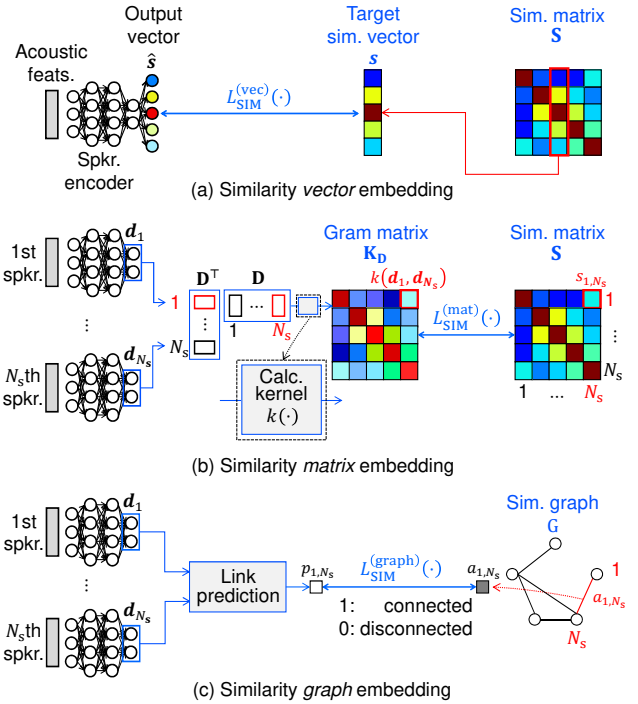

図 4 (a) 類似度スコアベクトル埋め込み，(b) 類似度スコア行列埋め込み，(c) 類似度グラフ埋め込みに基づく DNN 話者埋め込みの学習

2.2 主観的話者間類似度に基づく学習法

我々はこれまでに，類似度スコア行列の表現形式として，(1) スコア行列のベクトル，(2) スコア行列全体，そして (3) スコア行列から導出されるグラフの構造を用いる DNN 話者埋め込みの学習法を提案している [7], [8]．

2.2.1 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，音声特徴量を入力とし，当該話者の類似度スコアベクトルを予測するように speaker encoder を学習する．学習時の損失関数は，次式で与えられる．

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

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 4(a) に損失関数 $L_\mathrm{SIM}^{(\mathrm{vec})}(\cdot)$ の計算手順を示す．

2.2.2 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 speaker encoder を学習する．$\boldsymbol{d}_i = [d_i(1), \cdots, d_i(N_\mathrm{d})]^\top$ 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．

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

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

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

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次式で計算される Gram 行列である．

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

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．図 4(b) に損失関数 $L_{\mathrm{SIM}}^{(\mathrm{mat})}(\cdot)$ の計算手順を示す．

2.2.3 類似度グラフ埋め込み

類似度グラフ埋め込みでは，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 5）を構築し，話者埋め込みの対から類似度グラフの辺の有無を予測するように speaker encoder を学習する．学習時の損失関数は，次式で与えられる．

$$L_{\mathrm{SIM}}^{(\mathrm{graph})}(\mathbf{D}, \mathbf{A}) = -\sum_{i,j=1, i\neq j}^{N_\mathrm{s}} a_{i,j} \log p_{i,j} - \sum_{i,j=1, i\neq j}^{N_\mathrm{s}} (1 - a_{i,j}) \log(1 - p_{i,j}), \tag{6}$$

ここで，$a_{i,j}$ 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．本稿では，類似度スコア行列の値を [0,1] の範囲に正規化することで類似度グラフの隣接行列を定義する．$p_{i,j}$ 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 [12] を参考に $p_{i,j} = \exp(-||\boldsymbol{d}_i - \boldsymbol{d}_j||_2^2)$ とする．図 4(c) に損失関数 $L_{\mathrm{SIM}}^{(\mathrm{graph})}(\cdot)$ の計算手順を示す．

3. 提案する active learning

前述した DNN 話者埋め込み学習法における主観スコアリングの作業数は，学習話者数 N_s の 2 乗に比例する．本稿では，この作業コストを削減しつつ，効率的に DNN 話者埋め込みを学習するための active learning を提案する．

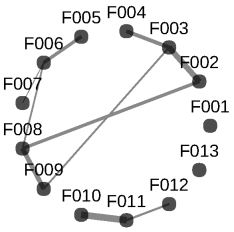

図 5 図 3(b) に示す 13 名の類似度スコア行列から導出される類似度グラフ．各節点が各話者を表し，主観的に類似した話者対（即ち，$s_{i,j} > 0$）の間に辺が張られている．ここでは，より主観的に類似した話者対により太い辺が張られている．

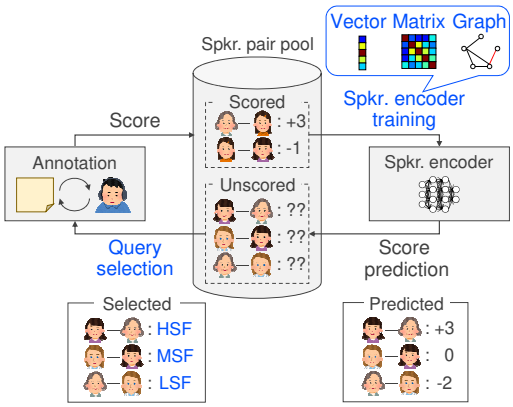

図 6 主観的話者間類似度を考慮した DNN 話者埋め込みのための active learning

Active learning [13] は少数のラベル付きデータと多数のラベルなしデータを用いて機械学習モデルを逐次的に学習するための枠組みであり，(1) ラベル付きデータを用いたモデル学習と (2) 学習後のモデルを用いて次にラベル付けすべきデータを決定するクエリ選択を交互に反復する．

図 6 に主観的話者間類似度を考慮した DNN 話者埋め込みのための active learning の概念図を示す．ここでは，${}_{N_\mathrm{s}}\mathrm{C}_2$ の話者対をスコア付けされた群 $\mathcal{D}_\mathrm{s}$ とそうでない群 $\mathcal{D}_\mathrm{u}$ の 2 群に分割し，$\mathcal{D}_\mathrm{u}$ に含まれる話者対の類似度スコアは学習開始時には観測されていないと仮定する．

3.1 スコア付けされた話者対を用いた話者埋め込み学習

話者埋め込み学習では，スコア付けされた話者対 $\mathcal{D}_\mathrm{s}$ のデータを用いて speaker encoder を学習する．この学習では，2.2 節で述べたいずれかの損失関数を最小化する．ここで，各 active learning の反復で，speaker encoder のモデルパラメータはリセットされずに逐次的に更新される．

3.2 スコア付けされていない話者対からのクエリ選択

クエリ選択では，まず，逐次的に学習された speaker encoder を用いて，次にスコア付けすべき話者対を選択する際の基準となるクエリ（即ち，$\mathcal{D}_\mathrm{u}$ に含まれる話者対に対する仮の類似度スコア）を生成する．次に，oracle（例えば，人間のアノテータ）がより高い優先度の話者対に対してスコア付けを行う．この枠組みでは，スコア付けすべき話者

対の優先度を決めるクエリ戦略が重要な役割を持つ．本稿では，クエリ戦略として (1) 予測された類似度が最も小さい（即ち，$-v$ に近い）話者対を優先する Lower-Similarity First (LSF)，(2) LSF の逆に対応する Higher-Similarity First (HSF)，そして (3) 予測された類似度が最も 0 に近い話者対を優先する Middle-Similarity First (MSF) の 3 つを比較する．

3.3 考察

提案する active learning は，Human-In-The-Loop (HITL) [14] の DNN 話者埋め込み学習として解釈できる．この観点から，人間の知覚評価に基づく音声合成の学習（例えば，DNN ベースの識別器を人間で置き換えた敵対的生成ネットワーク [15]）の実現も期待できる．

4. 実験的評価

4.1 実験条件

本稿では，我々の従来法 [7], [8] と同様に，JNAS コーパス [16] の 153 名の日本人女性話者間の類似度スコア行列 **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 [17] 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 [18] を用いた．DNN 学習時には，図 3(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．

4.1.1 DNN 話者埋め込み学習の条件

本稿では，2.2 節で述べた 3 つの学習法（類似度 { スコアベクトル，スコア行列，グラフ } 埋め込み）を比較した．Speaker encoder の DNN 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．Speaker encoder 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．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．学習時の最適化には，学習率を 0.01 とした AdaGrad [19] を用いた．

類似度スコアベクトル埋め込みに基づく学習 (“Sim. (vec)”) では，類似度スコア行列 **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．類似度スコア行列埋め込みに基づく学習 (“Sim. (mat)”) では，カーネル関数に sigmoid カーネルを利用し，同様のスコア正規化を行った．類似度グラフ埋め込みに基づく学習 (“Sim. (graph)”) では，スコアを $[0, 1]$ 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．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．

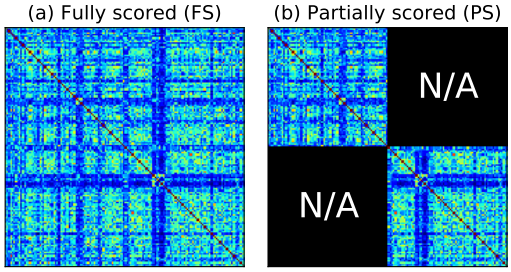

図 7 Active learning における (a) Fully Scored (FS) と (b) Partially Scored (PS) の設定

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 (Phonetic PosteriorGram: PPG) [20] と話者埋め込みで条件付けた VAE [5] を用いた．PPG 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，speaker encoder の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 PPG を推定するように学習した．学習のエポック数は 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 PPG 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，PPG，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．学習のエポック数は 25 とした．合成音声波形の生成には，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．

4.1.3 Active learning の条件

Active learning では，図 7(b) に示すように，140 名の学習話者を 2 群（前半 70 名と後半 70 名）に分割し，異なる群の間で話者間類似度スコアは観測されていないという設定から学習を開始した．本稿では，(1) スコア付けされた話者対を用いた 1 エポックの DNN 話者埋め込み学習と (2) 学習された話者埋め込みを用いたクエリ選択を反復した．Active learning の反復回数は 115 とし，各反復毎のクエリ数は実験的に 43 と設定した．

本稿では，提案する active learning の有効性を，2.2 節で述べた各学習法 (“Sim. (*)”) 毎に独立に評価した．ここで，3.2 で述べた 3 つのクエリ戦略 (“LSF”, “HSF,” “MSF”) に加え，active learning を用いずに，図 7(a) と (b) に示すス

コア付けでそれぞれ speaker encoder を 115 エポック学習させた Fully Scored (“FS”) と Partially Scored (“PS”) も比較した．

4.2 客観評価

客観評価として，話者埋め込みの対を用いて当該話者対が主観的に類似しているかどうかを判定する 2 値分類モデルの Area Under the ROC Curve (AUC) [21] を計算した．AUC は 0.5 から 1.0 の値を取り，値が 1 に近ければ近いほどより精度が高い 2 値分類モデルであることを意味する．本稿では，提案する active learning の反復によりこの AUC がどのように変化するかを評価した．

図 8 に評価結果を示す．この図中の赤線と青線はそれぞれ “FS” と “PS” の状態から active learning を用いずに speaker encoder を 115 エポック学習させた後の最終的な AUC を意味する．ここで，“LSF,” “HSF,” そして “MSF” のクエリ戦略を用いて active learning を行った最終結果の AUC は，“FS” と必ずしも一致しない．これは，最終的に観測される類似度スコアはすべての手法で一致するが，active learning により speaker encoder は異なる順番で話者間の主観的な類似度を学習したためである．評価結果から，まず，active learning におけるクエリ戦略は，AUC の改善に大きく影響することが確認できる．特に，“MSF” を用いた active learning は，学習法の違いに依らず，AUC の改善に有効であることが確認できる．また，“Sim. (vec)” と “Sim. (graph)” における active learning は，“PS” よりも高い AUC を少ない反復回数で達成しており，この傾向は学習話者同士の対 (“Seen-Seen”) と学習話者-未知話者の対 (“Seen-Unseen”) の両方で共通していることがわかる．一方で，“Sim. (mat)” は “Seen-Seen” のケースで AUC を改善しているが，“Seen-Unseen” のケースで劣化していることも確認でき，この学習法が用いるデータに強く依存する傾向にあることを示している．

4.3 主観評価

Active learning 後の speaker encoder から得られる DNN 話者埋め込みを用いた VAE ベース多話者音声生成モデルの主観評価を実施した．本稿では，未知話者 13 名 (“F001”–“F013”) の音声を用いた話者適応の合成音声の自然性と話者類似性を，5 段階の Mean Opinion Score (MOS) テストと Degradation MOS (DMOS) テストによりそれぞれ評価した．この評価では，“FS” と “PS” の学習後の DNN 話者埋め込みと，クエリ戦略として “MSF” を用いて，異なる反復回数 (30, 60, 90) で終了させた active learning 後の DNN 話者埋め込みを比較した．この active learning の反復回数は，それぞれ全体の類似度スコアの 62.5%, 75%, そして 87.5%を観測したことに対応する．評価者はクラウドソーシングにより集められた 50 名であり，650（50 発話/話者 × 13 話者）個の音声サンプルの中からランダムに抽出された 20 サンプルの品質を評価した．合計の評価セット数は (MOS or DMOS) × 50（評価者数）= 100 であった．

表 1 合成音声の自然性に関する MOS 値．表の 2 列目は，スコア付された話者対の割合を意味する． **太字**の MOS 値は $p > 0.05$ で “FS” と有意差がないことを意味する

		Sim. (vec)	Sim. (mat)	Sim. (graph)
PS	50.0%	2.91±0.14	**2.98±0.13**	2.94±0.14
MSF	62.5%	2.99±0.12	**2.97±0.13**	**3.11±0.13**
	75.0%	**3.11±0.13**	**3.02±0.13**	**3.12±0.14**
	87.5%	**3.18±0.13**	**3.04±0.13**	**3.13±0.14**
FS	100%	3.19±0.13	3.02±0.12	3.18±0.13

表 2 合成音声の話者類似性に関する DMOS 値．表の 2 列目は，スコア付された話者対の割合を意味する． **太字**の DMOS 値は $p > 0.05$ で “FS” と有意差がないことを意味する

		Sim. (vec)	Sim. (mat)	Sim. (graph)
PS	50.0%	2.85±0.14	**2.90±0.13**	2.86±0.13
MSF	62.5%	**2.95±0.14**	**2.93±0.13**	**3.03±0.13**
	75.0%	**3.04±0.14**	**3.00±0.13**	**3.02±0.13**
	87.5%	**3.05±0.14**	**3.03±0.13**	**3.06±0.13**
FS	100%	3.14±0.14	2.98±0.13	3.08±0.14

表 1 と表 2 にそれぞれ MOS テストと DMOS テストの結果を示す．評価結果より，“MSF” は “FS” と同程度の合成音声品質を，より少ない観測スコアで達成していることが確認できる．即ち，提案する active learning は，主観スコアリングの作業コストと DNN 話者埋め込み学習時の計算コストを削減しつつ，多話者音声生成モデリングに適した話者表現を学習できる可能性を示唆した．一方で，“Sim. (mat)” の評価結果に着目すると，“PS,” “FS,” そして “MSF” の間に有意差は確認できない．この原因として，図 8(b)(2) に示す過適合傾向が考えられる．

5. おわりに

本稿では，主観的話者間類似度を考慮した DNN 話者埋め込みをより効率的に学習するための active learning を提案し，実験的評価によりその有効性を示した．今後は，active learning におけるクエリ数などのハイパーパラメータの影響を調査する．

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] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, “Deep representation learning in speech processing: Challenges, recent advances, and future trends,” *arXiv*, vol. abs/2001.00378, 2020.

[2] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 4080—-4084.

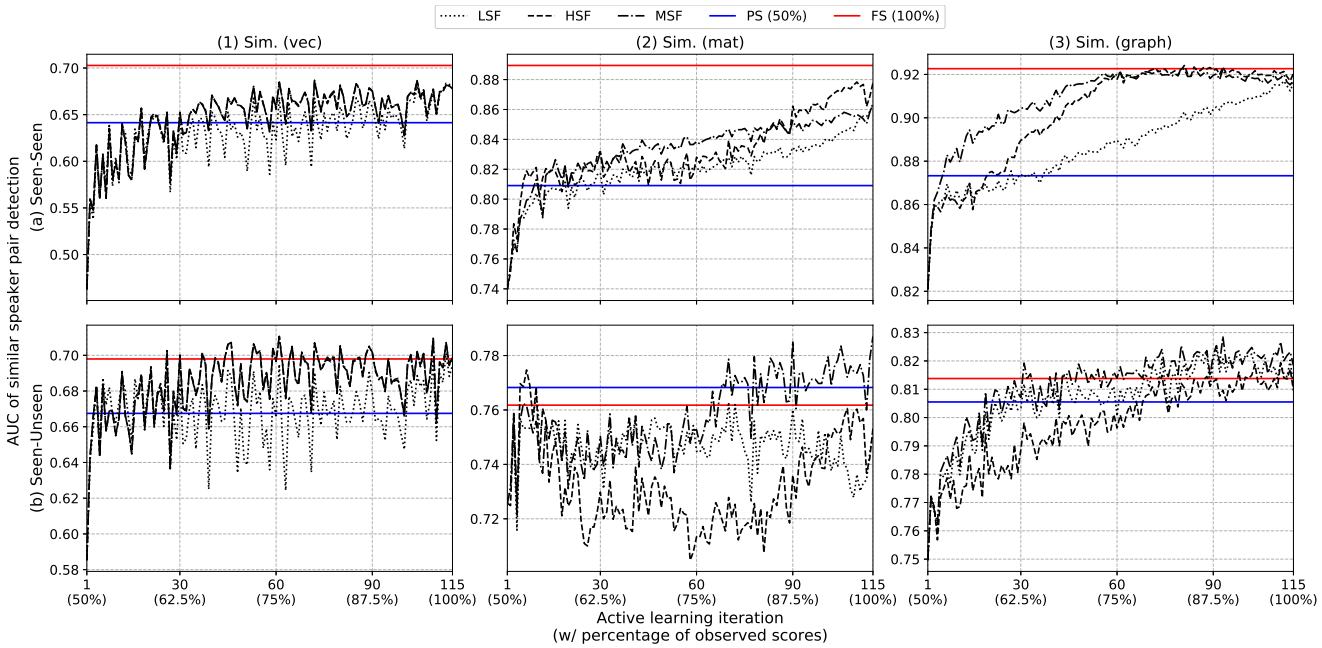

図 8 Active learning の反復に対する類似話者対検出モデルの AUC の変化

[3] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. ICASSP*, Alberta, Canada, Apr. 2018, pp. 5239–5243.

[4] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multi-speaker text-to-speech synthesis," in *Proc. NeurIPS*, Montreal, Canada, Dec. 2018, pp. 4480–4490.

[5] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *Proc. ICASSP*, Alberta, Canada, Apr. 2018, pp. 5274–5278.

[6] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 6184–6188.

[7] Y. Saito, S. Takamichi, and H. Saruwatari, "DNN-based speaker embedding using subjective inter-speaker similarity for multi-speaker modeling in speech synthesis," in *Proc. SSW*, Vienna, Austria, Sep. 2019, pp. 51–56.

[8] 齋藤佑樹, 高道慎之介, and 猿渡洋, "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," in **音講論（秋）**, 9 月 2020, pp. 697–698.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv*, vol. abs/1312.6114, 2013.

[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.

[11] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Cross-lingual text-to-speech synthesis via domain adaptation and perceptual similarity regression in speaker space," in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 2947–2951.

[12] J. Li, Y. Baba, and H. Kashima, "Simultaneous clustering and ranking from pairwise comparisons," in *Proc. IJCAI*, Stockholm, Sweden, Jul. 2018, pp. 1554–1560.

[13] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, Jan. 2009.

[14] F. M. Zanzotto, "Viewpoint: Human-in-the-loop artificial intelligence," *Journal of Artificial Intelligence Research*, vol. 64, pp. 243–252, Feb. 2019.

[15] K. Fujii, Y. Saito, S. Takamichi, Y. Baba, and H. Saruwatari, "HumanGAN: Generative adversarial network with human-based discriminator and its evaluation in speech perception modeling," in *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 6239–6243.

[16] K. Itou, M. Yamamoto, K.Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, May 1999.

[17] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, Apr. 1999.

[18] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.

[19] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, Jul. 2011.

[20] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, Seattle, U.S.A., Jul. 2016.

[21] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.