社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

話者照合のための話者性の変動量を考慮した声道長摂動による疑似話者生成

大野　史夏† 　若松　智花† 　塩田さやか†

† 東京都立大学

あらまし　性能の高い話者埋め込みに基づく話者照合システムを実現するためには，話者埋め込み抽出モデルを大規模な学習データを用いて学習する必要がある．十分な量の学習データを用意するための手法の一つとしてデータ拡張がある．従来の話者照合におけるデータ拡張では主に発話数の拡張が行われていたが，話者埋め込みに基づく話者照合システムでは，十分な話者数を用意することも精度向上のために重要であることが知られている．話者数の拡張においてはこれまでに，データ拡張として声道長摂動を用いた疑似話者生成を行うことの有効性も報告されている．一方で，疑似話者生成時の適切なパラメータの設定方法やデータ拡張に有効な疑似話者数が限られているなどの問題があった．そこで本研究では話者性の変動量を考慮して声道長摂動のパラメータを自動調整することで，従来の疑似話者選択法に比べてより多くの疑似話者を話者数の拡張に用いる手法を提案する．実験では，提案手法を用いた疑似話者生成を行い，最先端手法である ECAPA-TDNN に基づく話者照合システムを用いた評価を行った．実験結果から，話者埋め込みに基づく話者照合のためのデータ拡張において，1 話者あたりの発話数拡張と話者性の変動量を考慮した話者数拡張を併用することで話者照合システムの性能が最も向上したことを報告する．

キーワード　話者照合，疑似話者拡張，声道長摂動，ECAPA-TDNN

Pseudo-speaker augmentation based on vocal tract length perturbation considering speaker variability for speaker verification

Fumika ONO†, Tomoka WAKAMATSU†, and Sayaka SHIOTA†

† Tokyo Metropolitan University

Abstract　In order to construct a reliable speaker verification system based on speaker embeddings, it is necessary to train the speaker embedding extraction model using large-scale training data. Data augmentation is one method for preparing a sufficient amount of training data. Data augmentation in conventional speaker verification mainly involves increasing the number of utterances. However, in the speaker verification systems based on speaker embedding, providing sufficient speakers is also important for improving accuracy. As a method for increasing the number of speakers, the effectiveness of generating pseudo-speakers using vocal tract length perturbations (VTLP) has also been reported. On the other hand, there were problems, such as how to set appropriate parameters when generating pseudo-speakers and the number of effective pseudo-speakers. Therefore, in this paper, by automatically adjusting the parameters of VTLP in consideration of the amount of variation in speaker characteristics, we can increase the number of speakers by selecting more pseudo-speakers than in the conventional pseudo-speaker selection method. In the experiment, we generated pseudo-speakers using the proposed method and evaluated it using a speaker verification system based on ECAPA-TDNN, a state-of-the-art method. From the experimental results, we found that in data augmentation for speaker verification based on speaker embedding, we can improve the speaker verification system by using both augmentations of the number of utterances per speaker and the number of speakers that take into account the variation in speaker characteristics.

Key words　speaker verification，pseudo-speaker augmentation，VTLP，ECAPA-TDNN

1. はじめに

近年，様々な制度のデジタル化が進むにつれて人々のセキュリティに対する関心は高まりつつあり，生体認証技術に関する研究はその重要性を増している [1], [2]．生体認証に使用される生体情報は指紋，虹彩，静脈など多岐に渡っているが，その中で音声を用いる生体認証技術を話者照合という．話者照合はハードウェアとしてマイクがあればデータの入手が可能であり導入コストが低いことや，オンラインでの生体認証の需要が高まっていることから，生体認証技術としての話者照合の実用的な価値が期待されている．

話者照合システムの最先端技術とされているのが，x-vector [3] や ECAPA-TDNN （Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network） [4] に代表される，深層学習 （Deep Neural Network; DNN）を用いた話者埋め込みに基づく手法 [5]〜[8] である．この手法は DNN の中間層で得られる出力を話者性を表現するベクトルとして利用している．より表現力の高い話者埋め込みを抽出するためには DNN を大量の学習データを用いて学習する必要があることが知られている．また，既存の学習データだけでなくさらに学習データを増強する目的でデータ拡張という手法が用いられており，話者照合においてもその有効性は確認されている [9]〜[14]．

従来の話者照合においてデータ拡張として広く使用されている手法は，ノイズ重畳により 1 話者あたりの発話数を増やす手法である．さらに話者埋め込みを用いた話者照合システムにおいては学習データに含まれる話者数も重要性も高く，話者数を増やす手法として声道長正規化 （Vocal-Tract-Length Normalization） [15], [16] の技術を応用した声道長摂動（Vocal Tract Length Perturbation; VTLP） [17], [18] を用いて擬似的に話者を生成するものがある．VTLP を用いて音声の周波数軸を摂動させることにより擬似的な話者を生成し，学習データに加えることで全体の話者数を増やすことができるため，話者埋め込みの複雑性を高めることが可能となっている．先行研究では，疑似話者を単純に追加するのではなく話者性の変動量が十分に大きい疑似話者のみを選択することで，話者照合の性能が向上することが報告されている [19], [20]．ただし，話者性の変化が大きい疑似話者のみを選別して話者数を拡張すると学習データの質は向上するものの拡張に使える疑似話者が減るため，話者数を増やすという本来の目的を十分に達成できないという問題がある．

そこで本研究ではまず先行研究に則り声道長摂動による疑似話者生成を行い，コサイン類似度に基づいて話者性の変動量を算出した．そして一定値以上の変化があるデータのみを選別し，話者数の拡張に使用した．その後，さらに多くの疑似話者を話者数の拡張に利用するために話者性の変動量が小さいとみなされたデータに対してはパラメータの再調整を行い，類似度を計算する手順を繰り返した．実験では，話者照合における最先端手法の 1 つである ECAPA-TDNN を用いて話者照合システムを構築し，学習データとして様々なデータ拡張を行ったときの性能を評価した．実験結果から，ノイズ重畳による 1 話者あたりの発話数の拡張，および VTLP を用いた話者数の拡張の両方を行い，話者性の変動量を考慮しつつ最大限のデータ数を確保したデータ拡張手法において最もシステムの性能が向上し，多様な話者特徴を学習データに含めることの重要性が確認できた．

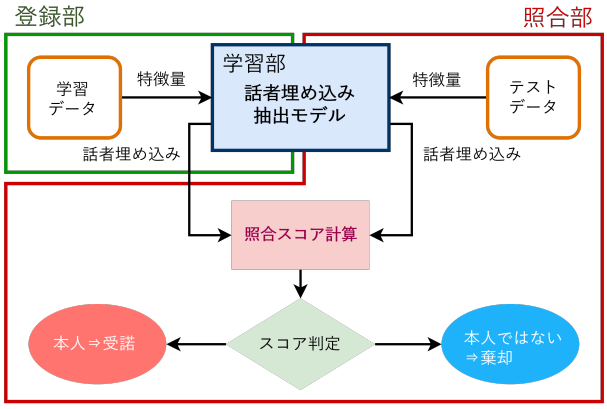

図 1　話者埋め込みに基づく話者照合システムのフロー
Fig. 1　Flow of speaker embedding-based speaker verification

2. 話者埋め込みに基づく話者照合

話者認識は次の 2 つのタスクに大別される．1 つは入力音声に対して複数の登録話者のうちから最も近い話者を識別する多値分類タスクの話者識別である．そして，もう 1 つは入力音声が登録話者本人によるものであるか否かを判定する二値分類タスクの話者照合である．図 1 に近年の主流である話者埋め込みに基づく話者照合システムのフローを示す．話者照合のシステムは，登録部，照合部，及びその両方で用いられる話者埋め込み抽出モデルの学習部によって構成されている．システムを構築するためにはまず，大規模な学習データを用いて話者埋め込み抽出モデルの学習を行う．話者埋め込み抽出モデルのタスクは話者識別となっている．より多くの話者を正確に識別可能なモデルは中間層で適切な話者性を抽出できていると仮定し，話者識別モデルの中間出力を話者埋め込みとして抽出し，登録部と照合部で利用する．登録部においては，登録話者の音声データを特徴量に変換し，その特徴量を話者埋め込み抽出モデルに入力することで話者埋め込みを抽出する．照合部においても同様の手順でテスト話者の話者埋め込みを抽出する．このようにして得られた話者埋め込み同士をコサイン類似度などを用いて類似度を計算し，しきい値との比較で本人か他人かを判定するようになっている．

3. 話者照合のためのデータ拡張

精度の高い話者埋め込みに基づく話者照合システムの構築には話者埋め込み抽出モデルの高性能化が不可欠であり，高性能な話者埋め込み抽出モデルの学習には大量な学習データが必要となる．また，既存のデータベースを用いるだけでなく，シミュレーションによって学習データを擬似的に増やすデータ拡張も広く用いられている．データ拡張の適用は学習データの量と多様性の増強に役立ち，話者照合に限らず様々な音声分野に

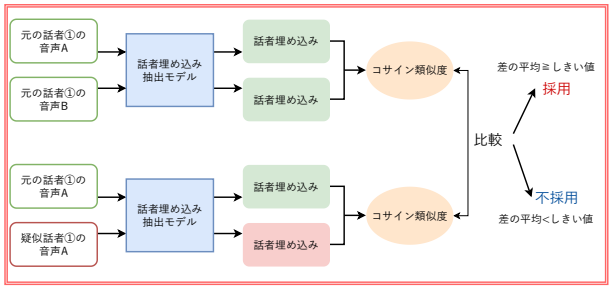

図 2　話者性の変動量の算出フロー

Fig. 2　Flow of calculating the amount of speaker variability

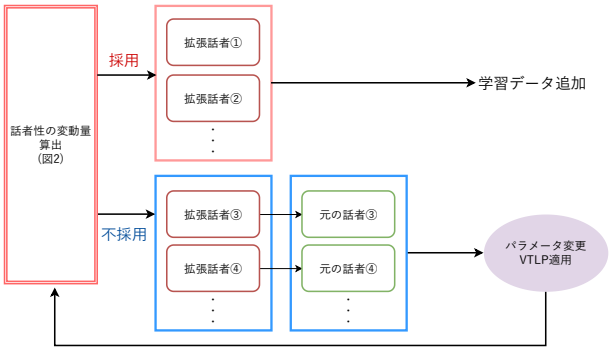

図 3　話者性の変動量を考慮したパラメータ調整フロー

Fig. 3　flow of Parameter adjustment considering the amount of speaker variability

おいても精度向上に有効であることが知られている．話者照合モデルの学習におけるデータ拡張は，ノイズや音楽を重畳する手法などによる発話数の拡張が主流であるが，話者照合においては VTLP を用いた手法などによる話者数の拡張も精度向上に繋がることが報告されている [21], [22].

3.1　ノイズ重畳による発話数拡張

ノイズ重畳は音声に雑音データを重畳することで発話数を拡張するデータ拡張手法である．大規模な学習データを必要とする DNN の学習において学習データの増強に使用されるほか，雑音の多い自然な環境に近づけることで実環境における頑健性を向上させる目的でも使用されることが多い [23]．話者照合における発話数の拡張は話者識別を行う話者埋め込み抽出モデルの学習を安定させることに繋がる．

3.2　声道長摂動による疑似話者生成

VTLN は音声認識分野などにおいて，話者ごとの音響的特徴のゆらぎを正規化する目的で用いられる技術である．この手法では，音声の短時間フーリエ変換を通して得られる対数振幅スペクトルの周波数軸を周波数伸縮係数に基づいて標準的な話者の対数振幅スペクトルに変換させている．元の音声の正規化周波数を ω，摂動後の周波数を ω'，周波数伸縮係数を α とすると，数式 （1）で表される．

$$\omega' = \omega + 2\arctan\frac{\alpha\sin(\omega)}{1-\alpha\cos(\omega)} \tag{1}$$

また，周波数伸縮係数のパラメータを調整することでピッチ変更の高低のレベルを調整することができる．音声認識などでは話者性の影響を除くために声道長を正規化するという目的で VTLN と呼ばれているが，本研究では逆に声道長にばらつきを与えるという目的の VTLP としてこの式を扱う．これまでに VTLP により声道長摂動を加えることで話者照合におけるデータ拡張に利用可能であることが報告されている [17], [18]．これらの手法では，学習データに対して VTLP を適用することで元の音声とは異なる声の高さに加工した音声を，新しい話者の音声として用いることで疑似的に話者数を増やす拡張を行っている．

4.　話者性の変動量を考慮した疑似話者生成

4.1　声道長摂動を用いた疑似話者生成による話者照合

前章で述べた VTLP によるデータ拡張を用いた話者照合 [19], [20] では，周波数伸縮係数を一律に設定して話者数を増やす手法と，その中から話者性の変動量が大きい疑似話者だけを選択して話者数を増やす手法が提案されており，話者照合において話者性の変動量が大きい疑似話者だけを選択する手法の方が性能が高くなることが報告されている．これは，話者性の変動量が小さい話者を元話者と異なる話者とみなして話者埋め込み抽出ネットワークを学習すると，元話者と疑似話者の区別が難しくなり，ネットワークの精度が低下することが原因だと考えられている．話者性の変動量が大きい音声だけを疑似話者の音声として選択する手法は，話者性の変動量が小さい疑似話者を除外することで話者埋め込み抽出ネットワークの学習を安定化させることができる反面，学習データに加えることができる疑似話者が少なくなり本来の目的であるデータ拡張としては十分に話者数を増強できなくなるというトレードオフが起こっている．

4.2　話者性の変動量によるデータ選択

VTLP による疑似話者生成では，すべての音声データに対して，あらかじめ設定した周波数伸縮係数に基づいて VTLP を適用し，加工後の音声の話者に新しい話者ラベルを付与することで疑似話者を生成する．先行研究で提案されている手法ではその後，話者性の変動量が固定値よりも大きい話者のみを疑似話者として選択する．生成した疑似話者の話者性の変動量を算出するフローの概要を図 2 に示す．元話者の音声同士及び疑似話者の音声と元話者の音声間のそれぞれの話者埋め込みを用いてコサイン類似度を求め，さらに元話者同士と元話者と疑似話者それぞれを比較したコサイン類似度の差分を求める．この差分を話者性の変動量とみなしている．ここで元話者の類似度の算出方法に関して，元話者同士のコサイン類似度でも発話ごとに変動があるため，基準の発話に対して同一話者の他の発話を用意し，複数のコサイン類似度を算出しその平均を本人同士のコサイン類似度として用いることとした．対して疑似話者の音声と元話者の音声間の類似度を求める際は，元話者による基準の 1 発話と同一元話者の音声に VTLP を適用して生成した疑似話者の全発話との話者埋め込み同士のコサイン類似度を算出する．このように計算された話者性の変動量に対してあらかじめ設定したしきい値以上であれば疑似話者の音声として学習データに追加し，しきい値未満であれば疑似話者として用いないようにしている．

4.3 話者性の変動量を考慮した疑似話者選択

VTLP は話者照合において話者数を増強する目的で利用されるデータ拡張手法であるため，話者性の変動量が小さいことでデータの選別によって除外される疑似話者は最小限に抑えるのが望ましいと考えられる．しかしながら先行研究では，話者照合の性能改善のために増やした疑似話者数からデータ選択をした結果，4 割程度に疑似話者が減ってしまったことが報告されている．先行研究における疑似話者によるデータ拡張の問題点を解消するために本研究では，話者性の変動量が小さいと判定され使用されなかった疑似話者の元話者に対して，パラメータを変更した VTLP を用いることでさらに新たな疑似話者を生成し，話者数の拡張に用いられるデータを増やすことを目指す．提案手法の概要を図 3 に示す．まず，図 2 で示した手順により疑似話者を話者性の変動量が大きいか小さいかを判定し，話者性の変動量が小さい疑似話者については，その元話者の音声にパラメータを変更した VTLP を適用することで新たな疑似話者を生成する．声道長の変化がより大きくなるようにパラメータの変更を行うため，当初のパラメータよりも話者性の変動量が大きくなり，疑似話者の音声として話者数の拡張に採用されるデータが増えることが期待される．その後，新たに生成した疑似話者に対して改めて図 2 で示した手順を適用し，学習データに加える疑似話者を選別する．パラメータの変動量を適切な値に設定して図 3 の手順を繰り返すことで自動的により多くの疑似話者選択を行うことが可能となる．ただし，VTLP のパラメータは絶対値を大きくするほど音声の歪みが大きくなることから，音声として認識できる限界に達した時点で提案手法によるデータ拡張を終了することとした．

5. 実　　験

提案手法の有効性を検証するために，発話数と話者数に対するデータ拡張を様々な条件で適用し，話者照合実験を行った．

5.1 データベース

本実験では，話者照合のために構築された音声コーパスである，JTubeSpeech-ASV [24] を用いて話者照合システムの構築および評価を行った．JTubeSpeech-ASV は YouTube 動画から自動収集した 900 時間の音声から構成される音声コーパスである．1 動画の中に登場する話者が 1 名の音声データのみから構成されており，1 チャンネルを 1 話者とみなしている．本データセットは日本語音声を主としているが，日本語以外に英語，中国語，韓国語などの言語が含まれている．話者埋め込み抽出モデルの学習および評価に用いた JTubeSpeech-ASV の話者照合用サブセットのうち，学習用データセットは 1,792 話者 107,271 発話の音声で構成されている．話者照合のテストには JTubeSpeech-ASV のテスト用データセットを用いる．このデータセットは 92 話者 20,976 発話の音声で構成されており，20,976 トライアルが含まれている．そのうち本人同士のトライアルが 228 セット，他人同士のトライアルが 20,748 セットである．サンプリング周波数は 16kHz である．ノイズ重畳には MUSAN データベース [25] を用いる．MUSAN データベースには 42 時間の様々なジャンルの音楽，12 言語の 60 時間にわたる会話，900 以上のノイズが含まれる．本実験では MUSAN データベースのうちノイズのサブセットのみを用いた．使用したサブセットには，機械音や環境音などのノイズが合計で約 6 時間収録されている．

表 1　各データ拡張条件における話者数とデータ量

Table 1　Number of speakers and total speaking time for each data augmentation condition

条件	話者数	総データ量（hrs）	話者毎の平均データ量 (hrs)
（A） データ拡張なし	1,792	498	0.28
（B） ノイズ重畳	1,792	1,495	0.83
（C） 従来法 (all)	5,376	1,494	0.83
（D） 従来法 (select)	2,868	794	0.44
（E） 提案法	4,086	1,128	0.63
（F） 提案法 + ノイズ重畳	4,086	5,380	3.00

5.2 実 験 条 件

本実験では話者埋め込みに基づく話者照合システムを構築する．話者埋め込み抽出モデルには ECAPA-TDNN を用いた．従来法では x-vector を用いた話者照合システムの性能が検討されているが，ECAPA-TDNN モデルは x-vector に比べて高い性能が得られることが知られている．モデルの入力特徴量には，対数メルフィルタバンクを使用した．話者埋め込みは ECAPA-TDNN の中間層の出力を使用し，512 次元のベクトルとして抽出した．VTLP については，パラメータの設定値による話者性の変化と話者照合性能の関係性を検証するために，全ての話者に対して単一のパラメータで VTLP を適用する従来法（all），先行研究に則り話者性の変化に応じて従来法（all）から拡張話者の選別を行った従来法（select），そして提案手法により様々なパラメータで VTLP を適用し，拡張話者の選別を行った提案法の 3 通りを用意した．比較条件を以下に示し，各条件における学習用データセットのデータ数を表 1 にまとめる．

（A） データ拡張なし

JtubeSpeech-ASV の学習用データセットのみを用いて話者照合モデルを学習する．データ拡張は行わない．

（B） ノイズ重畳

JtubeSpeech-ASV の学習用データセットに対し，ノイズ重畳による発話数の拡張を行う．ノイズデータは MUSAN のノイズデータセットからランダムに選択し，SNR は 0 とした．各音声データにつき 2 種類の雑音データを重畳し，1 話者あたりの発話数を 3 倍に増やした．

（C） 従来法（all）

JtubeSpeech-ASV の学習用データセットに対し，VTLP による話者数の拡張を行う．従来法（all）では学習用データセットに含まれる 1,792 話者に対して VTLP を適用した．周波数伸縮係数は 0.1，および-0.1 に設定し，各音声について VTLP を適用して音声を 2 つずつ生成する．VTLP を適用した音声は原音声とは異なる話者による発話とみなして新しい話者ラベルを付与することで，3,584 人の疑似話者を追加した．

（D） 従来法（select）

条件（C）の従来法（all）で生成した疑似話者のうち，元話

表 2　話者照合の条件ごとの EER（%）
Table 2　EER (%) of speaker verification for each condition

条件	EER（%）
（A） データ拡張なし	6.984
（B） ノイズ重畳	6.203
（C） 従来法（all）	7.018
（D） 従来法（select）	6.978
（E） 提案法	6.039
（F） 提案法 + ノイズ重畳	**5.388**

者からの話者性の変動量が大きい疑似話者のみを選択して話者数の拡張を行う．これにより，条件 （C）で擬似的に増やした 3,584 話者のうち 1,076 話者を話者数の拡張に用いた．

（E） 提案法

条件 （D）で選別により除外された疑似話者に対して 4.3 節で述べた提案手法を適用し，より多くの疑似話者を生成して話者数の拡張を行う．周波数伸縮係数は 0.1 から 0.17 まで 0.01 ずつ増加させる場合，または-0.1 から-0.17 まで 0.01 ずつ減少させる場合の 2 通りを実行した．本実験では話者性の変動量が 0.20 以上の疑似話者のみを選択した．これにより，条件 （D）において増やした疑似話者と合わせて 2,294 話者を追加の疑似話者とした．

（F） 提案法 + ノイズ重畳

条件（E）の音声にノイズを重畳することで，話者数と発話数の両方に対するデータ拡張を行った．ノイズデータは MUSAN のノイズデータセットからランダムに選択し，SNR は 0 とした．各音声データにつき 2 種類の雑音データを重畳し，1 話者あたりの発話数を 3 倍に増やした．

評価指標には等価エラー率（equal error rate; EER）を用いた．EER は他人受入率と本人拒否率が等価となる点から求められ，値が小さいほど良い精度と評価される．

5.3 実 験 結 果

表 2 に，条件（A）〜（F）それぞれにおける話者照合結果の EER を示す．まず（A）と（B）の 2 条件を比較すると，（B）の方が EER が 0.781 ポイント低下しており，ノイズ重畳による発話数拡張の有効性を確認できた．次に従来法（C）と（D）を比較すると，どちらの EER もデータ拡張を行っていない（A）の EER とほぼ同じになっている．なお，（C）ではデータ量が（B）とほぼ同程度であるにも関わらず全体で一番 EER が高くなった．このことから，ECAPA-TDNN では従来法による精度改善はあまり見られないと考えられる．これに対して（C）〜（E）の 3 条件を比較すると，VTLP を用いた話者数の拡張を行った中で話者性の変動量を考慮した（D）と（E）の EER が（C）よりも低くなっている．どちらも（C）に比べて話者数およびデータ量が少ないにも関わらず EER が改善したことから，話者性の変動量を考慮することの必要性が確認できる．また，3 条件のうち提案手法である（E）の EER が最も低く，（A）と比較して 0.945 ポイント EER が低下している．どちらの従来法と比較しても大幅に精度が改善しており，（D）と比較すると 0.939 ポイント EER が低下している．（D）と（E）ではどちらも話者性の変動量を考慮して話者数を拡張しているものの，（E）は（D）よりも多くの疑似話者を拡張に用いている．このことから，パラメータを調整することでより多くの疑似話者を用いることの有効性が確認できた．そして，（F）では提案法を用いて話者数を拡張した （E）からさらに話者ごとの発話数を増強したことにより，（E）と比較して 0.651 ポイント EER が低下した．同時に，発話数だけを拡張した（B）と比較しても 0.815 ポイント EER が低くなっており，全条件の中で最も良い性能となった．以上の結果より，話者埋め込みに基づく話者照合のためのデータ拡張において，ノイズ重畳による発話数拡張と話者性の変動量を考慮した話者数拡張を併用すると性能の向上に非常に効果的であることが確認できた．

6. お わ り に

本論文では，話者照合モデルの学習における学習データの拡張において，声道長摂動による疑似話者生成を行い，話者性の変動量を考慮して選別を行うと同時に VTLP のパラメータ変更によりさらに多くの話者数を拡張する手法を提案した．実験では，JTubespeech-ASV の学習用データセットに対してノイズ重畳による発話数拡張，疑似話者生成による話者数拡張，および両手法を用いたデータ拡張を行い，話者埋め込みに基づく話者照合モデルを学習して話者照合システムの性能を評価した．実験の結果，話者性の変動量の大きい話者のみを選別しつつ，より多くの疑似話者を学習データに加えた場合に最も高い照合精度を確認できた．

今後の課題として，最も適切な周波数伸縮係数としきい値の設定について，より定量的な設定方法を模索する必要性が考えられる．また，同一話者による音声間でも話者性の変動量に差異が生じることが考えられるため，より適切な疑似話者の選別方法を検討していく必要がある．

文　　献

[1] 俵直弘，“話者認識システムとなりすまし対策，” vol.78，no.6，音響学会誌，2022.

[2] 黒岩眞吾, 越仲孝文, 篠田浩一, 小川哲司, 松井知子, 王 龍標, 西田昌史, 柘植 覚, 網野加苗, 長内 隆, 石原俊一，“小特集 「話者認識に関する研究の動向」 にあたって，” vol.69，音響学会誌，2013.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)IEEE, pp.5329–5333 2018.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” , arXiv preprint arXiv:2005.07143, 2020.

[5] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” 2016 IEEE Spoken Language Technology Workshop (SLT)IEEE, pp.165–170 2016.

[6] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification.,” Interspeech, vol.2017, pp.999–1003, 2017.

[7] D. Garcia-Romero, D. Snyder, G. Sell, A. McCree, D. Povey, and S. Khudanpur, “x-vector dnn refinement with full-length recordings for speaker recognition.,” Interspeech, pp.1493–1496, 2019.

[8] P. Matějka, O. Plchot, O. Glembek, L. Burget, J. Rohdin, H. Zeinali, L. Mošner, A. Silnova, O. Novotnỳ, M. Diez, et al., “13 years of speaker recognition research at but, with longitudinal analysis of nist

sre," vol.63, , Computer Speech & Language, 2020.
[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)IEEE, pp.5329–5333 2018.
[10] S. Wang, J. Rohdin, O. Plchot, L. Burget, K. Yu, and J. Černockỳ, "Investigation of specaugment for deep speaker embedding learning," ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp.7139–7143 2020.
[11] P.S. Nidadavolu, V. Iglesias, J. Villalba, and N. Dehak, "Investigation on neural bandwidth extension of telephone speech for improved speaker recognition," ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp.6111–6115 2019.
[12] C.-L. Huang, "Exploring effective data augmentation with tdnn-lstm neural network embedding for speaker recognition," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)IEEE, pp.291–295 2019.
[13] Y. Zhu, T. Ko, and B. Mak, "Mixup learning strategies for text-independent speaker verification.," Interspeech, pp.4345–4349, 2019.
[14] H. Yamamoto, K.A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding.," Interspeech, pp.406–410, 2019.
[15] K. Johnson, "Vocal tract length normalization," vol.14, no.1, UC Berkeley PhonLab Annual Report, 2018.
[16] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," vol.6, no.1, IEEE Transactions on speech and audio processing, 1998.
[17] N. Jaitly and G.E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," Proc. ICML Workshop on Deep Learning for Audio, Speech and Language, vol.117, p.21, 2013.
[18] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol.1IEEE, pp.346–348 1996.
[19] T. Wakamatsu, S. Shiota, and H. Kiya, "Vocal tract length perturbation-based pseudo-speaker augmentation for speaker embedding learning," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) IEEE, pp.2228–2232 2023.
[20] 若松智花，塩田さやか，貴家仁志（都立大），"話者照合のための声道長摂動に基づく疑似話者生成によるデータ拡張，"，信学技報, 音声研究会，2024.
[21] C.-P. Chen, S.-Y. Zhang, C.-T. Yeh, J.-C. Wang, T. Wang, and C.-L. Huang, "Speaker characterization using tdnn-lstm based speaker embedding," ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp.6211–6215 2019.
[22] Z. Wu, S. Wang, Y. Qian, and K. Yu, "Data augmentation using variational autoencoder for embedding based speaker verification.," INTERSPEECH, pp.1163–1167, 2019.
[23] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds, "Robust speaker recognition in noisy conditions," vol.15, no.5, IEEE Transactions on Audio, Speech, and Language Processing, 2007.
[24] 塩田さやか，永森輝，若松智花，高道慎之介, "Jtubespeech-asv: Youtube から構築された話者照合のための日本語を主とした音声コーパス，"，情報処理学会研究報告，2023.
[25] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," , arXiv preprint arXiv:1510.08484, 2015.