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.

話者照合のための声道長摂動に基づく疑似話者生成によるデータ拡張

若松　智花†　　塩田さやか†　　貴家　仁志†

† 東京都立大学

あらまし　近年主流となっている深層学習モデルを用いた話者照合システムにおいて，信頼性の高いモデルを構築するためには大量の学習データが必要であり，データ拡張の適用が必要不可欠である．一般的に，ノイズを重畳することで発話数を増加させるデータ拡張手法が用いられることが多く，ほとんどの既存研究でその有効性が示されている．一方で，話者照合システムの性能は話者埋め込みベクトルの抽出器として用いる話者埋め込みネットワークの分類性能に依存することが知られており，学習データに含まれる話者数も性能の向上のために重要な要因の一つであるといえる．そこで本稿では，声道長の伸縮による話者性の変化に基づく疑似話者の生成を用いた話者数に対するデータ拡張手法を提案する．また，話者性の変化量が不十分な疑似話者の生成がかえって精度の低下に繋がる可能性を考慮し，話者性の変化量が十分な話者を選択するという手法についても検証した．x-vector に基づく話者照合において提案するデータ拡張手法の有効性を検証した結果，全ての疑似話者を使用する場合と比べて話者性の変化量を考慮した場合に話者照合の精度が向上したことが確認できた．さらに，提案する疑似話者拡張手法と従来のノイズ重畳による発話数の拡張を組み合わせることでデータ拡張の効果がより高くなり，JTubeSpeech-ASV データベースのテストセットにおいて 5.7% の EER を記録したことを報告する．

キーワード　話者照合，x-vector，データ拡張，声道長摂動

Vocal tract length perturbation-based pseudo-speaker augmentation for automatic speaker verification

Tomoka WAKAMATSU†, Sayaka SHIOTA†, and Hitoshi KIYA†

† Tokyo Metropolitan University

Abstract　In recent years, deep neural network (DNN)-based automatic speaker verification (ASV) systems have become mainstream. Data augmentation is an essential technique, as large amounts of training data are required to construct reliable ASV systems. It is well known that data augmentation increases the number of utterances by adding noise and is effective in most methods. However, it is known that the performance of ASV systems depends on the performance of the speaker embedding network used as the extractor of speaker embeddings. For this reason, the number of speakers in the training data is also an important factor for improving performance. Thus, we propose a method called pseudo-speaker augmentation, which utilizes a technique called vocal tract length (VTL) warping. Considering the possibility that generating pseudo-speakers with insufficient change in speaker characteristics may lead to a decrease in accuracy, we also examined the case where only using some speakers with a sufficient change of characteristics. In our experiments, the performance of the ASV system based on x-vector was improved when changes in speaker characteristics were taken into account. Furthermore, the highest performance was achieved in the system applying both the proposed pseudo-speaker augmentation and conventional utterance augmentation method, achieving an equal error rate of 5.7% on the JTubeSpeech-ASV database.

Key words　automatic speaker verification，x-vector，data augmentation，vocal tract length perturbation

1. はじめに

近年，音声対話システムの普及に伴い，音声を用いた生体認証技術である話者照合の必要性が高まっている．様々な分野で深層学習に基づく研究が進む中，話者照合においても x-vector [1] や ECAPA-TDNN（Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network）[2] に代表される DNN（Deep Neural Network）に基づくシステムが主流となって

きており，高い性能を達成している．これらの手法では，DNN の埋め込み層から話者埋め込みと呼ばれる固定長の特徴ベクトルを抽出し，話者の特徴を表現するベクトルとして用いる．このような話者埋め込みによる話者照合手法において精度の高いモデルを構築するためには，話者埋め込みを抽出する DNN の学習に用いる大規模な話者ラベル付き音声データが必要となる．また，より高い精度を実現するためには，大規模データだけではなく更にデータ拡張を用いることが一般的である．音声を扱う分野においては，ノイズ重畳によって発話数を増加させるデータ拡張手法が用いられることが多く，その有効性が示されている [3],[4]．また，データ拡張によってデータ量が増えるだけではなくデータのバリエーションも拡張されることで，モデルの頑健性の向上にも繋がる．

話者照合タスクでは，ノイズや残響音の追加 [1],[5]，Specaugment [4] など，発話数を増やすためのデータ拡張手法が一般的に使用されている．一方で，話者埋め込みネットワークに基づく話者照合システムにおいては，話者埋め込みの性能がシステムの性能向上のために重要な要因の一つである．話者埋め込みの表現能力は，埋め込みを抽出する DNN が分類可能な話者数が多いほど高くなることが知られており，これは学習データに含まれる話者数に依存する．そのため，話者照合システムを構築する際には，学習データの合計のデータ量だけではなく，十分な話者数が含まれていることも重要な要素であると考えられている [6]．しかしながら，話者ラベルを含むラベル付きデータを大量に収集することは非常にコストが高く，十分な量のデータの利用が困難な場合が多い．文献 [6] では，簡単な手順で話者数を増加させるために，Speed perturbation [7] を用いたデータ拡張手法が提案されている．データ拡張によって新しい話者を生成することで，埋め込み空間における話者特徴の表現範囲が拡張され，話者照合システムの性能が向上することが報告されている．

声道長正規化（Vocal Tract Length Normalization; VTLN） [8] は，周波数伸縮係数と呼ばれるパラメータに基づいて音声のスペクトルの周波数軸を伸縮する手法で，音声認識分野における声道長のばらつきによる認識性能への影響を打ち消すために使用されている．この手法を逆方向のアプローチとして適用した声道長摂動（Vocal Tract Length Perturbation; VTLP）[9] では，あえて音声データの特徴にばらつきを与えてモデル学習を行うことで高性能なモデルを構築したことが報告されている．声道長は異なる話者が異なる話者性を持つための要因の一つであることから，声道長の違いが生む話者性の変化を用いた話者埋め込みに基づく話者照合のための疑似話者拡張手法を提案した [10]．話者埋め込みネットワークは話者の分類を目的として学習されるため，一般的に学習データに含まれる話者数が多いほどモデルの頑健性が向上する．先行研究 [10] では，簡易的な構造の畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）において提案手法による話者数に対するデータ拡張の有効性を示したが，x-vector をはじめとした話者照合の最先端手法における有効性については検証がされていない．そこで本論文では，x-vector に基づく話者照合システムにおいて，声道長の伸縮に基づく話者性の変化を用いた話者数に対するデータ拡張手法の有効性を検証する．実験では，JTubeSpeech-ASV データベース [11] に対して話者数に対するデータ拡張を行い，話者照合システムの性能を評価した．また，VoxCeleb1 データベース [12] を用いて，提案手法によって生成された疑似話者の新しい話者としての妥当性を検証した．実験の結果，両方のデータベースにおいて，話者数に対するデータ拡張が従来の発話数に対するデータ拡張の効果をより向上させることが確認された．

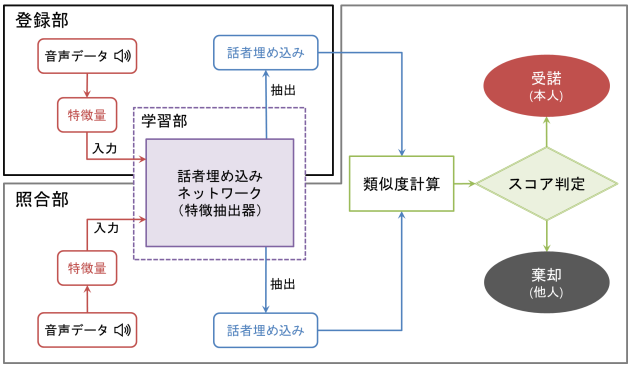

図 1　話者埋め込みに基づく話者照合システムのフロー

Fig. 1　Flow of speaker embedding-based ASV system

2.　x-vector に基づく話者照合

話者照合は，クエリとなる音声が事前に登録された話者本人の音声であるか，そうでないかを判別する二値分類タスクである．近年主流となっている深層学習による話者埋め込みネットワークを用いた話者照合システムのフローを図 1 に示す．話者照合システムは登録部と照合部の 2 つのフローと，両方のフローで用いられる話者埋め込みネットワークの学習部で構成されている．システム構築の際には，まず話者埋め込みと呼ばれる話者の特徴表現を抽出するための話者埋め込みネットワークを大量のラベル付きデータを用いて学習する．話者埋め込みネットワークは話者分類を行うニューラルネットワークモデルであり，学習したモデルの埋め込み層から話者埋め込みと呼ばれる特徴量を抽出するための特徴抽出器として用いる．次に，登録部において登録話者の音声データを音響特徴量に変換して話者埋め込みネットワーク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にテスト話者の音声から話者埋め込みを抽出する．最後に，抽出した 2 つの話者埋め込みの類似度を計算し，設定された閾値に対してスコア判定を行う．

話者照合の従来技術の一つに x-vector に基づく手法 [1] がある．可変長の発話を x-vector と呼ばれる固定次元のベクトルにマッピングする DNN を構築し，DNN の埋め込み層を用いて話者埋め込みを抽出するものである．DNN は学習データに含まれる話者を分類するために学習され，学習済みモデルの埋め込み層の出力から x-vector を抽出する．学習には，音響特徴量と対応する話者ラベルから構成される話者ラベル付きデータを使用し，高い精度のモデルを構築するためには大量のラベル付き学習データが必要となる．

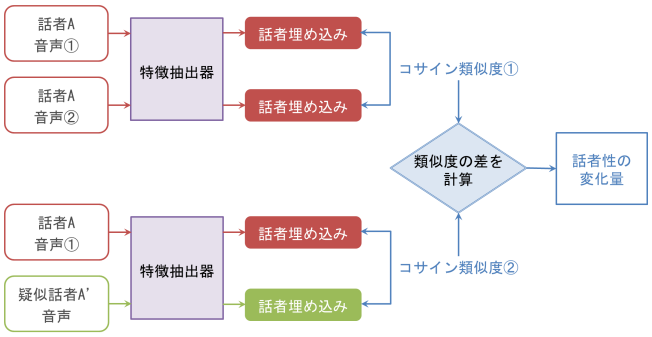

図 2　話者性の変化量に基づく疑似話者選択のフロー

Fig. 2　Flow of selecting pseudo-speaker based on the degree of changing speaker characteristics

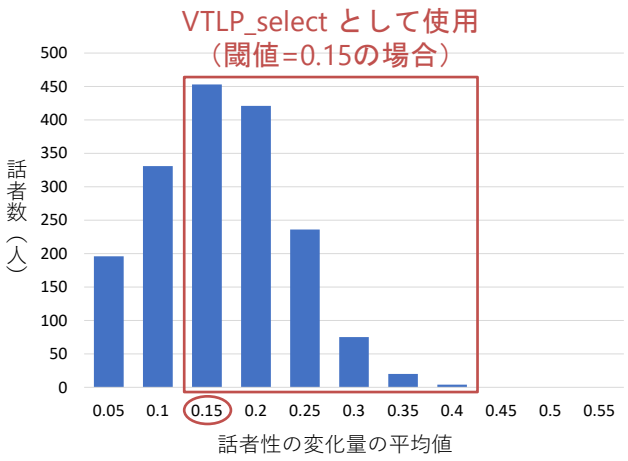

図 3　VTLP（$\alpha = 0.1$）による話者性の変化量の分布

Fig. 3　Distribution of the degree of changing speaker characteristics by VTLP ($\alpha = 0.1$) on JTubeSpeech-ASV

3.　疑似話者の生成によるデータ拡張

3.1　ノイズ重畳を用いた発話数のデータ拡張

モデルを学習する際には，テスト発話と類似した環境の音声が学習データに含まれていることが求められる．また，異なる環境や未知の環境にも適応可能な頑健性の高いシステムを構築するためには，より多くの環境を想定した様々な学習データを用いてモデルの学習を行う必要がある．しかしながら，音声データを大量に収集することはコストが高く，実際に様々な条件下で音声を収録することは難しい場合も多い．このような問題点に対処するため，様々なタスクにおいてデータ拡張は必要不可欠な技術となっており，多くの既存研究で有効性が報告されている．データ拡張は，単純にデータ量を増加させるだけではなく，データのバリエーションを増やすことも可能である．データの分布が広がることにより，モデルの過学習を防ぎ，モデルの頑健性や汎用性の向上に繋がる．一般的なデータ拡張の手法として，ノイズや背景音を元のデータに重畳する手法が用いられることが多い．多くの話者照合システムにおいて，実環境ではテスト発話に背景雑音が含まれることが想定されるため，異なる条件のノイズを学習データに含ませることで実環境に近い条件となることが期待され，より高い性能を発揮することが出来る．

3.2　声道長摂動を用いた話者数のデータ拡張

異なる話者が異なる音声特徴を持つための要因の一つに声道長がある．音声認識の分野では，声道長の長さの違いが認識精度に影響を及ぼすことが知られており，この影響に対処するために VTLN [8] と呼ばれる手法が一般的に用いられている．VTLN 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と呼ばれるパラメータに基づいて伸縮することで，声道の長さが異なる二人の話者のスペクトルをスケーリングする．周波数伸縮係数は話者毎に最適なパラメータが設定され，これによって話者間の声道長のばらつきを除去することが可能である．式（1）に，VTLN におけるスペクトルのスケーリングを示す．

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

ここで，元の音声の正規化周波数を ω，伸縮後の周波数を ω'，周波数伸縮係数を α とする．

一方，VTLN を正規化とは逆の方向性で適用することで音声データにバリエーションを与える VTLP [9] と呼ばれる手法がある．先行研究 [10] では，声道長の違いが話者性に影響を与えるという点に着目し，VTLP に基づく疑似話者生成による話者数のデータ拡張を行った．式（1）に基づいて音声を変換し，生成されたそれぞれの音声に元の話者とは異なる話者ラベルを付与することで疑似的に新しい話者を生成する．2 つの異なる α を用いた場合には，各話者から新たに 2 人の疑似話者を生成することになるため，話者数は 3 倍に拡張される．

3.3　話者性の変化を考慮した疑似話者拡張

疑似話者の生成において，VTLP 適用後の話者性が元の話者と大きく異なることが重要である．そこで，提案手法をより効果的に用いるために，話者性が特に大きく変化した疑似話者のみを選択する手法についても検討をした [10]．図 2 に，疑似話者選択のフローを示す．学習済みの話者分類モデルの一つを用いて元の話者と疑似話者のそれぞれの話者埋め込みを抽出し，これら 2 つの話者埋め込みのコサイン類似度の差を話者性の変化量として定義する．図 3 に，本研究で使用した JTubeSpeech-ASV [11] の学習データセットに対して周波数伸縮係数 $\alpha = 0.1$ で疑似話者を生成した際の，各話者における話者性の変化量の平均値を示す．この分布から，全ての話者に対して同じ周波数伸縮係数の値を用いた場合の話者性の変化量は話者によって異なることが確認できる．また，元の話者と疑似話者の話者性がほとんど変わっていない話者の場合，疑似話者を新しい話者として扱うためには話者性の変化量が不十分であると考えた．そこで，図 3 のように分布のピークとなる値を閾値として設定し，話者性の変化量が閾値以上の疑似話者のみを使用した．

4.　実　　験

本実験では，先行研究 [10] では評価を行っていなかった

表 1　各比較条件におけるデータ量
Table 1　Data amount of the training dataset in each condition

	データ拡張手法	JTubeSpeech-ASV			VoxCeleb1_400spk		
		話者数（人）	データ量（時間）	データ量 / 話者数	話者数（人）	データ量（時間）	データ量 / 話者数
（A）	None	1,792	498	0.28	400	110	0.28
（B）	Noise	1,792	1,495	0.83	400	331	0.83
（C）	VTLP	5,376	1,494	0.28	1,200	330	0.28
（D）	VTLP_select	3,296	916	0.28	827	227	0.27
（E）	VTLP_select+Noise	3,296	2,748	0.83	827	682	0.82
（F）	Noise_large	1,792	2,991	1.67	400	662	1.66

x-vector に基づく話者照合システムにおいて，提案する疑似話者拡張手法の性能評価を行う．

4.1　データベース

話者照合モデルの学習及びテストには JTubeSpeech-ASV [11] および VoxCeleb1 [12] を用いた．JTubeSpeech-ASV は，YouTube にアップロードされた動画から作成された話者照合用のデータベースであり，主に日本語話者の音声から構成されている．動画は自動で収集されているため，英語や中国語をはじめとする日本語以外の言語も含まれている．データセットは単一話者の音声のみで構成されており，動画のチャンネル ID が話者 ID として作成されている．学習用データセットには 1,792 話者による 107,271 発話 498 時間のデータが含まれている．評価用データセットには 92 話者による 20,976 トライアルのデータが含まれており，本人同士の 228 ペアと他人同士の 20,748 ペアで構成されている．評価用データセットに含まれる話者は，全て日本語話者となるように手動でアノテーションされている．VoxCeleb1 は，YouTube にアップロードされたインタビュー動画から抽出された，英語話者の著名人による音声データセットである．開発用データセットには 1,211 話者による 148,642 発話 340 時間のデータが含まれている．本実験では，提案手法によって生成した疑似話者が新しい話者として妥当であるかを検証するため，VoxCeleb1 の開発用データセットに含まれる 400 話者による 49,009 発話 110 時間のデータを使用した．本論文ではこのデータセットを VoxCeleb1_400spk と表記する．評価用データセットには 40 話者による 37,720 トライアルのデータが含まれており，全体の半数が本人同士の 18,860 ペア，残り半数が他人同士のペアで構成されている．

発話数に対するデータ拡張で用いるノイズデータには，MUSAN [13] を用いた．MUSAN には，42 時間の様々なジャンルの音楽，12 言語による 60 時間の会話，機械音，非機械音，環境音などを含む 900 種類以上のノイズが含まれる．本実験では，ノイズのサブセットのみを用いてノイズ重畳を行った．実験に用いる全てのデータのサンプリング周波数は 16kHz である．

4.2　実 験 条 件

本実験では，JTubeSpeech-ASV と VoxCeleb1_400spk の学習データセットに対して，発話数および話者数のデータ拡張を行い，2 章で述べた x-vector に基づく話者照合システムの学習を行った．従来手法としてノイズ重畳による発話数に対するデータ拡張と，提案手法として話者数に対するデータ拡張を適用し，話者照合システムにおける各データ拡張手法の有効性を検証する．表 1 は比較する各データ拡張手法におけるデータ量を示しており，各条件の詳細を以下に示す．

（A）　None

学習データに対してデータ拡張を行わない．この条件をベースラインシステムとする．

（B）　Noise

ノイズ重畳による発話数のデータ拡張を行う．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，重畳する際の SNR（Signal-to-Noise Ratio）は［-5，0，5，10，15］からランダムに設定した．この条件では，学習データに含まれる各発話毎に 2 種類のノイズと SNR の組み合わせでノイズ重畳を行い，発話数を 3 倍に拡張した．

（C）　VTLP

VTLP を用いて疑似話者を生成し，話者数のデータ拡張を行う．VTLP のパラメータである周波数伸縮係数 α は，+0.1 と -0.1 の 2 つの値に設定し，式（1）に基づいて音声を変換した．この条件では，学習データに含まれる全ての話者に対して疑似話者拡張を適用し，話者数を 3 倍に拡張した．

（D）　VTLP_select

3.3 節で説明した手順に基づいて，条件（C）で生成した疑似話者のうち話者性の変化量が閾値以上の疑似話者のみを用いて話者数のデータ拡張を行い，VTLP による話者性の変化量が話者照合システムの精度に影響するかどうかを検証する．最終的な話者数はベースラインの約 2 倍となり，条件（C）よりも少なくなっている．

（E）　VTLP_select+Noise

条件（D）と条件（B）を組み合わせることにより，話者数と発話数の両方に対してデータ拡張を行う．この条件では，条件（D）で学習に用いた各発話に対して条件（B）と同様の手順でノイズ重畳を行い，条件（D）に対して発話数を 3 倍に拡張した．

（F）　Noise_large

ノイズ重畳のみを用いて，条件（E）と総データ量が同程度になるように発話数のデータ拡張を行う．SNR は［-5，0，5，10，15］の全ての値を使用する．条件（B）と同様にノイズデータをランダムに選択してノイズ重畳を行い，発話数を 6 倍に拡張した．

条件（D）と（E）における VTLP_select の閾値は，JTubeSpeech-ASV が 0.14，VoxCeleb1_400spk が 0.2 とした．閾値の選び方に関しては，各データセットおよびパラメータにおける話者性の

表 2　JTubeSpeech-ASV で学習した際の，各比較手法における EER（%）
Table 2　EERs (%) for each augmentation method on JTubeSpeech-ASV

データ拡張手法		EER（%）
（A）	None	7.02
（B）	Noise	6.42
（C）	VTLP	7.46
（D）	VTLP_select	7.02
（E）	VTLP_select+Noise	5.70
（F）	Noise_large	6.14

変化量の分布のピークとなる値を参考にし，適当な値を設定した．また，VoxCeleb1 の学習データセットに対してはデータ拡張を行わずに話者照合モデルの学習を行った．モデル学習時には，フレーム長 25ms，フレームシフト 10ms，240 次元のフィルタバンクを入力特徴量として使用し，最適化関数に AdamW，損失関数に AM-Softmax を用いて最大 300,000 イテレーションの学習を行った．

話者照合のテストには学習時と同一言語の評価用データセットを使用し，話者照合モデルの埋め込み層から 512 次元の話者埋め込みベクトルを抽出して照合を行った．システムの性能は，等価エラー率（Equal Error Rate; EER）を用いて評価した．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，値が小さいほど話者照合精度が高いと評価される．

4.3 実 験 結 果

表 2，3 に，それぞれ JTubeSpeech-ASV と VoxCeleb1 で話者照合モデルを学習した際の，各比較条件における EER を示す．はじめに，表 2 の JTubeSpeech-ASV データセットにおける各比較手法の結果を比較する．（A）と（B）を比較すると，ベースラインの（A）に比べて（B）の方が EER が低くなっていることから精度が高くなっており，従来手法の発話数に対するデータ拡張が照合精度の向上に効果的であることが確認できた．（C）と（A）を比較すると，全ての話者に対して疑似話者拡張を行った場合には EER が高くなっており，精度が低くなっている．また，（C）と（D）を比較すると，（D）の方が話者数は少なくなっているのにも関わらず EER が 0.44 ポイント低下しており，話者照合の精度は高くなっている．この結果から，（C）では話者性が元の話者から十分に変化していない疑似話者が含まれていたことがモデル学習に悪影響を及ぼしていたが，（D）では話者性の変化量が大きい疑似話者のみを選んで使用したことでデータ拡張として適切に機能するようになったと考えられる．一方で，従来手法の発話数に対するデータ拡張と比べると，提案手法の話者数に対するデータ拡張を適用した場合の精度は低い．この理由の一つとして，提案手法により話者のバリエーションは増えたが，各話者の発話数が十分ではなかったため精度がそれほど向上しなかったと考えられる．そこで，この問題に対処するために（E）では話者数と発話数の両方に対してデータ拡張を適用した．（A）から（D）までの結果と（E）を比較すると，（E）の EER が最も低くなっており精度が高くなった．しかしながら（E）の結果だけでは，精度向上の要因はデータ量の総量が最も多いためである可能性も考えられ，各話者の発話数の増強によるものである根拠にはならない．そこで，発話数に対するデータ拡張のみを用いてデータ量が（E）と同程度になるようにデータ拡張を行った（F）との比較を行った．（E）と（F）を比較すると，（E）の方が EER が 0.44 ポイント低下しており，全ての条件の中で最も高い精度を記録した．このことから，話者照合システムの学習においてはデータの総量だけではなく話者数および各話者ごとの発話数がどちらも十分であることが重要であり，発話数の拡張と疑似話者拡張を組み合わせることでデータ拡張の効果が最も高くなることが確認された．

表 3　VoxCeleb1 で学習した際の，各比較手法における EER（%）
Table 3　EERs (%) for each augmentation method on VoxCeleb1

学習データセット	データ拡張手法		EER（%）
VoxCeleb1	None		8.35
VoxCeleb1_400spk	（A）	None	10.52
	（B）	Noise	10.67
	（C）	VTLP	11.74
	（D）	VTLP_select	11.37
	（E）	VTLP_select+Noise	10.01
	（F）	Noise_large	9.58
	（G）	VTLP_select+Noise_large	9.31

次に，表 3 の VoxCeleb1 データセットにおける結果を比較する．この実験では，JTubeSpeech-ASV と同様に提案手法である疑似話者拡張の有効性を確認するとともに，VoxCeleb1_400spk データセットにデータ拡張を適用することで本来の精度に近づけることを目標としている．まず，（A）から（E）までの結果を比較すると，JTubeSpeech-ASV の結果と同様の傾向が確認できた．（E）と（F）を比較すると，発話数に対するデータ拡張のみを行った（F）の方が EER が低下し精度が高くなっている．この理由として，学習データ全体のデータ量が不足している場合には，話者数よりも発話数の方がモデルの性能に影響を与えやすい可能性が考えられる．また，学習データセットが VoxCeleb1 の場合と VoxCeleb1_400spk の場合を比較すると，JTubeSpeech-ASV においても効果的であった（E）や（F）では，データ拡張を行うことで VoxCeleb1 の結果に近づいている．さらに VoxCeleb1 の結果に近づけるため，（D）に対して（F）と同様のノイズ重畳によるデータ拡張を行うことで，3,296 話者，5,496.66 時間のデータセットとなる VTLP_select+Noise_large を追加条件（G）として用意した．（G）の EER は 9.31% で，VoxCeleb1 の結果にはまだ精度は及ばないが，（A）から（F）の結果と比較すると最も高い精度となった．これらの結果から，提案手法で生成した疑似話者は本来の異なる話者間の話者特徴を再現するためには不十分であったと考えられるが，元のデータ量が極めて少ない場合にも JTubeSpeech-ASV と同様に各データ拡張手法の有効性を確認することが出来た．

5. ま と め

本稿では，近年主流となっている x-vector に基づく話者照合システムにおいて，話者性の変化を利用した疑似話者生成による話者数に対するデータ拡張の有効性を検証した．実験の結果，

JTubeSpeech-ASV と VoxCeleb1_400spk の両方のデータセットにおいて，提案手法による話者特徴の変化量が大きい話者のみを使用することでデータ拡張の有効性が高くなったことを確認した．また，発話数と話者数の両方に対してデータ拡張を行うことで，より高い精度を記録した．一方で，VoxCeleb1 の開発用データセットの一部を用いてモデルの学習を行った結果から，本来の異なる話者のデータを用いて学習したモデルに精度を近づけるためには，提案手法によるデータ拡張の効果が十分ではないことも示された．

今後の展望として，提案手法におけるパラメータの検討や，話者毎に適切なパラメータを用いた疑似話者の生成などが挙げられる．

文　　献

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

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," Proc. Interspeech 2020, pp.3830–3834, 2020.

[3] Y. Yang, S. Wang, M. Sun, Y. Qian, and K. Yu, "Generative adversarial networks based x-vector augmentation for robust probabilistic linear discriminant analysis in speaker verification," Proc. 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp.205–209, 2018.

[4] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, and Q.V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," Proc. Interspeech 2019, pp.2613–2617, 2019.

[5] T. Ko, V. Peddinti, D. Povey, M.L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp.5220–5224 2017.

[6] H. Yamamoto, K.A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," Proc. Interspeech 2019, pp.406–410, 2019.

[7] T.Ko, V.Peddinti, D.Povey, and S.Khudanpur, "Audio augmentation for speech recognition," Sixteenth annual conference of the international speech communication association, 2015.

[8] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," IEEE Transactions on speech and audio processing, vol.6, no.1, pp.49–60, 1998.

[9] 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.

[10] 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.

[11] 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，vol.2023-MUS-137, pp.No.4,1–4，2023.

[12] A. Nagrani, J.S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017.

[13] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," arXiv preprint arXiv:1510.08484, 2015.