多言語音声データベースを用いた話者照合のための声道長正規化によるデータ拡張 *

☆若松智花, 塩田さやか, 貴家仁志 (都立大)

1　はじめに

近年，生体認証技術の研究は非常に活発化しており，人々の生活に身近なものとなっている．特に指紋，静脈，顔などを用いた技術が実用化されてきている．その中で，音声を生体情報として用いる認証技術を話者照合と呼ぶ．マイクがあれば生体情報を入手することができるため実用化のハードルが低いことや，インターネットの普及に伴うオンラインや電話上での生体認証の必要性から，生体認証技術の一つとして話者照合の需要が高まってきている．

話者照合に関する最先端技術として，x-vector や ECAPA-TDNN に代表される深層学習 (Deep Neural Network; DNN) に基づく手法 [1–4] がある．このような最先端の話者照合システムにおけるモデルの学習には大規模なデータベースが必要ととなることが知られており，多くの既存研究では大規模な音声データベースが公開されている英語音声が用いられている．また，大量のデータを扱うための手法として，データ拡張が一般的である．データ拡張はデータ量や多様性を拡張することが可能であり，精度向上のために非常に効果的であることが知られている [2,5,6]．一方で，話者照合システムにおける言語依存性の問題が解消されていないことも知られている [7]．大規模なデータベースが存在しない言語を用いてシステムを構築する場合には，ファインチューニングやデータ拡張等の手法を用いる必要がある．

これまでのデータ拡張ではノイズを重畳することで各話者の発話数を増やしていた．しかし，近年主流となってきている話者埋め込みによる話者照合では，学習データに含まれる話者数も非常に重要な要素とみなされている．そこで本研究では，複数の言語を含むデータ量が限られているデータベースに対し，発話数および話者数の拡張という観点からデータ拡張の有効性について検討する．話者数を増やす方法として，声道長正規化 (Vocal-Tract-Length-Normalization; VTLN) を用いた．VTLN による話者性の変化を利用し，VTLN 適用前後の音声を別の話者とみなすことで疑似的に話者数の拡張を行った．実験では，従来のデータ拡張手法であるノイズ重畳を用いた話者毎の発話数の拡張，提案法である VTLN を用いた話者数の拡張を行い，話者埋め込みネットワークに基づく話者照合システムの性能を評価した．実験結果より，ノイズ重畳による発話数のデータ拡張と VTLN による話者数のデータ拡張の両方を行ったシステムにおいて最も良い性能となり，話者数の拡張という観点におけるデータ拡張の有効性が確認できたことを報告する．

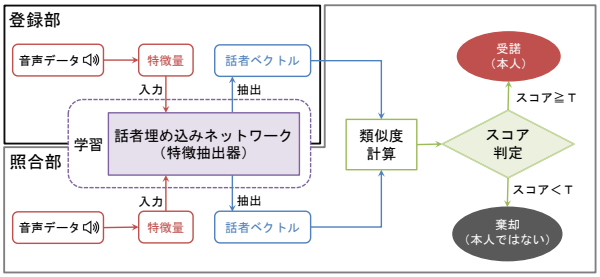

図 1　話者照合システムのフロー

2　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 2 値分類タスクである．近年主流となってい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 つの大きなフローと，両方で用いられるモデルの学習部で構成されている．システムの構築の際にはまず大量の学習データを用いて話者の特徴を抽出するための話者埋め込みネットワークを学習する．基本的に話者埋め込みネットワークは話者識別を行うモデルであり，学習したモデルの埋め込み層から特徴を抽出する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特徴量に変換して特徴抽出器に入力することで話者ベクトルを抽出する．照合部においても同様にテスト話者の音声の特徴量を特徴抽出器に入力して話者ベクトルを抽出する．その後，抽出した 2 つの話者ベクトルの類似度を計算し，設定された閾値 (T) に対してスコア判定を行う．

話者埋め込みネットワークに基づく話者照合では，話者ベクトルを抽出する特徴抽出器である話者識別モデルの識別性能が照合の性能に大きく影響する．そのため，モデル学習時点の識別性能を向上させるためにデータ拡張が重要であると考えられている．また，話者識別モデルが識別可能な話者数の多さが特徴抽

* Voice-Tract-Length-Normalization-based data augmentation for speaker verification with mixed language speech database. by Tomoka Wakamatsu, Sayaka Shiota, Hitoshi Kiya(Tokyo Metropolitan University)

出器の性能に大きく依存することが知られている．

3　VTLN によるデータ拡張

本研究では，ノイズ重畳と VTLN の 2 種類の手法によるデータ拡張を行った．

3.1　ノイズ重畳によるデータ拡張

大量の学習データを扱うためのデータ拡張手法については，ノイズや音楽の重畳や，リバーブなどといった様々な手法 [8–11] が報告されている．本研究では従来法としてノイズ重畳による話者毎のデータ数の拡張を行った．ノイズ重畳は，クリーンな音声にノイズ音声を重ね合わせることによるデータ拡張手法であり，大量のデータを必要とする DNN の学習において頻繁に使用されるデータ拡張手法の 1 つである．また，データ数を拡張するだけではなく，歪みのあるデータを学習データに加えることによってモデルのロバスト性の向上にも繋がる．

3.2　VTLN を用いた話者数に対するデータ拡張

VTLN は，主に音声認識分野において話者間の声道長の違いによって生じる歪みを取り除くために用いられる手法である [12]．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．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 とすると式 (1) で表される．

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

本研究では，声道長の伸縮による話者性の変化を利用して，VTLN によって加工された音声に原音声とは別の話者 ID を付けることで，疑似的に話者数を増やすデータ拡張手法として用いた．

4　実験

本実験では，話者照合における話者数に対するデータ拡張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適用し，話者照合を行った．また，特徴抽出器のモデル学習と話者照合のテストに同一言語のデータベースを用いることで話者照合の性能が最良となることを確認するため，モデル学習を行った言語とは異なる言語で話者照合を行い，言語依存性の有無を確認する．

4.1　データベース

本実験では，JTubeSpeech [13] という日本語音声データベースを用いて話者埋め込みネットワークに基づく話者照合システムの構築および評価を行った．JTubeSpeech は動画共有プラットフォームである YouTube に投稿されている動画の音声で構成された日本語音声データベースであり，音声認識と話者照合を行うことを前提として作成されている．話者照合で用いるサブセットは 1 動画内に登場する話者が 1 名だけのもので構成されており，YouTube における 1 チャンネルを 1 話者とみなす．本データセットは日本語音声データベースとして作成されているが，自動収集によって作成しているため実際には日本語以外に英語や中国語，韓国語などの言語も混在している．

ノイズ重畳で用いるノイズのデータベースには，MUSAN データベース [14] を用いる．MUSAN 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 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た．MUSAN のノイズデータセットには，機械音，非機械音，環境音などの約 6 時間の多様な種類のノイズが含まれている．

表 1　本実験で用いた CNN の各層の構築

	Layer	Kernel size	Input x Output
1	Conv1d	5	40x128
2	Conv1d	3	128x128
3	Conv1d	3	128x128
4	Conv1d	3	128x64
5	Fc	-	(64x3)x(emb)
6	Fc	-	(emb)x($n\_classes$)

4.2　実験条件

JTubeSpeech を用いてデータ拡張の有効性を検証する際のモデル学習に用いた畳み込みニューラルネットワークの構造を表 1 に示す．本実験における話者照合システムでは，第 5 層の出力である emb を話者ベクトルとして抽出し，話者照合に用いる．モデルの入力音声特性には，フレーム長 25ms，40 次元の MFCC を使用した．モデルの学習および話者照合のテストには JTubeSpeech の話者照合用サブセットを用いている．学習用データセットには 1,795 話者 502.49 時間，検証用データセットには 1,795 話者 138.21 時間の音声データを用いた．学習用データセットに対してノイズ，VTLN，および両手法を用いたデータ拡張を行い，話者照合の精度を比較する．VTLN については，VTLN による話者性の変化量の違いが話者照合の精度に影響するかどうかを検証するため，全ての話者に対して VTLN を適用する VTLN(all) と，話者性の変化が閾値より大きい話者に対してのみ VTLN を適用する VTLN(select) の 2 通りを行った．

比較条件を以下に示し，各条件における学習用デー

表 2　各条件における話者数とデータ量

条件	データ拡張	話者数	総発話時間 (hour)
(A)	なし	1,795	502.49
(B)	ノイズ	1,795	1,510.17
(C)	VTLN(all)	5,385	1,506.32
(D)	VTLN(select)	3,681	957.59
(E)	VTLN(select) + ノイズ	3,681	2,875.49

タセットのデータ数を表 2 にまとめる．

(A) ベースライン
学習データには JTubeSpeech の学習用データセットのみを用い，データ拡張は行わない．

(B) ノイズ
JTubeSpeech の学習用データセットに対し，ノイズによるデータ拡張を適用する．MUSAN のノイズデータセットからランダムにノイズを選択し，[-5，0，5，10，15] の中からランダムに選択した SN 比に従って音声に重畳する．

(C) VTLN(all)
JTubeSpeech の学習用データセットに対し，VTLN によるデータ拡張を適用する．VTLN(all) では，学習用データセットに含まれる 1,795 話者に対して VTLN を行う．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 VTLN でを適用して音声を 2 つずつ生成する．VTLN を適用した音声はクリーンな音声とは話者性が異なる音声とみなして新しい話者ラベルを付与することで，3,590 話者の疑似話者を追加した．

(D) VTLN(select)
条件 (C) の VTLN(all) のデータのうち，元の話者と VTLN 適用後の話者性の変化が大きい話者のみを対象として話者数のデータ拡張を行う．話者の選択には VTLN を適用する前後の音声の話者ベクトルを抽出し，それらのコサイン類似度が一定値以下のもののみを選択した．これにより，(C) において疑似的に増やした 3,590 話者のうち，1,886 話者を追加の疑似話者とした．

(E) VTLN(select) + ノイズ
条件 (D) の VTLN(select) 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．

さらに，学習データの話者数の増加によって、抽出する話者ベクトルの適切な次元数に影響があるかどうかを調べる為、各条件における話者埋め込み層の次元数を 512 次元，1,024 次元，2,048 次元の 3 通りに設定をしてモデルの学習を行った．

表 3　x-vector モデルを用いて話者照合を行った際の英語および日本語の EER(%)

条件	学習データ	テストデータ	EER(%) 男声	EER(%) 女性
(a)	英語 (Librispeech)	英語 (VCTK)	2.07	4.89
(b)		日本語 (JTubeSpeech)	10.68	

話者照合のテストには JTubeSpeech のテスト用データセットを用いる．このデータセットには，92 話者による 20,976 トライアルが含まれており，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．音声は全て日本人話者のものであることを実際に音声を聞いて確認した上で使用した．また，比較するモデルとして，VoicePrivacy 2020 (VP2020) [15] にて配布された x-vector に基づく話者照合モデルを用いた．このモデルは大量の英語音声で構成されている Librispeech [16] によって事前学習されている．英語話者の照合用のテストデータは，VP2020 で配布された VCTK [17] のものを用いた．

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．話者照合モデルを用いて登録発話とテスト発話の話者ベクトルを抽出した後，2 つの話者ベクトルのコサイン類似度をスコアとして用い EER を求めた．

4.3　実験結果

表 3 に，x-vector モデルを用いて英語話者および日本語話者それぞれの評価データを用いて話者照合を行った際の EER(%) を示す．条件 (a) はモデルの学習と照合に用いる言語が同一言語である場合，条件 (b) はモデルの学習と照合に用いる言語が異なる言語の場合を表している．(a) と (b) の結果を比べると，(b) の方が平均して 7.20 ポイントと大幅に精度が低下していることがわかる．この結果から，JTubeSpeech が雑音や圧縮を含んでおり Librispeech とはドメインの異なるデータであることだけではなく，言語の依存性も要因の一つであると考えられる．

次に，条件 (A)〜(E) の EER(%) を表 4 に示す．はじめに，話者ベクトルの次元数が 512 次元の場合における各条件の傾向を確認する．(A)〜(D) の 4 条件を比較すると，データ拡張を 1 種類だけ用いている (B)〜(D) の中では (B) が最も良い性能となった．続いて，(C)，(D) の 2 条件を比較すると，(D) の方が 0.391 ポイント精度が高くなったことが確認できる．

表 4　各データ拡張手法と話者ベクトルの次元数毎の EER(%)

	データ拡張	EER(%)		
		話者ベクトルの次元数		
		512	1,024	2,048
(A)	なし	6.637	7.456	7.456
(B)	ノイズ	5.422	**5.056**	5.702
(C)	VTLN(all)	8.030	8.333	7.181
(D)	VTLN(select)	7.639	7.340	5.620
(E)	VTLN(select)+ノイズ	**4.193**	5.702	**4.984**

これは，(C) の学習データには VTLN による話者性の変化が小さい音声も含まれているため，これらの話者特徴が元の音声の話者特徴と類似してしまうことで話者照合の精度が低くなったと考えられる．(D) では話者性の変化が特に大きい話者のみを選択して用いたことで効果的に話者数の拡張を行うことができたと考えられる．しかしながら，話者数を増やした一方で各話者毎の発話数は十分ではなかったと考えられるため，データ拡張なしの (A) よりも精度が低いことがわかる．一方，(E) では話者数を増やした (D) からさらに話者毎の発話数を増やしたことにより，発話数だけを拡張した (B) と比較して 1.229 ポイント精度が向上し，全ての条件の中で最も良い性能となった．他の次元数においても同様の傾向が見られ，特に (C)〜(E) の提案法による拡張法を比較すると，どの次元数についても (C)，(D)，(E) の順に精度が高くなった．以上の結果より，話者埋め込みネットワークを用いた話者照合のためのデータ拡張において，ノイズの重畳による発話数の拡張だけではなく話者数を増やすことが効果的であることが確認できた．

5　まとめ

本論文では，話者照合システムにおいて話者数の拡張という観点からデータ拡張の有効性について検証した．話者数の疑似的な拡張のために VTLN を適用し，また，話者の選択についても検討を行った．実験では，JTubeSpeech に対してノイズ，VTLN，および両手法を用いたデータ拡張を行い，話者埋め込みネットワークを用いて話者照合性能を評価した．実験の結果，全ての条件の中でノイズと VTLN の両方を同時に用いた場合が最も高い照合精度となった．

今後の課題として，VTLN 以外の手法を用いた話者数の疑似的な拡張の実装，最先端な話者埋め込みネットワークを用いた実験などが挙げられる．

謝辞　本研究は，JSPS 科研費 21H04900 と ROIS-DS-JOINT(021RP2022)，セコム財団挑戦的研究助成の助成を受けたものである．

参考文献

[1] David Snyder, *et al.* Deep neural network embeddings for text-independent speaker verification. In *Proc. INTERSPEECH*, pp. 999–1003, 2017.

[2] David Snyder, *et al.* X-vectors: Robust dnn embeddings for speaker recognition. In *Proc. ICASSP*, pp. 5329–5333, 2018.

[3] Daniel Garcia-Romero, *et al.* x-vector dnn refinement with full-length recordings for speaker recognition. In *Proc. INTERSPEECH*, pp. 1493–1496, 2019.

[4] Brecht Desplanques, *et al.* Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[5] Shuai Wang, *et al.* Investigation of specaugment for deep speaker embedding learning. In *Proc. ICASSP*, pp. 7139–7143, 2020.

[6] Phani Sankar Nidadavolu, *et al.* Investigation on neural bandwidth extension of telephone speech for improved speaker recognition. In *Proc. ICASSP*, pp. 6111–6115, 2019.

[7] Abhinav Misra and John HL Hansen. Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora. In *Proc. SLT*, pp. 372–377, 2014.

[8] Chia-Ping Chen, *et al.* Speaker characterization using tdnn-lstm based speaker embedding. In *Proc. ICASSP*, pp. 6211–6215, 2019.

[9] Hitoshi Yamamoto, *et al.* Speaker augmentation and bandwidth extension for deep speaker embedding. In *Proc. INTERSPEECH*, pp. 406–410, 2019.

[10] Zhanghao Wu, *et al.* Data augmentation using variational autoencoder for embedding based speaker verification. In *Proc. INTERSPEECH*, pp. 1163–1167, 2019.

[11] David Snyder, *et al.* Speaker recognition for multi-speaker conversations using x-vectors. In *Proc. ICASSP*, pp. 5796–5800, 2019.

[12] Li Lee and Richard Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on speech and audio processing*, Vol. 6, No. 1, pp. 49–60, 1998.

[13] Shinnosuke Takamichi, *et al.* Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification. *arXiv preprint arXiv:2112.09323*, 2021.

[14] David Snyder,*et al.* Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[15] VoicePrivacy 2020. `https://www.voiceprivacychallenge.org/vp2020/`.

[16] Vassil Panayotov, *et al.* Librispeech: an asr corpus based on public domain audio books. In *Proc. ICASSP*, pp. 5206–5210, 2015.

[17] Christophe Veaux, *et al.* Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research*, 2017.