Vol.2020-SLP-131 No.5
2020/2/13

深層学習に基づく話者照合システムのための非学習型帯域拡張法を用いたデータ拡張

宮本 春奈[1,a)]　塩田 さやか[1,b)]　貴家 仁志[1,c)]

概要：本論文では，深層学習に基づく話者照合システムのために非学習型帯域拡張法を適用して生成した広帯域 (wideband; WB) 音声を用いたデータ拡張を提案する．深層ニューラルネットワーク (deep neural network; DNN) を用いた手法の 1 つである x-vector に基づく話者照合システムの学習には大量のデータが必要となる．アメリカ国立標準技術研究所では話者照合のための狭帯域 (narrowband; NB) 音声データベースを多く提供しているが，WB 音声データベースはあまり公開されていない．これまでに，様々なノイズの重畳や帯域拡張データを混ぜ合わせてモデル学習に用いることで x-vector に基づく話者照合システムの性能向上を行う手法が報告されており，DNN に基づく帯域拡張を用いたデータ拡張についても報告されている．しかしながら，DNN に基づく帯域拡張法で生成された高帯域部の情報は少なく，多くの学習データを必要としながらも非学習型の帯域拡張法と品質はあまり変わりがなかった．筆者らはこれまで非学習型の帯域拡張法を NB 音声に適用することで機械学習に有効であることを報告してきた．そこで本論文では，NB 音声データに対して非学習型帯域拡張法を適用した音声を拡張データとして使用した場合の x-vector に基づく話者照合システムの性能評価を行った．実験結果より，データ拡張を行ったシステムはデータ拡張をしないシステムと比べて 22.7%のエラー改善率を得たことを報告する．

キーワード：話者照合，x-vector，帯域拡張，データ拡張

Data augmentation using non-learning-based bandwidth extension for automatic speaker verification based on deep-learning

Abstract: In this research, we propose a data augmentation scheme using wideband (WB) speech generated by non-learning-based bandwidth extension (BWE) methods for deep learning-based automatic speaker verification (ASV). Deep neural network (DNN)-based ASV systems require a large amount of training data for constructing the systems. The national institute of standards and technology provides a large amount of narrowband (NB) speech databases, however, only few WB speech databases are provided for ASV. There are some methods adopting data augmentation with adding noise or BWE for DNN-based ASV systems so far. One of those systems uses a DNN-based BWE method. However, although the DNN-based BWE method requires a large amount of training data, the qualities of generated speeches are almost same as those generated by non-learning-based BWE methods. The authors have been reported that applying the non-learning-based BWE methods to NB speech is effective for machine learning systems. Therefore, in this study, we evaluated the performance of the x-vector-based ASV system adopting the non-learning-based BWE methods as data augmentation. Experimental results showed that the proposed system provided the error reduction of 22.7%, compared with our baseline system.

Keywords: automatic speaker verification, x-vector, bandwidth extension, data augmentation

1. はじめに

近年，音声を利用した生体認証技術である話者照合システムの実用化が進んできている．話者照合システムは，スマートフォンやスマートスピーカなどの音声対話システム

[1] 首都大学東京
6-6 Asahigaoka, Hino-shi, Tokyo 191–0065, Japan
[a)] miyamoto-haruna@ed.tmu.ac.jp
[b)] sayaka@tmu.ac.jp
[c)] kiya@tmu.ac.jp

のアプリケーションと簡単に組み合わせることができるため，話者照合システムの商用目的での使用が特に期待されている．これまで話者照合に関する研究として，因子分析に基づく手法 [1]，深層ニューラルネットワーク (DNN) に基づく手法 [2–4]，確率的線形判別分析 (PLDA) [5] に基づく手法などがこれまで研究されており，話者照合システムの性能向上が報告されている．特に，DNN を用いた手法の 1 つである x-vector に基づく話者照合では，システム性能の大幅な改善を得られたことが報告されている [6]．しかし，x-vector に基づく手法では，高い性能を実現するために大量の学習データが必要となる．話者照合の研究において電話音声などの狭帯域 (narrowband; NB) 環境下で収録された音声データが大量に存在しているが，より高い照合性能を得るために近年は広帯域 (wideband; WB) 音声を使うことが主流となってきている．より多くの WB 音声を用意するために NB 音声に帯域拡張を適用して作った WB 音声を活用する研究がいくつか報告されている [7–10]．その中の一手法である，DNN に基づく帯域拡張法によって生成された WB データを x-vector に基づく話者照合システムのデータ拡張として適用する方法がある [7]．しかしながら，DNN に基づく帯域拡張法で生成された高帯域部の情報は少なく，多くの学習データを必要としながらも非学習型の帯域拡張法と品質はあまり変わりがなかった．

筆者らはこれまでに非学習型の帯域拡張法を NB 音声に適用し生成した WB 音声を用いることが機械学習に有効であることを報告してきた．しかしながら話者照合システムに対して原データが NB 音声を拡張して使用した場合の影響を評価していなかった．そこで本稿では，非学習型の帯域拡張法を x-vector に基づく話者照合システムのデータ拡張として適用することを提案する．システムの構築には非学習型の帯域拡張法を使用することで NB データから生成した WB データと，オリジナルの WB データと組み合わせた学習データを用いる．実験ではベースラインを Voxceleb [11, 12] および Speakers In The Wild (SITW) [13] データベースを用いた x-vector に基づく話者照合システムとし，NB 音声である National institute of standards and technology speaker recognition evaluation (NIST SRE) 2005 と NIST SRE 2006 のデータに対して非学習型の帯域拡張法を用いることで生成した WB データを拡張データとして用いてシステムの構築を行い評価した．実験結果より，データ拡張を行ったシステムはベースラインシステムと比べて 22.7%のエラー改善率を得たことを報告する．

2. x-vector に基づく話者照合

2.1 x-vector とデータ拡張

最新のシステムとして x-vector に基づく話者照合に関する研究が活発に行われている [6]．これは，可変長の発話から固定次元の話者ベクトルにマッピングする DNN を構築し，埋め込み層を用いて話者表現を抽出するものである．x-vector に基づく話者照合システムは，高い照合性能を達成するために大量の学習データが必要となる [6, 7]．そこで x-vetor に基づく手法に用いる学習データを拡張する手法として様々な研究が報告されている [8–10, 14].

2.2 確率的線形判別分析 (Probabilistic linear discriminant analysis; PLDA)

PLDA は抽出された話者ベクトルから話者性に寄与しない情報を低減する手法でありチャネル変動等を軽減することが知られている．また，i-vector や x-vector に基づく手法の識別器としても有効であることが報告されている．x-vector に基づく手法において PLDA のモデルは不特定話者データから次のように求められる．まず発話 u から抽出された x-vector ω_u をその生成過程を無視して式 (1) のように生成されたと考える．

$$\omega_u = \bar{\omega} + \Phi\delta + \Gamma\zeta_u + \epsilon_u. \tag{1}$$

ここで，Φ と Γ は話者とチャネルの部分空間を張る基底行列であり，δ と ζ_u は話者及びチャネル因子を表しており，それぞれ標準正規分布に従う．ϵ_u は残差成分を表し，平均ベクトル $0 \in R^{CD_F}$, 対角共分散行列 $ma \in R^{CD_F \times CD_F}$ のガウス分布に従う．$\bar{\omega}$ は x-vector 空間におけるオフセットである．式 (1) から確率生成モデルを考える．

$$p(\omega_u|\delta, \zeta_u) = N(\bar{\omega} + \Phi\delta + \Gamma\zeta_u,\ \Sigma). \tag{2}$$

式 (2) より登録話者の x-vector ω_1 と照合話者の x-vector ω_2 を用いて ω_1, ω_2 が同一話者モデルから生成されたか (H_1) 否か (H_0) に関する仮説に対して対数尤度比

$$\log \frac{p(\omega_1, \omega_2|H_1)}{p(\omega_1|H_0)p(\omega_2|H_0)} \tag{3}$$

を計算し，照合時のスコアとして用いて評価する．

3. 帯域拡張によるデータ拡張

本稿では，x-vector に基づく話者照合システムのための非学習型帯域拡張法によるデータ拡張の有効性を調査する．

3.1 データ拡張

x-vector に基づく話者照合システムでは DNN を使用するため，大量のデータが必要となる．特に，x-vector に基づく話者照合システムにおいて高い性能を実現するためには，大量の WB の学習データが必要である．しかしながら公開されているデータベースでは WB データの量と種類が十分でないため，性能が制限されてしまうという問題がある．NB データが学習データの一部として利用可能な場合は，x-vector に基づく話者照合システムにおけるデータ量と多様性の問題は緩和されることを除く．NB データと WB データを一緒に使用する，つまり，サンプリング周波数を

揃えるためには NB データをアップサンプリングする必要がある. しかし，アップサンプリングされたデータには高周波帯域の情報が含まれていないため，単純なアップサンプリングと WB データの情報量の違いが大きくなるという課題が挙げられる. 近年，x-vector に基づく話者照合システムにおけるデータ拡張として，NB 音声をアップサンプリングしたデータ，またアップサンプリングしたデータと帯域拡張データとを混ぜ合わせて学習に用いるデータ拡張により照合性能が向上することが報告されている [7]．その際に用いられる帯域拡張法は，DNN に基づく手法となっていた．しかし，DNN による帯域拡張で生成された高帯域部の情報は少なく，また多くの学習データを必要としながらも非学習型の帯域拡張法を品質はあまり変わりがなかった．また，筆者らはこれまでに非学習型の帯域拡張法が機械学習に有効であることを示してきた．本論文では，機械学習に有効である非学習型の帯域拡張法 LPAS と N-BWE をデータ拡張に用いる．

3.2 線形予測分析合成 (Linear prediction based analysis-synthesis; LPAS)

LPAS [15] は，非学習型の帯域拡張法の 1 つであり，線形予測分析を用いて高周波成分を生成する手法である．低周波成分からスペクトルエンベロープおよび残差誤差情報を抽出することで高周波成分を生成している．LPAS は，パワースペクトログラムの不連続性を緩和でき，生成された音声の自然性と明瞭度が高くなることが報告されている [16].

3.3 非線形帯域拡張法（Non-linear bandwidth extension; N-BWE)

非学習型の帯域拡張法の 1 つとして非線形帯域拡張法 (N-BWE) が提案されている [16]．N-BWE は，学習を行わないため処理が非常に軽く，また任意のサンプリング周波数に対応できることである．これまでに，i-vector に基づく話者照合システムと客観評価尺度の 1 つである RMS-LSD において，他の非学習型である帯域拡張法と比べて N-BWE では高い性能を示すことが報告されている [17].

4. 実験

4.1 データベース

本実験では Kaldi-toolkit [18] の SITW データベース [13] を用いた x-vector に基づく話者照合システムの構築，評価を行った．x-vector の抽出器である DNN の構築及び PLDA の推定のための開発用データベースには Voxceleb [11, 12] を用いた．全データのサンプリング周波数は 16 kHz であり，言語は英語である．Voxceleb データベースは 2 つのデータセットで構成されている．1 つ目の Voxceleb1 [11] は話者数 1,251，発話数は 153,516，もう 1 つのセットで

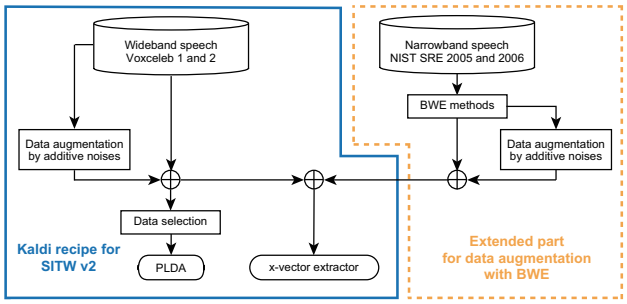

図 1 x-vector に基づく話者照合システムの学習部と帯域拡張によるデータ拡張のブロック図

ある Voxceleb2 [12] は話者数 5,994，発話数は 1,092,009 となっている．これらのデータセットは様々な民族や職業，年齢，アクセントを含むように構成されている．特定話者用のデータベースには SITW を用いた．SITW は収録状況やノイズを後から重畳するなどの制御を行わず，本来の背景ノイズを含む，より実環境に近いデータベースとなっている．SITW は登録データが話者数 199，発話数 1,958 となっており，テストデータは話者数 180，発話数 2,883 がそれぞれ含まれている．SITW と Voxceleb は別々の環境で収録または収集されているが，2 つのデータベースには話者 60 名が重複しているため，学習前に Voxceleb のデータベースから削除した．また，データ拡張の一種として重畳するノイズのデータベースには MUSAN [19] と RIRNOISE [20] を用いた．MUSAN データベースは 900 以上のノイズと 42 時間の様々なジャンルの音楽，12 言語の 60 時間にわたる会話が含まれている．RIRNOISE は部屋の残響ノイズである．

データ拡張用には，サンプリング周波数が 8 kHz の NB データである NIST SRE 2005 [21] と NIST SRE 2006 [22] データベースを使用した．NIST SRE 2005 の発話数は 1,492 文，NIST SRE 2006 では 10,468 文が含まれている．

4.2 実験条件

音響特性にはフレー長が 25ms，フレームシフトが 20ms から得られた対数エネルギーを含む 30 次元の MFCC を使用した．図 1 の青い線で囲まれた部分に，Kaldi-toolkit を使用した x-vector に基づく話者照合システムを学習するためのオリジナルレシピのブロック図を示す．オリジナルレシピでは，Voxceleb データベースから 1,245,525 発話を含む WB データが x-vector の抽出器の学習データとして使用される．また，Voxceleb データにノイズを付加した拡張データが生成される．ノイズ付加によるデータ拡張によって 4,000,000 発話以上のデータが生成され，その内の 1,000,000 発話をランダムに選択し学習データとして使用する．x-vector の抽出器のための学習データとして合計 2,245,525 発話を使用した．また PLDA の学習において，ベースラインシステムでは学習データの合計 2,245,525 発話から発話時間の長い順で整列した場合の先頭 200,000 発

話を用いた．

図 1 の黄色の点線で囲まれた部分には，帯域拡張によるデータ拡張のブロック図を示す．これはオリジナルレシピの拡張部であり，NIST SRE 2005 および NIST SRE 2006 を NB データとして使用した例を示している．拡張部では，帯域拡張による拡張データにもノイズ付加によるデータ拡張を適用する．各比較条件を以下に示す．

(A) 16k

全ての Voxceleb および SITW データを使用して，Kaldi-toolkit のオリジナルレシピで実行した．このシステムを大量の WB データが使用可能な場合と見なした．

(B) 16k (quarter)

オリジナルレシピと同じ手順で WB データが十分に得られない場合のシステムを構築した．WB データの量を 1,245,525 発話から，4 分の 1 の 311,381 発話に減らし，ノイズ付加によるデータ拡張によって 250,000 発話に増やすことで学習データ数は合計 561,381 発話となった. 本実験ではこのシステムをベースラインシステムと見なした.

(C) DA(UP)

x-vector の抽出器の学習の除くシステムの構築方法は (B) と同じとした．x-vector の抽出器を学習するために，NB データに対してアップサンプリングのみを適用した拡張データをベースラインシステムの学習データに追加した．学習データ数は合計 621,181 発話となった．アップサンプリングは，2 点アップサンプリングとローパスフィルタを使用した．

(D) DA(N-BWE)

システムの構築手順は (C) と同じとした． アップサンプリングの代わりに，N-BWE [16] による拡張データをベースラインシステムの学習データに追加し，学習データ数は合計 621,181 発話となった．N-BWE のパラメータ設定は [16] と同じとした．

(E) DA(LPAS)

システムの構築手順は (C) と同じとした．アップサンプリングの代わりに，LPAS [15] による拡張データをベースラインシステムの学習データに追加し，学習データ数は合計 621,181 発話となった．LPAS で使用する各パラメータは [15] と同じとした．

(F) DA(UP&N-BWE)

システムの構築手順は (C) と同じとした．拡張部分では，(C) と (D) における帯域拡張によるデータ拡張を学習データに使用し，学習データ数は合計 742,362 発話となった．

(G) DA(UP&LPAS)

システムの構築手順は (C) と同じとした．拡張部分では，(C) と (E) における帯域拡張によるデータ拡張を学習データに使用し，学習データ数は合計 742,362 発話となった．

(H) DA(N-BWE&LPAS)

システム構築の手順は (C) と同じとした．拡張部分では，(D) と (E) における帯域拡張によるデータ拡張を学習データに使用し，学習データ数は合計 742,362 発話となった．

(I) DA(UP&N-BWE&LPAS)

システム構築の手順は (C) と同じとした．拡張部分では，(C)，(D) および (E) における帯域拡張によるデータ拡張を学習データに使用し，学習データ数は合計 802,162 発話となった．

表 1 各システムで使用された学習データ数 (発話)

	x-vector extractor		PLDA
	Voxceleb	NIST SRE	Voxceleb
(A)	2,245,525	0	200,000
(B)	561,381	59,800	200,000
(C)	561,381	59,800	200,000
(D)	561,381	59,800	200,000
(E)	561,381	119,600	200,000
(F)	561,381	119,600	200,000
(G)	561,381	119,600	200,000
(H)	561,381	119,600	200,000
(I)	561,381	179,400	200,000

表 1 に各システムで使用されるデータ数をまとめる．PLDA の学習に使用した文章数は，(A) と (B) で同じになっているが含まれる発話文は異なる．帯域拡張によるデータ拡張を適用した全システムは，(B) と同じ PLDA 学習モデルを使用した．(C) から (I) のシステムでは NB データベースである NIST SRE 2005 と NIST SRE 2006 データセットを使用した. ノイズ付加による拡張データの合計は，各システムで使用される NB データ数の 4 倍となる．

全てのシステムは，等価エラー率 (equal error rate; EER) と最小検出コスト関数 (minimum detection cost function; minDCF) [23] によって評価した．EER は，false negatives rate (FAR) と false positives rate (FRR) に等しい重みを割り当てスコアを計算する．minDCF は一般的に，低い FAR を達成するよりも低い FRR を達成することが重要であるという考えに基づきシステムの性能を評価する．minDCF IE-2 と minDCF IE-3 との違いは，パラメータ P-target が 0.01 か 0.001 かである．これらのパラメータは NIST SRE evaluation plan によって定義されている．

4.3 実験結果

表 2 に，比較条件ごとの EER と minDCF を示す．(A) 16k と (B) 16k (quarter) を比較すると，学習データが (A) の四分の一の量しかない (B) では EER と minDCF のスコアが大幅に増えてしまっている．このことから，学習データ量が x-vector に基づく話者照合システムの性能に大きく影響することが分かる．(B) 16k (quarter) とデータ拡張を適用し

表 2　各システムの EER (%) と minDCF

x-vector systems / conditions	SITW Core task / Evaluation set		
	EER	minDCF IE-2	minDCF IE-3
(A) 16k	3.554	0.3636	0.5296
(B) 16k (quarter)	6.616	0.5722	0.7862
(C) DA(UP)	5.221	0.4943	0.7139
(D) DA(N-BWE)	5.358	0.5031	0.7249
(E) DA(LPAS)	**5.112**	0.4932	**0.6838**
(F) DA(UP&N-BWE)	5.139	0.4817	0.6961
(G) DA(UP&LPAS)	**5.112**	**0.4556**	0.6913
(H) DA(N-BWE&LPAS)	5.221	0.4787	0.6979
(I) DA(UP&N-BWE&LPAS)	5.522	0.5287	0.7564

たシステム (C)~(I) とを比較すると，データ拡張を適用した全てのシステムの EER が (B) の EER よりも低くなった．帯域拡張によるデータ拡張を適用することで，x-vector に基づく話者照合システムの性能を改善できたことが確認された．次に，(C) DA(UP)，(D) DA(N-BWE)，(E) DA(LPAS) を比較すると，学習データ量はいずれも同じであるが性能は異なり，(E) の LPAS を適用したシステムの EER が最も低くなった．このときのベースラインシステムからのエラー改善率は 22.7%であった．次に (F) DA(UP&N-BWE), (G) DA(UP&LPAS), (H) DA(N-BWE&LPAS) で比較すると，いずれも学習データ量は同じであるが (F) と (G) の性能が良いことがわかる．生成される WB データで比較すると LPAS と N-BWE は高周波帯域に情報が生成されるが，単純なアップサンプリングでは生成されないという違いがある．(F) の UP と N-BWE もしくは (G) の UP と LPAS の組み合わせはデータのバリエーションとしてほぼ同等だと考えられるため性能もほぼ同様になったと考えられる．一方，N-BWE および LPAS によるデータ拡張システムでは性能の改善幅が (F) と (G) より小さい．これは N-BWE と LPAS のデータのバリエーションが近いためであると考えられる．(I) DA(UP&N-BWE&LPAS) は (B) のベースラインシステムよりは EER が低くなったが，データ拡張を行ったシステムの中では最も改善が見られなかった．この結果から，データ拡張がシステムに与える影響はデータ量だけでなく，データのバリエーションに依存することが分かる．さらに一番 EER の低い (E) と (G) を minDCF IE-2, minDCF IE-3 に関して比較すると，(E) は minDCF IE-3 が最も低くなったが minDCF IE-2 の改善は少ない．一方 (G) では minDCF IE-2 は最も低くなり，minDCF IE-3 においてもデータ拡張を行ったシステムの中で 2 番目に低いことから，性能が安定していると考えられる．このことから，バリエーションを考慮し，かつデータ量を増やすことで性能が良くなることが分かる．

データ拡張を PLDA の学習に応用することも試したが，性能の改善は見られなかった．また，(A) のシステムにデータ拡張を施す簡易実験を行ったところ，EER の改善が得られた．

5. まとめ

本論文では，深層学習に基づく話者照合システムに対して非学習型帯域拡張法を適用して生成した WB 音声を用いたデータ拡張の効果を調査した．x-vector に基づく話者照合システムの学習には大量のデータが必要となる．NIST では話者照合のための NB 音声データベースを多く提供しているが，WB 音声データベースはあまり公開されていない．そこで本研究では，NB 音声データに対して非学習型帯域拡張法を適用した音声を拡張データとして使用した場合の x-vector に基づく話者照合システムの性能評価を行った．実験結果より，データ拡張を行ったシステムはデータ拡張をしないシステムと比べて 22.7%のエラー改善率を得たことを報告する．

今後の課題として，今回実験で使用したデータベース以外の NIST SRE によって公開されている NB データベースを x-vector に基づく話者照合システムに適用し，学習データのバリエーションを変えるなどモデル学習の方法を検討することが挙げられる．

謝辞　本研究の一部は JSPS 科研費若手研究 JP19K20271 と ROIS-DS-JOINT(021RP2019) の助成を受けたものである．

参考文献

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” *in Proc. INTERSPEECH*, pp. 999–1003, 2017.

[3] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[4] W. Hsu, Y. Zhang, R. J. Weiss, Y. Chung, Y. Wang, Y. Wu, and J. Glass, “Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[5] S. JD Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” *in Proc. IEEE 11th International Conference on Computer Vision*, pp. 1–8, 2007.

[6] D. Snyder, D. Garcia-Romero, G. Shell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[7] P. S. Nidadavolu, V. Iglesias, J. Villalba, and N. Dehak, “Investigation on neural bandwidth extension of telephone speech for improved speaker recognition,” *in Proc. IEEE International Conference on Acoustics, Speech and Signal Pro-*

cessing(ICASSP), pp. 6111–6115, 2019.

[8] C. Chen, S. Zhang, C. Yeh, J. Wang, T. Wang, and C. Huang, "Speaker characterization using tdnn-lstm based speaker embedding," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

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

[10] Z. Wu, S. Wang, Y. Qian, and K. Yu, "Data augmentation using variational autoencoder for embedding based speaker verification," *in Proc. INTERSPEECH*, pp. 1163–1167, 2019.

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

[12] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[13] M. Mitchell, F. Luciana, C. Diego, and L. Aaron, "The speakers in the wild (sitw) speaker recognition database," *in Proc. INTERSPEECH*, pp. 818–822, 2016.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[15] P. Bachhav, M. Todisco, and N. Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis," *in Proc. IEEE International Conference on Acoustics, Speech and Signal*, pp. 5429–5433, 2018.

[16] H. Miyamoto, S. Shiota, and H. Kiya, "Non-linear harmonic generation based blind bandwidth extension considering aliasing artifacts," *in Proc. APSIPA Annual Summit and Conference*, pp. 1868–1874, 2018.

[17] R. Kaminishi, S. Shiota, and H. Kiya, "Evaluation on non-linear artificial bandwidth extension using i-vector/plda speaker verification," *SIG Technical Reports*, , no. 14, pp. 1–6, 2018.

[18] P. Daniel, G. Arnab, B. Gilles, B. Lukas, G. Ondrej, G. Nagendra, H. Mirko, M. Petr, Q. Yanmin, S. Petr, et al., "The kaldi speech recognition toolkit," *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

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

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

[21] "Nist (2005) the nist year 2005 speaker recognition evaluation plan," `https://catalog.ldc.upenn.edu/docs/LDC2011S01/sre-05_evalplan-v5.pdf`, 2004.

[22] "The nist year 2006 speaker recognition evaluation plan," `https://catalog.ldc.upenn.edu/docs/LDC2011S09/sre-06_evalplan-v9.pdf`, 2006.

[23] "Nist 2016 speaker recognition evaluation plan," `https://www.nist.gov/sites/default/files/documents/itl/iad/mig/SRE16_Eval_Plan_V1-0.pdf`, 2016.