Vol.2019-MUS-123 No.28
Vol.2019-SLP-127 No.28
2019/6/22

話者照合のための非線形帯域拡張法を用いたデータ拡張の検討

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，x-vector に基づく話者照合システムにおいて帯域拡張法を用いて生成した広帯域音声によるデータ拡張に着目する．x-vector に基づく話者照合システムにおけるデータ拡張には，様々なノイズを加えるだけでなく，狭帯域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ものがこれまでに報告されており，さらに DNN による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．一方近年，帯域拡張法の一つとして非線形帯域拡張法 (N-BWE) が提案されている．N-BWE はモデル学習を行わず，計算量が非常に軽い手法として提案された．N-BWE は単純な非線形関数とフィルタのみで構成されているにも関わらず，話者照合の等価エラー率 (EER) と二乗平均平方根対数スペクトル歪みそれぞれにおいて高い性能を得られることが報告されている．そこで本論文では，x-vector に基づく話者照合システムを構築する際に，N-BWE を適用した音声を拡張データとして使用して実験を行った．実験結果より，アップサンプリングした音声と N-BWE で帯域拡張した音声を拡張データとして加えて学習を行った結果，アップサンプリングした音声のみを拡張データとして用いたシステムと比較して EER のエラー改善率は 24.5%を達成した．

キーワード：話者照合，x-vector, 非線形帯域拡張，データ拡張

Investigation on data augmentation using non-linear bandwidth extension for automatic speaker verification

Abstract: This paper focuses on the performance of x-vector based automatic speaker verification (ASV) systems using bandwidth extension (BWE) methods for data augmentation. For the x-vector-based ASV system, data augmentation methods have been reported so far. These reports consider to use large amount of narrowband data. And, upsampling operation and BWE methods are applied to expand the training data. Additionally, deep neural network-based BWE method was used for data augmentation. On the other hand, non-linear bandwidth extension (N-BWE) method has been proposed as one of bandwidth extension methods. N-BWE was proposed as method with light-weight computational cost and non-learning. Although, N-BWE consists only of simple non-linear function and filters, it has been reported that the N-BWE method obtained low equal error rate and small values of root mean square-log spectral distance in some ASV systems. Comparing the performance of x-vector-based ASV systems, some conditions of data augmentation which includes the N-BWE method were carried out. From the experimental results, the method using both upsampled and N-BWE speech as additional training data achieved to 24.5% error reduction.

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

1. はじめに

近年, 声を用いた生体認証技術である話者照合の実用化が進んできている. また, 携帯電話や PC などの普及により音声を入力インターフェースとしたシステムの稼働が容易になってきていることから, 話者照合のさらなる普及が期

[1] 現在，首都大学東京 システムデザイン研究科 情報科学域
Presently with Tokyo Metropolitan University, Faculty School of Systems Design, Department of Computer Science
[a)] miyamoto-haruna@ed.tmu.ac.jp
[b)] sayaka@tmu.ac.jp
[c)] kiya@tmu.ac.jp

待されている．話者照合は i-vector に基づく手法 [1] や，深層学習 (Deep Neural Network; DNN) に基づく手法 [2]，また，probabilistic linear discriminant analysis (PLDA) に基づく手法 [3] などによりその認証精度が非常に向上してきている．

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．これまでに話者照合のタスクとして NIST SRE などから公開されているデータベースの多くが 8 kHz でサンプリングされた音声 (narrowband; NB) データとなっていたが近年は 16 kHz でサンプリングされた音声 (wideband; WB) データを用いる機会が増えてきている．x-vector に基づく手法も WB データを用いて報告がされているが，x-vector の学習には大量のデータが必要となるため，データ拡張手法を使用した研究が報告されている [4, 5]．そこで NB データを DNN に基づく帯域拡張して用いるものも報告されている [6]．DNN による帯域拡張では，高周波成分を予測し生成することで照合性能の向上が確認されているが，音声データのための学習時間を多く要するという課題が挙げられる．近年，帯域拡張法の一つとして非線形帯域拡張法 (Non-linear bandwidth extension; N-BWE) [7] が提案されている．N-BWE は単純な非線形関数とフィルタのみで構成されているにも関わらず，話者照合の等価エラー率 (Equal error rate; EER) と二乗平均平方根対数スペクトル歪みそれぞれにおいて高い性能が得られることが報告されている [8]．

そこで本稿では，NB データに N-BWE を用いて生成した WB データを拡張データとして用い，x-vector に基づく話者照合システムを構築することを検討する．実験結果より，アップサンプリングした音声と N-BWE で帯域拡張した音声を拡張データとして加えて学習を行った結果，アップサンプリングした音声のみを拡張データとして用いたシステムと比較して EER のエラー改善率は 24.5%を達成した．

2. x-vector に基づく話者照合システム

2.1 x-vector とデータ拡張

近年，話者照合における state-of-the-art な手法の一つとして x-vector に基づく手法 [9] が広く用いられている．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．DNN の学習には大量のデータが必要であることが知られており，これまでにも x-vector に基づく手法の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 [4, 5]．

2.2 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 [3]．また，i-vector や x-vector に基づく手法の back-end 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 $\Sigma \in R^{CD_F \times CD_F}$ 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．

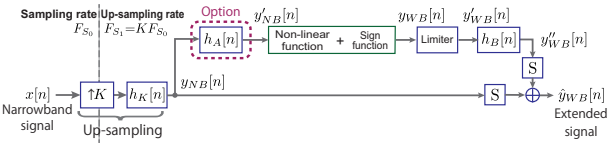


図 1 非線形帯域拡張法のフロー図

3. 非線形帯域拡張法とデータ拡張

3.1 非線形帯域拡張法 (N-BWE)

付帯情報を用いない手法でかつ学習を行わない帯域拡張法として非線形帯域拡張法 (N-BWE) が提案されている [7]．N-BWE の利点は，学習を行わないため処理が非常に軽く，任意のサンプリング周波数に対応できることである．図 1 に N-BWE のブロック図を示す．F_{S_0}Hz でサンプリングされた狭帯域音声 $x[n]$ に対して，インターポレータ K，およびローパスフィルタを用いたアップサンプリングを適用することで，高周波数成分を持たない $y_{NB}[n]$ を生成する．ここで，n は離散時間を表す変数である．次に，アップサンプリングされた信号 $y_{NB}[n]$ に対して式 (4) で表される非線形関数を用いることで高周波数成分が生成される．

$$y_{WB}[n] = \mathrm{sgn}(y'_{NB}[n]) \cdot |y'_{NB}[n]^{\alpha}| \times \beta, \tag{4}$$

ただし，

$$\mathrm{sgn}(a) = \begin{cases} 1 & (a > 0) \\ 0 & (a = 0) \\ -1 & (a < 0) \end{cases}. \tag{5}$$

ここで, α と β は非線形性制御のための任意のパラメーターであり, a は実数である．また, 図 1 の Limiter は以下の式で与えられる．

$$y''_{WB}[n] = \begin{cases} y'_{WB}[n], & y'_{WB}[n] \leq T_h \\ M, & y'_{WB}[n] > T_h \end{cases}. \quad (6)$$

ここで, T_h は閾値, M は定数である．また図 1 の $h_A[n]$ と $h_B[n]$ はそれぞれフィルタを示している．$h_A[n]$ は非線形関数を適用する帯域を選択するためのフィルタであり, $h_B[n]$ は非線形処理を施した音声に生じる低周波成分へのまわりこみなどによるノイズを取り除く目的がある．まわりこみを取り除くことで $y_{NB}[n]$ との足し合わせの際に元の音声を傷つけないためノイズが低減される．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一つである 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 [8].

3.2 データ拡張への応用

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，狭帯域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 [6]．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 による帯域拡張では，ネットワークの学習に大量のデータと計算コストが必要となる．そこで本稿では，非学習型の N-BWE を用いたデータ拡張を行うことを検討する．N-BWE は元音声のフォルマント構造を保持したままの帯域拡張が可能であり，信号の歪みも少ないため N-BWE で性能が改善されればデータ拡張のコストが非常に軽くなると期待できる．

4. 実験

帯域拡張法をデータ拡張に適用した場合の有効性を調査するために，N-BWE 及びアップサンプリングした音声を学習データに加えて x-vector に基づく話者照合システムを構築し，それぞれの EER と音声の客観評価尺度について評価した．

4.1 データベース

本実験では Kaldi-toolkit [10] の SITW データベース [11] を用いたレシピの core タスクを用いて x-vector に基づく話者照合システムの構築を行った．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 [12, 13] 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．Voxceleb1 [12] は話者数 1251, 発話数は 100,000 以上，Voxceleb2 [13] は話者数 6112, 発話数は 1,000,000 以上となっている．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．上記の全データを 8 kHz にダウンサンプリングしたものを NB データとし，それらの音声にローパスフィルタ補間器によるアップサンプリングを施した音声及び N-BWE による帯域拡張音声を拡張データとした．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 と Voxceleb は別々で収集されているが，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付加するノイズのデータベースには MUSAN [14] と RIRNOISE [15] を用いた．MUSAN データベースは 900 以上のノイズと様々なジャンルの音楽，12 言語の会話が含まれている．RIRNOISE は部屋の残響ノイズである．

4.2 実験条件

音響的特徴量には 19 次元の MFCC とその動的特徴量とその 2 次微分を含む 60 次元のベクトルを用いた．フレーム長は 25ms, フレームシフトは 10ms である．x-vector の次元数は 512 次元であり，PLDA の次元数は 150 次元とした．DNN と PLDA の学習に用いるデータ量は kaldi-toolkit のベースラインシステムに則り，DNN は約 2,000,000 発話，PLDA は約 200,000 発話とした．拡張データを使用したシステムでは，ベースラインシステムと比較し学習データの全体数は増加するが，実際に使用される発話数はベースラインシステムと同数となるようランダムに選択し DNN の学習を行った．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約 2,000,000 発話から発話時間の長い順で整列した場合の先頭 200,000 発話を用いており，拡張データを使用したシステムでも同様の条件下でデータの選択を行い，実際に使用されるデータの割合を同じとした．各比較条件を以下に示す．

(A) 8k
16 kHz の原音声から 8 kHz にダウンサンプリングした狭帯域音声 $x[n]$ を学習及びテストデータに用いた．

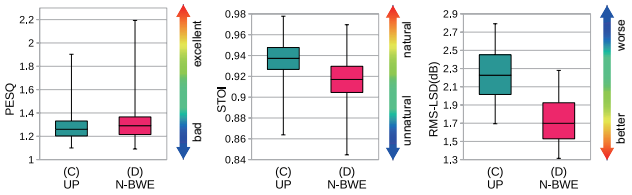
(B) 16k
全ての音声データに 16 kHz の原音声を用いた．

(C) Add(UP)
学習データに原音声と，NB データに対してアップサンプリングのみ行った音声 ($y_{NB}[t]$) を追加した約 3,000,000 発話をデータの全体数としてベースラインシステムの条件と揃うよう選択されたデータで DNN 及び PLDA の学習をしている．

(D) Add(N-BWE)
学習データに原音声，NB データに対して N-BWE を

表 1 x-vector に基づく話者照合実験結果

x-vector systems	評価タスク					
	dev			eval		
	EER	DCF IE-3	DCF IE-2	EER	DCF IE-3	DCF IE-2
(A) 8k	5.198	0.4870	0.6864	5.44	0.5285	0.7510
(B) 16k	3.235	0.2987	0.5039	3.554	0.3636	0.5296
(C) Add(UP)	4.236	0.3777	0.5779	4.593	0.4355	0.6217
(D) Add(N-BWE)	**3.389**	0.3349	0.5495	**3.663**	0.3726	0.5440
(E) Add(UP&N-BWE)	**3.196**	0.3230	0.5359	**3.581**	0.3713	0.5348



(a) PESQ (b) STOI (c) RMS-LSD

図 2 客観評価実験結果

適用した音声 ($y_{WB}[t]$) を追加した約 3,000,000 発話をデータの全体数としてベースラインシステムの条件と揃うよう選択されたデータで DNN 及び PLDA の学習をしている．この際のフィルタ $h_B[n]$ の設計を式 (7) に示す．$h_B[n]$ にはバンドパスフィルタ [7] を用いた．α，β の値は，1.8，100 を用いた．

$$h_A[n] = \begin{cases} 1 & (n = 0) \\ 0 & (n \neq 0) \end{cases}. \quad (7)$$

(E) Add(UP&N-BWE)

学習データに原音声，NB データに対してアップサンプリングのみ行った音声 ($y_{NB}[t]$) と N-BWE を適用した音声を追加した約 4,000,000 発話をデータの全体数としてベースラインシステムの条件と揃うよう選択されたデータで DNN 及び PLDA の学習をしている．

話者照合実験の評価には EER を用いた．客観評価には PESQ [16], STOI [17], RMS-LSD [18] を用いた．PESQ と STOI は原音声と劣化音声を比較することにより，劣化音声の自然性を評価している．PESQ は 0 (bad) から 4.5 (best) で表現され，STOI は 0 (bad) から 1 (best) で表現される．RMS-LSD は原音声と劣化音声間の対数スペクトル距離を示しており，値が低いほど原音声に類似していることを表している．

4.3 実験結果

表 1 に比較条件ごとの EER を示す．まず (A) 8k と (B) 16k を比較すると EER は (B) 16k の方が低い．これより高いサンプリング周波数のデータを用いる方が照合性能が高くなることがわかる．次に (A) 8k と (C) Add(UP) を比較する．(C) Add(UP) は (A) 8k と同じ情報量しか持っていないが，アップサンプリングを行い WB データと混ぜて学習することで照合性能が向上することがわかる．また，(C) Add(UP) と (D) Add(N-BWE) を比較すると，二つの違いは高帯域成分に信号が生成されているか否かであるが，(D) Add(N-BWE) の方が照合性能が高いため，この結果から N-BWE で生成された高帯域成分がモデルの頑健性向上に繋がることを示している．この理由としてアップサンプリングだけの音声は原音声とチャネル情報が大きく異なるため PLDA で緩和できるものの，PLDA の学習があまり上手くいっていないことが考えられる．一方，N-BWE により高帯域成分まで情報を生成した方が PLDA 変動の吸収が上手く働きより頑健になったと考えられる．次に，(C) Add(UP) と (D) Add(N-BWE) と (E) Add(UP&N-BWE) を比較する．これらの違いは拡張データがアップサンプリングのみか N-BWE のみか二つのデータが半分ずつ混ざっているかである．結果より，(E) Add(UP&N-BWE) では (C), (D) より EER が低く，また dev タスクにおいては (B) 16k の原音声のみで学習した場合よりも EER が低くなった．これはアップサンプリングも混ぜることでチャネル変動の頑健性が向上するからだと考えられる．これらの結果より (E) Add(UP&N-BWE) の手法はデータ拡張に貢献できると考えられる．

図 2 は，PESQ，STOI と RMS-LSD を用いた客観評価実験の結果を箱ひげ図で表したものである．箱の上辺と底辺は全結果のうち 5 %~95 %における結果の四分位範囲を，箱の中の線はデータの中央値を示している．箱の上下に伸びる線は全データの最大値と最小値を示す．図 2 (a) PESQ の結果では，帯域拡張することで値が多少高くなったがそもそもの値が低く，図 2 (b) STOI の結果では (D) N-BWE では (C) UP よりも値が低かったことがわかる．N-BWE では，振幅情報を生成するが位相情報を考慮しないために WB 音声を生成した際に，自然性が低下したと考えられる．一方，図 2 (c) RMS-LSD の結果では，(D) N-BWE の誤差が (C) UP よりも小さいという結果になった．この結果からも (E) Add (UP&N-BWE) の EER が改善したのは N-BWE がアップサンプリングと原音声のチャネル変動の中間に位置したからだと考えられる．

5. 結論

本論文では，x-vector に基づく話者照合システムのための N-BWE を用いたデータ拡張を検討した．N-BWE とは帯域拡張法の一つであり，モデル学習を行わず，計算量が非常に軽い手法として提案されている．本稿では，x-vector に基づく話者照合システムにおいて，学習データの母数に NB データをアップサンプリングしただけのデータと組み合わせるデータ拡張及び N-BWE を適用したデータでデータ拡張を行い，EER と客観評価尺度を用いて評価した．実験結果より，原音声である WB データ，NB データをアップサンプリングしたデータと N-BWE を適用したデータの三つ全てのデータを用いたデータ拡張を用いることで，

WB データとアップサンプリングしたデータのみを用いたデータ拡張よりも EER が 24.5%改善することを確認した．

謝辞 本研究の一部はJSPS 科研費若手研究JP19K20271の助成を受けたものである．

参考文献

[1] Najim Dehak, Patrick J Kenny, Rda Dehak, Pierre Dumouchel and Pierre Ouellet: Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing, (2011).
[2] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur: Deep neural network embeddings for text-independent speaker verification, INTERSPEECH, (2017).
[3] Simon JD Prince and James H Elder: Probabilistic linear discriminant analysis for inferences about identity, IEEE 11th International Conference on Computer Vision, (2007).
[4] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey and S. Khudanpur: Speaker Recognition for Multi-speaker Conversations Using X-vectors, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2019).
[5] C. Chen, S. Zhang, C. Yeh, J. Wang, T. Wang and C. Huang: Speaker Characterization Using TDNN-LSTM Based Speaker Embedding, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2019).
[6] Phani Sankar Nidadavolu, Vicente Iglesias, Jess Villalba and Najim Dehak: Investigation on Neural Bandwidth Extension of Telephone Speech for Improved Speaker Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2019).
[7] H. Miyamoto and S. Shiota and H. Kiya: Non-linear harmonic generation based blind bandwidth extension considering aliasing artifacts, in Proc. APSIPA Annual Summit and Conference, (2018).
[8] 上西 遼大，塩田 さやか，貴家 仁志: i-vector/PLDA に基づく話者照合における非線形帯域拡張法の評価, 情報処理学会　音声言語情報処理研究会, (2018).
[9] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur: Spoken language recognition using x-vectors, in Proc. Odyssey, (2018).
[10] P. Daniel, G. Arnab, B. Gilles, B. Lukas, G. Ondrej, G. Nagendra, H. Mirko, M. Petr, Q. Yanmin, S. Petr and others: The Kaldi speech recognition toolkit, IEEE 2011 workshop on automatic speech recognition and understanding, (2011).
[11] M. Mitchell, F. Luciana, C. Diego and L. Aaron: The Speakers in the Wild (SITW) Speaker Recognition Database, INTERSPEECH, (2016).
[12] A. Nagrani, J. S. Chung, and A. Zisserman: Voxceleb: a large-scale speaker identification dataset, INTERSPEECH, (2017).
[13] J. S. Chung, A. Nagrani, and A. Zisserman: Voxceleb2: Deep speaker recognition, INTERSPEECH, (2018).
[14] D. Snyder, G. Chen, and D. Povey: Musan: A music, speech, and noise corpus, arXiv preprint arXiv:1510.08484, (2015).
[15] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur: A study on data augmentation of reverberant speech for robust speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2017).
[16] AW. Rix, J. Beerends, M. Hollier and A. Hekstra: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation, (2001).
[17] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen: An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech, IEEE Trans. Audio, Speach, Language. Process., (2011).
[18] D. Zaykovskiy and B. Iser: Comparison of neural networks and linear mapping in an application for bandwidth extension, in Proc. of SPECOM, (2005).