話者照合のための非学習型帯域拡張法を用いたデータ拡張の検討

Study on data augmentation using non-learning-based bandwidth extension methods for speaker verification

首都大学東京システムデザイン研究科
情報科学域
18860620　宮本 春奈

目 次

あらまし

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．

Summary

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.

1 はじめに

近年，携帯端末のログインや入室管理などにおける本人認証を行う機会が増えている．本人認証では主に，パスワードやIDカードを用いる場合が多いが，IDカードの所持やパスワードの保管など利用者の負担が大きいという問題がある．そのため，指紋や虹彩，声紋などといった人の身体的特徴を用いて個人認証を行う生体認証システムが注目を集めている．身体的特徴の1つである声は，その発話内容や言語に依存せず，発声器官の形状の違いなどの身体的特徴や話し方の癖といった行動的特徴を併せ持っているため，生体認証に用いる特徴量として非常に魅力的である．

音声を鍵として用いる生体認証技術を話者照合と呼ぶ．話者照合はマイクを用いるだけで入力データを取得可能なため実用性が高く，携帯電話などの通信を経由することに関しても親和性が高い．また，機械学習を用いた話者照合として因子分析に基づく手法 [1] や深層ニューラルネットワーク (Deep neural network; DNN) に基づく手法 [2–4]，確率的線形判別分析 (Probabilistic linear discriminant analysis; PLDA) [5]，x-vectorに基づく手法 [6] など数多く研究され，そのシステム性能の改善も報告されている．そのためネットバンキングや携帯電話などのセキュリティシステムとして，またスマートスピーカなどの音声対話システムにおけるユーザ個別サービスを実現するための技術としての使用が期待されている．

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．これは，可変長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．これまでに話者照合のタスクとしてアメリカ国立標準技術研究所 (National institute of standards and technology; NIST) などから公開されているデータベースの多くが8kHz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16kHz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 に基づく話者照合において WB データを用いた手法が報告がされているが，x-vector における DNN 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も多く報告されている．その多くがノイズ重畳等によるデータ拡張を考えており，NB データを活用するものは少ない．NB データを活用する手法の 1 つに DNN を用いた帯域拡張法を適用した方法が報告されている [7]．しかし DNN に基づく帯域拡張法では，NB 音声から生成された WB 音声の品質が学習データに依存し DNN の学習にも多くのデータが必要となる．

帯域拡張法はサンプリング周波数が異なるときにサンプリング周波数が高い方に合わせるために必要となる技術であり，音質改善の点からも重要な技術である．これまで様々な帯域拡張法が提案されてきているが，大きく分けると non-blind 法もしくは blind 法になる．non-blind 法とは，低周波成分と符号化された高周波成分を付帯情報として用いて失われた高帯域成分を再現するもので，blind 法とは低周波成分のみから失われた高周波成分を生成するものである．付帯情報を必要とする non-blind 法より blind 法の方が主に研究されている．また，帯域拡張法は学習型と非学習型の手法に分類することもできる．近年，blind 法でかつ非学習型の帯域拡張法として非線形帯域拡張法 (Non-linear bandwidth extension; N-BWE) [8–10] や Linear prediction basedanalysis-synthesis (LPAS) [11] 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ため任意のサンプリング周波数に対応でき，学習を行わないため処理が非常に軽いにも関わらず，GMM-UBM に基づく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]．LPAS も計算量が多少多いものの明瞭性は高い手法となっている．

そこで本論文では，NB データにこれらの非学習型帯域拡張法を用いて生成した 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でより頑健なシステムを実現する

ことを検討する．実験では，x-vector に基づく話者照合システムのための学習データの影響を調査するために，WB 原音声のみを学習データに使用するベースラインシステムと，NB データに非学習型の帯域拡張法を適用した拡張データを使用したシステムとの比較を行う．システムの有効性を評価するために，Voxceleb [12, 13] および Speaker In The Wild (SITW) [14] データベースを使用した．WB データが少ない場合を想定したシステムを構築するために Voxceleb データセットの 4 分の 1 を使用し，これをベースラインシステムとした．NIST SRE で公開されている話者照合のための NB データセットを用いたシステムを構築し比較実験を行った．また，それぞれの帯域拡張法に関して客観評価実験を行い比較した．客観評価実験では，遅延量，RMS-LSD [15]，PESQ [16]，STOI [17] の四つの尺度を用いた．遅延量は各手法で使用されるフィルタや FFT の点数に依存するものである．実験結果より，帯域拡張による拡張データを学習に加えたシステムは原音声の WB データのみを学習に用いたシステムと比べて 22.7% のエラー改善率を達成したことを報告する．

2 話者照合

2.1 話者照合システム

近年，声を用いた生体認証システム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，特に音声対話システムのセキュリティとしての普及が期待されている．

一般的な話者照合システムは学習部，登録部，照合部とに分かれている．学習部でシステムの大元となる不特定話者モデルの学習を行い，登録部では登録ユーザの音声から特定話者モデルを学習する．照合部では入力音声と登録部で学習した特定話者モデルから照合スコアを計算し，任意で設定した閾値と比較したときに棄却または受理の判断をする．話者照合システムの基本構成を図 1 に示す．

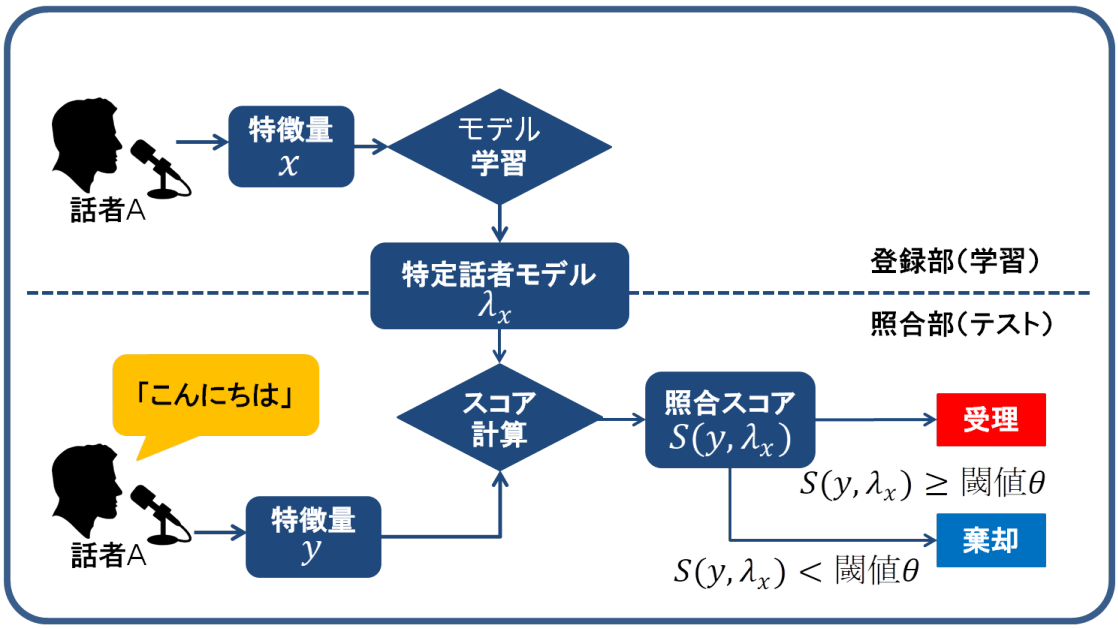


図 1: 話者照合システム

2.2 x-vector に基づく話者照合

話者照合は因子分析に基づく手法 [1] や，DNN に基づく手法 [2]，確率的線形判別分析 (probabilistic linear discriminant analysis; PLDA) に基づく手法 [5] 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．最新のシステムとして 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]．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のフロー図を図 2 に示す. 図 2 の DNN が埋め込み層を含む不特定話者モデルになっており，そこから抽出される話者表現が登録部及び照合部にある x-vector を示している．DNN の学習には大量のデータが必要であり, これまでにも x-vetor 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 [7, 18]. また，照合部では一般的に PLDA を用いている．

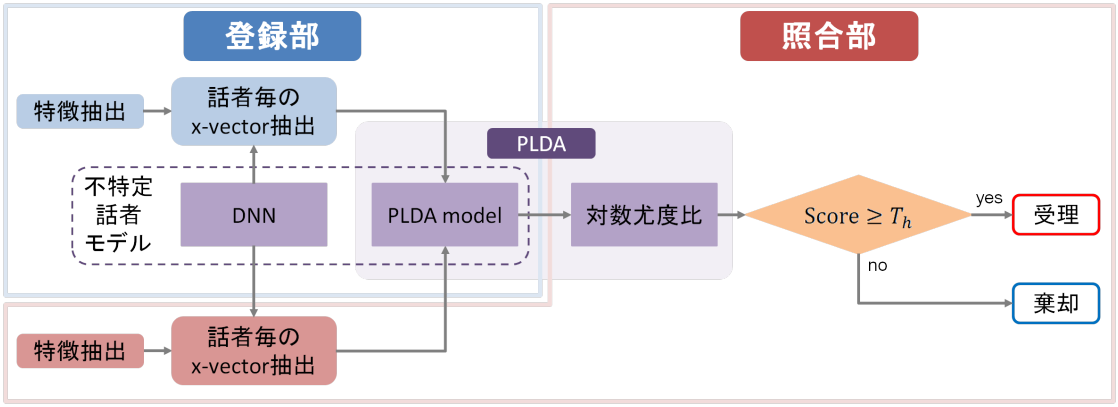


図 2: x-vector に基づく話者照合システム

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

PLDA [5] 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 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 $ma \in R^{CD_F \times CD_F}$ 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 データ拡張と帯域拡張

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 [6, 7, 19]．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と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 [11] は、blind および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 [8].

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

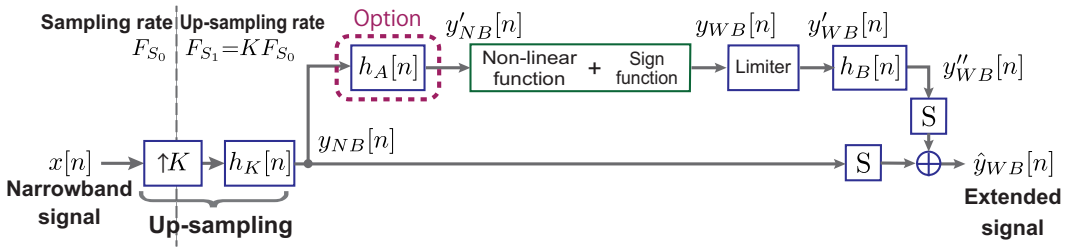


図 3: 非線形帯域拡張法のフロー図

blind かつ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 [8–10]．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 3 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 3 の Limiter は以下の式で与えられる.

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

ここで, T_h は閾値, M は定数である. また図 3 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 [20].

4 帯域拡張の評価

4.1 スペクトログラム

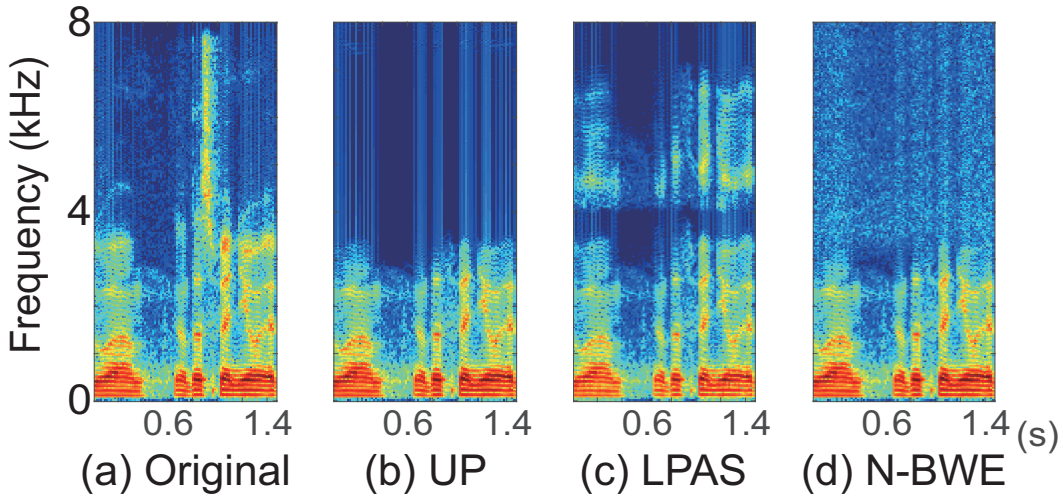


図 4: スペクトログラム図

図 4 に，16 kHz でサンプリングされた原音声 (a) Original，NB から WB にアップサンプリングされた音声 (b) UP，LPAS による WB 音声 (c) LPAS，N-BWE により生成された WB 音声 (d) N-BWE それぞれのスペクトログラムを示す．図より，(c) と (d) では高周波成分が生成されていることが分かる．これらの手法から生成された WB 信号は NB 信号から生成されるため，明瞭度だけでなく話者性も向上する．

4.2 遅延

非学習型の帯域拡張法である N-BWE と LPAS に関してアップサンプリングした音声との品質を比較するために，遅延量に関する評価と客観評価実験を行った．

近年，ネットワーク環境の普及により，ビデオ通話やリアルタイム通信を行うアプリケーションが広く使用されるようになった．ビデオ電話のようなリアルタイム通信を行うアプリケーションには，送

信される信号のサンプリング周波数に合わせた帯域拡張法が必要となる．また，帯域拡張された音声は自然性だけでなく，音声認識や話者照合などの音声信号処理システムにおいても高い精度を維持することが求められる．実際に，視覚情報と聴覚情報との差が 10 ms 以上になると使用者が不快感を感じるという報告がある [21]．そのため，音声強調や帯域拡張等の音声の品質改善手法をユーザに提供するためには，非常に低遅延な手法が求められる．

表 1 に各手法ごとの遅延量を示す．アルゴリズムとしての遅延量は使用するフィルタの次数に大きく依存する．結果より，(C) N-BWE が最も遅延量が少なかった．これは (C) では使用しているフィルタが，次数の小さいバンドパスフィルタ 1 つのみであり，遅延があまり生じないためである．(B) LPAS はフィルタだけではなく FFT の点数が遅延に影響するために遅延量が 10 ms を超えてしまっている．

表 1: 各手法ごとの遅延量

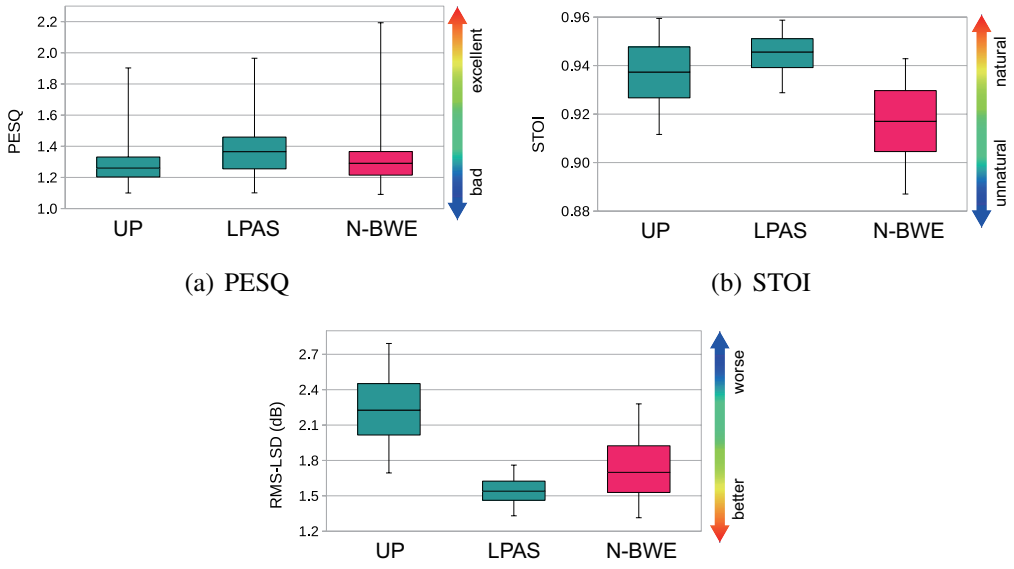
Compared method	Latency (ms)
(A) UP	0.068
(B) LPAS	14.187
(C) N-BWE	**0.443**

4.3 客観評価尺度

客観評価実験には Perceptual evaluation of speech quality (PESQ) [16], Short-time objective intelligibility (STOI) [17], Root mean square - log spectral distance (RMS-LSD) [15] の 3 つの客観評価尺度を用いた．また，各手法ごとの遅延量についても評価を行った．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)~4.5 (best), STOI は 0 (bad)~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．

図 5 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 5(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 5(b) STOI の結果では N-BWE が 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 5(c) RMS-LSD の結果では，N-BWE の誤差が UP よりも小さいという結果になった．

x-vector に基づく話者照合実験では，RMS-LSD が改善する N-BWE と自然性と明瞭性が良くなる LPAS の 2 つの非学習型帯域拡張法を用いた．



(a) PESQ

(b) STOI

(c) RMS-LSD

図 5: 客観評価実験結果

5 実験

5.1 データベース

本実験では Kaldi-toolkit [22] の SITW データベース [14] 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 [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．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 [12] は話者数 1,251，発話数は 153,516，もう 1 つのセットである Voxceleb2 [13] 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 [23] と RIRNOISE [24] 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．source-noises，real-rirs-isotropic-noises, simulated-rirs の 3 つのデータベースから構成されている．本実験では simulated-rirs のみ使用した．

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 [25] と NIST SRE 2006 [26] データベースを使用した．National institute of standards and technology speaker recognition evaluation (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．

5.2 実験条件

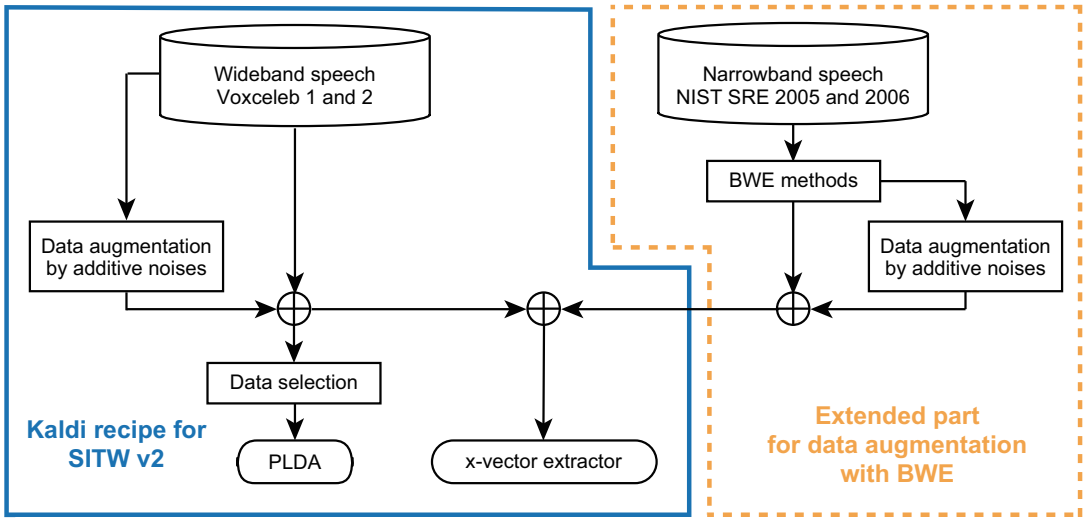


図 6: x-vector に基づく話者照合システムの学習部と帯域拡張によるデータ拡張のブロック図

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 6 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 6 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 の抽出器の学習を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 [8] 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 [8] 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 [11] 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 [11] 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 2: 各システムで使用されたデータ数 (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 2 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) [27] 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．

5.3 実験結果

表 3: 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

表 3 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．

6 まとめ

本論文では，非学習型帯域拡張法に関する研究及びその応用法として，話者照合のためのデータ拡張に関する研究を行った．DNN を使用した x-vector に基づく話者照合システムでは大量の WB 学習データが必要となる．NIST では話者照合のための NB データベースを多く提供しているが，WB データベースはあまり公開されていない．帯域拡張法をデータ拡張として使用することで NB データを有効活用することが可能となれば，話者照合のためのデータ拡張における課題を緩和することが可能となる．そこで本論文では，NB 音声データに対し非学習型の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．

7 謝辞

本研究は，著者が首都大学東京システムデザイン研究科情報科学域において，多くの方々のご指導，ご協力のもとで行われたものです．

まず，指導教員である貴家仁志教授，塩田さやか助教には，本研究の全般にわたり，その進行，執筆，発表に関して詳細なご指導をしていただきました．特に塩田さやか助教には，本研究のみならず，研究に対する姿勢から資料作成において多くの貴重な助言，ご指導をいただきました．ここに深く感謝いたします．また，小野順貴教授，高間康史教授には本論文の審査を通して貴重なご助言とご指導を賜り，深く感謝の意を表します．そして，本研究を進めるにあたりお世話になった先輩，同輩方にも感謝いたします．

最後に，これまでの学生生活を理解し，支援してくださった両親に，厚く御礼申し上げます．

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. 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. ICASSP*, 2019.

[5] S. JD Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *in Proc. ICASSP*, 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. 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. ICASSP*, pp. 6111–6115, 2019.

[8] 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.

[9] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Investigation on blind bandwidth extension with a non-linear function and its evaluation on x-vector-based speaker verification," *in Proc. INTERSPEECH*, pp. 4055–4059, 2019.

[10] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Blind bandwidth extension with a non-linear function and its evaluation on automatic speaker verification," *IEICE trans. Inf. Sys*, vol. E103-D, 2019.

[11] 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.

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

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

[14] 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.

[15] D. Zaykovskiy and B. Iser, "Comparison of neural networks and linear mapping in an application for bandwidth extension," *in Proc. of SPECOM*, pp. 1–4, 2005.

[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*, vol. 862, 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.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[18] 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.

[19] 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.

[20] 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.

[21] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Jounal pf the American Academiy Audiology 11*, pp. 330–336, 2000.

[22] 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.

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

[24] 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. ICASSP*, pp. 5220–5224, 2017.

[25] “The nist year 2005 speaker recognition evaluation plan,” https://catalog.ldc.upenn.edu/docs/LDC2011S01/sre-05_evalplan-v5.pdf, 2004.

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

[27] “Nist 2016 speaker recognition evaluation plan,” https://www.nist.gov/system/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf, 2016.

発表論文

1. 宮本 春奈, 塩田 さやか, 貴家 仁志, “ 話者照合のための低周波成分への影響を考慮した非線形帯域拡張とその客観評価, ” 日本音響学会春季大会, no.2-8-2, pp.31–34, 2018 年 3 月 14 日.

2. 宮本 春奈, 塩田 さやか, 貴家 仁志, “超広帯域音声のための低周波成分への影響を考慮した非線形帯域拡張法に基づく話者照合の検討, ” 電子情報通信学会 音声研究会, vol.117, no.517, (no.SP2017-93), pp.51–55, 2018 年 3 月 19 日.

3. Haruna MIYAMOTO, Sayaka SHIOTA, and Hitoshi KIYA, “ Nonlinear Harmonic Generation Based Blind Bandwidth Extension Considering Aliasing Artifacts, ” Proc. APSIPA Annual Summit and Conference, Honolulu, Hawaii, USA, 15th November, 2018.

4. 宮本 春奈, 塩田 さやか, 貴家 仁志, “ 非線形帯域拡張法における客観評価尺度と 遅延時間の評価, ” 日本音響学会春季大会, no.2-P-25, pp.1011–1014, 2019 年 3 月 6 日.

5. 宮本 春奈, 塩田 さやか, 貴家 仁志, “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, ” 情報処理学会 音声言語情報処理研究会, vol.2019-SLP-127, no.28, pp.1–5, 2019 年 6 月 22 日.

6. Ryota KAMINISHI, Haruna MIYAMOTO, Sayaka SHIOTA, and Hitoshi KIYA, “ Blind bandwidth extension with a non-linear function and its evaluation on x-vector-based speaker verification , ” Proc. ISCA International Conference on Interspeech, 15th September, 2019.

7. Haruna MIYAMOTO, Sayaka SHIOTA, and Hitoshi KIYA, “ Investigation on Latency Issues and Objective Measurements of Non-Linear Blind Bandwidth Extension, ” Proc. IEEE Global Conference

on Consumer Electronics, pp.712–714, Osaka, Japan, 17th October, 2019.

8. Ryota KAMINISHI, Haruna MIYAMOTO, Sayaka SHIOTA, and Hitoshi KIYA, “ Blind bandwidth extension with a non-linear function and its evaluation on automatic speaker verification, ” IEICE Trans. Inf. & Sys., vol.E103-D, no.1, January 2020.

9. 宮本 春奈, 塩田 さやか, 貴家 仁志, “ 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,”情報処理学会 音声言語情報処理研究会, 2020 年 2 月 13 日.

著者紹介

宮本 春奈

平　8 (1996) 年 3 月　神奈川県生まれ.
平　23 (2011) 年 3 月　稲城市立稲城第一中学校卒業.
平　26 (2014) 年 3 月　私立八王子学園八王子高等学校卒業.
平　30 (2018) 年 3 月　首都大学東京 システムデザイン部卒業.
令　2 (2020) 年 3 月　首都大学東京大学院 システムデザイン研究科 情報科学域 博士前期課程 修了見込.