i-vector/PLDA に基づく話者照合による非線形帯域拡張法の評価

上西 遼大[1,a)]　塩田 さやか[1]　貴家 仁志[1]

概要：本論文は, i-vector/PLDA に基づく話者照合システムを用いて非線形帯域拡張（N-Bwe）法を評価することを目的としている. N-Bwe 法とは帯域拡張法の一つで, 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かかわらず, GMM-UBM に基づく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（RMS-LSD）において高い性能を得られることが報告されている. PLDA に基づく話者照合は話者とチャネルの依存性を分離することに焦点を当てているが, 帯域制限による劣化音声を用いた場合については議論されていない. そこで本論文では, PLDA に基づく話者照合システムを構築し, N-Bwe や他の帯域拡張法を用いることでサンプリング周波数の違いによる帯域制限のかかった音声がシステムに与える影響について調査し, 帯域拡張を適用した音声の客観評価と EER の関係を考察した. 実験結果より, N-Bwe で生成された音声は低い RMS-LSD を得られ, かつアップサンプリングのみを行なった音声と比較して EER が 1.78 ポイント改善したことを報告する.

キーワード：話者照合, i-vector, PLDA, 非線形帯域拡張, SITW

EVALUATION ON NON-LINEAR ARTIFICIAL BANDWIDTH EXTENSION USING I-VECTOR/PLDA SPEAKER VERIFICATION

Kaminishi Ryota[1,a)]　Shiota Sayaka[1]　Kiya Hitoshi[1]

Abstract: This paper aims to evaluate an effect of a non-linear bandwidth extension (N-Bwe) method by using i-vector/PLDA-based automatic speaker verification (ASV) systems. The N-Bwe method has been reported as a blind, non-learning and light-weight BWE approach. Although the N-Bwe method consists of a simple non-linear function and filters, it has archived high accuracy in terms of speaker individuality and root mean square log-spectral distortion (RMS-LSD). Recently, i-vector/PLDA-based ASV systems become one of the state-of-the-art ASV systems. While the PLDA-based ASV approaches focus on removing speaker and channel dependency, there are few discussions about speeches which degraded by band limits. Thus, this paper investigates the influence of the speech degradation by band limits toward the PLDA-based ASV systems. In the experiments, the N-Bwe and shift-based BWE methods were evaluated by the PLDA-based ASV systems. From the results, the N-Bwe method improved 1.78 points of equal error rate (EER) from the simply up-sampled situation.

Keywords: Speaker verification, i-vector, PLDA, Non-linear bandwidth extension, SITW

1. はじめに

話者照合とは, ユーザーの声を用いた生体認証技術のことである. 最先端の話者照合システムとして i-vector に基づく手法 [1–3], probabilistic linear discriminant analysis (PLDA) に基づく手法 [4–6], x-vector に基づく手法 [7–9] がなどが提案されている. これらの手法はアメリカ国立標準技術研究所 (NIST) から公開されている speaker recog-

[1] 首都大学東京
[a)] kaminishi-ryota@ed.tmu.ac.jp

nition evaluation (SRE) シリーズや Speaker In the Wild (SITW) と呼ばれる世界標準のデータベースを用いて評価され, 高い性能を得られることが示されている. PLDA に基づく手法は話者とチャネルの依存性を分離することに焦点を当てた手法である. これらの話者照合システムは機械学習に基づいているため, 学習データとテストデータのサンプリング周波数が同じであると想定している. 学習及びテストデータのサンプリング周波数が異なる場合, 一般的にはサンプリング周波数が高い方をダウンサンプリングさせ, 低いサンプリング周波数に合わせる. しかし, 入力されるテストデータのサンプリング周波数が低いため, 全ての学習データをダウンサンプリングさせて話者照合システムを再び構築しなおすには高いコストがかかるという問題点がある. テストデータのサンプリング周波数が低い場合, アップサンプリングを適用してサンプリング周波数を学習データに合わせることも可能である. しかしアップサンプリングだけでは、帯域制限の影響が残るため話者照合性能が低下してしまうことが知られている [10,11].

帯域拡張法は帯域制限などにより高周波数成分が欠落しているデータに対して高周波数を復元する技術の一である [12–16]. これまでに多くの帯域拡張法が提案されているが大まかには, 付帯情報を用いる手法と用いない手法に分類される. 付帯情報を用いない手法は低周波数成分のみを用いて高周波数成分を推定するものである. 近年, 付帯情報を用いず, 学習を行なわない, かつ計算量が軽い手法として非線形帯域拡張法 (N-Bwe) が提案された [10]. N-Bwe は単純な非線形関数で構成されているにもかかわらず, GMM-UBM に基づく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（RMS-LSD）において高い性能を得られたことが報告されている.

近年, 複数の帯域制限が混合しているデータを用いてモデル学習を行う話者照合システムが報告されている [11,17]. しかし, 帯域拡張法を用いた場合の話者照合システムの影響を調査したものはほとんどない. そのため本論文では付帯情報を用いない帯域拡張法に焦点を当て, 最先端の話者照合システムへの影響を調査する. 実験では, PLDA に基づく話者照合システムを構築し, N-Bwe や他の帯域拡張法を用いた場合にシステムに与える影響について調査し, 帯域拡張を適用した音声の客観評価と EER の関係を考察した. 実験結果より, N-Bwe で生成された音声は低い RMS-LSD を得られ, かつアップサンプリングのみを行なった音声と比較して EER が 1.78 ポイント改善したことを報告する.

2. i-vector/PLDA に基づく話者照合システム

2.1 i-vector

近年, i-vector に基づく話者照合システムは最新のシステムの一つとしてみなされている [1–3]. i-vector における話者モデルは式 (1) によって定義される.

$$M_u = m_{ubm} + T\omega_u \tag{1}$$

ここで, $m_{ubm} \in R^{CD_F}$, $T \in R^{CD_F \times D_T}$ はそれぞれ universal background model (UBM) の GMM スーパーベクトル, 全変動 (TV) 行列呼ばれており, また, C は混合数であり, D は音響特徴量の次元数を表す. $\omega_u \in R^{D_T}$ は発話ごとに与えられる確率変数であり, 平均ベクトルが $0 \in R^{D_T}$ で共分散行列が単位行列 $I \in R^{D_T \times D_T}$ のガウス分布 $N(\omega; 0, I)$ に従う. 照合時には, 登録話者と照合話者の i-vector ω_1, 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否かの対数尤度比を計算することで照合性能を評価する.

2.2 PLDA [4]

この章では i-vector に基づく話者照合のための PLDA について説明する. PLDA は, 発話から抽出された i-vector をその生成過程を無視し, 確率変数とみなして式 (2) のように表現することを考える.

$$\omega_u = \bar{\omega} + \Phi\delta + \Gamma\zeta_u + \epsilon_u \tag{2}$$

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 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. 話者照合システムにおいて話者性とチャネル変動は照合性能に大きく影響があるため, PLDA を用いてそれらを分けることで性能を改善できることが知られている.

3. 非線形帯域拡張法

3.1 付帯情報を用いない帯域拡張法

帯域拡張法としてこれまでに多くの手法が報告されている. これらの手法は付帯情報を用いるか用いないかに分類することができる. 本論文では付帯情報を用いず, かつ学習を行わない帯域拡張法に焦点を当てる. 一般的な帯域拡張ではインターポレータとローパスフィルタによるアップサンプリングを狭帯域音声に適用し, 高周波域を持たないアップサンプリング音声を生成する. 付帯情報を用いない帯域拡張法ではアップサンプリングによりできた空の高周波成分を低周波数成分のみで補うことを目的としている.

3.2 線形予測分析合成法 (LPAS)

これまでに低周波数帯域のスペクトルを高周波数帯域へシフトさせる帯域拡張法が提案されている [18,19]. これらの手法では, 広帯域信号は 4 kHz 未満の信号を変調し高周波数に移動することで生成される. また, シフトベースの手法の品質を改善するためにシフトに基づく手法を拡張した LPAS [20] が提案された. LPAS は狭帯域信号からスペクトル包絡線および残留誤差情報から抽出された高周波数成分を用いて広帯域信号生成する手法である. 生成された

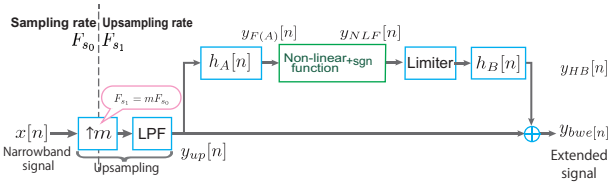

図 1: 非線形帯域拡張法のブロック図

高周波成分は単純にシフトされたものよりも自然なものになることが報告されている. 本論文は LPAS を帯域拡張法の従来法とみなす.

3.3 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に非線形帯域拡張法 (N-Bwe) が提案されている [10].

図 1 は N-Bwe 法のブロック図を示している. 図に示すように, 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 m, 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 $y_{up}[n]$ を生成する. ここで, n は離散時間変数である. 次に, アップサンプリングされた信号に対して式 (3) 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.

$$y_{NLF}[n] = \mathrm{sgn}(y_{F(A)}[n]) \cdot |y_{F(A)}[n]^{\alpha}| \times \beta. \qquad (3)$$

ただし,

$$\mathrm{sgn}(a) = \begin{cases} 1 & (a > 0) \\ 0 & (a = 0) \\ -1 & (a < 0) \end{cases}. \qquad (4)$$

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.

$$y_{HB}[n] = \begin{cases} y_{NLF}[n], & y_{NLF}[n] \le T_h \\ M, & y_{NLF}[n] > T_h \end{cases}. \qquad (5)$$

ここで, T_h は閾値, M は定数である.

3.4 PLDA に基づく話者照合と N-Bwe

帯域拡張法は帯域制限による高周波成分を補うことを目的としている. PLDA に基づく話者照合システムは話者とチャネル変動を取り除くことに重点を置いている. しかし, これまでに帯域制限によって失った変動については議論されていない. そこで本論文では, 帯域制限により劣化した音声を用い帯域拡張法と PLDA システムの性能について調査し, 議論する.

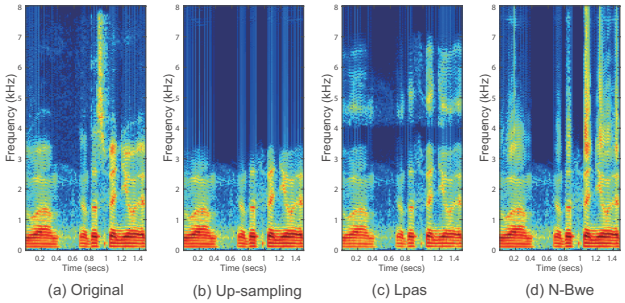

図 2: スペクトログラム ($m = 2$; $F_{S_0} = 8$kHz, $F_{S_1} = 16$kHz)

3.5 帯域拡張法のスペクトログラムによる比較

図 2 は, 原音声, アップサンプリング, LPAS, N-Bwe による音声信号のスペクトログラムを示している. まず, 16 kHz でサンプリングされた原音声の信号 (a) は 0 kHz から 8 kHz までの周波数成分を有していることがわかる. 次, に原音声のサンプリング周波数を 16 kHz から 8 kHz に落とし, また 8 kHz から 16kHz にアップサンプリングした音声が図 2 の (b) である. 図からもわかるように 4 kHz 以上の高い周波数成分を含んでいない. 信号 (c) は LPAS [21] によって生成された音声, 信号 (d) は N-Bwe で生成された音声である. (c), (d) から帯域拡張法によってアップサンプリングではなかった高周波数成分が生成されていることがわかる.

4. 実験

N-Bwe の有効性を評価するために i-vector/PLDA に基づく話者照合実験行い, また, その EER と生成した音声を客観評価尺度で評価したスコアとの関係について調査した.

4.1 データベースの詳細

本実験では Kaldi-toolkit [22] と SITW データベース [23] を用いて i-vector/PLDA に基づく話者照合システムの構築を行なった. その際, UBM, PLDA, TV 行列を推定するために Voxceleb データベースを用いた. Voxceleb データベースは二つのデータセット Voxceleb1 [24], Voxceleb2 [25] で構成されており, どちらのデータセットも Youtube にアップロードされた著名人のインタビュービデオから収集されている. Voxceleb1 は話者数 1251, 発話数は 100,000 以上, Voxceleb2 は話者数 6112, 発話数は 1,000,000 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, アクセントで構成されている. 登録及びテスト用のデータベースには SITW を用いた. SITW は収録状況を制御したデータベースではなく, 本来の背景ノイズ等を含み, より実環境に近いデータベースとなっている. また, ノイズ用のデータベースとして MUSAN [26] と RIRNOISE [27] 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

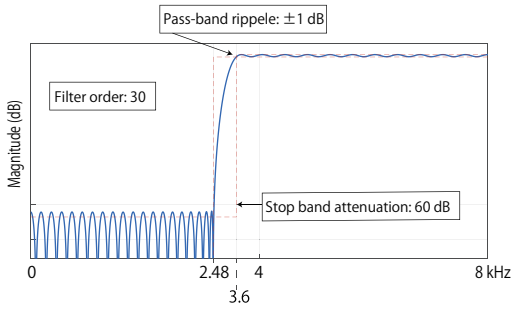

図 3: N-Bwe における $h_B[n]$ のフィルタ設計

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. ノイズデータベース以外の全てのデータベースの言語は英語であり, 16 kHz でサンプリングされている. 本実験でサンプリング周波数が 8 kHz となっている狭帯域音声は全て原音声の 16 kHz から 8 kHz へのダウンサンプリングしたものを表す.

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. UBM の混合数は 1024, i-vector の次元数は 400 次元であり, PLDA の次元数は 150 次元である. 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 1,000,000 以上の発話を学習することは非常に時間を要するため, 1,000,000 のうち 100,000 発話を用いて UBM と TV 行列を学習した. PLDA も同様に Voxceleb データセットにノイズを付与した音声を用いて学習した. 話者照合実験の評価には等価エラー率 (EER) を用いた. 表 1 に比較手法をまとめた. 詳細は以下の通りである.

(A) Up (train)
狭帯域音声 (8 kHz サンプリング) に対してアップサンプリングのみを行なった音声 ($y_{up}[n]$) を登録及びテストデータとして用いた.

(B) Lpas (train)
狭帯域音声に LPAS [20] を適用した音声を登録及びテストデータとして用いた.

(C) N-Bwe (train)
狭帯域音声に N-Bwe [10] を適用した音声を登録及びテストデータとして用いた. フィルタ $h_A[n]$ には以下の式 (6) を用いた. フィルタ $h_B[n]$ は図 3 のように定義した.

$$h_A[n] = \begin{cases} 1 & (n = 0) \\ 0 & (n \neq 0) \end{cases}. \qquad (6)$$

非線形関数 (式 (3)) の α と β はそれぞれ 2 と 100,000 とした.

(D) Up (test)
狭帯域音声 (8 kHz サンプリング) に対してアップサンプリングのみを行なった音声 ($y_{up}[n]$) をテストデータとして用いた. 登録データは 16kHz の原音声である.

(E) Lpas (test)
狭帯域音声に LPAS [20] を適用した音声をテストデータとして用いた. 登録データは 16kHz の原音声である.

(F) N-Bwe (test)
狭帯域音声に N-Bwe [10] を適用した音声をテストデータとして用いた. フィルタ $h_A[n]$ には上記の式 (6) を用いた. フィルタ $h_B[n]$ は図 3 のように定義した. 非線形関数式 ((3)) の α と β はそれぞれ 2 と 100,000 とした. 登録データは 16kHz の原音声である.

(G) Down
16 kHz の原音声から 8 kHz にダウンサンプリングされた狭帯域音声 $x[n]$ を登録及びテストデータとして用いた.

(H) Org
全ての音声データは 16 kHz の原音声である.

表 1: 比較手法

	登録データ	テストデータ
(A)Up (train)	アップサンプリング	アップサンプリング
(B)Lpas (train)	Lpas	Lpas
(C)N-Bwe (train)	N-Bwe	N-Bwe
(D)Up (trial)	原音声 (16k)	アップサンプリング
(E)Lpas (trial)	原音声 (16k)	Lpas
(F)N-Bwe (trial)	原音声 (16k)	N-Bwe
(G)Down	ダウンサンプリング	ダウンサンプリング
(H)Org	原音声 (16k)	原音声 (16k)

手法毎に UBM, TV 行列, PLDA を学習し直すことはコストが非常にかかってしまうため現実的ではない. そのため, 本実験では UBM, TV 行列, PLDA の推定には 16 kHz でサンプリングされた原音声を用いた. (G) Down に関してのみ UBM, TV 行列, PLDA に用いた音声データは 8 kHz にダウンサンプリングされたデータを用いた. 本論文では二つのシナリオを調査した. 一つ目は登録データ, テストデータ共にサンプリング周波数が異なる場合であり, 二つ目はテストデータのみがサンプリング周波数が異なる場合である.

客観評価には PESQ [28], STOI [29], RMS-LSD 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) までで表現され, 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

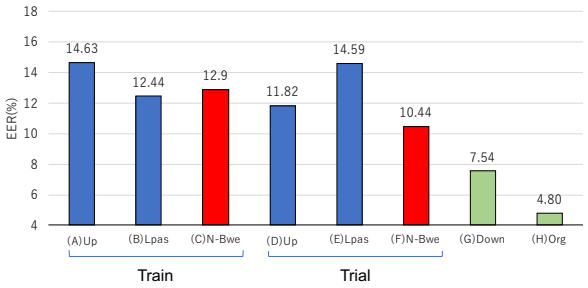

図 4: 話者照合実験結果 (Development タスク)

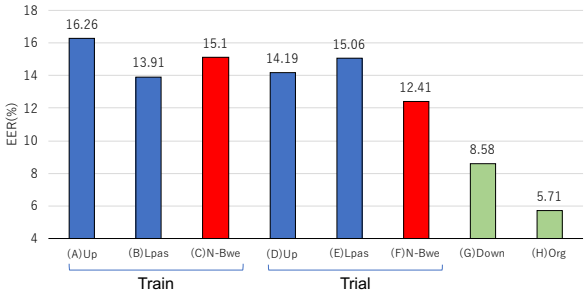

図 5: 話者照合実験結果 (Evaluation タスク)

ペクトル距離を示しており, 値が低いほど距離が近いことを表している.

4.3 実験結果

図 4, 5 に手法ごとの EER を示す. 図 4, 5 では評価タスクが異なるものの, ほぼ同じ傾向が得られた. そこで図 4 を用いて結果を考察する. まず (G) Down (8k) と (H) Org (16k) を比較すると EER は (H) Org (16k) の方が低い. これよりサンプリング周波数が高い方が照合性能が高いことがわかる. 次に (G) Org (16k) と (A) Up (train) を比較する. (A) Up (train) はアップサンプリングによりサンプリング周波数は原音声と揃えたものの, 高帯域成分に情報を持っていない. このことから高帯域成分の有無が話者照合の照合性能に大きく影響を与えることを確認した. また, (A) Up (train) と (B) Lpas (train) 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, (B) Lpas (train) の方が照合性能が高いため, この結果から話者照合において高帯域成分が重要であるいえる. 次に, (A) Up (train) と (C) N-Bwe (train) を比較すると (B) Up のときと同様に (C) N-Bwe (train) の方が照合性能が改善されていることが確認できる. 次に, (A) Up (train) と (D) Up (test), (C) N-Bwe (train) と (F) N-Bwe (test) を比較する. これらの違いはテストデータのみに処理を施したか, テストデータ及び登録データ両方に処理を施したかの違いであるが, 図 4, 5 どちらの場合においてもテストデータのみの方が照合性能が良い. これより UBM や TV の学習データと特定話者モデルの登録データで大幅な劣化がある場合にモデルの学習がうまくいかず EER が低下すると考えられる. 次に, (D) Up (test) と (E) Lpas (test) を見てみると, (E) Lpas (test) よりも (D) Up (test) の方が精度が良い. また (D) Up (test) と (F) N-Bwe (test) を見てみると, (D) Up (test) の手法よりも (F) N-Bwe (test) の方が精度が良い. これらの結果より (F) N-Bwe (test) の手法は照合性能を改善できていると考えられる.

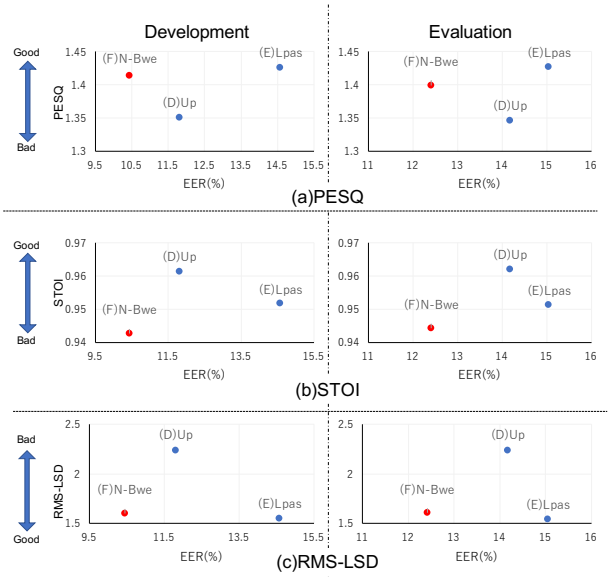

図 6: EER と客観評価値の比較

図 6(a), (b) は EER と PESQ 及び STOI の平均値の関係を示したものである. PESQ と STOI は主観評価を客観的に表すための尺度であり, スコアが高い方が自然性が高いことを示している. 図 6 より, EER と PESQ, STOI のスコアは相関が低いことがわかる. 次に原音声との距離を測る RMS-LSD と EER の関係を図 6(c) に示す. RMS-LSD は値が低いほど原音声に近いことを表すため, (F)N-Bwe は EER, RMS-LSD ともに低いことがわかる. 以上のことより, 客観評価値と EER に強い相関はなかったが, N-Bwe は EER, RMS-LSD ともに低いことから性能のいい手法であると言える.

5. 結論

本論文では, i-vector/PLDA に基づく話者照合システムによる N-Bwe の効果を評価することを目的とした. N-Bwe は, 学習を行わず計算量が軽い帯域拡張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のにもかかわらず, GMM-UBM に基づく話者照合と RMS-LSD において高い性能を得られることが報告されている. 本論文では帯域制限された音声に対して N-Bwe などを適用した場合の PLDA に基づく話者照合システムへの影響を調査した. 実験結果より, アップサンプリングした音声と比較して N-Bwe を適用した音声の方が話者照合実験において EER が 1.78 ポイント改善することを確認した. 今後の課題としては, x-vector に基づく手法での評価などが挙げられる.

謝辞 本研究の一部は科学研究費基盤 (B)2628006 による.

参考文献

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

[2] Wan, L., Wang, Q., Papir, A. and Moreno, I. L.: Generalized end-to-end loss for speaker verification, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4879–4883 (2018).

[3] Rohdin, J., Silnova, A., Diez, M., Plchot, O., Matějka, P. and Burget, L.: End-to-end DNN Based Speaker Recognition Inspired by i-vector and PLDA, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4874–4878 (2018).

[4] Prince, S. J. and Elder, J. H.: Probabilistic linear discriminant analysis for inferences about identity, *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, IEEE, pp. 1–8 (2007).

[5] Brummer, N., Silnova, A., Burget, L. and Stafylakis, T.: Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model, *arXiv preprint arXiv:1802.09777* (2018).

[6] Bahmaninezhad, F. and Hansen, J. H.: i-vector/PLDA speaker recognition using support vectors with discriminant analysis, *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, IEEE, pp. 5410–5414 (2017).

[7] Snyder, D., Garcia-Romero, D., Povey, D. and Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification, *Proc. Interspeech*, pp. 999–1003 (2017).

[8] Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D. and Khudanpur, S.: Spoken language recognition using x-vectors, *Odyssey: The Speaker and Language Recognition Workshop, Les Sables d'Olonne* (2018).

[9] Shon, S., Tang, H. and Glass, J.: Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model, *arXiv preprint arXiv:1809.04437* (2018).

[10] Miyamoto, H., Shiota, S. and Kiya, H.: Non-linear harmonic generation based blind bandwidth extension considering aliasing artifacts, *in Proc. APSIPA Annual Summit and Conference* (2018).

[11] Nidadavolu, P. S., Lai, C.-I., Villalba, J. and Dehak, N.: Investigation on Bandwidth Extension for Speaker Recognition, *Proc. Interspeech 2018*, pp. 1111–1115 (2018).

[12] Pulakka, H., Laaksonen, L., Vainio, M., Pohjalainen, J. and Alku, P.: Evaluation of an Artificial Speech Bandwidth Extension Method in Three Languages, *IEEE Trans. Audio, Speech, and Language. Process.*, Vol. 16, No. 6, pp. 1124–1137 (2008).

[13] Sriskandaraja, K., Sethu, V., Le, P. N. and Ambikairajah, E.: Investigation of Sub-Band Discriminative Information Between Spoofed and Genuine Speech., *INTERSPEECH*, pp. 1710–1714 (2016).

[14] Seo, H., Kang, H. and Soong, F.: A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise, *in Proc. ICASSP 2014*, pp. 6087–6091 (2014).

[15] Le, P. N., Ambikairajah, E., Choi, E. H. and Epps, J.: A nonuniform subband approach to speech-based cognitive load classification, *in Proc. ICICS 2009*, pp. 1–5 (2009).

[16] Sak, H., Senior, A. and Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling, *Fifteenth annual conference of the international speech communication association* (2014).

[17] Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J. and Ramabhadran, B.: Efficient Knowledge Distillation from an Ensemble of Teachers, *Proc. Interspeech 2017*, pp. 3697–3701 (2017).

[18] Larsen, E., Aarts, R. M. and Danessis, M.: Efficient high-frequency bandwidth extension of music and speech, *Audio Engineering Society Convention 112*, Audio Engineering Society (2002).

[19] Thiruvaran, T., Sethu, V., Ambikairajah, E. and Li, H.: Spectral shifting of speaker-specific information for narrow band telephonic speaker recognition, *Electronics Letters*, Vol. 51, No. 25, pp. 2149–2151 (2015).

[20] Bachhav, P., Todisco, M. and Evans, N.: 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).

[21] Un, C. and Magill, D.: The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 kbits/s, *IEEE Trans. on Comm.*, Vol. 23, No. 12, pp. 1466–1474 (1975).

[22] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al.: The Kaldi speech recognition toolkit, *IEEE 2011 workshop on automatic speech recognition and understanding*, No. EPFL-CONF-192584, IEEE Signal Processing Society (2011).

[23] McLaren, M., Ferrer, L., Castan, D. and Lawson, A.: The Speakers in the Wild (SITW) Speaker Recognition Database., *Interspeech*, pp. 818–822 (2016).

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

[25] Chung, J. S., Nagrani, A. and Zisserman, A.: VoxCeleb2: Deep Speaker Recognition, *arXiv preprint arXiv:1806.05622* (2018).

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

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

[28] Rix, A., Beerends, J., Hollier, M. and Hekstra, A.: 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).

[29] Taal, C. H., Hendriks, R. C., Heusdens, R. and Jensen, J.: An algorithm for intelligibility prediction of time–frequency weighted noisy speech, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 7, pp. 2125–2136 (2011).