Vol.2017-SLP-115 No.4
2017/2/17

非線形帯域拡張法に基づく話者照合の検討

中西 亮介[1,a)]　塩田 さやか[1]　貴家 仁志[1]

概要：声を用いた生体認証技術である話者照合は実用化が進みつつある．今後のさらなる展開として携帯電話などの電話音声を用いたセキュリティシステムとしての利用が期待されている．しかしながら電話での通話音声は通信速度の確保のため帯域制限がかかっていることが多い．帯域制限のかかった音声は明瞭性に欠け，音質や話者性が大きく低下することが知られている．本研究では非線形帯域拡張法を電話音声などの帯域制限のかかった音声に適用し，話者照合における帯域拡張の有効性を評価する．帯域拡張法は狭帯域音声から広帯域音声を作る技術としてこれまでいくつか提案されている．しかし，これまで話者照合への適用例はほとんど報告されていない．提案法は狭帯域音声に非線形関数を用いることで広帯域音声を生成し，狭帯域音声と加算合成するため非常に処理が軽いという特徴を持つ．提案法の性能評価は話者照合の精度で評価するために話者照合実験により行われた．その結果，学習データとテストデータそれぞれに提案法を適用し 8kHz から 16kHz に帯域拡張した場合に帯域拡張を行う前に比べエラー改善率が 27.7% 改善した．

キーワード：非線形帯域拡張法，超解像，話者照合，GMM-UBM

Non-linear artificial bandwidth extension of narrowband speech for speaker verification

Nakanishi Ryôsuke[1,a)]　Shiota Sayaka[1]　Kiya Hitoshi[1]

Abstract: Speaker verification is expected to be in practical use as a biometric authentication system using speech. Speaker verification systems are particularly expected to be performed on telephone networks. It is well known that the bandwidth limitation speeches lack clarity and drastically degrade the speech quality and the speaker individuality. This paper proposes a non-linear bandwidth extension method for adapting it to the narrowband speeches, and evaluates it for a speaker verification system. Several artificial bandwidth expansion methods have been proposed to generate a wideband signal from a narrowband signal. However, most the conventional expansion methods have not been applied to speaker verification systems. In the proposed method, a wideband speech is generated from a narrowband one by using a non-linear bandwidth expansion method, so that a light-weight bandwidth extension is given. The proposed method is evaluated under some speaker verification experiments to confirm the performance of the speaker verification. As a result, the proposed method has an Error Reduction of 27.7% compared to the use of narrowband speeches, where the bandwidth of the training data and the test data are respectively expanded from 8kHz to 16kHz.

Keywords: non-linear artificial bandwidth extension, super resolution, speaker verification, GMM-UBM

1. はじめに

近年，声を用いた生体認証システムである話者照合技術の性能が向上してきており，実際にセキュリティシステムとしての実用化が進んできている．今後の展開として期待

[1] 首都大学東京大学院システムデザイン研究科
Department of Information and Communication Systems Engineering, Tokyo Metropolitan University, 6–6, Asahigaoka, Hino-shi, Tokyo 191–0065, Japan
[a)] nakanishi-ryousuke@ed.tmu.ac.jp

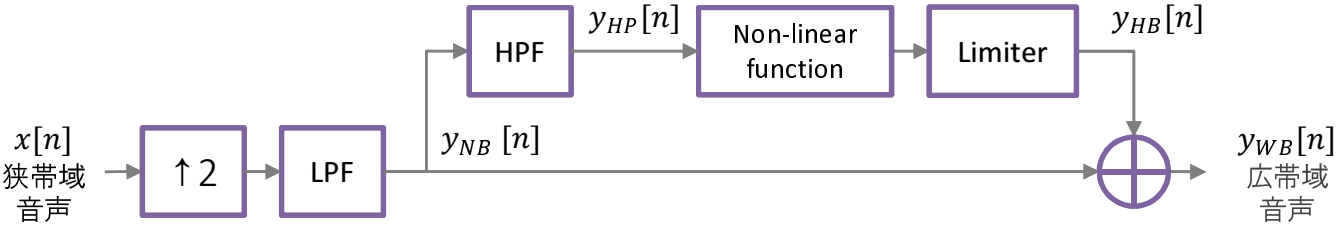


図 1: 非線形帯域拡張法のフロー

されているのが電話回線を通したセキュリティシステムの運用である．携帯電話などの音声通話では，通信速度を確保するために 300–3400Hz に帯域制限がかかった音声を用いた通信が多く行われている．しかし，帯域が制限された音声は人間の耳にも明瞭性が欠け，音質や話者性が低下してしまい，また音声認識や話者認識の観点からも帯域制限がかかった音声は広帯域音声に比べて認識性能が低下してしまうなど様々な問題を引き起こしてしまうことが知られている．帯域制限により失われた広帯域成分を復元する技術，帯域拡張法がそれらの問題に有効であることが広く知られている．これまでに帯域拡張を実現するため様々なアルゴリズムが提案されているが，大別すると分析合成のように入力信号を様々な要素に分解してから再び合成することで拡張音声を得るものとアップサンプリングした狭帯域音声に加工した高域を整形して加算合成する方法がある．分析合成型は学習が必要となることから音質が良くなる一方で計算量が大きくリアルタイム処理に向かないという問題がある．加算合成型は計算量が少ない一方で分析合成型ほど音質が良くならないと言われている．また，帯域拡張法の性能評価には，原音声とどれくらい近いか，自然性がどれくらいなのかなどの尺度を用いてきた．本研究では，計算量が少なくかつ話者照合の精度を上げることを目的として非線形帯域拡張法を提案する．非線形帯域拡張法では，ハイパスフィルタをかけた音声に非線形関数を用いることで広帯域音声を生成する．生成した広帯域音声と狭帯域音声を加算することで広帯域音声を得る．非線形帯域拡張法の有効性を確認するために話者照合実験を行い，狭帯域音声と提案法を用いた広帯域音声の照合性能を比較したところ，非線形帯域拡張法を用いることで，照合性能が大幅に改善することを確認した．

2. 帯域拡張法

帯域制限により失われた広帯域成分を復元するための帯域拡張法として，様々な手法が提案されている．本章では，これまでに提案されている主な帯域拡張法について簡単にまとめる．

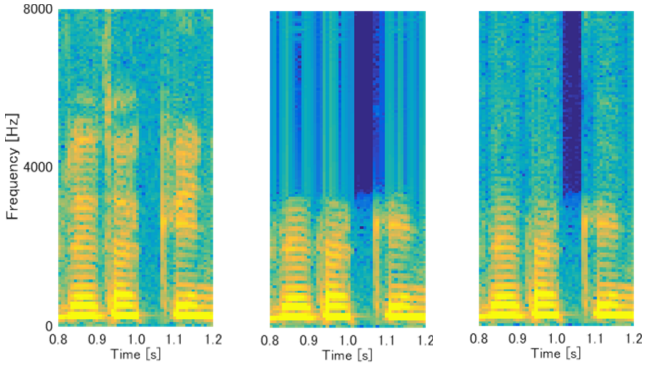
帯域拡張法の例として，低帯域成分を広帯域成分に複製するような比較的処理の軽い手法 [1–3] やピッチ抽出により基本周波数成分を生成する手法 [4–6]，低帯域成分から広帯域スペクトルエンベロープを推定する手法 [7–9]，準結合型辞書学習 (SCDL) に基づく帯域拡張 [10] などが挙げられる．また，LPC や線形周波数スペクトル (LFS)，MFCC など様々な特徴量表現をもとに低帯域成分と広帯域成分のマッピングをとる手法も多い [11, 12]．モデルベースの手法としては，GMM に基づく手法 [13] やニューラルネットワークによる関数変換 [14] や適応型スプラインニューラルネットワークを用いたディープニューラルネットワークを用いた広帯域スペクトルの推定 [15]，対数パワースペクトルを用いたディープニューラルネットワーク (DNN) に基づく帯域拡張法 [16]，LSTM-RNN を用いた帯域拡張 [17]，DNN から得られたボトルネック特徴を用いた LSTM-RNN による帯域拡張 [18]，双方向型 LSTM-RNN とスパース表現を組み合わせた帯域拡張 [19]，共同辞書を用いた帯域拡張 [20]，CRBM に基づく帯域拡張 [21] などが挙げられる．これらの手法の性能評価として，処理にかかる計算量や音声認識率，MOS 値による主観評価，PESQ やスペクトル歪みなどを用いた客観評価などが広く用いられている．

3. 非線形帯域拡張法

画像信号処理の分野において報告された非線形信号処理による超解像画像処理の手法がある [22]．この手法は低解像度の画像から高解像度の画像，つまりナイキスト周波数を超える高周波成分を疑似的に生成する手法である．基本的な手順はアンシャープマスキング (鮮鋭化フィルタ) とほぼ等しいが，途中で非線形関数を用いることで高精度な超解像画像が生成できる手法となっている．本研究で非線形帯域拡張法として扱うのは，上記の超解像技術を音声の帯域拡張に用いたものである．図 1 に非線形帯域拡張法のフローを示す．はじめに狭帯域信号 $x[n]$ をアップサンプリングした信号 $y_{NB}[n]$ にハイパスフィルタ (HPF) を適用し，$y_{HP}[n]$ を得る．次に $y_{HP}[n]$ に非線形関数により広帯域成分 $y_{HB}[n]$ を生成する．広帯域成分 $y_{HB}[n]$ は

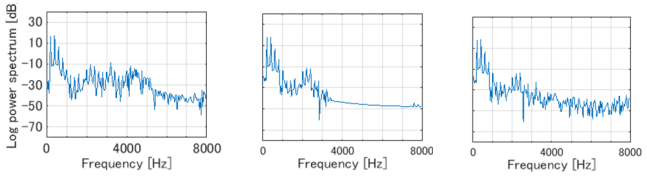
$$y_{HB}[n] = y_{HP}[n]^{\alpha} \times \beta \qquad (1)$$

により計算される． ここで，n はサンプリング点，α および β はユーザ指定のパラメータを表す．HPF を適用した信号 $y_{HP}[n]$ は正弦波 $\sin k\omega_0$ の組合せで表現できる．このとき，$\omega = 2\pi f_s$ である．f_s はサンプリングレートを，k は整数値 ($k = 0, \pm 1, \pm 2, ...$) をそれぞれ表す．三角関数の倍角公式より，式 (1) の非線形関数を用いることでナイキス



(a) 原音声 (16kHz)　(b) 4kHz 帯域制限音声 $y_{NB}[n]$　(c) 帯域拡張された音声 $y_{WB}[n]$

図 2: スペクトログラムによる比較



(a) 原音声 (16kHz)　(b) 4kHz 帯域制限音声 $y_{NB}[n]$　(c) 帯域拡張された音声 $y_{WB}[n]$

図 3: 対数パワースペクトル (1 フレーム) による比較

ト周波数より高い周波数成分を生成することができる．つまり，式 (1) により生成された広帯域成分 $y_{HB}[n]$ は原音声に存在しない広帯域の成分を持つ．非線形関数により生成された信号 $y_{HB}[n]$ の振幅の絶対値が大きくなりすぎるとクリッピングやエイリアシングの問題が起こるため，リミッタによる丸め込みを行う．最後に，以下の式のように広帯域成分 $y_{HB}[n]$ と狭帯域成分 $y_{NB}[n]$ を加算することで帯域拡張された信号 $y_{WB}[n]$ を得る．

$$y_{WB}[n] = y_{NB}[n] + y_{HB}[n]. \qquad (2)$$

図 2 (a) に原音声 (16kHz サンプリング)，(b) 帯域幅を 4kHz に制限した音声 $y_{NB}[n]$ および (c) 提案法により帯域拡張された音声信号 $y_{WB}[n]$ のスペクトログラムを示す．図 2 (b) と (c) を比較すると，図 2 (b) では帯域制限により 4 kHz より高い周波数には信号が現れていないが，図 2 (c) は非線形帯域拡張法を適用することで 4kHz より高い周波数部にも信号が生成されることが確認できる．次に同サンプルの 1 フレームの対数パワースペクトルを比較する (図 3)．図 2 と同様に提案法 (c) では広帯域にもパワーが生成されていることがわかる．一方で，提案法は加算合成型の手法であり，本来の広帯域成分を生成することを目指してはいないため，パワースペクトルが原音声と近くなっているわけではないことも確認できる．前章で述べたようにこれまでの帯域拡張法は原音声に近づけることや自然性向上を目的としてきているが，提案する非線形帯域拡張法は広帯域成分の生成による音質向上と合わせて，機械学習手法に対する性能向上を目指しており，本論文でも評価には実際に話者照合実験における精度について言及する．

表 1: 実験条件

UBM 用データベース	JNAS（女性のみ） 16kHz サンプリング
UBM 学習データ	23657 文章
登録話者データベース	VLD データベース [24] （ヘッドセット，フィルタあり） 48kHz サンプリング
学習データ （特定話者モデル）	70 文章 × 17 名 （時期 01） （計 1190 文章）
テストデータ	30 文章 × 17 名 （時期 01，02） （計 510 文章/時期）
GMM 混合数	1024
フレーム長	25 msec
フレームシフト	10 msec
特徴量	MFCC 19 次+ Δ + ΔΔ

表 2: 比較する条件

(A) 8k → 16k アップ サンプリング	学習データ (UBM，特定話者モデル) に 16kHz の音声を使用し， テストデータは 8kHz の音声を 16kHz に アップサンプリングした音声を使用
(B) 8k → 16k 帯域拡張 (テストのみ)	(A) のテストデータに 提案法を適用し，帯域拡張
(C) 8k	学習データ，テストデータともに サンプリングレート 8kHz の音声を使用
(D) 8k → 16k 帯域拡張 (学習・テスト)	(C) の学習データとテストデータ それぞれに提案法を適用し，帯域拡張
(E) 16k	学習データ，テストデータともに サンプリングレート 16kHz の音声を使用

表 3: 非線形帯域拡張法で使用したパラメータ

手法	HPF の 阻止域端周波数	α	β
(B) 8k → 16k 帯域拡張 (テストのみ)	4kHz	2.0	20000
(D) 8k → 16k 帯域拡張 (学習・テスト)	4kHz	2.0	20000

4. 実験

非線形帯域拡張法に基づく話者照合の有効性を確認するために，GMM-UBM に基づく話者照合実験を行った [23]．

4.1 実験条件

表 1 に主な実験条件を示す．登録話者の特定話者 GMM は UBM から MAP 適応を用いて推定した．VLD データベースでは同一話者の発話を約 3 週間の間隔をあけて 2 回

音声収録を行っている．本実験では学習データに 1 回目の収録 (時期 01) を用い，テストデータには学習データと同時期のもの (時期 01) と 2 回目の収録 (時期 02) の 2 時期を用いた．

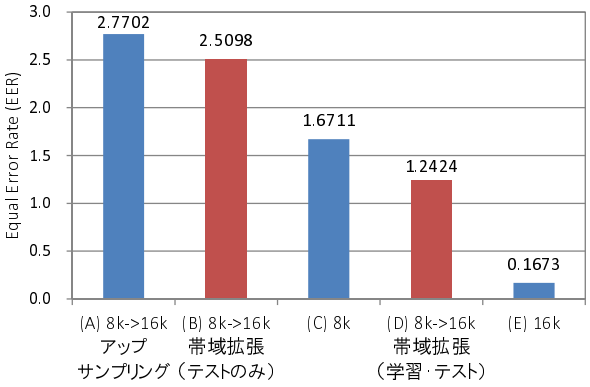
表 2 に話者照合実験で比較する各条件を示す．(A) はサンプリングレート 16kHz の学習データ (UBM および特定話者モデル) を用いてモデルを学習し，サンプリングレート 8kHz のテストデータをアップサンプリングしてサンプリングレート 16kHz にして実験を行ったものを表す (図 1 の $y_{NB}[n]$)．(B) は (A) の音声に非線形帯域拡張法を適用したものを表す．(A)，(B) においては，VLD データベース本来のサンプリングレートは 48kHz であるため 8kHz および 16kHz になるようにダウンサンプリングをしている．(C) では，学習データおよびテストデータのサンプリングレートを 8kHz に合わせた音声を用いて実験を行った．(C) において，JNAS データベース本来のサンプリングレートは 16kHz であるため，8kHz になるようにダウンサンプリングをしている．(D) は (C) の音声を学習データも含めてアップサンプリングしてサンプリングレート 16kHz にしたあとに，非線形帯域拡張法を適用して帯域拡張したものを表す．また，表 3 に (B) および (D) で用いた HPF のフィルタ係数および非線形関数のパラメータ α，β を示す．これらのパラメータは予備実験により手法ごとに決定した．(E) は学習データおよびテストデータにサンプリングレート 16kHz の音声を使用したものを表す．

また，VLD データベース以外のデータベースを用いた比較として，登録話者用データベースに NTT-VR データベース [25] の女性音声のみを用いた場合でも実験を行った．学習データとテストデータはどちらも 1990 年 8 月に収録されたものを用いた．話者数は 13 名であり，話者一人につき学習データは 116 文章を，テストデータは学習にも用いられた 116 文章の中から選択された 30 文章を使用してクローズドテストを行った．NTT-VR データベース本来のサンプリングレートは 16kHz であるため，(A) と (B)，(C) においては VLD データベースと同様に 8kHz にダウンサンプリングしたものを使用した．また，HPF の阻止域端周波数および非線形関数のパラメータ α，β は予備実験によりそれぞれ 4kHz，2.0，50000 とした．

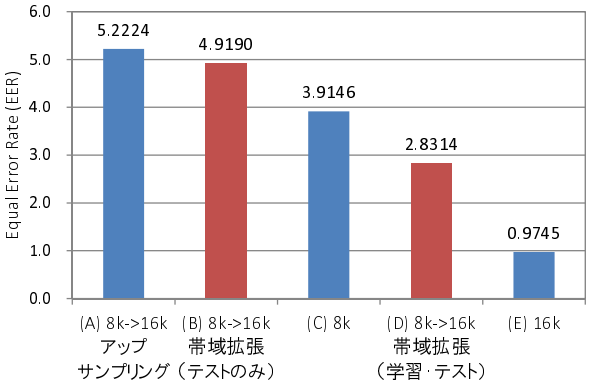
4.2 実験結果

4.2.1 VLD データベース

図 4 (a) に VLD データベースにおける特定話者モデル用学習データとテストデータの収録時期が同じ場合の手法ごとの等価エラー率 (EER) を示す．まず，(A) 8k→16k アップサンプリングと (E) 16k を比較する．(A) と (E) は学習モデルが共通のものでテストデータに帯域制限があるかないかの違いだけであるが，照合性能が大幅に低下している．このことより帯域制限が照合性能に大きく影響を与えることが確認できる．次に，(A) 8k→16k アップサンプリング と (B) 8k→16k 帯域拡張 (テストのみ) を比較すると，(A) と (B) もモデルは同じであるが (B) の EER は (A) の EER よりも低くなっている．このことから非線形帯域拡張法により生成された広帯域成分が，話者照合システムの性能を向上させることがわかる．次に (C) 8k と (A) 8k→16k アップサンプリング および (B) 8k→16k 帯域拡張 (テストのみ) それぞれを比較すると，(C) の EER は (A) および (B) のどちらよりも高い．つまりアップサンプリングや提案法による帯域拡張をテストデータにのみ用いる場合より学習データも低サンプリングレートで学習しなおした方が性能が高いことがわかる．しかし，(D) 8k→16k 帯域拡張 (学習・テスト) と (C) 8k を比較すると，(D) の EER は (C) の EER よりも低くなっている．このことから，学習データのサンプリングレートを一度下げたあとに，非線形帯域拡張法により帯域拡張したデータによりモデル学習をすることで話者照合システムの精度がさらに改善することがわかった．



(a) 学習データとテストデータが同時期の場合



(b) 学習データとテストデータが異なる時期の場合

図 4: 各条件における EER(%)

図 4 (b) に VLD データベースにおける特定話者モデル用学習データとテストデータの収録時期が異なる場合の手法ごとの EER を示す．図 4 (a) と比較すると，収録時期が異なることで全体的に EER が高くなっているが，収録

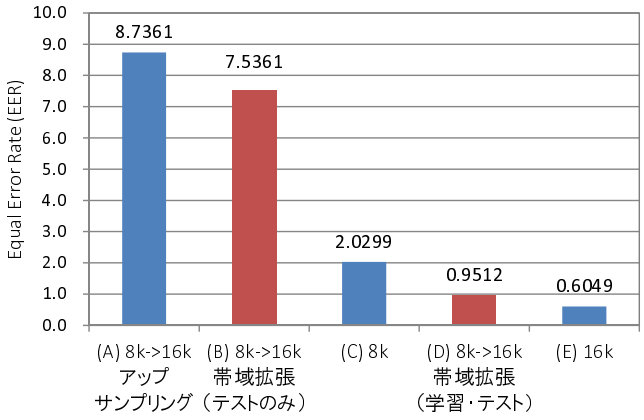


図 5: NTT-VR データベースを用いた場合の各条件における EER(%)

時期が同じ場合と同じ傾向になっている．学習データとテストデータの収録時期が異なるという現実的な状況においても提案法が有効であり時期差に依存しにくい頑健な手法であることがわかる．

4.2.2 NTT-VR データベース

図 5 に NTT-VR データベースにおける手法ごとの EER を示す．図 4 と比較すると，各条件における EER の傾向は VLD データベースを用いた実験と同じ傾向となった．このことから，提案法のデータベースに関する汎用性が確認できた．

NTT-VR データベースの場合，特に (A) と (B) の EER が (E) と比べて大幅に上昇しており，サンプリング周波数が低いときの問題が顕著に出ている．モデルも低周波数に変えて学習することで大幅な改善が得られているが学習をしなおすという観点から見ても (D) の帯域拡張をモデルにも行うことで (B) から比べて EER が 6.6 % も改善し，もとの 16kHz サンプリングの性能に非常に近い結果となっている．提案法はもとの音声に復元しようとする手法ではないもののこのような結果になるのは非常に興味深いといえる．

5. おわりに

本稿では非線形帯域拡張法に基づく話者照合を提案した．非線形帯域拡張法は非線形関数を用いることで狭帯域成分から広帯域成分を生成する手法である．提案法の有効性を調査するために，話者照合実験を行い狭帯域音声と提案法を用いた広帯域音声の照合性能を比較した．実験結果より，非線形帯域拡張法により生成された広帯域成分は話者照合の性能を向上させることが確認できた．また，学習データとテストデータを高いサンプリングレートに合わせて照合を行うよりも，一度低いサンプリングレートに合わせた後に非線形帯域拡張法を適用することで話者照合システムの性能が向上することがわかった．

今後の課題としては，他の手法との比較および i-vector など他の手法への適応，MOS 値などの主観評価実験やノイズを含む音声での検討などが挙げられる．

謝辞 本研究の一部は科学研究費若手 (B)93008552 による．

参考文献

[1] Carl, H.: Untersuchung verschiedener Methoden der Sprachcodierung und eine Anwendung zur Bandbreitenvergröerung von Schmalband-Sprachsignalen, Dissertation, Ruhr-Universität Bochum (1994).

[2] Enbom, N. and Kleijn, W. B.: Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients, *1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No.99EX351)*, pp. 171–173 (1999).

[3] Jax, P. and Vary, P.: Wideband extension of telephone speech using a hidden Markov model, *2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No.00EX421)*, pp. 133–135 (2000).

[4] 藤敦　渉, 関本英彦, 戸田智基, 猿渡　洋, 鹿野清宏：GMM に基づく最尤変換法による携帯電話音声の帯域拡張，情報処理学会研究報告音声言語情報処理 (SLP)， Vol. 2007, No. 75, pp. 63–68 (2007).

[5] Uysal, I., Sathyendra, H. and Harris, J. G.: Bandwidth extension of telephone speech using frame-based excitation and robust features, *2005 13th European Signal Processing Conference*, pp. 1–4 (2005).

[6] Miet, G., Gerrits, A. and Valiere, J. C.: Low-band extension of telephone-band speech, *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, Vol. 3, pp. 1851–1854 vol.3 (2000).

[7] Kornagel, U.: Spectral widening of the excitation signal for telephone-band speech enhancement, *Proc. International Workshop on Acoustic Echo and Noise Control*, pp. 215–218 (2001).

[8] Fuemmeler, J. A., Hardie, R. C. and Gardner, W. R.: Techniques for the regeneration of wideband speech from narrowband speech, *EURASIP Journal on Applied Signal Processing*, Vol. 2001, No. 1, pp. 266–274 (2001).

[9] Jax, P. and Vary, P.: On artificial bandwidth extension of telephone speech, *Signal Processing*, Vol. 83, No. 8, pp. 1707–1719 (2003).

[10] Sreeram, G. and Sinha, R.: Semi-Coupled Dictionary Based Automatic Bandwidth Extension Approach for Enhancing Children' s ASR, *Interspeech 2016*, pp. 2577–2581 (2016).

[11] Cheng, Y. M., O'Shaughnessy, D. and Mermelstein, P.: Statistical recovery of wideband speech from narrowband speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 544–548 (1994).

[12] Qian, Y. and Kabal, P.: Dual-mode wideband speech recovery from narrowband speech., *Proc. 8th European Conf. Speech, Commun. Tech.*, pp. 1433–1437 (2003).

[13] Wang, Y., Zhao, S., Yu, Y. and Kuang, J.: Speech Bandwidth Extension Based on GMM and Clustering Method, *2015 Fifth International Conference on Communication Systems and Network Technologies*, pp. 437–441 (2015).

[14] Kontio, J., Laaksonen, L. and Alku, P.: Neural Network-Based Artificial Bandwidth Expansion of Speech, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 873–881 (2007).

[15] Uncini, A., Gobbi, F. and Piazza, F.: Frequency recovery of narrow-band speech using adaptive spline neural networks, *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, Vol. 2, pp. 997–1000 vol.2 (1999).

[16] Li, K. and Lee, C. H.: A deep neural network approach to speech bandwidth expansion, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4395–4399 (2015).

[17] Tachioka, Y. and Ishii, J.: Long short-term memory recurrent-neural-network-based bandwidth extension for automatic speech recognition, *Acoustical Science and Technology*, Vol. 37, No. 6, pp. 319–321 (2016).

[18] Gu, Y., Ling, Z.-H. and Dai, L.-R.: Speech Bandwidth Extension Using Bottleneck Features and Deep Recurrent Neural Networks, *Interspeech 2016*, pp. 297–301 (2016).

[19] Liu, B. and Tao, J.: A Novel Research to Artificial Bandwidth Extension Based on Deep BLSTM Recurrent Neural Networks and Exemplar-based Sparse Representation, *Interspeech 2016*, pp. 3778–3782 (2016).

[20] Sadasivan, J., Mukherjee, S. and Seelamantula, C. S.: Joint dictionary training for bandwidth extension of speech signals, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5925–5929 (2016).

[21] Wang, Y., Zhao, S., Qu, D. and Kuang, J.: Using conditional restricted Boltzmann machines for spectral envelope modeling in speech bandwidth extension, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5930–5934 (2016).

[22] Gohshi, S. and Echizen, I.: Limitations of super resolution image reconstruction and how to overcome them for a single image, *2013 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, pp. 71–78 (2013).

[23] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B.: Speaker verification using adapted Gaussian mixture models, *Digital signal processing*, Vol. 10, No. 1, pp. 19–41 (2000).

[24] Shiota, S., Fernando, V., Yamagishi, J., Ono, N., Echizen, I. and Matsui, T.: Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification, *Proc. Interspeech*, pp. 239–243 (2015).

[25] Matsui, T. and Furui, S.: Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 3, pp. 456–459 (1994).