社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

i-vectorを用いた話者照合のための非線形帯域拡張法及びフィルタ設計に関する検討

上西遼大† 塩田さやか† 貴家 仁志†

† 首都大学東京大学院システムデザイン研究科
E-mail: †kaminishi-ryota@ed.tmu.ac.jp

あらまし 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いる. 今後さらに携帯電話などの通信を介し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, 通信を介した音声信号は通信速度を維持するために帯域制限がかけられていることが多い. 帯域制限のかかった信号は明瞭性にかけ, 音質や話者性が大きく低下するだけでなく, 話者照合精度が落ちてしまうことも報告されている. そのため帯域制限により失われた広帯域成分を復元, または生成する様々な帯域拡張法が提案されている. 先行研究において, 狭帯域音声信号に非線形関数を適用し広帯域成分を生成する手法を用いることで GMM-UBM における話者照合の精度が向上することが報告された. 本研究では非線形帯域拡張法を拡張した手法を提案し, さらに i-vector を用いた話者照合システムに適用することを報告する. 話者照合実験において, 提案法は従来法の非線形拡張法よりも高い識別性能が得られた.

キーワード i-vector, 非線形帯域拡張法, 話者照合

Nonlinear artificial bandwidth extension and filter design for i-vector based speaker verification

Ryota KAMINISHI†, Sayaka SHIOTA†, and Hitoshi KIYA†

† Department of Information and Communication Systems Engineering, Tokyo Metropolitan University, 6-6, Asahigaoka, Hino-shi, Tokyo 191–0065 Japan
E-mail: †kaminishi-ryota@ed.tmu.ac.jp

Abstract Recently, speaker verification, which is a biometric authentication technology using voice, comes to be in practical use. Especially, the speaker verification systems are expected to be security systems via communication networks. However, it is often the case that bandwidth limitation is applied to speech signals in order to keep the communication speed. The band-limited speech signals degrade naturalness and intelligibility. It is also reported that speaker verification accuracy is significantly degraded by the band limitation. Thus, various bandwidth extension methods have been reported to restore broadband components. In the previous research, a nonlinear bandwidth extension method has been proposed. At the method, a nonlinear function is applied to narrowband speeches and is generated wideband components. This paper proposes an expanded nonlinear bandwidth extension. In addition, the proposed method is evaluated by using i-vector based speaker verification systems. The experimental results show that the proposed method improves the performance than the conventional method.

Key words i-vector, nonlinear artifical bandwidth extension, speaker verification

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. 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, 話者照合システムのさらなる普及が期待されている. しかし, 収録環境によってはシステムが想定しているサンプリング周波数と入力音声のサンプリング周波数が必ずしも一致しているとは限らない. 特に通信を介した場合には低いサンプリング周波数かつ帯域制限によって音声の明瞭性や話者性が大幅に低下してしまいシステムの性能にも大きな影響を与えてしまうことが知られている.

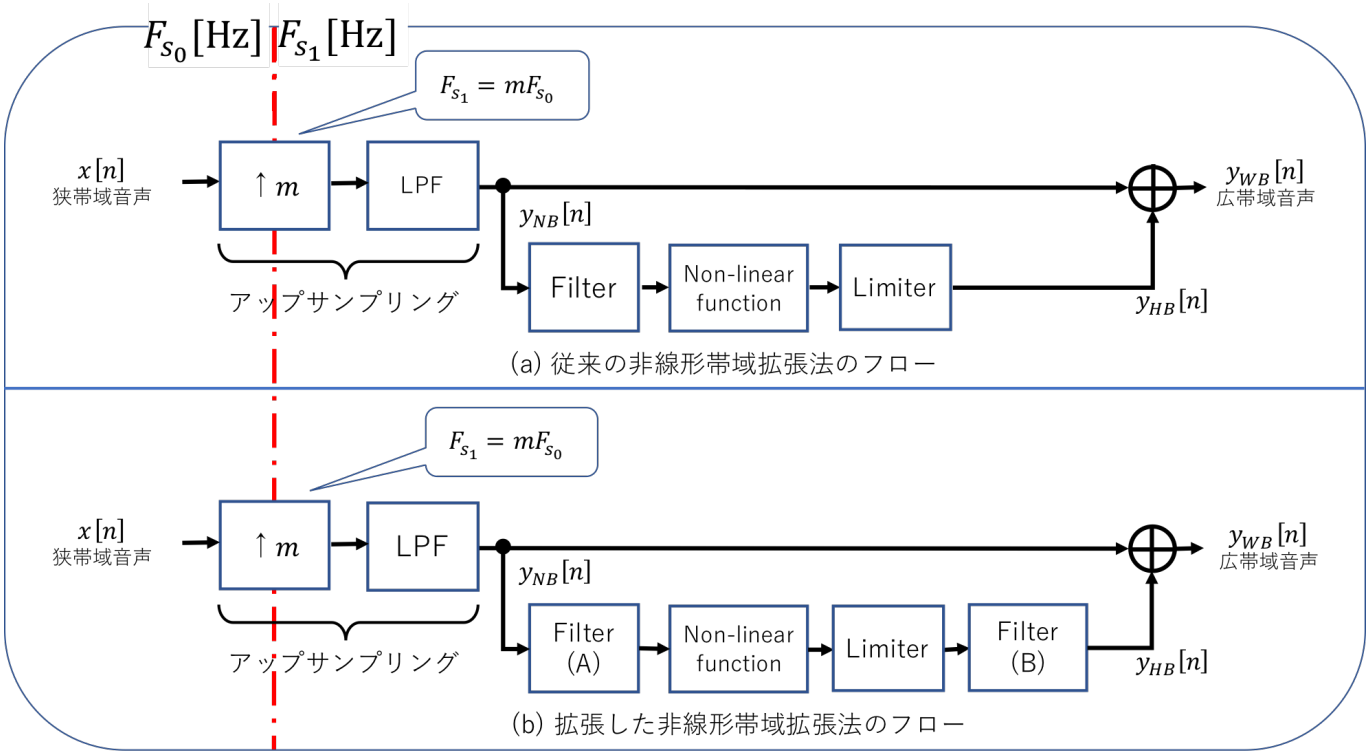

図 1: (a) 従来の非線形帯域拡張のフロー　(b) 拡張した非線形帯域拡張法のフロー (LPF=Low Pass Filter)

この問題に対して, 帯域拡張という広帯域成分を復元・生成する技術を用いることで, 音声の明瞭性や話者性が改善されることが報告されている [1] [2].

先行研究として非線形関数を用いた帯域拡張法が提案されている [3]. この非線形帯域拡張法は, 学習を行わないため処理が非常に軽く, かつ任意のサンプリング周波数に対応できること, また, GMM-UBM に基づく話者照合 [4] において照合性能を大幅に改善できる手法となっている. しかし, 非線形処理により低周波数成分にもノイズがまわり込んでしまうという問題があった. そこで本研究では非線形帯域拡張法の拡張を行い, またフィルタ設計についても調査した. さらに i-vector に基づく話者照合システム [5] による評価実験を行ない照合精度が上がることを確認した.

2. 非線形帯域拡張法の拡張

2.1 非線形帯域拡張法 [3]

従来の非線形帯域拡張法のフローを図 1(a) に示す. サンプリング周波数 F_{s_0}[Hz] の狭帯域音声 $x[n]$ にアップサンプリングを施した信号を $y_{NB}[n]$ とする. 得られた信号をフィルタに通し, 式 (1) のような非線形処理を施す.

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

ここで n はサンプリング点, α, β はユーザ指定のパラメータを表す. 非線形処理の前にフィルタ処理を施す理由として, 狙った帯域のみを用いて非線形帯域拡張が可能になることが挙げられる. 式 (1) によって生成された信号は原音声にはない広帯域成分を含む音声となる. 非線形関数のパラメータによっては信号の振幅が大きくなりすぎ, クリッピングやエイリアジングによるまわりこみなどの問題が生じてしまうため, リミッタによる丸め込みを行う. 最後に式 (2) 狭帯域成分 $y_{NB}[n]$ と広帯域成分 $y_{HB}[n]$ を加算することにより帯域拡張されたサンプリング周波数 F_{s_1}[Hz] の信号 $y_{WB}[n]$ を生成する.

$$y_{WB}[n] = y_{NB}[n] + y_{HB}[n] \tag{2}$$

この非線形帯域拡張法は, 学習を行わないため処理が非常に軽く, かつ任意のサンプリング周波数に対応できるという利点がある.

2.2 提　案　法

本節では提案法である非線形帯域拡張法の拡張について説明する. 図 1(b) に拡張した非線形帯域拡張法のフローを示す. 拡張された点は, 非線形関数を通した後の信号を Filter(B) に通すことで足し合わせる広帯域成分の周波数の選択を行なっていることである. Filter(B) にはハイパスフィルタ (HPF) やバンドパスフィルタを想定しており, 特に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と期待できる.

図 2 に各条件での音声スペクトログラムを示す. 図 2 の (a) が原音声 (サンプリング周波数 16kHz), (b) が帯域制限された音声 ($y_{NB}[n]$), (c), (d) はそれぞれ従来法と提案法を用いて帯域拡張した際の広帯域音声 ($y_{HB}[n]$) を示す. また (e) には (b) と (d) を足し合わせた音声 ($y_{WB}[n]$) を示す. 図 2(b) から帯域制限により高周波成分がなくなっていることがわかる. 次に, 従来法における $y_{HB}[n]$ のスペクトログラム (c) と提案法における $y_{HB}[n]$ のスペクトログラム (d) を比較すると, 従来法で

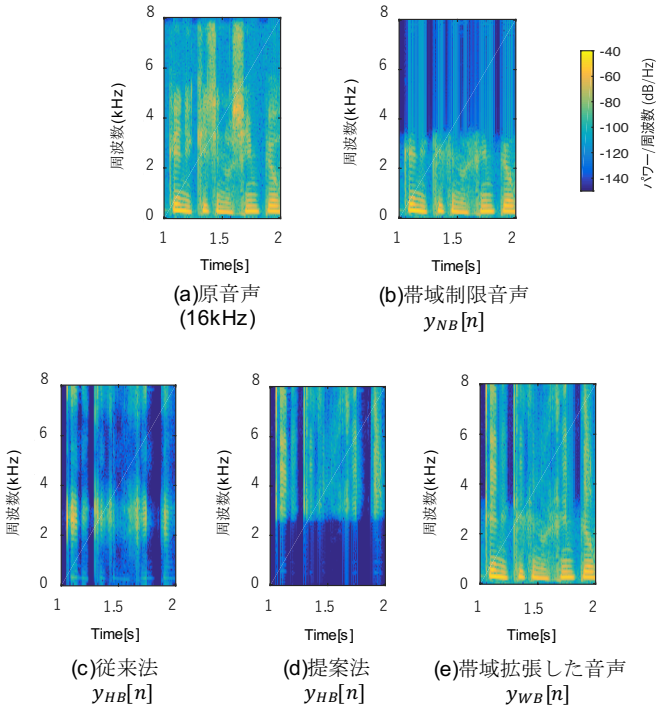

図 2: スペクトログラムによる比較

は最後にフィルタ処理をしていないため, 低域の周波数成分が残ってしまっているが, 提案法では最後にフィルタ処理をしているため, 低域に周波数成分が残っていないことが確認できる. 従来法においても低周波成分のパワーは低いため大幅に劣化させる要因にはならないが影響が残ってしまうことが図 2 よりわかる. さらに (a) と (e) を比較すると (e) の帯域拡張された部分のスペクトログラムは (a) に近い周波数成分ではないことがわかる. これは提案法が, 本来の広帯域の音声を生成することを目指しているわけではないためである. このため, 本稿では評価を話者照合実験における精度を用いて言及する.

3. 実　　験

i-vector に基づく話者照合システムにおいて提案法の有効性を評価した.

3.1 実 験 条 件

表 1 に i-vector に基づく話者照合システムの主な実験条件を示す. VLD データベースはサンプリング周波数 48kHz で収録されているが, 本実験では 16kHz にダウンサンプリングしたものを 16kHz の原音声として扱う. 表 2 に本実験における比較条件を示す. (A),(B),(C) ではサンプリング周波数 16kHz の原音声を用いて UBM, TV 行列, i-vector の学習を行なっている. テストデータは原音声から 8kHz にダウンサンプリングしたものがシステムの入力音声だと想定し, (A) ではアップサンプリングのみを施したもの (図 1 の $y_{NB}[n]$), (B) では従来の非線形帯域拡張法を施したもの (図 1(a)), (C) では提案法を施したもの (図 1(b)) を用いてる. (D) は入力音声が低いサンプリング周波数 (8kHz) であることを想定し, テストデータに合わせて学習データをダウンサンプリングし, 全てのデータのサンプリング周波数が 8kHz である状態でモデルの学習をし, 評価したものである. (E),(F) は (D) の全データに対して従来法及び提案法を用いてサンプリング周波数を 16kHz にあげ, 学習と評価を行なったものである. (G) はすべてのデータが原音声 (16kHz) であった場合のシステムの上限値を表している. また従来法で使用したパラメータを表 3 に, 提案法で使用したパラメータを表 4 に示す. 話者照合の評価尺度には本人棄却率と他人受理率が等しくなる点である等価エラー率 (Equal error rate; EER) を用いた.

表 1: 実験条件

UBM, TV 用データベース	JNAS(女性) サンプリング周波数 16kHz
UBM, TV 用学習データ	23657 文章
UBM 学習回数	30 回
TV 行列学習回数	10 回
i-vector 次元数	400 次元
登録話者データベース	VLD データベース [6] サンプリング周波数 48kHz
学習データ	70 文章 × 17 名 (女性) 1190 文章
テストデータ	30 文章 × 17 名 (女性) 510 文章
フレーム長/フレームシフト	25ms/10ms
特徴量	MFCC19 次 + Δ + ΔΔ

表 2: 比較条件

	学習データ	テストデータ
(A) アップサンプリング	原音声 16k[Hz]	アップサンプリング 8k → 16k[Hz]
(B) 帯域拡張 (テストのみ) (従来法)	原音声 16k[Hz]	非線形帯域拡張 (従来法) 8k → 16k[Hz]
(C) 帯域拡張 (テストのみ) (提案法)	原音声 16k[Hz]	非線形帯域拡張 (提案法) 8k → 16k[Hz]
(D) ダウンサンプリング	ダウンサンプリング 8k[Hz]	ダウンサンプリング 8k[Hz]
(E) 帯域拡張 (学習, テスト) (従来法)	非線形帯域拡張 (従来法) 8k → 16k[Hz]	非線形帯域拡張 (従来法) 8k → 16k[Hz]
(F) 帯域拡張 (学習, テスト) (提案法)	非線形帯域拡張 (提案法) 8k → 16k[Hz]	非線形帯域拡張 (提案法) 8k → 16k[Hz]
(G) 原音声	原音声 16k[Hz]	原音声 16k[Hz]

3.2 実 験 結 果

図 3 に手法ごとの EER を示す. まず, (A) アップサンプリングと (G) 原音声を比較する. (A) と (G) の違いはテストデータが帯域制限されているか, いないかのみの違いであるが, (A) の照合性能が大幅に低下している. このことにより, 音声の帯域制限は i-vector に基づく話者照合の照合性能に大きく影響を与

表 3: 従来法で使用したパラメータ

手法	Filter/ 阻止域 周波数	α	β	limiter
(B)	HPF/ 1.6kHz	2	1	0.001
(E)	HPF/ 1.6kHz	2	100000	0.0001

表 4: 提案法で使用したパラメータ

手法	Filter(A)/ 阻止域 周波数	Filter(B)/ 阻止域 周波数	α	β	limiter
(C)	オールパス フィルタ	HPF/ 1.6kHz	2	1	0.001
(F)	HPF/ 0.01kHz	HPF/ 2.48kHz	2	100000	0.0001

えているということが確認できる. 次に (A) アップサンプリングと (B) 従来法を用いた帯域拡張 (テストのみ) を比較してみる. (A) と (B) の違いはテストデータが帯域制限されているか, 帯域拡張されているかである. 照合性能は僅かながらであるが, (A) アップサンプリングと (C) 提案法を用いた帯域拡張 (テストのみ) を比較してみると, (C) の方が精度が良い. これはまわりこみによるノイズが除かれたため, i-vector の精度が若干であるが改善したためであると考えられる. しかし, (D) ダウンサンプリングを (A) アップサンプリング, (B) の従来法の帯域拡張 (テストのみ), と比較すると, アップサンプリングや帯域拡張したものよりもダウンサンプリングした音声で学習しなおしたものの方が性能が良いということがわかる. 次に, (E) の従来法による帯域拡張 (テスト, 学習) と (D) ダウンサンプリングを比較してみると, (E) の方が照合性能が悪い. i-vector では特定話者を表現するために非常に高次元な GMM スーパーベクトルを因子分析により次元圧縮を行なっている. そのため, 低周波数成分の若干の歪みもモデル化する際に大きく影響を与えてしまっていると考えられる. 実際に (D) ダウンサンプリング, (F) の提案法の帯域拡張 (テスト, 学習), とを比較してみると, (F) の EER が若干低くなっている. これらのことより拡張した非線形帯域拡張法の提案手法が有効であり, サンプリングレートを落として学習しなおすより, 非線形帯域拡張法を適応して学習しなおす方が照合性能が良いということがわかった. ただし, GMM-UBM での性能改善幅よりも i-vector の改善幅が小さいのはモデル化が GMM-UBM よりも難しいからであると考えられる.

最後にフィルタについて言及する. 表 3, 表 4 より, 従来法では, 非線形関数の前段でハイパスフィルタによる帯域選択をすることで性能が向上した. しかし, 提案法では非線形関数の後段でフィルタをかけることから前段による帯域選択はあまり本実験のようなクリーンな音声に対しては影響を及ばさないことがわかった.

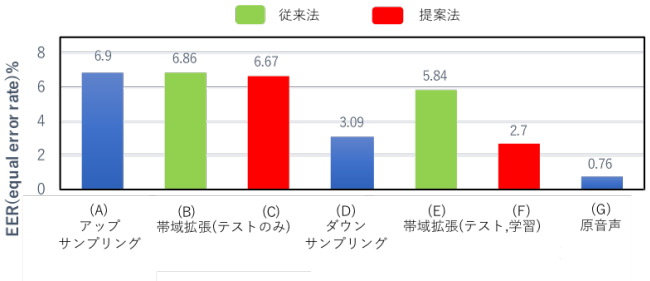

図 3: 実験結果

4. おわりに

本稿では従来の非線形帯域拡張法を拡張した非線形帯域拡張法とし, i-vector に対応できるような手法を提案した. 従来の手法では非線形関数の前にハイパスフィルタをかけ広域成分を生成することにより, 精度の向上が見られたが, i-vector にもとづく話者照合では, 拡張した非線形帯域拡張法を用いることで, 照合性能の改善が確認できた. 今後の課題は, 短い発話を非線形帯域拡張法を適応し, i-vector を抽出することにより照合性能の変化を確認することなどがあげられる.

謝辞 本研究の一部は科学研究費基盤 (B)2628006 に よる.

文 献

[1] Chih-Wei Wu and Mark Vinton. “Blind bandwidth extension using k-means and support vector regression”. Vol. Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pp. 721–725, 2017.

[2] Pramod.B Bachhav, Massimiliano Todisco, Moctar Mossi, Christophe Beaugeant, and Nicholas Evans. “Artificial bandwidth extension using the constant q transform”. Vol. Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pp. 5550–5554, 2017.

[3] 中西涼介ら. “非線形帯域拡張法に基づく話者照合とその応用”. 日本音響学会春季大会, pp. 131–134, 2017.

[4] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, Vol. 10, No. 1-3, pp. 19–41, 2000.

[5] Najim Dehak, Patrick. J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. “Front-end factor analysis for speaker verification”. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798, 2011.

[6] S. Shiota, V. Fernando, N. Echizen I. Yamagishi, J.and Ono, and T. Matsui. “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification”. *Proc. Interspeech*, pp. 293–243, 2015.