

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

宮本 春奈 塩田 さやか 貴家 仁志

† 首都大学東京 システムデザイン学部

E-mail: †miyamoto-haruna@ed.tmu.ac.jp

あらまし 本論文では、エイリアシングの影響を考慮した非線形帯域拡張法を提案し、その有効性を客観評価および話者照合において評価している。帯域拡張法には電話音声のような狭帯域音声に適用するものだけでなく、次世代通信のための超広帯域に対応させるものなどがある。統計モデルを用いた手法によりどちらの条件においても高い品質を得ることが報告されている。一方、学習のプロセスを必要としない帯域拡張法である非線形帯域拡張法も近年報告されている。非線形帯域拡張法では非線形関数を用いることで広帯域音声を生じ可能であるが一方で扱う信号がデジタル信号であるためにエイリアシングの影響を受けるといった問題もあった。そこで、本研究ではエイリアシングを回避するためのフィルタを加えた非線形帯域拡張法を提案する。提案法を評価するための実験として PESQ および RMS-LSD という客観的評価尺度による比較と話者照合実験を行った。特に話者照合実験において信号を 16 kHz から 32 kHz へ拡張した場合に提案法は従来法と比較して 29.7%のエラー削減率を得たことを報告する。

キーワード 非線形帯域拡張, エイリアシング, 話者照合, PESQ, RMS-LSD

Speaker verification based on non-linear bandwidth extension considering aliasing artifacts for super-wideband applications

Haruna MIYAMOTO, Sayaka SHIOTA, and Kiya HITOSHI

† Faculty of System Design, Tokyo Metropolitan University

E-mail: †miyamoto-haruna@ed.tmu.ac.jp

Abstract This paper has two aims that are to propose a novel bandwidth extension (BWE) method considering aliasing artifacts, and to apply various BWE methods to speaker verification to evaluate the effectiveness of the BWE ones. BWE methods enable us not only to enhance narrowband signals but also to adapt signals to super-wideband systems. It has been reported that statistical based BWE approaches can estimate clear wideband and super-wideband signals. Recently, a non-linear BWE method has also been reported as a resynthesis approach. In this paper, it is first pointed out that digital signals generated by the non-linear BWE method include some aliasing artifacts due to the band limitation to be decided according to the sampling frequency. Next, a new non-linear artificial BWE method, which allows us to avoid the influence of aliasing artifacts, is proposed. Moreover, to evaluate the proposed framework, speaker verification experiments and objective tests, i.e. PESQ and RMS-LSD, are conducted. Especially, experimental results show that speech signals extended to 32 kHz by the proposed framework provide the error reduction of 29.7%, compared with conventional methods.

Key words non-linear artificial bandwidth extension, aliasing artifacts, speaker verification, PESQ, RMS-LSD

1. ま え が き

帯域拡張法には電話音声のような狭帯域音声を広帯域へ拡張するものだけでなく広帯域から超広帯域へ拡張させるものなどがある。特に通信網の発展に伴い超広帯域を用いる通信網も主

流になりつつある。これまでに広帯域および超広帯域への帯域拡張法として統計モデルを用いることで高い品質を得られることが報告されている [1-6]。一方、学習のプロセスを必要としない帯域拡張法である非線形帯域拡張法も提案されている [7]。非線形帯域拡張法は狭帯域音声に非線形関数を適用するだけで

広帯域音声を生成できる手法となっている。しかし、非線形帯域拡張法ではデジタル信号を扱うために特に低周波成分がエイリアシングの影響を受け若干音声が歪むという問題があった。そこで、本研究ではフィルタを加えることでエイリアシングの影響を緩和させる新しい非線形帯域拡張法を提案する。音声の客観評価に用いられる、PESQ および RMS-LSD という尺度によりエイリアシングの影響を受けていた従来法と提案法を比較した結果、従来法より自然性の向上および歪みの減少が確認できた。また、話者照合実験において提案法では、従来法と比較して 8kHz から 16kHz に拡張した際には 31.8%、16kHz から 32kHz に拡張した際に 29.7%のエラー削減率を得た。

2. 非線形帯域拡張法

本章では従来法である非線形帯域拡張法について説明する。図 1 に非線形帯域拡張法のフローを示す。はじめに F_{S_0} [Hz] でサンプリングされた狭帯域信号 $x[t]$ ($t = 1, 2, \dots, T$) を F_{S_1} [Hz] サンプリングへあげるためにゼロ値補間によるアップサンプリングを行った後に、ナイキスト周波数を遮断周波数とするローパスフィルタ (LPF) に通した信号 $y_{NB}[t]$ を用意する。次に、アップサンプリングされた信号 $y_{NB}[t]$ に対して Filter(A) をかけることで特定の周波数帯域だけを含む信号 $y_{HP}[t]$ を得る。さらに、 $y_{HP}[t]$ に対し以下のように定義された非線形関数をかけることで高周波成分を含む信号 $y_{HB}[t]$ を生成する。

$$y_{HB}[t] = y_{HP}[t]^\alpha \times \beta. \quad (1)$$

高周波成分が生成される理由について述べる。非線形関数の入力信号である $y_{HP}[t]$ を時間フレームで切り出した信号として逆フーリエ変換で表すと、式 (1) は以下ようになる。

$$y_{HB}[m, n]^\alpha \times \beta = \left\{ \frac{1}{N} \sum_{k=0}^{N-1} Y_{HP}(m, k) e^{j2\pi kn/N} \right\}^\alpha \times \beta, \quad (n = 0, 1, \dots, N-1). \quad (2)$$

ここで、 m はフレームインデックス、 N はフレーム長、 k は離散時間インデックス、 $Y_{HP}(k)$ は信号 $y_{HP}[t]$ の DFT 係数である。例えば非線形関数のパラメータ α の値が 2 の場合、2 乗の正弦波は 2 倍角の公式より以下ようになり、元の信号より高い周波数成分が生成されることがわかる。

$$\sin^2(2\pi kt/N) = \frac{1 - \cos(2(2\pi kt/N))}{2}. \quad (3)$$

非線形関数に用いるパラメータ α 、 β は任意に設定できることから、関数より出力される信号 $y_{HB}[t]$ はクリッピングが起こる可能性がある。そのため、リミッタによる丸め込みを行った信号 $y_{HB}[t]$ と狭帯域成分のみの信号 $y_{NB}[t]$ を足し合わせることで広帯域信号 $y_{WB}[t]$ を生成する。ただし、 α の値によっては符号情報が打ち消されるため非線形関数をかけた後に信号 $y_{NB}[t]$ の符号情報に合わせることにした。

図 2 に周波数 1.5kHz、2kHz、5kHz、6kHz の正弦波を足し合わせた混合信号をサンプリング周波数 16kHz の狭帯域信号とし、アップサンプリングの倍率 l を 3 とし非線形帯域拡張

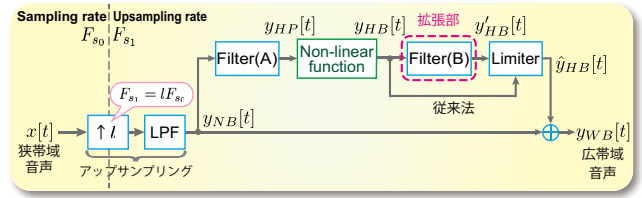


図 1 非線形帯域拡張法のフロー図 (従来法および提案法)

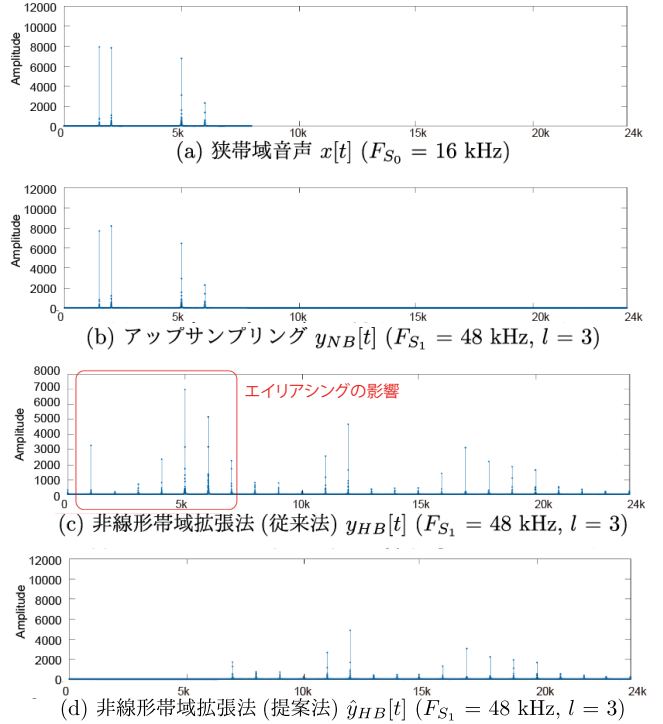


図 2 振幅スペクトルの例 (1.5 kHz, 2 kHz, 5 kHz, 6 kHz の正弦波を合成)

法を行った際の振幅スペクトルを示す。アップサンプリングされた信号 $y_{NB}[t]$ は図 2(b) に示すように帯域が広がっても元の狭帯域信号 (図 2(a)) と同じ帯域にのみ成分をもつ。従来の非線形帯域拡張法 ($\alpha = 2, \beta = 2$) を用いた場合の信号 $y_{HB}[t]$ は、図 2(c) に示されるとおり元の信号よりも高い周波数成分を含んでいる。しかし、高周波成分だけでなく低周波成分にも信号が回り込んでいることがわかる。これは、離散時間信号が周期的な周波数特性を有するために発生するエイリアシングの影響によるものである。そのため、従来法では、エイリアシングの影響を受けた信号と狭帯域音声を足し合わせるため広帯域音声 $y_{WB}[t]$ の低周波数帯域に歪みが生じる可能性がある。

3. 回り込みを考慮した非線形帯域拡張法

非線形帯域拡張法におけるエイリアシングの影響を低減するための回り込みを考慮した手法について考える。提案法のフローとしては図 1 の拡張部に示すように、非線形関数をかけた後に生成される広帯域信号 $y_{HB}[t]$ に Filter(B) としてハイパスフィルタまたはバンドパスフィルタを適用する。このフィルタをかけることで図 2(c) に示したエイリアシングの影響を低減することができる。信号が図 1 の拡張部を通ることで図 2(d)

表 1 GMM-UBM システムの実験条件

データベース (UBM)	JNAS (女性)
学習データ数 (UBM)	23657 文章
データベース (特定話者モデル)	VLD データベース [13] (ヘッドセット)
話者	17 名 (女性)
学習データ数 (特定話者モデル)	70 文章 / 話者 (全 1190 文章)
テストデータ	30 文章 / 話者 (全 510 文章)
GMM 混合数	1024
フレーム長/フレームシフト	25 msec / 10 msec
特徴量	MFCC 19 次+ Δ + $\Delta\Delta$

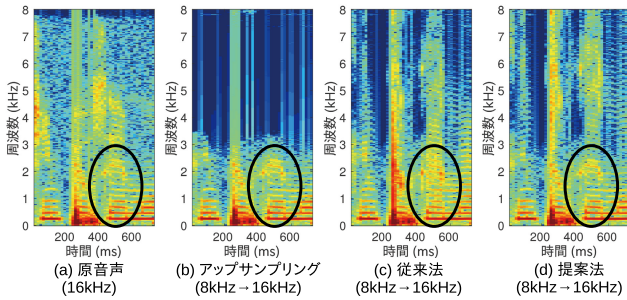


図 3 スペクトログラム ($F_{s_0} = 8\text{kHz}$, $F_{s_1} = 16\text{kHz}$)

のようなエイリアシングの影響を受けていない高周波成分のみを持つ信号 $\hat{y}_{HB}[t]$ が生成され、元の狭帯域音声信号をアップサンプリングした信号 $y_{NB}[t]$ と足し合わせることで広帯域信号 $y_{WB}[t]$ を得ることができる。

図 3 に、16kHz でサンプリングされた原音声 (a)、原音声のサンプリング周波数を 8kHz に下げた後から 16kHz にアップサンプリングした音声 (b)、音声 (b) に対して従来の非線形帯域拡張法を行った広帯域音声 (c)、音声 (b) に対して提案法を行った広帯域音声 (d) のスペクトログラムを示す。アップサンプリングされた音声および広帯域音声は、それぞれ図 1 の $y_{NB}[t]$ および $y_{WB}[t]$ に対応する。図より、従来の広帯域信号 (a) および非線形帯域拡張法を用いた広帯域信号 (c)、(d) は、高周波数帯域 (4~8kHz) において高周波成分を含んでいることがわかる。さらに、図 3(c) と図 3(d) の低周波成分を比較すると、提案法では Filter (B) を加えることによりエイリアシングの影響が低減されることが確認できる。特に楕円で囲まれている部分では差が顕著に現れており、提案法では歪みの少ない音声となることを示している。

4. 実験

提案法の有効性を評価するために、PESQ, RMS-LSD による客観評価実験および GMM-UBM に基づく話者照合実験 [10] を行った。

4.1 実験条件

表 1 に、GMM-UBM に基づく話者照合システムを構築するための実験条件を示す。本稿では、サンプリング周波数を 8kHz から 16kHz へ帯域拡張する場合 ($l = 2$, $F_{s_0} = 8\text{kHz}$, $F_{s_1} = 16\text{kHz}$) と 16kHz から 32kHz へ帯域拡張する場合 ($l = 2$, $F_{s_0} = 16\text{kHz}$, $F_{s_1} = 32\text{kHz}$) の 2 つの条件で実験を行った。JNAS デー

表 2 フィルタ設計 (HPF : ハイパスフィルタ, BPF : バンドパスフィルタ)

手法	従来法 2 および提案法		従来法 1	
	Filter (A)	Filter (B)	Filter (前)	Filter (後)
フィルタ名	HPF	HPF	BPF	BPF
フィルタタイプ	HPF	HPF	BPF	BPF
通過域端周波数 ω_{s_1}	0.45	0.45	0.25	0.47
阻止域端周波数 ω_{p_1}	0.30	0.30	0.20	0.43
通過域端周波数 ω_{s_2}	-	-	0.50	0.98
阻止域端周波数 ω_{p_2}	-	-	0.55	1.00
阻止域減衰量 A_{s_1} [dB]	60	60	40	40
阻止域減衰量 A_{s_2} [dB]	-	-	40	40
通過域リップル δ_p [dB]	1	1	1	1

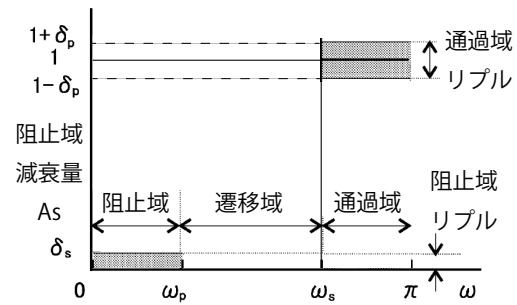


図 4 フィルタ仕様 (HPF)

データベースのサンプリング周波数は 16kHz であるため、帯域拡張を行う場合については 16kHz からダウンサンプリングを行い 8kHz にした音声信号を狭帯域音声 $x[t]$ として用いた。また、VLD データベースのサンプリング周波数は 48kHz であるため、一度全てのデータを 16kHz にダウンサンプリングした後に、JNAS データベースと同様の処理を行った。実験の比較条件は以下の通りである。ただし、比較条件で述べるフィルタの仕様は図 4 に示す。

(A) UP

入力音声に対してアップサンプリングのみ行った音声 ($y_{NB}[t]$) を使用した。

(B) 従来法 1

スペクトルシフティング [11] を用いた帯域拡張法を従来法とした。スペクトルシフティングでは狭帯域のスペクトル成分をコサイン関数によってシフトすることで広帯域成分を生成しており、処理のフロー図やフィルタ設計は文献 [12] に則っている。また、スペクトルシフティングの前後に用いる Filter (前), Filter (後) の設定は表 2 の通りである。

(C) 従来法 2

2 章で述べた非線形関数を用いた非線形帯域拡張法による帯域拡張を行った際の Filter (A) の設定は表 2 の通りである。Filter (B) にはオールパスフィルタを用いた。アップサンプリング後のサンプリング周波数 F_{s_1} が 16kHz, 32kHz どちらの場合も α, β はそれぞれ 1.8, 100 とした。

(D) 提案法 1

狭帯域音声に対して提案法の非線形帯域拡張法による帯域

拡張法を行った。Filter (A) にはオールパスフィルタを用い、Filter (B) の設定は表 2 の通りとした。帯域拡張後のサンプリング周波数 F_{s1} が 16kHz, 32kHz どちらの場合も α , β はそれぞれ 1.8, 100 とした。

(E) 提案法 2

(C) 提案法 1 と同様に狭帯域音声に対して提案法の非線形帯域拡張法を適用し生成した音声データ ($y_{WB}[t]$) を使用した。Filter(A), Filter(B) の設定は表 2 の通りである。 α , β は, F_{s1} が 16kHz の場合は 1.5, 100 とし, また, 32kHz の場合は 2.5, 15500000 とした。

(F) 提案法 3

(C) 提案法 1 と同様に狭帯域音声に対して提案法の非線形帯域拡張法を適用し生成した音声データ ($y_{WB}[t]$) を使用した。Filter(A), Filter(B) の設定は表 2 の通りである。帯域拡張後のサンプリング周波数 F_{s1} が 16kHz, 32kHz どちらの場合も α , β はそれぞれ 1.5, 100 とした。

(G) 8k

サンプリング周波数 8kHz の音声を使用した。

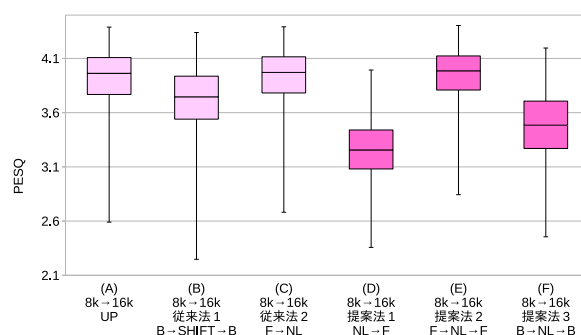
(H) 16k

サンプリング周波数 16kHz の音声を使用した。

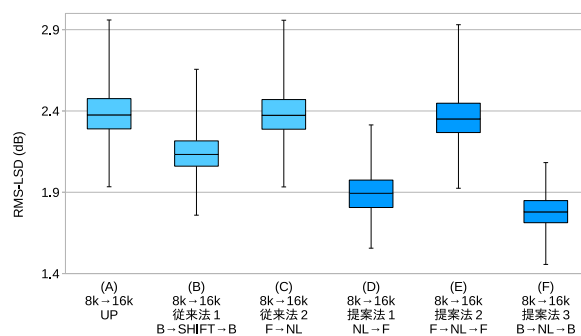
客観評価実験では、音声における自然性を表す客観的な尺度である PESQ [14] と原音声と処理された音声との平均対数スペクトル距離を表す RMS-LSD (Root Mean Square - Log Spectral Distance) [15] の 2 つを使用した。評価に用いたデータは VLD データベースの 1700 文章で、サンプリング周波数を 16kHz, 32kHz にそれぞれ下げたものをリファレンスとして各スコアを計った。また、話者照合の評価には、本人棄却率と他人受理率が等価となる、等価エラー率 (Equal Error Rate; EER) を用いた。

4.2 実験結果

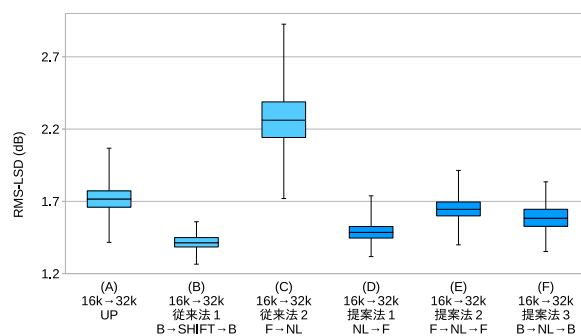
図 5 は、PESQ と RMS-LSD を用いた客観評価実験の結果を箱ひげグラフで表したものである。箱の上辺と底辺は全結果の四分位範囲を、箱の中の線はデータの中央値を示している。箱の上下に伸びる線は全データの最大値と最小値を示す。まず図 5(a) の PESQ についてみる。ここで、PESQ は値が高いほど入力音声の自然性が高いことを意味する。従来法および提案法どれにおいてもアップサンプリングのみの (A) と同等かそれ以下という結果になった。従来法 1 はスペクトルをシフトしているだけであるため、スペクトルの滑らかさが維持されない場合に音の劣化に繋がるためだと考えられる [16]。また、提案法も調音構造は維持できるものの自然性の向上を担保した手法ではないため、PESQ の改善は保証されていない。次に RMS-LSD について比較する (図 5(b), (c))。RMS-LSD は値が低いほど、比較する 2 つの信号の誤差が小さいことを意味する。図 5(b) および (c) より、(D) 提案法 1, (E) 提案法 2, (F) 提案法 3 はどれも (C) 従来法 2 よりも低い値となっている。このことから非線形帯域拡張法によりエイリアシングの影響が提案法によって緩和されていることがわかる。一方、(B) 従来法 1 も低い値となっており、特に 32kHz に上げた時のスコアが一番低



(a) PESQ (Reference: original 16 kHz sampling data)



(b) RMS-LSD (16 kHz)



(c) RMS-LSD (32 kHz)

図 5 客観評価結果 (VLD データベース)

くなっている。(B) 従来法 1 もスペクトルシフティング後にエイリアシングを防ぐためのフィルタをかけていることが理由の 1 つだと考えられる。

次に、話者照合実験における結果について述べる。図 6 に $F_{s0} = 8 \text{ kHz}$, $F_{s1} = 16 \text{ kHz}$ としたときの各手法における EER を示す。(G) 8k と (H) 16k を比較すると、帯域制限された音声データを用いた場合では、音質および話者性が劣化するため EER が大幅に悪くなっている。アップサンプリングのみ行った音声 (A) UP も (G) 8k と同様に悪い EER となっており、また、(B) 従来法 1 および (C) 従来法 2 の EER についても改善がない。一方、(D) 提案法 1 および (E) 提案法 2, (F) 提案法 3 は、(B) 従来法 1 と (C) 従来法 2 よりも低い EER となった。これは、提案法が話者性を維持したまま、エイリアシングの影響を緩和することができたためだと考えられる。ここで、(D) 提案法 1 と (E) 提案法 2 の EER の違いについて考える。入力音声のサンプリング周波数が 8kHz の場合、4kHz までの帯域にのみ周波数成分があるため、なるべく多くの情報を残した

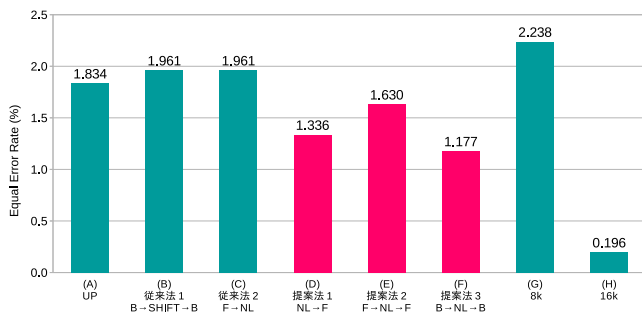


図 6 等価エラー率 ($F_{s_0} = 8 \text{ kHz}$, $F_{s_1} = 16 \text{ kHz}$)

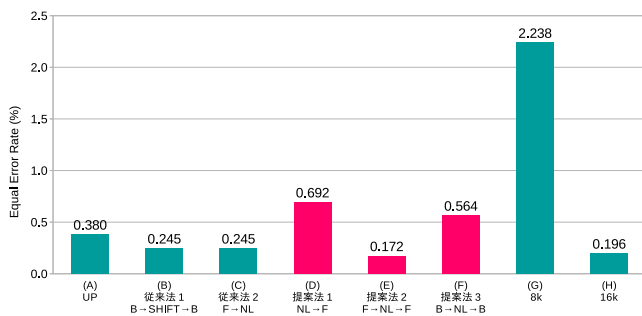


図 7 等価エラー率 ($F_{s_0} = 16 \text{ kHz}$, $F_{s_1} = 32 \text{ kHz}$)

まま非線形関数にかける方が話者性をより正しく維持可能であると考えられる。そのため非線形関数の前にフィルタをかけない (D) 提案法 1 の方が EER が低くなったと言える。(F) 提案法 3 の場合、非線形関数の前にバンドパスフィルタをかけているもののほぼ成分は残っているため性能が良かったと言える。

次に、 $F_{s_0} = 16 \text{ kHz}$, $F_{s_1} = 32 \text{ kHz}$ としたときの結果 (図 7) について述べる。図 7 において、(A) ~ (F) は入力音声として (H) 16k と同じ周波数成分を持っていることになる。しかし、(A) UP はサンプリング周波数が 32 kHz と上がっているため (H) 16k と MFCC のフィルタバンクのかかり方が変わるため同じ情報をもっているものの EER が悪くなっている。一方、(A) UP と (B) 従来法 1, (C) 従来法 2 を比較すると (H) 16k までは及ばないものの EER の改善が見られる。次に (D) 提案法 1 を見ると EER がかなり高くなってしまっているが、一方でこれは (E) 提案法 2 は原音声である (H) 16k よりも EER が低くなっている。図 6 の狭帯域音声を拡張する場合と異なり、入力音声に十分な情報が含まれているため非線形関数にかける成分をフィルタによって制限する方法がより適切に話者性を表現できたためだと考えられる。(F) 提案法 3 の EER は悪化していることも同様の理由だと言える。

図 6, 7 と比較すると F_{s_1} が 16kHz の場合に非線形関数を用いた手法 (C) ~ (F) の RMS-LSD の傾向と EER が対応しており PESQ と EER は逆の傾向となっていることから、帯域拡張法の精度を計る指標として EER と RMS-LSD は関係があると考えられる。 $F_{s_1} = 32 \text{ kHz}$ の場合は別の傾向となっているためフィルタとの関係についてより詳しい調査が必要である。

5. むすび

従来の帯域拡張法によって生成された信号にはエイリアシン

グの影響が含まれることから本稿では、回り込みを考慮した非線形帯域拡張法を提案した。話者照合実験の結果より、提案法は、エイリアシングの影響が緩和されたため従来法よりも高い話者性を表現できることを示した。また、RMS-LSD による評価において、提案法はエイリアシングを含む従来の帯域拡張法よりも高い評価が得られた。今後の課題としてはフィルタ設計との関連の調査や別のデータベースの使用などがあげられる。

謝辞

本研究の一部は科学研究費基盤 (B) 26280066 による。

文 献

- [1] H. seo, *et al.*, "A maximum a posteriorbased reconstruction approach to speech bandwidth expansion in noise," in Proc. ICASSP 2014, 6087–6091, 2014.
- [2] G.B. Song, *et al.*, "A study of HMM-based bandwidth extension of speech signals," Signal Processing, 89, 10, 2036–2044, 2009.
- [3] K. Li, *et al.*, "A deep neural network approach to speech bandwidth expansion," in Proc. ICASSP 2015, 4395–4399, 2015.
- [4] Y. Gu, *et al.*, "Waveform Modeling Using Stacked Dilated Convolutional Neural Networks for Speech Bandwidth Extension," in Proc. Interspeech 2017, 1123–1127, 2017.
- [5] Y. Wang, *et al.*, "Superwideband Extension For AMB-WB Using Conditional Codebooks," Acoustics, Speech and Signal Processing, 2014. ICASSP 2014. IEEE International Conference on, 3695–3698, 2014.
- [6] M. Tammi, *et al.*, "Scalable Superwideband Extension For Wideband Coding," in Proc. ICASSP, 161–164, 2009.
- [7] 宮本春奈 ら, "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法とその客観評価," 日本音響学会講演論文集 (春), 2018.
- [8] C. K. Un *et al.*, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 kbits/s," IEEE Trans. on Comm., 1466–1474, 1975.
- [9] S. Gohshi, *et al.*, "Limitations of super resolution image reconstruction and how to overcome them for a single image," in Proc. SIGMAP 2013, 71–78, 2013.
- [10] D.A. Reynolds, *et al.*, "Speaker verification using adapted gaussian mixture models, Diigital Signal Processing," 10, 19–41, 2000.
- [11] T. Thiruvaran, *et al.*, "Spectral shifting of speaker-specific information for narrow band telephonic speaker recognition," 51, 25, 2149–2151, 2015.
- [12] E. Larsen, *et al.*, "Efficient high-frequency bandwidth extension of music and speech," in Audio Engineering Society Convention 112., 23, 5627, 1–5, 2002.
- [13] Sayaka Shiota, *et al.*, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic apeaker verification," in Proc. Interspeech 2015, 239–243, 2015.
- [14] A. W. Rix, *et al.*, "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, 862, 2001.
- [15] R. M. Gray, *et al.*, "Distortion measures for speech processing," Acoustics, Speech and Signal Processing, IEEE Transactions, 28, 4, 367–376, 1980.
- [16] E. Larsen, *et al.*, "Reproducing Low-Pitched Signals through Small Loudspeakers," Journal of the Audio Engineering Society. Audio Engineering Society, 50, 147–164, 2012.