話者照合のための低周波数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 *

☆宮本春奈，塩田さやか，△貴家仁志　(首都大東京)

1　はじめに

電話などの通信を介する音声は通信速度を維持するために周波数帯域が制限されている．帯域制限のかかった音声は広帯域成分を失うことから個人性や明瞭性が低下し，また音声認識などのシステム性能も低下してしまう．この問題に対処するためにこれまで様々な帯域拡張法が提案されてきている [1–4].

筆者らはこれまでに処理量が非常に少ない帯域拡張法として非線形帯域拡張法を提案してきた．非線形帯域拡張法とは狭帯域音声に非線形関数を適用することで高周波成分を生成し，狭帯域成分と足し合わせることで広帯域音声を作成するものである．本研究では，非線形関数により生成される高周波成分がエイリアシングにより低周波成分に影響を与えてしまい音質が低下してしまう問題に着目した，非線形帯域拡張法の改善手法を提案する．話者照合実験において提案法では，従来法と比較し 31.9%のエラー削減率を得た．また，PESQ と RMS-LSD を用いた客観的音声評価においても，提案法が従来法よりも本来の広帯域音声に近くなったことを報告する．

2　非線形帯域拡張法 [5]

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．近年，画像信号処理の分野において非線形処理による超解像画像処理の手法が提案された [6]. この手法は低解像度の画像から高解像度の画像，つまり帯域制限のかかった画像から失われた高周波成分を擬似的に生成する手法である．本報告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 Fs_0 [Hz] でサンプリングされた狭帯域信号 $x\,[t]\,(t = 1, 2, \ldots, T)$ を Fs_1 [Hz] へアップサンプリングした信号 $y_{NB}[t]$ 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信号 $y_{HB}[t]$ を生成する．

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

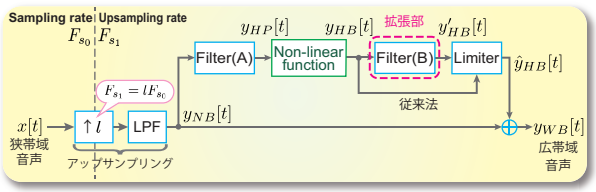


図 1　非線形帯域拡張法のフロー図 (従来法および拡張法)

ここで非線形関数の入力信号である $y_{NB}[t]$ 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) は以下のようになる．

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

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のようになり，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．

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

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．

図 2 に周波数 1.5 kHz，2 kHz，5 kHz，6 kHz の正弦波を足し合わせた混合信号を入力する狭帯域信号 (サンプリング周波数: 16 kHz) とし，アップサンプリングの倍率 l を 3 として非線形帯域拡張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．一方で，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

* Non-linear artificial bandwidth extension considering aliasing artifacts for speaker verification and its objective evaluation. by MIYAMOTO, haruna and SHIOTA, sayaka and KIYA, hitoshi (Tokyo Metropolitan University)

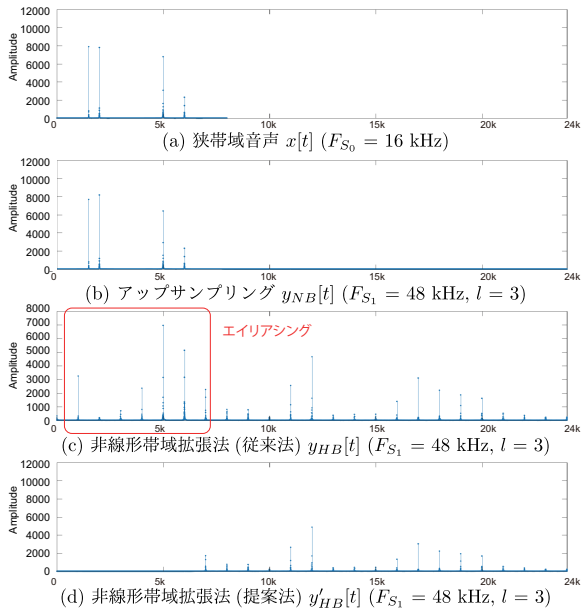


図 2　振幅スペクトルの例 (1.5 kHz, 2 kHz, 5 kHz, 6 kHz の正弦波を合成)

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 $y_{WB}[t]$ にノイズが乗る可能性がある．

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) に示すような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エイリアシングの影響のない高周波成分のみを持つ信号 $y_{WB}[t]$ が生成され (図 2 (d))，元の狭帯域音声信号 $y_{NB}[t]$ と足し合わせることで広帯域信号 $\hat{y}_{HB}[t]$ を得ることができる．

図 3 に，16 kHz でサンプリングされた原音声 (a)，サンプリング周波数を 8 kHz から 16 kHz にアップサンプリングされた音声 (b)，従来の帯域拡張法による広帯域音声 (c)，提案法により生成された広帯域音声 (d) それぞれ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提案法による広帯域音声は，それぞれ図 1 の $y_{NB}[t]$ および $y_{WB}[t]$ に対応する．図より，従来の広帯域信号および提案法の広帯域信号は，高周波数帯域 (4~8 kHz) において高周波成分が生成

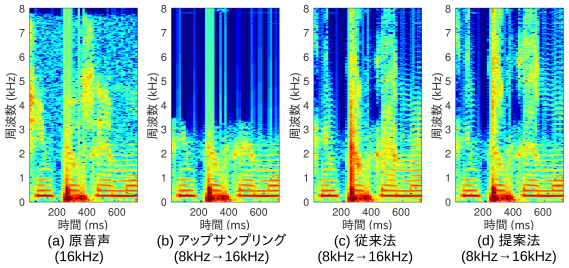


図 3　スペクトログラム (F_{s_0} = 8kHz, F_{s_1} = 16kHz)

されていることがわかる．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．提案法によって生成された広帯域信号は狭帯域信号 $y_{NB}[t]$ から生成されるため，明瞭度だけではなく話者性も向上する．

4　実験

提案法の有効性を評価するために，GMM-UBM に基づく話者照合実験 [7] および客観評価実験を行った．

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 8 kHz から 16 kHz へ帯域拡張する場合 ($l = 2, F_{s_0}$ = 8 kHz, F_{s_1} = 16 kHz) と 16 kHz から 32 kHz へ帯域拡張する場合 ($l = 2, F_{s_0}$ = 16 kHz, F_{s_1} = 32 kHz) の 2 つの条件で実験を行った．JNAS データベースのサンプリング周波数は 16 kHz であるため，帯域拡張を行う場合については 16 kHz からダウンサンプリングを行い 8 kHz にした音声を狭帯域音声 $x[t]$ 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 48 kHz であるため，一度全てのデータを 16 kHz 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．比較方法を以下に詳細に示す．

(A) UP
狭帯域音声に対してアップサンプリングのみ行った音声 ($y_{NB}[t]$) を学習およびテストデータとして使用した．

(B) 従来法
狭帯域音声に対して従来の非線形帯域拡張法を適用し生成した音声データ ($y_{WB}[t]$) を学習およびテストデータとして使用した．Filter(A) にはハイ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 16 kHz, 32 kHz どちらの場合も α , β はそれぞれ 1.8, 100 とした．

(C) 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

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 (UBM)	23657 文章
データベース (特定話者モデル)	VLD データベース [8] (ヘッドセット)
話者	17 名 (女性)
学習データ (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 + ΔΔ

を適用し生成した音声データ ($y_{WB}[t]$) を学習およびテストデータとして使用した．Filter(A) にはオールパスフィルタを，Filter(B) にはハイパスフィルタを用いた．帯域拡張後のサンプリング周波数 F_{s_1} が 16 kHz, 32 kHz どちらの場合も α ，β はそれぞれ 1.8，100 とした．

(D) 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 ($y_{WB}[t]$) を学習テストデータとして使用した．ただし Filter(A)，Filter(B) どちらにもハイパスフィルタを用いた．α，β は，F_{s_1} が 16 kHz の場合は 1.5，100 とし，また，32 kHz の場合は 2.5，15500000 とした．

(E) 8k

学習およびテストデータとしてサンプリング周波数 8 kHz の音声を使用した．

(F) 16k

学習およびテストデータとしてサンプリング周波数 16 kHz の音声を使用した．

パラメータ β は信号の振幅にのみ関係するため信号と α の値によって適切に設定されている．話者照合の評価には，等価エラー率 (EER) を用いた．

客観評価実験では，PESQ [9] と RMS-LSD (Root Mean Square - Log Spectral Distance) [10] 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 16 kHz に下げたものをリファレンスとして各スコアを計った．

4.2　実験結果

図 4 に $F_{s_0} = 8$ kHz, $F_{s_1} = 16$ kHz としたときの各手法における EER を示す．(E) 8k と (F) 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，話者性が劣化したため 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

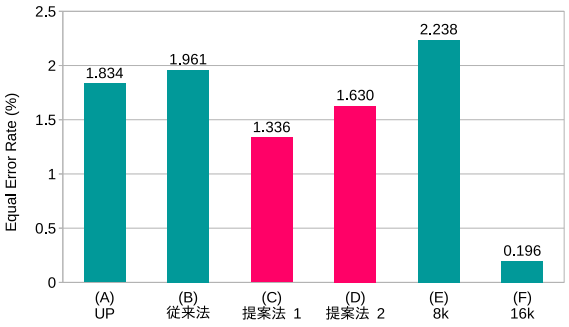


図 4　等価エラー率 ($F_{s_0} = 8$ kHz, $F_{s_1} = 16$ kHz)

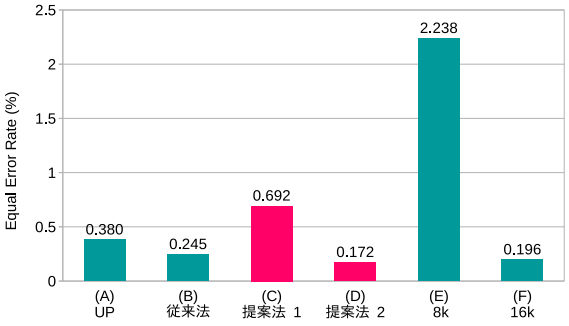
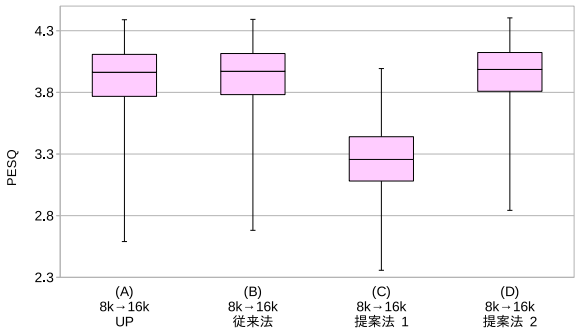


図 5　等価エラー率 ($F_{s_0} = 16$ kHz, $F_{s_1} = 32$ kHz)

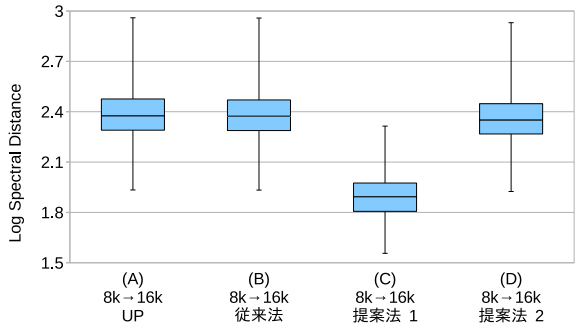
も (E)　8k と同様に悪い EER となっており，また，(B) 従来法の EER も改善がない．一方，(C) 提案法 1 および (D) 提案法 2 は，(B) 従来法よりも低い EER となった．これは，エイリアシングの影響を緩和することができたためだと考えられる．(C) 提案法 1 と (D) 提案法 2 の違いについて考えると，サンプリング周波数が 8 kHz の場合，4 kHz までの帯域にのみ周波数成分があるため，なるべく多くの情報を残したまま非線形関数にかける方が話者性の表現能力を向上させることが可能であると考えられる．

次に，$F_{s_0} = 16$ kHz, $F_{s_1} = 32$ kHz としたときの結果 (図 5) について述べる．図 5 において，(A) ~ (D) は狭帯域音声として (F) 16k と同じ周波数成分を持っていることになる．(A) UP は (F) 16k と比べて，MFCC のフィルタバンクのかかり方が変わってしまうため同じ情報をもっているものの EER が高くなっている．(A) UP と (B) 従来法を比較すると (F) 16k までは及ばないものの EER の改善が見られる．ここで (C) 提案法 1 を見ると EER がかなり高くなってしまっていることがわかる．一方，(D) 提案法 2 は (F) 16k よりも EER が低くなっている．図 4 の結果と比べ，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と考えられる．

図 6 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



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



(b) RMS-LSD (16 kHz)

図 6　客観評価結果 (VLD データベース)

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．PESQ は音の自然性を表す尺度で値が高いほど入力音声が自然であることを意味する．(A), (B), (D) の中央値はほぼ等しかったが，(C) の中央値は明らかに低かった．提案法は，音声の自然性向上を目的としていないため，PESQ の改善は保証されない．実際，図 4 において，(C) の EER が最も低いことから PESQ の改善が話者照合精度の改善に繋がるわけではないことを示している．次に各手法の RMS-LSD を比較する (図 6 (B))．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．(C) 提案法 1 および (D) 提案法 2 は，両方の条件下での (B) 従来法よりも低い値を達成した．また，図 4 と比較すると 16 kHz の場合，RMS-LSD が低いものは等価エラー率も低く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等価エラー率と RMS-LSD は関係があると期待できる．

5　おわりに

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．実験結果より，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再現できることを示した．さらに PESQ と RMS-LSD により評価した結果，提案法は従来の帯域拡張法よりも高い評価を得られることが確認できた．

今後の課題としては，実用的通信方式を用いて評価を行うことや他の帯域拡張手法と比較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) 26280066 による．

参考文献

[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] 塩田さやか ら, “低周波成分への影響を考慮した非線形帯域拡張法と音声認識への応用,” 日本音響学会講演論文集 (秋), 159–160, 2017.

[6] S. Gohshi, et. al., “Limitations of super resoltion image reconstruction and how to overcome them for a single image,” in Proc. SIGMAP 2013, 71–78, 2013.

[7] D.A. Reynolds, *et al.*, “Speaker verification using adapted gaussian mixture models, Diigital Signal Processing,” 10, 19–41, 2000.

[8] Sayaka Shiota, *et al.*, “Voice liveness detection algorithms based on pop noise caused by human breath for automatic apeaker verfication,” in Proc. Interspeech 2015, 239–243, 2015.

[9] 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.

[10] R. M. Gray, *et al.*, “Distortion measures for speech processing,” Acoustics, Speech and Signal Processing, IEEE Transactions, 28, 4, 367–376, 1980.