非線形帯域拡張法における客観評価尺度と遅延時間の評価*

☆宮本春奈，塩田さやか，貴家仁志　(首都大学東京)

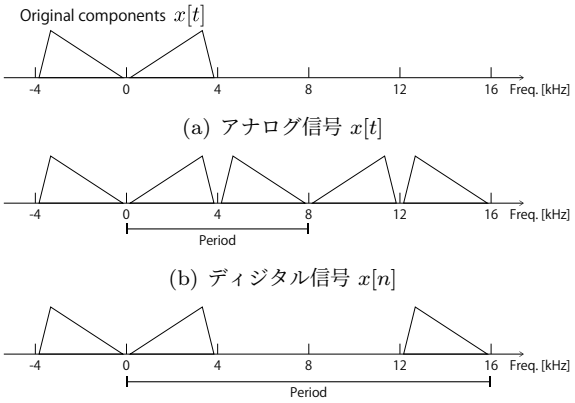
1　はじめに

近年，テレビ電話のようなリアルタイム通信を行うアプリケーションが広く利用されている．しかし，画像と音声両方の情報を用いる通信の場合，2つの情報にずれが生じるとユーザは不快感を感じてしまうことが知られている [1]．そのため，音声強調や帯域拡張等の音声の品質改善手法をユーザに提供するためには，非常に低遅延な手法が求められる．

帯域拡張法はサンプリング周波数が異なるときにサンプリング周波数が高い方に合わせるために必要となる技術であり，音質改善の点からも重要な技術である．これまで様々な帯域拡張法が提案されてきているが，大きく分けると non-blind 法もしくは blind 法になる．non-blind 法とは，低周波成分と符号化された高周波情報を付帯情報として用いて失われた高帯域成分を再現するもので，blind 法とは低周波成分のみから失われた高周波成分を生成するものである．付帯情報を必要とする non-blind 法より blind 法の方が汎用性が高いため blind 法が主に研究されている．また，帯域拡張法は学習型と非学習型の手法に分類することもできる．筆者らはこれまでに blind かつ非学習型の非線形帯域拡張法を提案してきた [2]．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手法である．これまでに従来の帯域拡張法よりも高い話者性が得られることを報告してきたが，遅延量に関する比較実験を行っていなかった．そこで本稿では，複数の blind かつ非学習型である帯域拡張法をアルゴリズムとしての遅延量に関して，比較評価する．また，3つの客観評価尺度を用いた評価も行い，遅延量と音質についても考察を行った．実験結果より，非線形帯域拡張法はサンプリング周波数に依存せず，低遅延で誤差の小さい音声を生成することを報告する．

2　帯域拡張法

帯域拡張法とは，低いサンプリング周波数の音声信号である狭帯域音声信号を目的のサンプリング周波数まで上げる際に必要な技術であり，失われた高周波成分を生成する方法である．これまでに提案された手法を大きく分けると blind 法と non-blind 法に分類することができる．本稿は blind かつ非学習型である帯域拡張法に着目する．blind かつ非学習型の帯域拡張法では基本的にアップサンプリングを行い，空いた周波数領域へ高周波成分を生成する．図 1 に周波数領域でのアップサンプリング処理の過程を示す．



(a) アナログ信号 $x[t]$

(b) ディジタル信号 $x[n]$

(c) アップサンプリング信号 $y_{NB}[n]$

図 1　周波数領域におけるアップサンプリング処理の過程

図 1 (a) はアナログ信号であるため帯域制限および周期性をもたない．このアナログ信号がデジタル信号に変換されると，信号は離散時間信号となるため，図 1 (b) に示すように，信号がある一定の帯域幅に制限され，かつ周期性をもつ．図 1 (c) は図 1 (b) の信号に対し補間器とローパスフィルタによるアップサンプリングを行うことで一定の帯域に情報を持たないアップサンプリングされた信号を示している．blind 法ではこの空いた周波数成分にどのように信号を生成するかで様々な手法が提案されている．

2.1　スペクトルシフティング法 (SHIFT) [5, 6]

blind かつ非学習型の帯域拡張法の一つである SHIFT では，4 kHz 未満の周期を変調することによって高周波成分を生成し，その成分をアップサンプリングにより空いた周波数領域にシフトすることで広帯域音声を生成している．単純な処理のため処理量が非常に少ないという利点がある．

2.2　線形予測分析合成法 (LPAS) [7]

LPAS は，線形予測分析を用いて高周波成分を生成する手法である．低周波数成分からスペクトルエンベロープおよび残差誤差情報を抽出することで高周波成分を生成している．生成された高周波成分は，SHIFT により帯域拡張された音声よりも自然性が高い．しかし，処理の過程に FFT を含むためアルゴリズムとしての遅延量が大きいという課題がある．

2.3　リアルタイム通信を行うアプリケーションにおける遅延問題

近年，ネットワーク環境の普及により，ビデオ通話やリアルタイム通信を行うアプリケーションが広く

*Evaluation on objective measurements and latency of non-linear bandwidth extension method.　by MIYAMOTO, haruna and SHIOTA, sayaka and KIYA, hitoshi (Tokyo Metropolitan University)

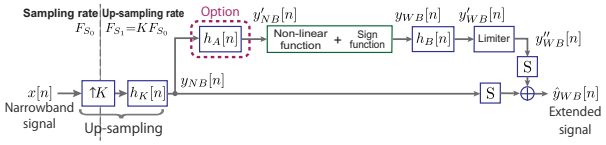


図 2　非線形帯域拡張法のフロー図

使用されるようになった．そのため，音声信号処理においても低遅延で高品質な音声を生成可能な技術が求められている．使用者がアプリケーションを使用する際に得られる視覚情報と聴覚情報との間に時間差が生じてしまうと違和感を感じる可能性があるためである．実際に，視覚情報と聴覚情報との差が 10 ms 以上になると使用者が不快感を感じるという報告がある [1]．しかし，これまで帯域拡張法と遅延量について比較評価されていなかった．

3　非線形帯域拡張法と遅延問題

3.1　非線形帯域拡張法 (N-BWE)

blind かつ非学習型であり，遅延量の少ない手法として非線形帯域拡張法 (N-BWE) の改善法が報告されている [2]．この章では，その手順と遅延問題および応用されるアプリケーションについて説明する．

図 2 に N-BWE のフロー図を示す．アップサンプリングされた信号 $y_{NB}[n]$ は，補間器 K と線形デジタルフィルタ $h_K[n]$ によって生成されており (図 1 (c))，n は離散時間変数を表す．$y_{NB}[n]$ のサンプリング周波数は F_{S_1} であり，$y_{NB}[n]$ は広帯域成分が失われている．ここで，非線形関数を用いることで連続時間信号における広帯域成分を生成することができる．非線形関数は次のように定義される．

$$y_{WB}[n] = \mathrm{sgn}(y'_{NB}[n]) \cdot |y'_{NB}[n]^{\alpha}| \times \beta, \qquad (1)$$

ただし，

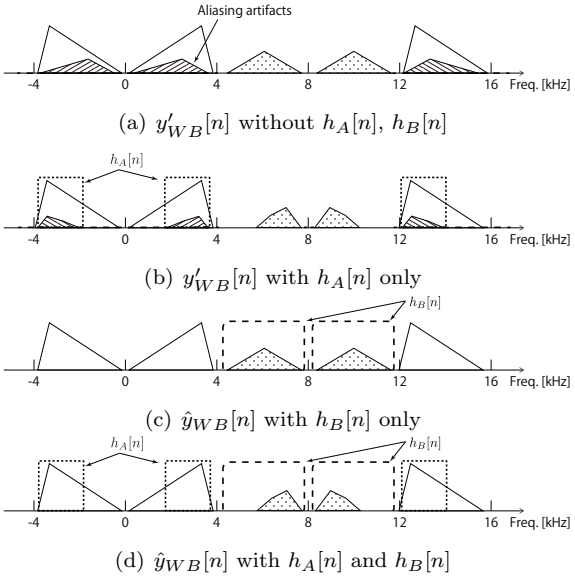
$$\mathrm{sgn}(a) = \begin{cases} 1 & (a > 0) \\ 0 & (a = 0) \\ -1 & (a < 0) \end{cases}. \qquad (2)$$

ここで，パラメータ α と β は非線形性を調整するもので，a は実数である．図 2 に示すリミッターは以下のように定義されている．

$$y''_{WB}[n] = \begin{cases} y'_{WB}[n], & y'_{WB}[n] \le T_h \\ M, & y'_{WB}[n] > T_h \end{cases}. \qquad (3)$$

T_h は閾値を示し，M は一定値とする．図 2 の S では低周波数成分および高周波数成分の生成する過程におけるそれぞれの遅延を補正している．

図 3 に周波数領域における N-BWE で使用されるフィルタの効果を示す．図 3 (a) に示すように，フィルタを用いない場合高周波成分は帯域全体に生成されエイリアシングが発生してしまう．図 3 (a) と (b) を比較すると，フィルタ $h_A[n]$ は非線形関数にかける信号の成分をフィルタリングすることで非線形関数により生成される信号を制限する役割を果たすことがわかる．これは特にノイズが含まれる音声において強調したくない帯域がある場合に有効なフィルタである．図 3 (a) と (c) を比較すると，(a) では周り込みによるエイリアシングが発生しているが，(c) ではフィルタ $h_B[n]$ を適用していることでその影響が緩和されていることがわかる．図 3 (d) では，(c) にフィルタ $h_A[n]$ を追加した方法となっているため，エイリアシングを含まず，かつ図 3 (b) と同様の効果が期待できる手法である．本稿では (c) と (d) に着目して調査を行った．



(a) $y'_{WB}[n]$ without $h_A[n]$, $h_B[n]$

(b) $y'_{WB}[n]$ with $h_A[n]$ only

(c) $\hat{y}_{WB}[n]$ with $h_B[n]$ only

(d) $\hat{y}_{WB}[n]$ with $h_A[n]$ and $h_B[n]$

図 3　周波数領域における非線形帯域拡張法のフィルタの影響 ($K = 2$)

3.2　非線形帯域拡張法の遅延問題

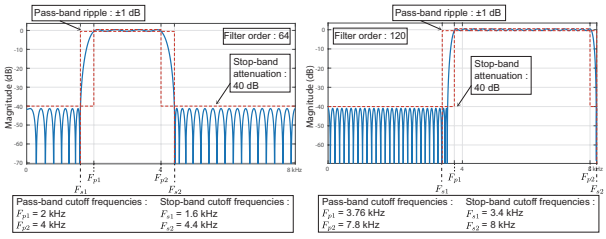
ビデオ電話のようなリアルタイム通信を行うアプリケーションには，送信される信号のサンプリング周波数に合わせた帯域拡張法が必要となる．また，帯域拡張された音声は自然性だけでなく，音声認識や話者照合などの音声信号処理システムにおいても高い精度を維持することが求められる．N-BWE では，個人性と RMS-LSD [8] の観点から高性能であることが報告されている [2]．しかし，アルゴリズムとしての遅延量は使用するフィルタの次数に大きく依存するため各帯域拡張法のと遅延量と客観評価を比較した報告が必要である．

4　実験

本実験では，様々な帯域拡張法の遅延量を比較評価し，さらに，客観評価尺度と遅延量の関係についても調査した．

4.1　実験条件

実験で用いる音声データには 2 種類の英語データベースを使用した．一つ目の CMU データベース [9] は，話者 3 名からなる計 3377 文章で構成されており，32 kHz でサンプリングされたものである．もう一つは TSP データベース [10] に含まれる 3gpp データで，



(a) BPF: $h_A[n]$　　(b) BPF: $h_B[n]$

図 4　フィルタ仕様

その中の英語話者 6 名からなる計 12 文章を使用した．TSP データベースのサンプリング周波数は 48 kHz である．実験結果は二つのデータベースを合わせた数値で示している．

近年，IP ベースの高速パケットアクセス回線による VoIP(Voice over IP) の普及により，電話音声信号の圧縮が重要視されなくなってきた．また，音声符号化に関しては，より広い帯域の音声信号を符号化する手法へと注目が移っており，広帯域以上 (超広帯域) の音声による通話サービスが普及しつつある．そのため本来の音声により近い音声を再現することが可能となってきたが，通信網が全て同一規格になったわけではないため，広帯域から超広帯域へ拡張可能な帯域拡張技術も必要となってきている．そこで本稿では二つのシナリオを想定し実験を行った．一つは，入力信号のサンプリング周波数が 8 kHz で，帯域拡張法により 16 kHz へ周波数を上げた音声を生成するシナリオ ($K = 2, F_{s_0} = 8$ kHz)．次に，入力信号がサンプリング周波数 16 kHz で超広帯域 (Super wide-band; SWB) サービスを前提とし 32 kHz へ周波数を上げた音声を生成するシナリオである ($K = 2, F_{s_0} = 16$ kHz)．以降，それぞれ WB シナリオ，SWB シナリオと呼ぶ．これらのシナリオに適用させるために，使用するデータを目的のサンプリング周波数までダウンサンプリングしてから実験を行った．本実験で比較に用いた手法は以下の 5 手法である．

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

(B) SHIFT
狭帯域音声に対してスペクトルシフト法 (SHIFT) [5] を適用し生成した音声データ ($y_{WB}[t]$) をテストデータとして使用した．$h_A[n]$，$h_B[n]$ にはバンドパスフィルタを用いた．

(C) LPAS
狭帯域音声に対して線形予測分析合成法 (LPAS) [7] を適用し生成した音声データ ($y_{WB}[t]$) をテストデータとして使用した．フィルタは [7] に記載されているものを使用した．

(D) N-BWE I
狭帯域音声に対して 3 章の非線形帯域拡張法を適用し生成した音声データ ($y_{WB}[t]$) をテストデータとして使用した．フィルタ $h_B[n]$ の設計を式 (4) に示す．$h_B[n]$ には図 4 に示すバンドパスフィルタを用いた．α，β の値は，WB，SWB 両シナリオともに 1.8，100 を用いた．

$$h_A[n] = \begin{cases} 1 & (n = 0) \\ 0 & (n \neq 0) \end{cases}. \tag{4}$$

(E) N-BWE II
(D) N-BWE I と同様に狭帯域音声に対して 3 章の非線形帯域拡張法を適用し生成した音声データ ($y_{WB}[t]$) をテストデータとして使用した．ただし $h_A[n]$，$h_B[n]$ には図 4 に示すバンドパスフィルタ (a)，(b) をそれぞれ用いた．α，β は，WB，SWB 両シナリオともに 1.5，100 を用いた．

実験で用いる $h_A[n]$，$h_B[n]$ および α と β は K の値にのみ依存するが，K は両シナリオにおいて同じ値を使用したため，両シナリオ共通の設定を用いた．

客観評価実験には RMS-LSD [8]，PESQ [3]，STOI [4] の 3 つの尺度を使用した．RMS-LSD は値が低いほど，生成された音声が原音声に近いことを示す．PESQ と STOI は生成音声の自然性を評価する尺度で，値が高いほど，自然性が高いことを意味する．PESQ と STOI は WB シナリオでのみ評価し，RMS-LSD は WB，SWB 両シナリオで評価を行った．

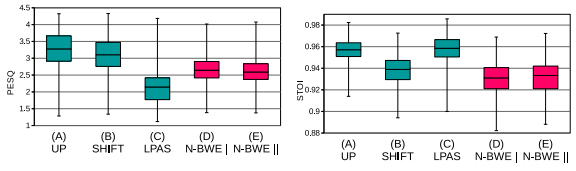
表 1　アルゴリズムとしての遅延量

手法	遅延 (ms)
(A) UP	0.068
(B) SHIFT	0.643
(C) LPAS	14.187
(D) N-BWE I	**0.443**
(E) N-BWE II	0.643

4.2　実験結果

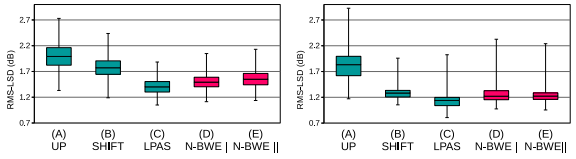
手法ごとの遅延量を比較するために，WB シナリオ ($K = 2, F_{s_0} = 8$ kHz, $F_{s_1} = 16$ kHz) で評価した結果を表 1 に示す．結果より，(D) N-BWE I が最も遅延量が少なかった．これは (D) では使用しているフィルタが，次数の小さいバンドパスフィルタ 1 つであり，遅延があまり生じないためである．(B) と (E) は (D) と比べるとフィルタが一つ追加されているものの次数があまり高くないため，(D) とほぼ同様の遅延量となった．(C) はフィルタだけでなく FFT の点数が遅延に影響するために遅延量が 10 ms を超えてしまっている．このことから，(C) 以外はリアルタイム通信を行うアプリケーションに十分適用可能であるといえる．

次に各手法を客観評価尺度を用いて評価した．図 5, 6 は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．



PESQ　　STOI

図 5　自然性評価 (WB シナリオ)



(a) WB シナリオ　　(b) SWB シナリオ

図 6　RMS-LSD

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.

図 5 に WB シナリオの場合の PESQ および STOI の結果を示す. PESQ の値は 0∼4.5 の範囲で, STOI の値は 0.0 ∼ 1.0 の範囲で表されており, どちらも値が大きい方が自然性が高いことを示している. PESQ, STOI どちらの結果からも帯域拡張を施した手法 ((B)∼(E)) は全て (A) よりも値が低かったことがわかる. 帯域拡張法では, 振幅情報を生成するが位相情報を考慮しないために広帯域音声を生成した際に, 自然性が低下したと考えられる. PESQ と STOI は自然性評価を行っているためほぼ同じ傾向になるが, (C) のみ PESQ が悪く, STOI が良い値となっている. PESQ はリファレンス音声との歪みを聴覚モデルを介して評価しているが, STOI は明瞭度を測る指標となっているため, LPAS は元音声と比較すると歪みが大きいが明瞭度は高いという結果だったことがわかる.

図 6 (a), (b) に WB, SWB シナリオそれぞれでの RMS-LSD の結果を示す. WB, SWB 両シナリオでの傾向はほぼ同様となった. 図 6 のどちらの結果でも (C) が最小の結果となったが, 非線形帯域拡張法である (D), (E) も (C) とほぼ同様の結果となった. (D), (E) と同程度の遅延量である (B) はスペクトルの誤差としては大きくなっている一方, (D), (E) は遅延量も非常に低く, スペクトル歪みも LPAS と同等であることから非線形帯域拡張法の有効性が確認できた.

以上の結果より, 非線形帯域拡張法はリアルタイム通信を行うアプリケーションにおいて効果的であるといえる.

5　おわりに

本稿では非線形帯域拡張法の遅延量を評価し, また客観評価尺度と遅延量との関係について調査を行った. 特に本稿では, blind かつ非学習型である低遅延な手法として非線形帯域拡張法と他の帯域拡張法との比較を行った. 実験結果より, 非線形帯域拡張法では比較した帯域拡張法の中で最も低遅延であり, また RMS-LSD 値から生成音声が原音声と比べ誤差が小さいことを示した.

今後の課題としては, ITU-T G712 [11] の実用方式に則った音声に対して非線形帯域拡張法を適用することが考えられる. また, 他の機械学習型のシステムと比較することが挙げられる.

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] J. Agnew, *et al.*, "Just noticeable and objectionable group delays in digital hearing aids," Jounal pf the American Academiy Audiology 11, 330–336, 2000.

[2] H. Miyamoto, *et al.*, "Non-linear harmonic generation based blind bandwidth extension considering aliasing artifacts," in Proc. APSIPA Annual Summit and Conference, 2018.

[3] AW. 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.

[4] C. H. Taal, *et al.*, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," IEEE Trans. Audio, Speach, Language. Process., 19, 7, 2125–2136, 2011.

[5] T. Thiruvaran, *et al.*, "Spectral shifting of speaker-specific information for narrow band telephonic speaker recognition," Electronics Letters, 51, 25, 2149–2151, 2015.

[6] E. Larsen, *et al.*, "Efficient high-frequency bandwidth extension of music and speech," 112th AES Convention, 23, 5627, 2002.

[7] P. Bachhav, *et al.*, "Efficient Super-Wide Bandwidth Extension Using Linear Prediction Based Analysis-Synthesis," in Proc. IEEE International Conference on Acoustics, Speech and Signal, 5429–5433, 2018.

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

[9] J. Kominek, *et al.*, "CMU ARTIC database for speach synthesis," `http://festvox.org/cmu_artic/index.html`, 2003.

[10] P. Kabal, "TSP Speach Database," `http://mmsp.ece.mcgill.ca/Documents/Data/`, 2002.

[11] ITU-T, Recommendation G.712., "Transmission Performance Characteristics for Pulse Code Modulation Channels," 1996.