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 *

◎塩田さやか，△貴家仁志（首都大）

1　はじめに

近年，Skype や IP 電話，スマートフォンの電話用アプリケーションなどといった音声の通話用帯域ではなく広帯域を用いた音声通信が普及しつつある．しかし，全ての音声通信がそのような広帯域の通信網を用いるための制度の移行には長い時間がかかるため，携帯電話などの通信速度を確保するための帯域制限がかかった音声での通信が今後も長らく使われることが想定されている．帯域制限が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てしまうため，これまでに失われた広帯域成分を復元するための手法として様々な帯域拡張法が提案されてきている [1, 2]．筆者らはこれまでに処理量が非常に少ない手法として非線形帯域拡張法を提案してきた [3]．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広帯域成分を生成し、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、非線形関数を適用して生成する広帯域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うことを考慮して、非線形帯域拡張法を拡張した手法を提案する．音声認識実験において従来法と比較し，提案法の単語正解精度が約 11 ポイント向上したことを報告する．

2　非線形帯域拡張法 [3]

はじめに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．ただし，従来法では非線形関数の出力は拡張部を通らずに Limiter 部に入力される．フローではまず，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 $y_{HP}[n]$ に非線形処理を施し広帯域成分 $y_{HB}[n]$ を生成する．用いる非線形関数は以下のように定義される．

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

ここで，n はサンプリング点，α および β は非線形関数を決定するためのパラメータである．非線形処理を施した信号 $y_{HP}[n]$ の振幅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ったものを $y'_{HB}[n]$ とする．最後に，生成した広帯域成分 $y_{HB}[n]$ と狭帯域成分のみの $y_{NB}[n]$ を足し合わせることで帯域拡張された信号 $y_{WB}[n]$ を生成する．非線形帯域拡張法は処理が非常に軽く，また，任意のサンプリング周波数に拡張することも可能であるという利点がある．

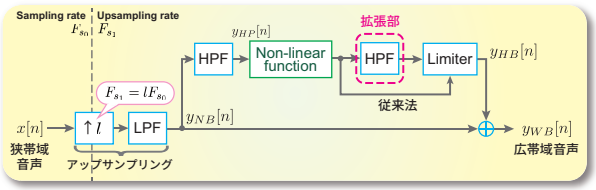

図 1　拡張された非線形帯域拡張法のフロー

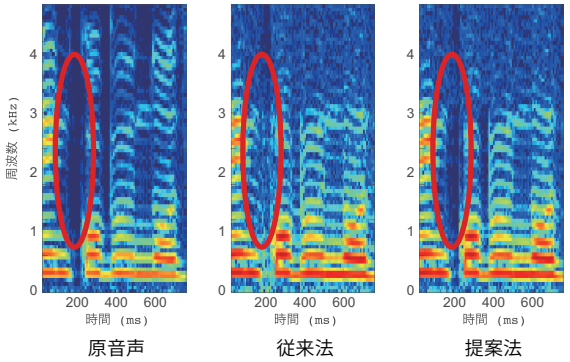

図 2　音声スペクトログラムによる低周波数領域の比較

3　提案法

従来の非線形帯域拡張法では，非線形関数を通すことで生成した広帯域成分がエイリアシングの影響で低周波成分にまで周りこむ可能性があった．そのため，元の狭帯域音声と足し合わせる際に低周波成分に影響を与えていると考えられる．そこで，提案法では図 1 の拡張部を従来の非線形帯域拡張法に加えることで回り込んでしまった周波数成分を除去してから元の狭帯域音声と足し合わせることを提案する．図 2 に原音声，従来法，提案法それぞれのスペクトログラムを示す．特に従来法の楕円で囲まれた部分が周りこみの影響をうけているが提案法ではその影響が緩和されていることが確認できる．音声は特に低周波領域に言語特徴が強く現れるため，低周波成分が影響を受けることは音声認識の性能低下に繋がる可能性があるが提案法ではその影響が緩和されると期待される．

* A non-linear bandwidth extension method considering effect of low frequency components and it's application for speech recognition, by SHIOTA Sayaka, KIYA Hitoshi (Tokyo Metropolitan University)

4　評価実験

提案法の音声認識に対する有効性を確認するために音声認識器として Julius (ver.4.4.2) およびディクテーションキット (ver.4.4) の DNN-HMM に基づく音声認識実験を行った．

4.1　実験条件

音声認識に用いる音響モデルおよび言語モデルは Julius と合わせて公開されているディクテーションキットに同梱されているものを用いた．音響モデルの概要は次の通りである．学習データに JNAS および『日本語話し言葉コーパス』模擬講演データを用いた性別非依存の DNN-HMM 音響モデル．DNN の構成は，入力層 1320 ノード，出力層 2004 ノード，中間層 2048 ノード，隠れ層 5．言語モデルの概要としては，国立国語研究所の「現代日本語書き言葉均衡 corpus」(BCCWJ) の全テキスト (約 1 億語) を用いた単語 Trigram モデルで語彙サイズは約 59000．音声データの特徴量としては 40 次元のフィルタバンクおよび動的特徴量および二次動的特徴量の 120 次元を用い，また静的特徴量においてはケプストラム平均正則化を適用している．DNN へは 11 フレームを連結した 1320 次元の特徴量として入力を行っている．その他の Julius に用いるパラメータは予備実験により調整したものを用いて音声認識実験を行った．テストデータは JNAS より学習データに含まれない男性 23 名，女性 23 名を選び，各話者 100 文章の合計 4600 文章を用いた．データベースのサンプリング周波数は 16kHz であり，帯域拡張を行う手法については 16kHz からダウンサンプリングを行い 8kHz にした音声を狭帯域音声として用いた．

次に比較手法について述べる．16kHz サンプリングで収録された原音声（Original）の音声を oracle なデータとし，原音声を 8kHz サンプリングにダウンサンプリングしてからアップサンプリングを行った音声（Upsampling）をベースラインとした．アップサンプリングした音声は図 1 の $y_{NB}[n]$ に該当する．アップサンプリングした音声に従来の非線形帯域拡張法を適用したものが従来法，提案手法を適用したものを提案法とした．また，非線形関数に用いるパラメータ α は 1.5 とした．

4.2　実験結果

図 3 に非線形関数のパラメータ β が $200, 300, 400$ のときの従来法と提案法の単語正解精度を示している．従来法はベースラインである Upsampling より非常に高い性能を得られているが，提案法は従来法よりも更に性能が向上していることがわかる．従来法で

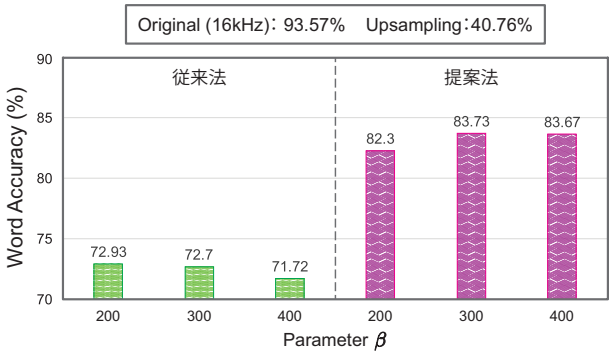

図 3　単語正解精度（$\alpha = 1.5$）

はパラメータ β の値が上がるにつれて単語正解精度が徐々に低下している．一方，提案法では β の値があがると精度も向上していることがわかる．これは β の値が大きくなると生成した広帯域成分のパワーが全体的に強くなることで低周波領域に回り込んでしまった成分も大きくなってしまい認識性能に影響を与えていたからだと考えられる．提案法では非線形関数をかけた後にフィルタを通すことで低周波成分の影響を緩和できるため性能の改善に繋がったと言える．

5　おわりに

本稿では，非線形帯域拡張法の拡張を行うことで低周波成分に回り込んでいた周波数成分の影響を緩和することを提案した．音声認識実験の結果より，提案法は従来法と比較して単語正解精度が約 11 ポイント向上し提案法の有効性を示すことができた．

今後の課題として，学習部においても提案法を用いた時の有効性についての検討及びノイズ環境下における性能の調査，他の帯域拡張法との比較，フィルタについての調査などが挙げられる．

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] Y. Wang, et. al., “Speech Bandwidth Extension Based on GMM and Clustering Method,” 2015 Fifth International Conference on Communication Systems and Network Technologies, pp. 437-441, 2015.

[2] K. Li, et. al., “A deep neural network approach to speech bandwidth expansion,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4395–4399, 2015.

[3] 塩田さやか, 貴家仁志, “非線形帯域拡張法に基づく音声認識の改善,” 日本音響学会春季研究発表会, 2017.