非線形帯域拡張法に基づく音声認識の改善*

◎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．しかし，このような帯域制限がかかった音声は広帯域成分を失ってしまうために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．帯域拡張法の主な例としては，ピッチ抽出による基本周波数成分生成やスペクトルエンベロープの推定，広帯域へのマッピング推定などが挙げられる．一方，筆者らは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．非線形帯域拡張法では狭帯域音声信号に非線形関数を適用することで広帯域成分を生成する．生成された広帯域成分と狭帯域成分を合わせた広帯域音声を用いることによって話者照合システムにおける性能改善が得られることをこれまでに報告してきた．本論文ではこの手法を用いることで帯域制限によって大幅に性能が低下してしまう音声認識についても性能を改善することを報告する．

2 帯域拡張法

帯域拡張法とは，帯域制限により失われた広帯域成分を復元するため手法であり，これまでに様々な手法が提案されてきている．本章では，それらの帯域拡張法について簡単に紹介する．

処理が軽い帯域拡張法の例としては，低帯域成分の一部を抽出し加工したものを広帯域成分に複製するような手法 [1, 2] やピッチ抽出により基本周波数成分を生成する手法 [3–5]，低帯域成分から広帯域スペクトルエンベロープを推定する手法 [6–8] などが挙げられる．また，LPC や線形周波数スペクトル (LFS)，MFCC など様々な特徴量表現をもとに低帯域成分と広帯域成分のマッピングをとる手法も多い [9, 10]．モデルベースの手法としては，ガウス混合モデル (GMM) に基づく手法 [11] やニューラルネットワークによる関数変換 [12]，適応型スプラインニューラルネットワークを用いたディープニューラルネットワークを用いた広帯域スペクトルの推定 [13]，対数パワースペクトルを用いたディープニューラルネットワークに基づく帯域拡張法 [14]，LSTM-RNN を用いた帯域拡張 [15] などが挙げられる．これらの手法の主な性能評価には，計算量や音声認識率，MOS 値による主観評価，PESQ やスペクトル歪みなどを用いた客観評価などが用いられている．

3 非線形帯域拡張法 [16]

本章では，提案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 [17]．この手法は低解像度の画像から高解像度の画像，つまり失われた高周波成分を疑似的に生成する手法である．基本的な手順はアンシャープマスキング (鮮鋭化フィルタ) とほぼ等しいが，途中で非線形関数を用いることで失われた広帯域に信号を生成させることで超解像画像を作る手法となっている．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．ここで用いる非線形関数は以下のように定義される．

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

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．このとき非線形処理を施した信号 $y_{HB}[n]$ 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．

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

具体的にどのような成分が広帯域として生成されるのかをスペクログラムで比較したものを図 2 に示す．図 2 の左から (a) 原音声 (16kHz サンプリング)，(b) 帯域幅を 4kHz に制限した音声 $y_{NB}[n]$ および (c)

*Non-linear artificial bandwidth extension of narrowband speech for speech recognition, by SHIOTA Sayaka, KIYA Hitoshi (Tokyo Metropolitan University)

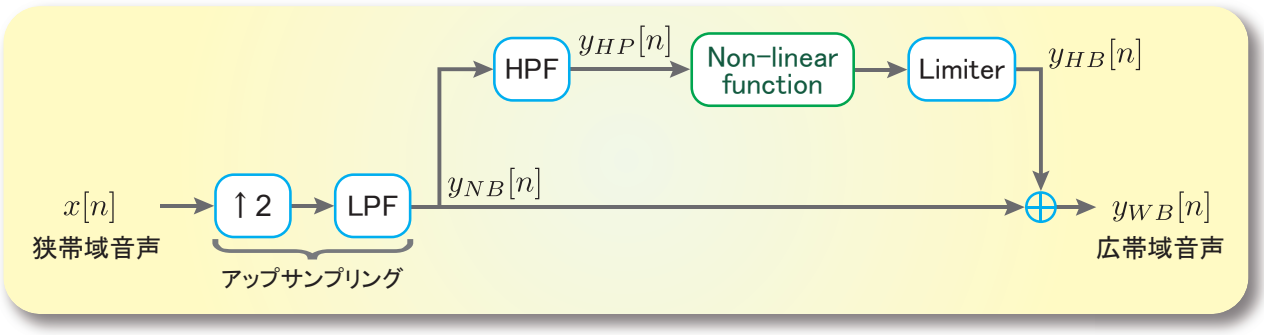

図 1　非線形帯域拡張法のフロー

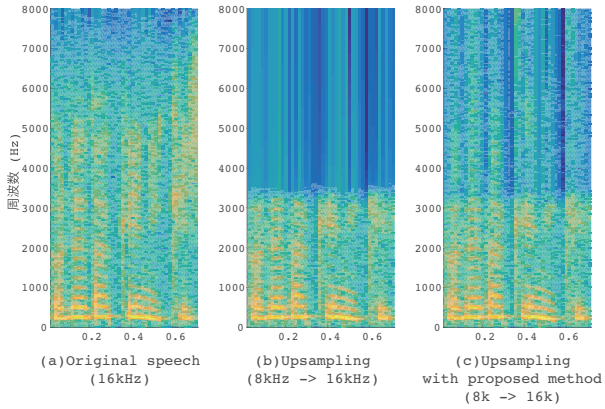

図 2　音声スペクトログラムでの比較

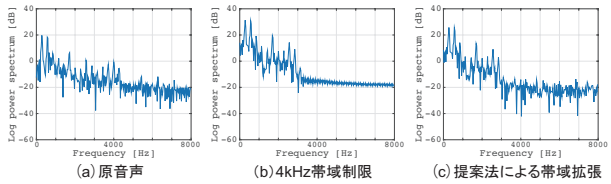

図 3　対数パワースペクトル（1 フレーム）での比較

提案法により帯域拡張された音声 $y_{WB}[n]$ となっている．(b) と (c) を比較すると (b) では 4kHz 以下の帯域にしか信号が現れていないが (c) では提案法を用いることで高い周波数帯域にも信号が生成されていることが確認できる．さらに細かく比較するために同サンプル内の 1 フレームの対数パワースペクトルを図 3 にて比較する．図 2 と同様に提案法 (c) では広帯域にもパワーが生成されていることがわかる．一方で，提案法は本来の広帯域成分を復元することを目指した枠組みではないため，パワースペクトルが原音声に近くなっているわけではないことも確認できる．前章で述べたようにこれまでの帯域拡張法は原音声に近づけることや自然性向上を目的としてきているが，筆者らが提案する非線形帯域拡張法は広帯域成分を生成することで機械学習手法に対する性能向上を目指しており，本論文でも評価には実際に音声認識実験における精度について言及する．

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) を用いた 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』模擬講演データに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事前実験により調整し，固定値として音声認識実験を行った．また，テストデータに関しての主な実験条件を表 1 に示す．

表 2 に比較する手法の各条件についてまとめた．本実験では帯域制限における音声認識システムの影響についても調べるために原音声（16kHz サンプリング）の音声を oracle なデータとして用意し，原音声に 4kHz で帯域制限をかけた音声を (B)，(B) をさらに値を間引くことでダウンサンプリングして 8kHz サ

表 1　テストデータの実験条件

データベース	JNAS データベース
性別/人数	男性 23 名，女性 23 名
文章数	男性 100 文章 女性 100 文章
サンプリング周波数	16kHz
フレーム長	25ms
フレームシフト	10ms

表 2　使用したテストデータの条件（全てサンプリング周波数は 16kHz）

(A) 原音声	原音声 (16kHz サンプリング)
(B)4kband 帯域制限	(A) を 4kHz で帯域制限をかけた音声
(C) アップサンプリング	(B) を 8kHz にダウンサンプリング後にアップサンプリングした音声
(D) 非線形帯域拡張法	(C) に提案法を用いた音声

ンプリングの音声を作り，そこからもう一度アップサンプリングを行った音声を (C) とした．(C) は図 1 の $y_{NB}[n]$ に該当する．最後に提案法として (C) の音声に非線形帯域拡張法を用いた音声を (D) として用意した．(D) は図 1 の $y_{WB}[n]$ である．(A)〜(D) 全て音声認識システムに入力する段階では 16kHz サンプリングとなっている．

4.2　実験結果

図 4 に各条件での単語正解率 (WER) 及び単語正解精度 (ACC) を示す．図 4 より，(A) の原音声では高い認識率が得られているが (B) の帯域制限がかかった音声および (C) のアップサンプリングされた音声では認識性能が大幅に低下している．これは，帯域制限により音声の明瞭性が低下することおよびデータのミスマッチが原因である．このことからも狭帯域音声が音声認識性能を大幅に低下させてしまうことがわかる．一方，提案法である非線形帯域拡張法を用いた音声 (D) は (B),(C) と比べて 30 ポイント以上認識性能が改善している．提案法は非常に計算量が低いにも関わらず，また，本来の音声を復元しているわけではないにも関わらず大幅な性能改善が見込めることがわかった．本実験では音響モデルの再作成を行っていないがすでに文献 [16], [18] の結果より，モデル自体も帯域拡張したデータを用いて学習を行うことで非常に高い識別性能が得られることが報告されて

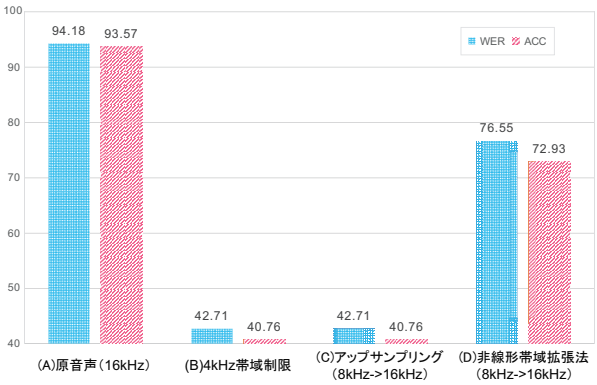

図 4　単語正解率 (WER) 及び単語正解精度 (ACC)

表 3　条件ごとの誤り傾向

	置換誤り	削除誤り	挿入誤り
(A)	4.61%	1.22%	0.61%
(B),(C)	24.43%	32.86%	1.94%
(D)	19.44%	4.01%	3.62%

いるため，音響モデルを作り直すことでより高い認識率が得られることが期待される．

次に，図 4 の各条件ごとに WER と ACC について比較すると，(A) であまり変化がないことから挿入誤りが少なく認識性能が高いことがわかる．一方，帯域制限がかかった音声である (B),(C) および提案法 (D) では (A) よりはそれぞれ WER と ACC との差が大きくなっている．そこでさらに，表 3 に比較手法それぞれにおける置換誤り，削除誤り，挿入誤りをまとめた．ただし，(B) および (C) の結果はまったく一緒だったのでまとめて表示している．表 3 の各誤りの傾向について分析すると特に帯域制限がかかった音声では削除誤りが (A) に比べて 30 ポイント以上増加していた．一方，(A) と (D) の削除誤りについては削除誤りの低下は 2.79 ポイントであり，大幅な改善が得られている．これらの結果から，非線形帯域拡張法が音声認識にもおいて有用であることが確認できた．

5　おわりに

本稿では，非線形帯域拡張法を用いることで狭帯域音声における音声認識の改善について報告した．音声認識において，学習に用いられた周波数帯域とは異なる狭帯域音声が入力されると音声認識性能が著しく低下してしまう．これまでに様々な帯域拡張法が提案されてきたが本研究ではアンシャープフィルタを用いた手法にさらに非線形関数を導入した非線形帯域拡張法を用い，広帯域音声を生成した．音声認識実

験により，提案法を用いて狭帯域音声から広帯域に拡張することで認識率が大幅に改善した．

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，計算量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] N. Enbom, et. al., “Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients,” 1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No.99EX351), pp. 171–173, 1999.

[2] P. Jax, et. al., “Wideband extension of telephone speech using a hidden markov model,” 2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium, pp. 133–135, 2000.

[3] 藤敦渉ら, “GMM に基づく最尤変換法による携帯電話音声の帯域拡張,” 情報処理学会研究報告音声言語情報処理 (SLP), vol. 2007, no. 75, pp. 63–68, 2007.

[4] I. Uysal, et. al., “Bandwidth extension of telephone speech using frame-based excitation and robust features,” 2005 13th European Signal Processing Conference, pp. 1–4, 2005.

[5] G. Miet, et. al., “Low-band extension of telephone-band speech,” 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), vol. 3, pp. 1851–1854, 2000.

[6] U. Kornagel, “Spectral widening of the excitation signal for telephone-band speech enhancement,” Proc. IWAENC, pp. 215–218, 2001.

[7] J. A. Fuemmeler, et. al., “Techniques for the regeneration of wideband speech from narrowband speech,” EURASIP Journal on Applied Signal Processing, vol. 2001, no. 1, pp. 266–274, 2001.

[8] P. Jax, et. al., “On artificial bandwidth extension of telephone speech,” Signal Processing, vol. 83, no. 8, pp. 1707–1719, 2003.

[9] Y. M. Cheng, et. al., “Statistical recovery of wideband speech from narrowband speech,” IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, pp. 544–548, 1994.

[10] Y. Qian, et. al., “Dual-mode wideband speech recovery from narrowband speech.” INTERSPEECH, 2003.

[11] 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.

[12] J. Kontio, et. al., “Neural network-based artificial bandwidth expansion of speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 3, pp. 873–881, 2007.

[13] A. Uncini, et. al., “Frequency recovery of narrow-band speech using adaptive spline neural networks,” 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258), vol. 2, pp. 997–1000, 1999.

[14] 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.

[15] Yuuki Tachioka, et. al., “Long short-term memory recurrent-neural-network-based bandwidth extension for automatic speech recognition”, Acoustical Science and Technology, vol. 37, pp. 319, 2016.

[16] 中西亮介ら, “非線形帯域拡張法に基づく話者照合の検討,” 音声言語情報処理研究会, vol. 2017-SLP-115, pp. 1–6, 2017.

[17] S. Gohshi, et. al., “Limitations of super resolution image reconstruction and how to overcome them for a single image,” 2013 International Conference on Signal Processing and Multimedia Applications (SIGMAP), pp. 71–78, 2013.

[18] 中西亮介ら, “非線形帯域拡張法に基づく話者照合とその応用,” 日本音響学会春季研究発表会, 2017.