Vol.2019-SLP-129 No.6
2019/10/11

HMMおよびEnd-to-End音声認識における非線形帯域拡張法の性能調査

今泉　遼[1,a)]　塩田　さやか[1]　貴家　仁志[1]

概要：本論文では非線形帯域拡張法を適用した音声が音声認識システムに与える影響について性能調査する. 音声認識システムと入力されるテストデータにサンプリング周波数の不一致がある場合, 高周波数に合わせる手法として帯域拡張法がある. 帯域拡張法をテストデータに用いることでサンプリング周波数ごとに音声認識システムを再構築する必要がないという利点がある. しかし, これまでに音声認識における帯域拡張法の影響については報告がされていなかった. そこで本論文では非学習型の帯域拡張法が GMM-HMM および End-to-End に基づく音声認識システムに与える影響を調査した. 実験結果より GMM-HMM に基づく手法に対しては, 明瞭性を損なわない帯域拡張法を用いた場合に高い精度が得られる傾向があるが, End-to-End に基づく手法に対しては, 原音声とのスペクトル距離が近くなる帯域拡張法を用いた場合に高い精度が得られることがわかった.

キーワード：音声認識, End-to-End, 非線形帯域拡張, 客観的評価尺度

Performance survey of nonlinear bandwidth extension method in HMM and end-to-end speech recognition

Imaizumi Ryo[1,a)]　Shiota Sayaka[1]　Kiya Hitoshi[1]

Abstract: In this paper, we investigate the performance of automatic speech recognition (ASR) systems using some nonlinear bandwidth extension (BWE) methods for narrowband evaluation data. When the sampling frequency of training data for ASR systems is different from that of an input utterance, BWE methods are used to generate harmonics frequencies lacked by band-limitation. The advantage of using BWE methods is not to require the reconstruction of ASR systems for each sampling frequency of input utterances. However, it has not been reported about the effects of BWE methods on ASR. Therefore, in this paper, GMM-HMM-based and End-to-End ASR systems are performed with band extended utterances by non-learning BWE methods in order to investigate the effects. From the experimental results, the GMM-HMM-based ASR system obtained high accuracy when the BWE methods provided high intelligibility. In contrast the End-to-End ASR obtained high accuracy when the BWE methods provided a close spectral distance to original speech.

Keywords: Automatic speech recognition, End-to-End, nonlinear bandwidth extension, objective evaluation

1. はじめに

音声認識とは話し言葉を文字列に変換して文字に変換する技術である. 音声認識が使われている主なアプリケーションにはスマートスピーカーなどの音声対話システムやテレビ番組などの文字起こしなどが挙げられる. 音声認識のシステムの従来手法として Gausian Mixture Model-Hidden Markov model (GMM-HMM) による音響モデルと n-gram などによる言語モデルが用いられてきた. しかし, 近年の深層学習の発展により Deep Neural Network (DNN) を用い

[1] 首都大学東京 システムデザイン学部
Tokyo Metropolitan University
[a)] imaizumi-ryo@ed.tmu.ac.jp

た手法である DNN-HMM [1] や End-to-End [2] を用いる音声認識システムが主流になってきている. 音声認識を使用するアプリケーションには Cliant Sarver System (CSS) を利用しているものが多くある. CSS ではユーザーの音声収録条件をコントロールできない場合もあるため, システムが想定しているサンプリング周波数と入力音声のサンプリング周波数が一致しないことがある. サンプリング周波数が一致しない場合, サンプリング周波数が高い方の音声 (広帯域音声) を低い方 (狭帯域音声) に合わせるためにダウンサンプリングを行うことが多い. しかしシステムのサンプリング周波数の方が高い場合, 入力音声のサンプリング周波数に合わせて再構築し直す必要がある. またダウンサンプリングされた音声では音声の明瞭性などが低下してしまうため音声認識自体の性能が低下してしまうことが知られている. 一方, 狭帯域音声を広帯域音声に拡張させる手法に帯域拡張法がある. 帯域拡張法は欠落した広帯域成分を復元・生成するものである. 帯域拡張法には機械学習などを用いる学習型の手法と非学習型の手法がある. 非学習型の帯域拡張法である非線形帯域拡張法の例として, スペクトルシフティング (SHIFT) [3][4], 線形予測分析合成法 (LPAS) [5], 非線形帯域拡張法 (N-BWE) [6] などの手法がある. これらの手法が話者照合システムにおいて有効であることが報告されている [7]. しかし, GMM-HMM および End-to-End に基づく音声認識システムの性能にどのような影響を与えるのかについては報告されていなかった. そこで本研究では広帯域音声で構築された音声認識システムに狭帯域音声が入力されることを想定し, 帯域拡張法の与える影響を調査する. 本実験では日本語話し言葉コーパス (Corpus of Spontaneous Japanese ; CSJ) [8] を用いて構築した GMM-HMM および End-to-End に基づく音声認識システムに対して帯域拡張した音声の WER を算出し, WER と客観的評価尺度との比較を行った. 実験結果より GMM-HMM に基づく手法では明瞭性の高くなる帯域拡張法において高い精度が得られる傾向があるが, End-to-End に基づく手法では RMS-LSD が小さくなる帯域拡張法において高い精度が得られることがわかった.

2. 音声認識システム

本章では GMM-HMM および End-to-End に基づく音声認識システムについて説明する.

2.1 GMM-HMM

従来の音声認識では GMM-HMM に基づく音響モデルが広く用いられてきた. HMM は, 遷移確率を含むことから時系列を扱うことが得意な確率モデルであり, 音声認識においては left-to-right 型で状態列を, GMM で各状態列の出力確率をモデル化している. GMM-HMM 音声認識システムは言語モデルの推定も必要であり様々な手法が用いられている.

2.2 End-to-End

深層学習の発展により DNN を用いた End-to-End 音声認識が提案され高い識別性能を得られることが報告されている. End-to-End の音声認識では Hybrid Connnectionist temporal classification Attention based encoder decoder[11] を用いていて, 音声を直接音素や単語に変換するネットワークを構築する. このネットワークは直接単語列を出力できるシステムである. 利点としては単語を出力するため言語モデルを分けて推定する必要がない. システムの構築が容易であるなどが挙げられる.

3. 帯域拡張法

帯域拡張法には多くの手法が報告されているが本実験では付帯情報を用いず, 学習も行わない非線形帯域拡張法に着目する.

3.1 スペクトルシフティング法 (SHIFT)

非学習型の帯域拡張法の 1 つである. 狭帯域の周期を変調して高周波成分を生成して, その成分を広帯域の周波数領域にシフトすることで帯域拡張する. 単純なため, 処理時間が短いという利点がある.

3.2 線形予測分析合成法 (LPAS)

LPAS は, 線形予測分析を用いて狭帯域信号からスペクトル包絡線, 残留誤差情報から抽出された高周波成分を用いて広帯域周波成分を生成する手法である. 自然性が高い音声が生成されるが, 処理時間がかかるという問題がある.

3.3 非線形帯域拡張法 (N-BWE)

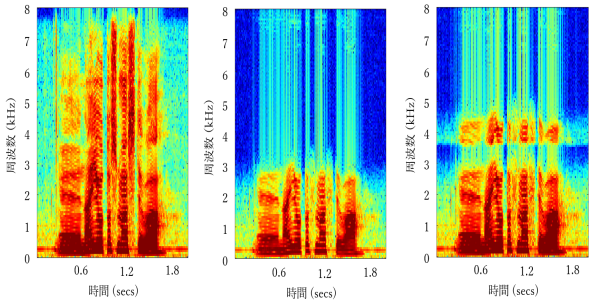
N-BWE は非線形関数をアップサンプリングされた音声にかけて高周波成分を生成する手法である. 狭帯域音声をアップサンプリングした音声を $y_{NB}[t]$, 広帯域音声を $y_{HB}[t]$ としたとき非線形関数は以下の式で表される.

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

ここで α および β は高周波成分の生成に影響を与えるパラメータであり, α や β の値により生成される広帯域音声の特徴を変化させることができる.

3.4 音声認識への帯域拡張法の影響

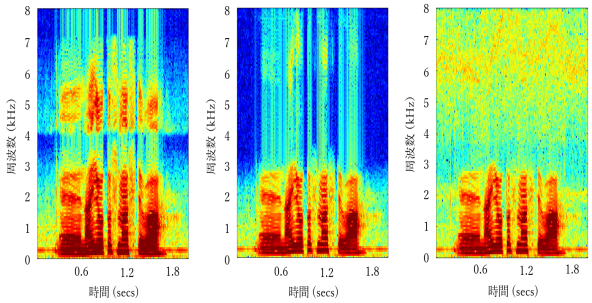
16 kHz でサンプリングされた原音声と 8 kHz にダウンサンプリングされた音声から各帯域拡張法を用いた音声のスペクトログラムを示す. (a) は原音声, (b) はアップサンプリングのみ, (c) は SHIFT, (d) は LPAS, (e) は N-BWE で STOI が良くなるように調整した音声 (N-BWE1), (f) が N-BWE で RMS-LSD が良くなるように調整した音声



(a) Original RMS-LSD [dB] /STOI

(b) Upsampling 2.335 [dB] /0.973

(c) SHIFT 2.120 [dB] /0.955



(d) LPAS 1.555 [dB] /0.966

(e) N-BWE1 2.217 [dB] /0.951

(f) N-BWE2 1.718 [dB] /0.946

図 1: スペクトログラムと客観評価値 (RMS-LSD/STOI)
発話内容 「えー内容としましては」

(N-BWE2) である. (d) から LPAS と N-BWE2 は高周波成分が生成されていることがわかる. また (e) , (f) から高周波成分の生成は RMS-LSD の値に大きく影響することがわかる. これまでに上記の手法の帯域拡張法は, 話者照合においては有効であるとされてきた. しかし, 音声認識において帯域拡張法を適応した際の影響は調査されていなかった. そこで本研究ではそれぞれの帯域拡張法の特徴と音声認識システムの性能について調査する.

4. 実験

3 章で述べた非学習型の帯域拡張法を適用して生成した音声の客観的評価尺度およびそれらの音声を GMM-HMM および End-to-End に基づく音声認識に用いた場合の性能調査を行った.

4.1 実験条件

表 1 に実験で用いた CSJ データベースの詳細を示す. CSJ データベースの学習セットを用い, GMM-HMM および End-to-End の音声認識システムを構築した. それぞれの構築には Kaldi ツールキット [12] および ESPnet [13] に含まれる CSJ レシピを用いた. 本実験ではサンプリング周波数の不一致は学習データとテストデータ間に存在し, テストデータがシステムの想定よりも低いサンプリング周波数であるとした. CSJ データのサンプリング周波数は 16 kHz であるため, 本実験では狭帯域音声を 8 kHz, 目標とする広帯域音声を 16 kHz とした. 8kHz の音声には原音声をダウンサンプリングしたものを使用した. 比較の条件を以下に記す.

表 1: 日本語話し言葉コーパス (CSJ)

	Train	Eval1	Eval2	Eval3
講演数	2672	10	10	10
話者数 (性別)	1383 (男 924 女 459)	10 (男 10)	10 (男 5 女 5)	10 (男 5 女 5)
内容	学会講演 ＋模擬講演	学会講演	学会講演	模擬講演
時間 (h)	600	2.28	2.42	1.71

Upsampling : 狭帯域音声にアップサンプリングのみを行った音声

SHIFT : 狭帯域音声に SHIFT を適用して帯域拡張した音声

LPAS : 狭帯域音声に LPAS を行い帯域拡張した音声

N-BWE1 : 狭帯域音声に対して先行研究による提案法の非線形帯域拡張法による帯域拡張法を行った. α , β は STOI の値が高くなるようにそれぞれ 2, 50 とした.

N-BWE2 : 狭帯域音声に対して先行研究による提案法の非線形帯域拡張法による帯域拡張法を行った. α , β は RMS-LSD の値が低くなるようにそれぞれ 0.8, 100 とした.

Original : サンプリング周波数 16 kHz の音声を使用した.

客観的評価尺度には, 音声の明瞭性を示す客観的評価尺度である STOI (Short-Time Objective Intelligibility) [9] と原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) [10] の 2 つを使用した. STOI は 0 から 1 の範囲で表され, 1 に近いほど明瞭性が高いことを示す. RMS-LSD は数値が低いほど 2 つの音声の誤差が小さいため原音声に近いことを示す. RMS-LSD の値 D は式 (2) で表される.

$$D = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{N}\sum_{i=0}^{N-1}|\log_{10} A_k(i) - \log_{10}\hat{A}_k(i)|^2} \quad (2)$$

$A_k(i)$ および $\hat{A}_k(i)$ はリファレンスの音声と比較音声の k 番目のフレームのパワースペクトルで, N と k はそれぞれフレーム長とフレーム番号を示している. また音声認識の性能評価には WER を用いた.

4.2 実験結果

図 2, 3 に Eval1 から Eval3 の全テストデータの STOI と RMS-LSD を測った結果を箱ひげ図で示す. 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

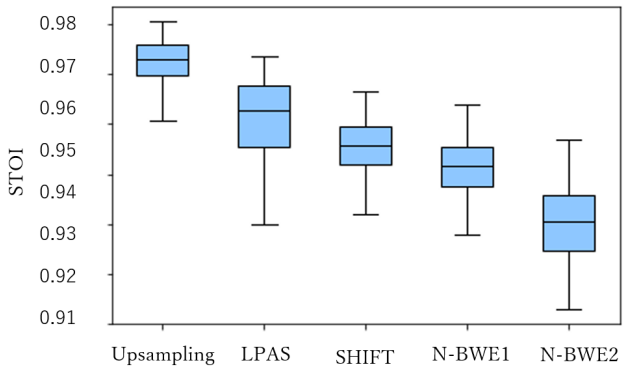


図 2: 客観的評価尺度 (STOI)

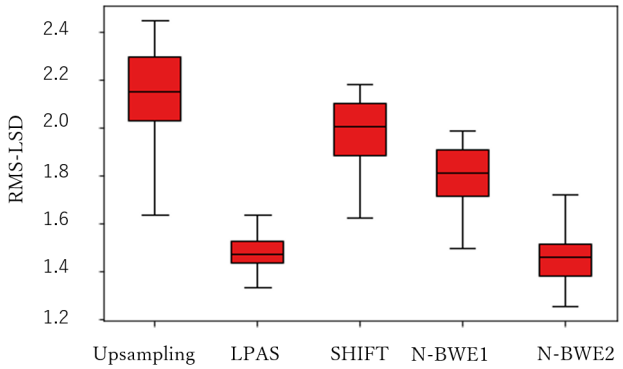


図 3: 客観的評価尺度 (RMS-LSD)

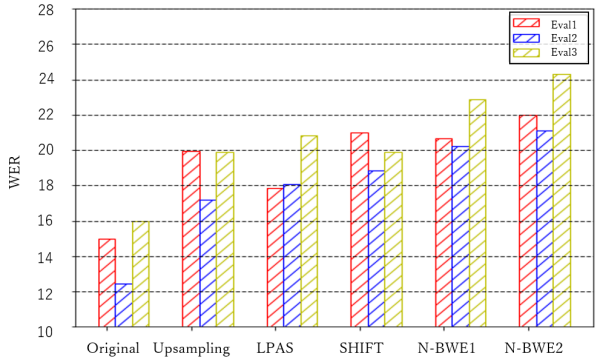


図 4: 単語誤り率 (GMM-HMM)

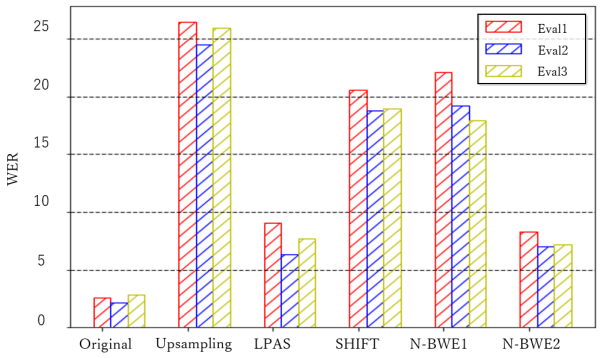


図 5: 単語誤り率 (End-to-End)

最小値を示す. 図 2 を見ると, Upsampling の値が最も高く, 帯域拡張法はどの手法も値は Upsampling より微減していることがわかる. 帯域拡張法はスペクトルの不連続性や位相を復元していないことから明瞭性が低下してしまうためだと考えられる. 帯域拡張法間で比較すると, LPAS は線形予測分析を用いることから他の手法より STOI の値が高い傾向にあることがわかる. また, N-BWE1, 2 を比較すると, STOI を高くなるようパラメータを調整した N-BWE1 の方が N-BWE2 より値が低いことが確認できる. SHIFT と N-BWE1 では大きな差はなかった. 次に RMS-LSD について比較する. 図 3 を見ると LPAS と N-BWE2 は値が小さく他の手法は Upsampling より少し値が小さいという結果であった. N-BWE2 は非線形関数のパラメータを RMS-LSD が小さくなるように調整しているため妥当な結果であるといえる.

次に音声認識の結果について述べる. 図 4 が GMM-HMM に基づいた音声認識システム, 図 5 は End-to-End に基づいた音声認識システムで認識をした結果の WER を示している. まず Original の WER を比較してみると GMM-HMM より End-to-End の方がどの評価セットにおいても WER が低いことがわかる. 特徴量の違いなどはあるが深層学習による手法の方が性能が高いことを示していると考えられる. Upsampling に注目すると GMM-HMM, End-to-End どちらの場合にも Original より WER が高くなっており, 特に End-to-End の方は WER の値が大幅に高くなっている. このことから帯域制限のかかった音声はシステムの性能に大きく影響を与えることがわかる. 次に客観的評価尺度と WER を比較する. STOI と GMM-HMM の WER の傾向をみると Eval セットの違いはあるものの Upsampling 以外は STOI の値が高いとき WER は下がっている. これは GMM-HMM に基づいた音声認識では音声の明瞭性が WER が影響を与えるからだと考えられる. 次に RMS-LSD と End-to-End を比較する. RMS-LSD の値が低い LPAS や N-BWE2 の WER は Upsampling と比べると Original に近い精度まで改善できている. RMS-LSD の値が高い SHIFT, N-BWE1 は WER も高くなっていることから End-to-End の音声認識では音の明瞭性よりもスペクトル距離が原音声に近いことが重要であると考えられる.

5. まとめ

本研究では帯域拡張法を適用して生成した音声の客観的評価尺度およびそれらの音声が GMM-HMM や End-to-End に基づく音声認識システムに与える影響について調査した. 実験結果より GMM-HMM に基づく音声認識は音の明瞭性が WER に影響しており, End-to-End に基づく音声認識では音の明瞭性よりも原音声と帯域拡張した音声のスペクトル距離が WER に影響を及ぼすことを確認した. 今後の課題としては別のデータベースによる評価, 異なる周波数での評価などが挙げられる.

謝辞 本研究の一部は JSPS 科研費若手研究 JP19K20271

及び ROIS-DS-JOINT (021RP2019) の助成を受けたものです.

参考文献

[1] G. Hinton, et al. : *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal Processing Magazine, 29, 6, pp. 82–97, 2012.

[2] J. Chorowski, et al. : *End-to-end continuous speech recognition using attention-based recurrent nn:First results*, arXiv preprint arXiv:1412.1602, 2014.

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

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

[5] 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, pp.1–5, 2018.

[6] H. Miyamoto, et al. : *Non-linear harmonic generation based blind bandwidth extension considering aliasing artifacts*, in Proc. APSIPA, pp.1868–1874, 2018.

[7] R. Kaminishi, et al. : *Blind bandwidth extension with a non-linear function and its evaluation on x-vector-based speaker verification*, in Proc. Interspeech, 2019.

[8] 前川喜久雄 (国立国語研究所) :「日本語話し言葉コーパス」の概観 version2.0, `https://pj.ninjal.ac.jp/corpus_center/csj`

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

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

[11] S. Watanabe, et al. : *Hybrid CTC/Attention Architecture for End-to-End Speech Recognition*, IEEE Journal of Selected Topics in Signal Processing, 11, 8 , 2017.

[12] D. Povey, et al. : *The kaldi speech recognition toolkit*, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2011.

[13] S. Watanabe, et al. : *ESPnet: End-to-End Speech Processing Toolkit*, arXiv preprint arXiv:1804.00015v1, 2018.