Vol.2013-SLP-97 No.1
2013/7/25

Denoising Autoencoder を用いた残響下大語彙音声認識の検討

小宮山 大樹[†1,a)]　石井 敬章[†1]　篠崎 隆宏[†2]　堀内 靖雄[†1]　黒岩 眞吾[†1]

概要：Denoising Autoencoder を用いて残響が重畳した対数パワースペクトルからその影響を除去した対数パワースペクトルを生成する手法を提案する．音声の時間変化をモデル化するため，提案法では連続した複数の短時間分析窓によるスペクトルフレームを連結したものをネットワークの入力として用いる．さらに，音声認識に必要なサブ音素レベルでの時間分解能を維持しながら時定数の大きな残響の影響をより正しく捕らえることを目的として，長さの異なる 2 つの分析窓長を併用する拡張手法を提案する．実験では，CENSREC-4 を用いた数字音声認識により提案法が従来手法よりも効果的であることを示す．さらに，JNAS を用いた音声認識を行い，提案法が大語彙連続音声認識においても耐残響フロントエンドとして有効であることを示す．

1. はじめに

残響とは音源からマイクロフォンに直接到達する直接音とともに，壁や天井，床などで反射した音が様々な時間遅れで到達し重なり合って観測される現象である．残響は音声の明瞭度を著しく下げ，音声認識の性能低下を招く要因になる．残響を抑制する手法は古くからの課題であり，研究が行われて来た．それらの手法は大きく，単一のマイクロフォンを用いる手法と複数のマイクロフォンを用いる手法に分類することができる．

一般に，単一のマイクロフォンで残響抑制を行うよりも，複数のマイクロフォンを用いた方が方が高い残響抑圧効果が得られる [1]．しかし，複数のマイクロフォンを用いた手法は一定の間隔を持ったマイクの配置が必要となり，利用面から制約を受ける．それに対して，単一のマイクロフォンを用いる手法は高い性能を得ることが技術的に難しくなる半面，非常に利用しやすいという長所がある．本論文では，これらのうち単一のマイクロフォンを用いる手法を対象とする．単一のマイクロフォンを用いる既存の耐残響音手法としては，長時間スペクトル減算 [2] や Frequency Domain Linear Prediction [3] などの長時間分析窓を用いた正規化手法が挙げられる．しかし，未だ十分な性能は得られていないのが現状である．

他方，ニューラルネットワークにおける近年の研究において，局所最適解に陥りにくい初期値決定法を用いた Deep Learning [4] が提案され，音声分野においても応用が研究されている．耐雑音音声認識への応用としては，Deep Learning の分野における一手法である Denoising Autoencoder (DAE) [5] を用いて加法性雑音を除去する手法が提案され，有効性が示されている [6][7]．

このような背景のもと本研究では，短時間スペクトル特徴量の一定長のフレーム系列を入力とする DAE を用いた残響除去手法を提案する．またその拡張として，音声認識に必要なサブ音素レベルでの時間分解能を維持しながら時定数の大きな残響の影響をより正しく捕らえることを目的として，短時間分析窓を用いたスペクトル特徴量とともに長時間分析窓による分析を併用した拡張手法を提案する．提案手法の性能評価は，残響下音声認識評価環境データベース (CENSREC-4[8]) を用いた残響下における連続数字音声認識と，新聞記事読み上げ音声コーパス (JNAS[9]) に残響を畳みこんだデータセットを用いた連続単語認識により行う．

本論文の章構成は以下のようになる．第 2 章において，Autoencoder および DAE の概要について説明する．第 3 章において，提案手法について説明する．第 4 章において CENSREC-4 を用いた実験について示し，第 5 章で JNAS を用いた実験について示す．最後に，第 6 章でまとめと今後の課題を示す．

[†1] 現在，千葉大学
Presently with Chiba Uniersity, Chiba, Chiba, 263–8522, Japan
[†2] 現在，東京工業大学
Presently with Tokyo Institute of Technology, Yokohama, Kanagawa, 226–8502, Japan
[a)] komiyama@chiba-u.jp

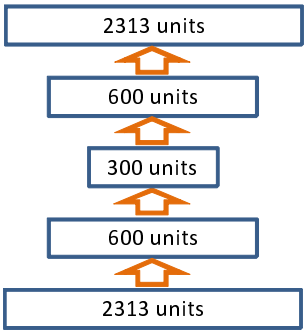

図 1 Autoencoder の例

2. Denoising Autoencoder（DAE）

Autoencoder は小さく絞られた中間層を持った，特徴量の次元圧縮を目的とした多層ニューラルネットワークである．ネットワークは中間層を通して入力から同じ出力が得られるように学習が行われ，窄められた中間層が圧縮された特徴量を表すと解釈される．多階層のネットワークをバックプロパゲーションにより直接最適化することは困難であることから，初期値を設定する Pre-training と最終的なパラメタを求める Fine-tuning の 2 つのフェーズにより学習を行う手法が提案されている．具体的には Pre-training では Restricted Boltzmann Machine（RBM）と呼ばれる二層構造のネットワークを順次積み上げる形で教師なし学習を行う．次に，積み上げた RBM のパラメタをコピーし入出力を反転させ，それらの出力と入力を接続することで図 1 に示すような上下対称のネットワークを得る．最後に Fine-tuning において，入出力に同じデータを与えたバックプロパゲーションにより教師あり学習を行う．DAE は Autoencoder の拡張であり，ネットワーク構造は Autoencoder と同様である．しかしその利用において，ノイズデータからクリーンデータを予測出力することを目的としている．学習手順もほぼ同様であるが，バックプロパゲーションの際に入力側に雑音重畳特徴量，出力側に対応するクリーン特徴量を設定する点が違いである．

3. 提案手法

音声認識のための耐残響音フロントエンドとして，DAE を応用した手法の提案を行う．提案法では図 2 に示すように，短時間窓による残響重畳音声の対数パワースペクトルの N フレーム長のセグメントを DAE に入力し，出力として対応する N フレーム長の残響が除去された対数パワースペクトルのセグメントを得る．入力音声に対してこの操作を 1 フレーム毎にシフトしながら適用することで，オーバーラップした N フレーム長の対数パワースペクトルのセグメントの系列を得る．その上で，図 3 に示すように，各時刻フレームで対数パワースペクトルの平均を求めることで，最終的な対数パワースペクトルフレームの時系列を得

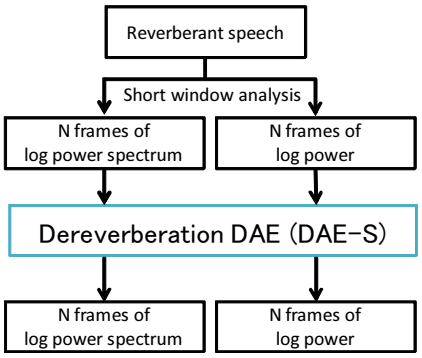

図 2 単一窓幅セグメント特徴量を用いた DAE による残響除去 (DAE-S)

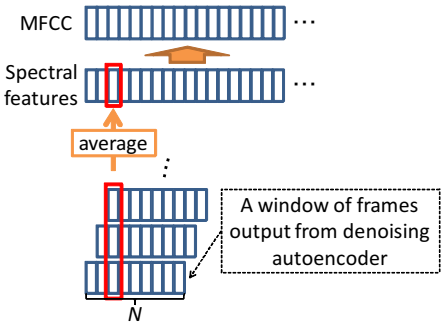

図 3 残響除去後のオーバーラップした対数パワースペクトルフレームセグメントの系列から単位フレームの系列を得る手順．各時刻フレームでの平均を用いる

る．音声認識で使用する MFCC [10] などの特徴量は，得られた対数パワースペクトルを元に求めることが出来る．この手法では，短時間分析窓による対数パワースペクトルの N フレームのセグメントを DAE の入力とすることで音声の時間変化をモデル化し，それにより音声の知識に基づいた残響除去が可能となると期待される．以下，本手法を DAE-S と表記することにする．なお評価実験では，MFCC とともに対数パワー項を特徴量とするため，対数パワースペクトルとともに対数パワー項を入力に加えて用いている．

従来の耐残響手法で長時間の分析窓長を用いているのは，時定数の長い残響に合わせて窓幅を大きくすることで正確な分析を行うためである．提案法においても，長時間分析窓により得られる残響特性情報を加えることで，性能がさらに向上すると期待される．そこで，図 4 に示すように短時間窓による対数パワースペクトルに加えて長時間窓を用いた分析結果を合わせて利用する DAE-S の拡張手法の提案を行う．長時間分析窓のスペクトルベクトルに対しては，メルフィルタバンクを適用することで次元数の縮小を行う．入出力が対称な DAE では出力側に残響除去された長時間スペクトルが得られることになるが，音声認識では使用されず不要となる．このため，バックプロパゲーションによる Fine-tuning および評価時にはそれらのユニットを取り除いて使用する．以下，本手法を DAE-SL と表記す

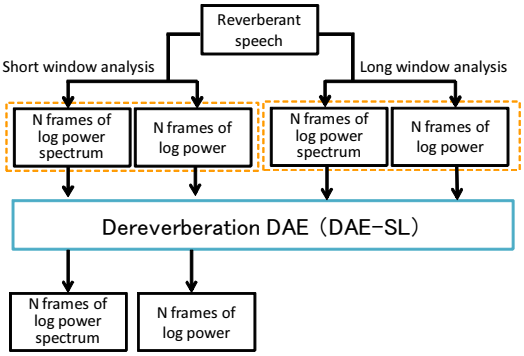

図 4 複数窓幅セグメント特徴量を用いた DAE による残響除去 (DAE-SL)

ることにする．

4. CENSREC-4 を用いた数字認識実験

4.1 実験条件

残響下での音声認識タスクを目的としたデータセット CENSREC-4 を用いて DAE の学習および評価を行う．CENSREC-4 は学習セットとして 8440 文のクリーンな連続数字読み上げ音声を含んでいる．マルチコンディション学習セットは，クリーン学習セットを 4 分割しそれぞれに異なる残響のインパルス応答を畳み込むことで定義されている．用いられている残響は，Office， Elevator hall, In-car および Living room である．音声のサンプリング周波数は 16kHz，量子化ビット数は 16bit である．認識語彙は数字 11 種類（1 から 9 および 0（マル）と Z（ゼロ））で，各発話は 1 から 7 桁の連続数字音声となっている．テスト音声には，マルチコンディション学習セットと同じ 4 環境のインパルス応答 (Office,Elevator hall, In-car, Living room) を畳み込んだ 4004 文（TestA）と，それらとは別の 4 環境のインパルス応答 (Lounge, Japanese style room, Meeting room, Japanese style bath) を畳み込んだ 4004 文（TestB）の 2 種類を用いる．音声認識は CENSREC-4 のデフォルトの設定で行なっており，HMM に使用する特徴量はフレーム幅 25ms，フレームシフト 10ms の窓を用いた MFCC12 次元と対数パワーおよびその Δ と ΔΔ の 39 次元である．

DAE の学習においては，計算量の問題から学習データの一部のみを使用した．Pre-training に用いたのは，クリーン学習セットからランダム選択した 2110 文とマルチコンディション学習セットからランダム選択した 2110 文の合計 4220 文 (2.4h) である．また Fine-tuning では入力信号に同じ 4220 文を用い，ネットワークの出力側にはインパルス応答を畳み込む前の対応するクリーン音声を用いた．

Pre-training においては，最初のレイヤーに Gaussian-binary RBM を用い，次のレイヤーには binary-binary RBM を使用した．Pre-training には Contrastive Divergence 法を用い，各層において学習の繰り返し数を 100，ミニバッチサイズを 100，learning rate を 0.002 とした．Fine-tuning は二乗誤差を目的関数として，ミニバッチサイズを 1000，learning rate を 0.02 として，共役勾配法 [11] に基づくバックプロパゲーションにより行った．

DAE-S において使用する短時間分析窓の窓幅は 25ms，フレームシフトは 10ms である．単位フレームの対数パワースペクトルベクトルの次元数は 256 である．ネットワークにはこのベクトルと対数パワーを N フレーム分連結したものを入力する．すなわち，ネットワークの入力に用いられるベクトルの次元数は，$(256+1)\times N$ である．DAE-SL においては，長時間分析窓として 500ms の窓幅を用いる．そこから対数パワーとメルフィルタバンクにより 24 次に次元数を落としたスペクトルを導出し，それらを DAE-S において用いている入力ベクトルに連結して用いる．フレームシフトは短時間分析窓と同じ 10ms である．ネットワークの入力に用いられるベクトルの次元数は，$(256+1+24+1)\times N$ となる．

以下の実験においては，認識性能とともに計算時間を考慮した事前実験より，特に断らない限りデフォルトの構成として 5 階層の DAE を用いる．また，セグメント長としては N=9 を用いる．この場合，DAE-S の入力層のサイズは $(256+1)\times 9 = 2313$ であり，DAE-SL の入力層のサイズは $(256+1+24+1)\times 9 = 2538$ である．また入力層の次の階層のデフォルトのユニット数は 600，その次の中央の階層のユニット数は 300 とした．すなわち，DAE-S のデフォルト構成は，図 1 の例に示したネットワークと同じである．デフォルトのバックプロパゲーションの繰り返し数は 30 とした．ネットワークの構造は入出力に対して対称であることから，以下では入力側の半分についてユニット数を表記することでその構造を表すことにする．すなわち，上記デフォルトの条件における DAE-S の構造は “ 2313-600-300 ”，DAE-SL の場合は “ 2538-600-300 ” と表記する．

HMM の学習では，学習セットのサブセットを用いた DAE の学習の後，得られた DAE を全学習セットに適用し得られた残響除去処理後のスペクトルから MFCC を作成したものを学習データとして用いる．認識評価は，同様に得られた DAE を評価データに適用し，残響除去処理後のスペクトルから MFCC を作成したものを用いて行う．学習および評価の手順を図 5 に示す．

4.2 実験結果

図 6 に DAE-S におけるバックプロパゲーションの繰り返し数と，クリーンスペクトルと残響除去処理後のスペクトルとの間の平均二乗誤差の推移を示す．平均二乗誤差は学習セット，評価セット TestA および TestB のそれぞれに対して評価を行った．TestB の誤差が学習セットや TestA と比べて大きいのは，オープンな残響環境であるためであ

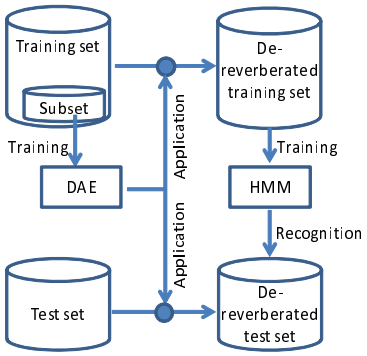

図 5 DAE および HMM の学習と評価の手順

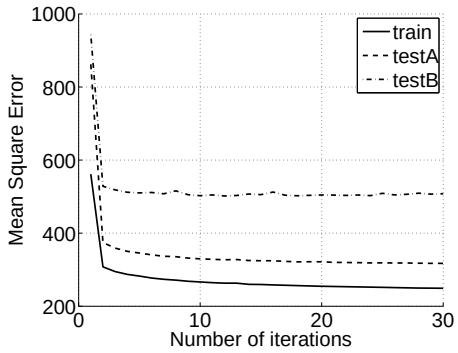

図 6 CENSREC-4 データを用いた DAE-S のバックプロパゲーション学習における繰り返し数と平均 2 乗誤差の推移

表 1 数字読み上げ音声を用いた場合の Pre-training の有無およびバックプロパゲーションの繰り返し数と，実行時間および単語正解精度の関係

CPU time(Acc)	iter=0	iter=10	iter=30
Pre-training なし	0h(-)	21.3h(85.9)	57.3h(92.5)
Pre-training あり	17.3h	41.5h(95.8)	72.3h(96.4)

る．いずれのデータセットに対しても，平均二乗誤差の大きな減少は繰り返しの 10 回目あたりまでに得られている．学習セットと TestA では，その後も僅かながら減少が見られた．

表 1 に DAE-S における学習時間と評価セット TestB における 正解精度の関係を示す．Pre-training を行わずバックプロパゲーションによる繰り返し数を 30 とした場合と，Pre-training を行いバックプロパゲーションの繰り返し数を 10 とした場合を比較すると，Pre-training を行った場合の方がトータルとして少ない学習時間で高い正解精度が得られていることが分かる．

表 2 に，DAE-S においてネットワーク構成を変えた場合の正解精度を示す．学習にかかった時間は Pre-training と Fine-tuning の全行程にかかった実行時間である．ユニット数が少ないうちはユニットが増えるにつれて正解精度も高くなる傾向がある．しかし，2313-900-300 より大きくしても正解精度はほぼ頭打ちになっていることが分かる．

表 3 に，DAE-S の入力に用いるセグメント長 N を変化させた際の正解精度を示す．フレーム数が少ない間は N の

表 2 ネットワークの構成と正解精度

Network structure	Time(hour)	Accuracy(%)	
		TestA	TestB
2313-300-150	35.9	96.6	93.2
2313-600-150	70.1	97.8	95.5
2313-600-300	72.3	97.9	96.4
2313-900-300	110.5	98.3	97.1
2313-900-600	121.0	98.5	97.1
2313-1200-600	156.7	98.2	96.8
2313-1500-600	214.4	98.5	97.5
2313-900-600-300	129.3	98.3	97.0

表 3 フレームセグメント長 N と正解精度

N	testA	testB
5	97.3	95.8
7	97.8	96.3
9	97.9	96.4
11	98.1	96.7
13	97.9	96.1

表 4 CENSREC-4 における各種手法との比較

Method	testA	testB
Baseline(Clean)	83.8	82.8
Baseline(Multi)	92.9	87.8
CMN(Clean)	86.5	88.6
CMN(Multi)	91.8	89.7
HDelta(Multi)	95.7	94.7
DAE-S	97.9	96.4
DAE-SL	98.4	97.0

増加とともに正解精度が向上するが，11 フレームで最大値となり，13 フレームでは減少することが分かる．

表 4 に各種従来法と提案手法の正解精度の比較を示す．“ Baseline ”は MFCC+Δ+ΔΔ を用いた場合の結果であり，“ CMN ”は cepstral mean normalization を適用した場合の結果である．“ Hdelta ” は Hybrid delta 法でデルタ特徴量を取得した場合の結果 [12] である．“ Clean ”はクリーン学習セットで HMM を学習した場合の結果，“ Multi ”はマルチコンディション学習セットで HMM を学習した場合の結果である．従来法と提案法 (DAE-S および DAE-SL) を比べると，提案法の方が高い正解精度を与えている．

2 つの提案法を比較すると，単一窓幅セグメント特徴量を用いた DAE-S（TestA：97.9%，TestB：96.4%）よりも複数窓幅セグメント特徴量を用いた DAE-SL（TestA：98.4%，TestB：97.0%）の方が高い正解精度を与えていることが分かる．残響環境クローズである TestA だけでなく残響環境オープンである TestB に対しても頑健な認識性能が得られており，提案法が効果的であることがわかる．

5. JNAS 新聞読み上げコーパスを用いた大語彙認識実験

前章の CENSREC-4 を用いた実験では，数字認識におけ

る提案手法の有効性を示した．本章では，大語彙連続単語音声認識をタスクとして評価実験を行う．

5.1 実験条件

JNAS の音声に CENSREC-4 の環境残響を畳み込んだデータセットを使用する．JNAS は日本語大語彙連続音声認識研究を目的としたコーパスであり，155 セット (約 100 文/セット，計 16176 文) の新聞記事を男女計 306 名の話者が読み上げた音声が含まれている．音声のサンプリング周波数は 16kHz，量子化ビット数は 16bit である．

実験では JNAS のデータのうち，196 名の話者からの 19895 文を学習セット，20 名の話者からの 840 文をテストデータセットとして用いる．学習セットには，前章の学習データで用いた 4 種類の残響インパルス応答 (Office,Elevator hall, In-car, Livingroom) のいずれかをランダムに畳み込んだものを用いた．評価セットには，前章の TestB で用いた学習セットとは独立の 4 種類のインパルス応答 (Lounge, Japanese style room, Meeting room,Japanese style bath) のいずれかをランダムに畳み込んだものを用いた．DAE の学習は，残響重畳音声データを入力としてクリーンデータを推定する「残響→クリーン」学習と，残響重畳音声データとクリーン音声の両方を入力してどちらの場合も対応するクリーン音声をターゲットとする「残響+クリーン→クリーン+クリーン」学習の 2 通りを行った．

計算量の制約から，DAE の学習は全学習セットから話者 80 名をランダムに選択し，さらに話者ごとに 15 文をランダムに選択した計 1200 文 (2.3h) を使用して行った．すなわち，「残響→クリーン」学習の Pre-training では残響重畳音声 1200 文を入力として用い，Fine-tuning では同じ残響重畳音声を DAE の入力側，対応するクリーン音声を出力側に用いて行った．「残響+クリーン→クリーン+クリーン」学習では，Pre-training に残響重畳音声 1200 文とそれに対応するクリーン音声 1200 文の計 2400 文を使用し，Fine-tuning では Pre-training で使用した 2400 文を入力側，対応するクリーン音声 2 セット分の 2400 文を出力側に使用した．DAE の構成は，前章の「デフォルト構成」と同じである．

DAE の学習の後音響モデルとして使用する HMM の学習では得られた DAE を全学習セットに適用し，得られた残響除去処理後のスペクトルから MFCC を作成したものを学習データとして用いる．認識評価は，同様に得られた DAE を評価データに適用し，残響除去処理後のスペクトルから MFCC を作成したものを用いて行う．HMM で使用する特徴量は前章と同じ MFCC12 次元と対数パワーおよびその Δ と ΔΔ の 39 次元であり，CMN を適用している．HMM は 1000 状態の状態共有トライフォンモデルで，各状態のガウス混合分布の混合数は 32 である．言語モデルは JNAS で学習した 3-gram を使用し，認識には WFST デコーダ T^3[13] を使用した．

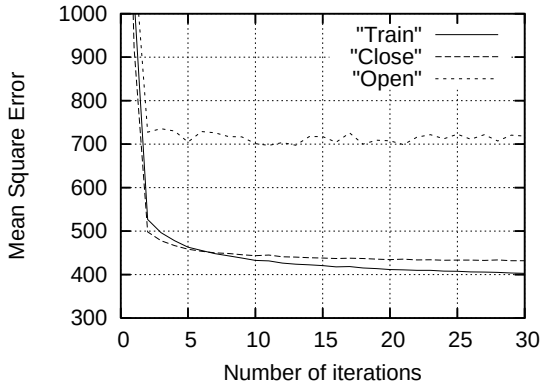

図 7 新聞読み上げ音声におけるバックプロパゲーションの繰り返し数と平均 2 乗誤差

表 5 JNAS データを用いた場合の，Pre-training の有無およびバックプロパゲーションの繰り返し数と，実行時間および単語正解精度の関係

CPU time(Acc)	iter=0	iter=10	iter=30
Pre-training なし	0h(-)	16.3h(53.3)	35.1h(58.6)
Pre-training あり	8.0h(-)	22.2h(63.9)	46.3h(64.7)

5.2 実験結果

図 7 にバックプロパゲーションの繰り返し数とクリーンスペクトルとの平均二乗誤差の推移を示す．二乗誤差は学習セットと，学習セットと同じ CLOSED な残響を畳み込んだ評価用データ，および学習セットとは異なる OPEN な残響を畳み込んだ認識評価に用いるのと同じ評価用データに対して求めた．図より前章の数字音声データを用いた場合とほぼ同様の傾向が見られる．

表 5 に Pre-training の有無およびバックプロパゲーションの繰り返し数と，実行時間および正解精度の関係を示す．本実験は並列スレッド数を約 10 で実行している．Pre-training なしの itter=30 と Pre-training ありの iter=10 を比較すると Pre-training を行った場合の方がより短い学習時間でより高い認識性能が得られていることが分かる．

表 6 にベースラインおよび提案法における単語正解精度を示す．残響評価セット (Reverberation set) は学習セットに対して Open な 4 種類の残響を畳み込んだ評価データ，クリーン評価セット (Clean set) は残響の重畳を行う前のクリーンな評価データである．“ Baseline(clean) ”は残響を畳み込んでいないクリーンな学習セットから学習した HMM を用いた場合，“ Baseline(reverb) ”は 4 種類の残響を畳み込んだ学習セットを用いた場合である．Baseline(reverb) は Baseline(clean) と比較して，クリーン評価セットに対する認識精度がやや低下するものの，残響評価セットに対する認識性能が大きく向上している．“ Reverb(DAE-S) ”および“ Reverb(DAE-SL) ”は，「残響→クリーン」学習を行った DAE-S および DAE-SL を用いた場合の結果である．ど

表 6 JNAS を用いた大語彙連続音声認識での認識精度の評価

Method	Reverberation set	Clean set
Baseline(clean)	44.3	79.4
Baseline(reverb)	59.8	76.1
Reverb(DAE-S)	64.7	66.1
Reverb(DAE-SL)	66.1	65.6
Reverb+clean(DAE-S)	63.3	73.9
Reverb+clean(DAE-SL)	64.8	73.8

ちらも残響評価セットに対する認識精度においてベースラインと比較して大きな向上が見られる．前章での数字認識の場合と同様，DAE-SL を用いた方が DAE-S を用いるよりも高い認識性能が得られている．一方で，クリーン評価セットに対する認識精度はベースラインと比較してどちらも大きく下がってしまっている．“ Reverb+clean(DAE-S) ”および“ Reverb+clean(DAE-SL) ”は，DAE の学習において入力側にもクリーンデータを混ぜて学習を行った「残響+クリーン→クリーン+クリーン」学習の結果である．これにより，残響評価セットでの正解精度が「残響→クリーン」学習と比べて若干低下するものの，クリーン評価セットに対する認識精度が大きく向上することが分かる．残響評価セットにおいて，ベースラインとして最良の結果が得られた残響重畳データからモデル学習した場合の単語正解精度が 59.8% であるのに対して，「残響+クリーン→クリーン+クリーン」学習を行った DAE-SL の単語正解精度は 64.8%であり，単語誤り率削減率では 12.4%の改善となった．

6. まとめと今後の展望

DAE を用いて残響が重畳した対数パワースペクトルから残響を除去する手法の提案を行った．音声の時間変化をモデル化するため，提案法では連続した複数の短時間分析窓によるスペクトルフレームを連結したものをネットワークの入力として用いる．さらに，音声認識に必要なサブ音素レベルでの時間分解能を維持しながら時定数の大きな残響の影響をより正しく捕らえることを目的として，長さの異なる 2 つの分析窓長を併用する拡張手法を提案した．実験では，CENSREC-4 を用いた数字音声認識により提案法が OPEN および CLOSED な残響環境の両方において従来手法よりも効果的であることを示した．さらに JNAS を用いた大語彙連続音声認識においても，提案手法が有効であることを示した．2 つの提案手法を比較すると，長さの異なる 2 つの分析窓長を併用した方が残響除去に対してより効果的に働くことを示した．JNAS における残響評価セットについて，残響重畳学習セットから学習したベースラインの単語正解精度が 59.8% であるのに対して，「残響+クリーン→クリーン+クリーン」学習を行った DAE-SL の単語正解精度は 64.8%であり，12.4%の単語誤り率削減率が得られた．今後の課題としては，実際の残響重畳音に対する提案手法の適用と評価が挙げられる．

謝辞 本研究は科研費 25280058 の助成を受けたものである．また本研究の一部は科研費 21300060 の助成を受けたものである．

参考文献

[1] Kumatani, K., McDonough, J. and Raj, B.: Microphone Array Processing for Distant Speech Recognition: From Close-Talking Microphones to Far-Field Sensors, *Signal Processing Magazine, IEEE*, Vol. 29, No. 6, pp. 127–140 (2012).

[2] Gelbart, D. and Morgan, N.: Evaluating Long-Term Spectral Subtraction For Reverberant ASR, *IN ASRU, MADONNA DI CAMPIGLIO*, pp. 103–106 (2001).

[3] Ganapathy, S., Pelecanos, J. and Omar, M.: Feature normalization for speaker verification in room reverberation, *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, IEEE, pp. 4836–4839 (2011).

[4] Hinton, G. and Salakhutdinov, R.: Reducing the dimensionality of data with neural networks, *Science*, Vol. 313, No. 5786, pp. 504–507 (2006).

[5] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P. A.: Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion, *J. Mach. Learn. Res.*, Vol. 11, pp. 3371–3408 (2010).

[6] Lu, X. and S. Matsuda, C. Hori, H. K.: Speech restoration based on deep learning autoencoder with layer-wised pretraining, *Proc. Interspeech* (2012).

[7] Maas, A., Le, Q., O'Neil, T., Vinyals, O., Nguyen, P. and Ng, A.: Recurrent Neural Networks for Noise Reduction in Robust ASR, *Proc. INTERSPEECH* (2012).

[8] Nishiura, T., Nakayama, M., Denda, Y., Kitaoka, N., Yamamoto, K., Yamada, T., Tsuge, S., Miyajima, C., Fujimoto, M., Takiguchi, T., Tamura, S., Kuroiwa, S., Takeda, K. and Nakamura, S.: Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments —Newest Part of the CENSREC Series—, *Proc. LREC'08* (2008).

[9] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, *Acoust Soc Jpn E*, Vol. 20, No. 3, pp. 199–206 (1999).

[10] Davis, S. B. and Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transaction on Acoustic Speech and Singal Processing*, Vol. 28, No. 4, pp. 357–366 (1980).

[11] Hestenes, M. R. and Stiefel, E.: Methods of Conjugate Gradients for Solving Linear Systems, *Journal of Research of the National Bureau of Standards*, Vol. 49, No. 6, pp. 409–436 (1952).

[12] Ichikawa, O., Fukuda, T. and Nishimura, M.: Dynamic Features in the Linear-Logarithmic Hybrid Domain for Automatic Speech Recognition in a Reverberant Environment, *Selected Topics in Signal Processing, IEEE Journal of*, Vol. 4, pp. 816–823 (2010).

[13] Dixon, P. R., Caseiro, D. A., Oonishi, T. and Furui, S.: The TITech large vocabulary WFST speech recognition system, *Proc. IEEE ASRU*, pp. 443–448 (2007).