Vol.2014-SLP-102 No.6
2014/7/25

ディープオートエンコーダとDNN-HMMを用いた残響下音声認識

三村 正人[1]　坂井 信輔[1]　河原 達也[1]

概要：本研究では、フロントエンドとバックエンドの両方でディープラーニングを用いた残響下音声認識システムについて述べる。このシステムでは、フロントエンドでディープオートエンコーダ（DAE）を用いて音響特徴量の強調 (残響除去) を行い、バックエンドで DNN-HMM 音響モデルにより音声認識を行う。提案手法の性能を、Reverb Challenge 2014 の音声認識タスクにより評価した。まず、マルチコンディションデータを用いて学習した DNN-HMM による認識精度は、すべての残響条件で、MLLR 適応したベースライン GMM-HMM を顕著に上回った。次に、DAE による特徴量強調を行うことにより、クリーン音声を用いて学習した DNN-HMM の残響下音声認識精度を大幅に改善した。さらに、マルチコンディション学習 DNN-HMM と DAE の組み合わせにより、より困難な条件での認識精度が顕著に改善した。これに加えて、DNN により得られる音素識別情報を DAE の入力に追加することで、残響除去の性能が向上した。強調された特徴量に対する DNN-HMM の教師なし適応により、すべての条件で認識精度が向上した。

1. はじめに

近年、統計的な手法に基づく音声認識は、学習データの増加や計算資源の高度化を背景にめざましい進歩を遂げ、音声検索などのアプリケーションも日常的に使われるようになってきた。しかし、実環境における音声認識はいまだ困難なタスクであり、残響や雑音の大きな条件では認識精度は大きく低下する。

音声認識技術がさらに幅広く用いられるための一つの鍵は、ハンズフリーマイク等による簡便な音声入力方式の実現である。このような条件では、残響の影響が不可避である。そのため、残響下音声認識の高精度化に多くの努力が傾けられている。残響下音声認識は、通常、フロントエンドでの音声・特徴量の強調と、バックエンドにおける音響モデル適応や認識システムの高度化によって実現される [1][2][3][4][5][6][7][8]。

残響除去の最も簡易な手法は CMN[1] であるが、通常、残響時間は音声分析窓長にくらべて大きいため、残響時間が長くなると対応できない。これに対して、より洗練された手法（[2] など）が提案されている。

音声強調には、逆フィルタによる逆たたみ込み [3][4][5] や、遅延の大きい残響成分の推定に基づくスペクトルサブトラクション [6][7] に基づくものがある。ただし、SNR を基準として音声強調を行うことが必ずしも音声認識精度を向上させるとは限らないため、バックエンドの音声認識システムの尤度を基準として音声強調を行う手法も提案されている [8]。バックエンドの処理としては、音響モデルを MLLR などを用いて残響に適応するのが典型的である。

本研究では、近年注目を集めているディープラーニングに基づくアプローチを提案する。すなわち、標準的な DNN-HMM [9] による音声認識とデノイジング・ディープオートエンコーダ (DAE) [10][11] による特徴量の強調 (残響除去) の組み合わせにより、残響下音声認識を行う。DNN と DAE の組み合わせは、非常に深い構造を持つ単一の DNN とも見なせるが、異なるターゲットにより学習したネットワークを接続することで相乗的な効果が期待できる。

2. DNN-HMM

ニューラルネットワークによるパターン認識は長い歴史を持つが [12]、近年、深い構造を持つニューラルネットワーク (DNN: Deep Neural Network) が、有効な事前学習法 [13] が確立されたことと、学習データ量および計算資源の増加を背景として、あらためて注目を集めている。音声認識においても、DNN と HMM(hidden Markov Models) の組み合わせにより、多くのタスクで従来の GMM(Gaussian Mixture Model)-HMM より有意に高い認識精度を示すことが報告されている [14]。

DNN を HMM と組み合わせる際、大きく二つのアプローチがある。一つは HMM の状態確率の計算を従来の GMM でなく DNN により直接行うハイブリッドアプローチであり [15][9]

[1] 京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

[16]、もう一つは DNN の出力を従来の GMM-HMM の入力として用いるタンデムアプローチである [17][18][19]。本研究では、構成が単純で多くのタスクで有用性が確認されている前者のハイブリッドアプローチにより音響モデルを構築する。このアプローチによる音響モデルを以降では DNN-HMM と呼ぶ。

DNN-HMM では、教師ラベルを用いた識別的な学習 (fine-tuning) を行う前に、教師なし生成学習 (pre-training)[13] によりネットワークの適切な初期値を設定する。まず、RBM(Restricted Boltzmann Machine) を一層ずつ独立に学習する。次に、これらの RBM を積み重ねて DBM(Deep Belief Network) を作成する。さらに、乱数によって初期化したソフトマックス層を追加することで、DNN の初期ネットワークを構成する。最後に、フレーム毎の正解ラベル（状態 ID）を用いて、誤差逆伝播法 (バックプロパゲーション:BP) による教師あり学習を行う。

DNN-HMM に組み込むためのニューラルネットワークは、HMM の他のコンポーネントと独立に学習するのが一般的である。DNN-HMM における DNN 以外のパラメータは、通常学習された GMM-HMM のものをコピーして用いる。また、教師あり学習に用いる状態ラベルも、この GMM-HMM による強制アラインメントにより生成する。

3. ディープオートエンコーダ (DAE) による特徴量強調

前節で述べたディープニューラルネットワークは、学習のターゲットを変えると回帰タスクのためのディープオートエンコーダ (DAE) として用いることができる [20]。DAE において、下層の数層は効率的な符号を得るためのエンコーダと見なされ、残りの数層は入力を復元するためのデコーダと見なされる。DAE は全体として上下対称な構造を持つネットワークとするのが一般的である。

このような DAE でも、RBM による初期化が重要である。ただし、DNN-HMM で用いるような識別のための DNN とは異なり、エンコーダの各層と対応するデコーダの各層で共通の RBM を用いる。なお、デコーダ層ではノード間の重みを転置して用い、バイアスの初期値には隠れ層でなく可視層のバイアスを用いる。

さらに、入力を雑音や残響の付加されたデータ、ターゲットを元のクリーンデータとすることで、このネットワークをデノイジングオートエンコーダとして学習することができる [21]。デノイジングオートエンコーダでは、雑音や残響の付加されたデータをクリーンなデータに復元するようなネットワークを学習する。入力とターゲットが異なる点以外、学習アルゴリズムは通常のオートエンコーダと同一である [22]。

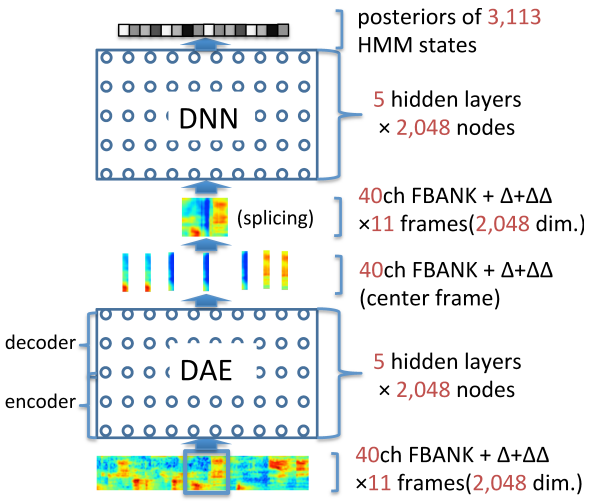

図 1 提案手法によるネットワークの構造

4. 提案手法

4.1 DAE と DNN のカスケード

本研究では、第 2 節と第 3 節で述べた 2 つのネットワークの組み合わせにより残響下音声認識を行う。これは、異なるターゲットを用いて学習したネットワークが相補的な効果をもたらすことを期待したものである。提案手法によるネットワークを、本研究の評価実験で用いた具体的なパラメータ数とともに図 1 に示す。

入力データの特徴量はまず DAE により残響除去される。次にこの強調された特徴量を DNN へ入力し、状態の事後確率を計算する。DAE、DNN ともに前後 11 フレームの特徴量を入力とする。これらの 2 つのネットワークは独立に学習する。

4.2 DNN 音素情報を用いた DAE による特徴量強調

最近、I-vector を用いた話者適応 [23] や noise-aware training [24] のように、付加的な情報により DNN の入力ベクトルを拡張することで認識精度の向上を試みるアプローチが注目されている。本研究では、DAE の入力ベクトルに音素情報を付加することで残響除去の性能向上を試みる。

当該フレームがどの音素に属するかによりクリーン音声特徴量の存在する領域が異なるため、この情報を用いることで、音響特徴量のみからクリーン音声を再現するより効果的と考えられる。音素情報には種々の表現が考えられるが、本研究では単純に DNN の出力として得られる音素 HMM 状態の事後確率を DAE の入力にフィードバックさせる。この状態事後確率は、当該フレームがどの音素状態に属しているかのソフトな表現になっている。以下では、この DAE を pDAE と呼ぶ。

5. 評価実験

5.1 実験データと条件

提案手法を、Reverb Challenge 2014 音声認識タスク [25] により評価した。学習データは、クリーンな WSJCAM0 音

表 1 残響下評価データに対する各システムの認識性能 (単語誤り率 (%))

		SimData							RealData		
		Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
		Near	Far	Near	Far	Near	Far		Near	Far	
(1)	Baseline (clean, w/o CMLLR)	18.26	25.60	41.87	82.20	53.59	87.99	51.73	89.91	87.58	88.74
(2)	Baseline (multicond, w/o CMLLR)	21.28	21.18	23.12	38.83	28.24	44.77	29.56	58.96	55.60	57.28
(3)	Baseline (multicond, w CMLLR)	16.57	18.21	20.31	32.43	24.86	39.28	25.27	50.37	48.01	49.19
(4)	DNN-HMM (clean)	11.79	17.21	24.17	55.11	31.55	68.79	34.82	74.64	71.98	73.35
(5)	DAE + DNN-HMM (clean)	9.03	10.13	10.89	19.84	11.35	22.32	13.92	41.90	42.74	42.32
(6)	pDAE + DNN-HMM (clean)	8.57	9.56	10.89	18.72	11.45	20.27	13.24	39.12	38.99	39.06
(7)	DNN-HMM (multicond)	9.25	9.93	12.35	19.26	12.52	20.25	13.92	40.11	40.32	40.22
(8)	DAE + DNN-HMM (multicond)	14.47	14.08	11.77	16.87	12.90	17.32	14.57	35.96	36.53	36.25
(9)	pDAE + DNN-HMM (multicond)	13.89	13.57	11.69	16.72	12.71	17.35	14.32	35.00	35.31	35.16
(10)	pDAE + DNN-HMM (multicond) + adap.	10.89	11.42	11.30	15.76	11.86	16.06	12.89	31.33	32.61	31.95

表 2 クリーン評価データに対する各システムの認識性能 (単語誤り率 (%))

		ClnData			
		Room 1	Room 2	Room 3	Ave.
(1)	Baseline (clean, w/o CMLLR)	13.01	12.69	12.23	12.64
(2)	Baseline (multicond, w/o CMLLR)	30.92	30.28	30.17	30.46
(3)	Baseline (multicond, w CMLLR)	16.25	15.28	15.37	15.63
(4)	DNN-HMM (clean)	6.73	7.09	6.81	6.88
(7)	DNN-HMM (multicond)	10.47	10.56	9.66	10.24

声 [26] に各種 RIR(Room Impulse Response) をたたみ込み、RIR と同一の部屋で収録した定常背景雑音を（SN 比が 20dB となるように）加えることで作成したものである（マルチコンディション学習データ)。データ量は 7,861 発話（17 時間）である。評価データは、模擬残響データ（SimData）と実残響データ（RealData）からなる。SimData は、クリーンな音声に RIR をたたみ込み、定常背景雑音を SNR が 20dB となるように加えることで作成したものである。RIR は 3 つの部屋（Room 1(small), Room 2(medium), Room3(large)）および 2 つのマイク距離（near=50cm, far=200cm）で収録されたものである。3 つの部屋の T60 は、それぞれ 0.25s、0.5s、0.7s である。これらの部屋は、学習データの RIR を収録した部屋と異なっている。RealData は SimData と異なる一つの部屋で収録された。この部屋の残響時間 T60 は 0.7s である。マイク位置は 2 種類である (near= 100cm, far= 250cm)。SimData と RealData のいずれの発話も、WSJCAM0 から選択された文である。すべての残響音声はマイクロフォンアレーの 8ch データで収録されているが、本研究の実験では、学習データおよび評価データのいずれも 1ch のみを用いる。

音声認識実験におけるデコーダは HTK-3.4 の HVite を用いた。ただし、DNN による認識を行うために一部のコードを改変した。言語モデルは、reverb challenge で提供されるベースライン言語モデルを用いた。ビーム幅などのデコーディングパラメータは、GMM-HMM と DNN-HMM で共通のものを用いた。ただし、GMM-HMM と DNN-HMM で尤度の存在範囲が異なるため、言語重みおよび挿入ペナルティは各々で最適化した。DNN・DAE の学習プログラムは Python により実装した。また、CUDAMat ライブラリを用いて GPGPU による学習の高速化を行った。

評価データに対するベースライントライフォン GMM-HMM による認識性能を表 1 の第 1 行 (クリーンデータで学習した GMM-HMM)、第 2 行（マルチコンディションデータで学習した GMM-HMM）および第 3 行（MLLR 適応したマルチコンディション GMM-HMM）に示す。ベースライン GMM-HMM の状態数は 3,113 であり、状態あたりの混合数は 10 である

5.2 DNN-HMM の評価

実験に用いた DNN-HMM の諸元は以下の通りである。

ネットワークの入力は対数メルフィルタバンク出力 40 次元+Δ+ΔΔ を 11 フレーム分つなげた 1320 次元とした。この入力特徴量は、発話毎に平均 0 となるような正規化を行った上、さらにグローバルに平均 0、分散 1 となるように正規化を行った。出力層はベースライン GMM-HMM のトライフォン共有状態 3113 とした。ネットワークは隠れ層 5 とソフトマックス出力層からなる 6 層である。隠れ層のノード数は 2048 とした。ネットワークはマルチコンディションデータを用いて学習した RBM により初期化した。DNN の教師あり学習は、フレームの状態ラベルとマルチコンディションデータを用いて、誤差逆伝播法により行った。損失関数にはクロスエントロピを用いた。確率的勾配降下法におけるミニバッチのサイズは 256 とした。学習係数の初期値は 0.08 とした。各エポック終了時

に開発セットのフレーム識別精度を計算し、前エポックからの改善が 0.2 %を下回ったとき学習係数を半減させた。学習はエポック数 20 で停止させた。モメンタムは 0.9 とした。学習に用いるフレームの状態ラベルは、クリーンデータの MFCC およびベースライン GMM-HMM を用いて生成した。なお、遷移行列など DNN 以外の HMM パラメータは、ベースライン GMM-HMM のものをコピーして用いた。

マルチコンディションデータを用いて学習した DNN-HMM による単語誤り率を表 1 の第 7 行に示す。SimData のすべての条件において、マルチコンディション DNN-HMM の認識精度は、MLLR 適応したベースライン GMM-HMM（第 3 行）を大幅に上回った。最も残響の影響が大きい条件（Room3、Far）で、単語誤り率は 19.03 ポイント軽減した。また、RealData に対して、いずれのマイク距離でも適応済ベースライン GMM-HMM の精度を顕著に上回り、平均で 8.97 ポイント単語誤り率が軽減した。

マルチコンディション DNN-HMM に加えて、クリーンな学習データ（WSJCAM0）により学習した DNN-HMM の評価も行った (表 1 の第 4 行)。クリーン DNN-HMM の認識精度は、すべての残響条件において、マルチコンディション DNN-HMM を大幅に下回った。この結果から、GMM-HMM と同様、DNN-HMM においてもマルチコンディション学習が有効であるといえる。

参考のため、残響の重畳を行わないクリーンな評価データ（”ClnData”）に対する認識実験も行った。クリーン評価データに対するベースライン GMM-HMM の単語誤り率を表 2 の第 1 行から第 3 行に、クリーン DNN-HMM の誤り率を第 4 行に、マルチコンディション DNN-HMM の誤り率を第 5 行に示す。

クリーン評価データに対する GMM-HMM の認識精度は、マルチコンディション学習により大幅に低下した（平均で 12.64 %から 30.46 %）。一方、マルチコンディション DNN-HMM の認識精度は、クリーン GMM-HMM を 2.4 ポイント上回った。また、クリーン DNN-HMM による認識精度はマルチコンディション DNN-HMM より高かったが、この学習データによる差（3.36 ポイント）は GMM-HMM における差（17.82 ポイント）ほど顕著ではなかった。

5.3 DAE の評価

DAE の入力として、DNN の入力と同じ 11 フレーム分の FBANK 特徴量 (1320 次元) を用いた。また、DAE の出力は中心フレームのみ (120 次元) とした。DAE の教師あり学習は、マルチコンディションデータを入力、クリーン音声をターゲットとして行った。なお、対応する入力フレームと出力フレームは、マルチコンディションデータの作成時にアラインメントされている。

DAE は、DNN に用いたのと同じ RBM のうち、最初の 3 つの RBM により初期化を行った。すなわち、これら 3 つの RBM とそれらを反転した 3 つのネットワークをカップリングして初期ネットワークとした。ただし、最後の隠れ層と出力層の間のリンクおよび出力層のバイアスについては、RBM パラメータのうち中心フレームの 120 次元に対応するもののみを用いた。結果として、DAE は 5 つの隠れ層を持つ計 7 層の構造を持つネットワークとなる。

DAE の教師あり学習は、自乗誤差を損失関数とする誤差逆伝播法により行った。ミニバッチサイズやモメンタム等のパラメータは DNN と同一である。ただし、学習係数の初期値は 0.001 と DNN より小さい値を用いた。また、各エポック終了時に、得られた DAE による開発セットの強調特徴量とクリーン DNN-HMM を用いてフレーム識別率を計算し、前エポックからの改善が 0.2 %を下回ったとき学習係数を半減させた。学習は 20 エポックで停止させた。

DAE の出力は中心フレームのみ（120 次元）であるため、DNN-HMM への入力はこれらを 11 フレーム分つなげた 1320 次元とする。

残響音声のスペクトルと DAE により残響除去したスペクトルの例を図 2 に示す。DAE による強調特徴量では、時間軸に沿ったスペクトラムの残響成分が軽減されているのがわかる。

5.3.1 DAE とクリーン DNN-HMM の組み合わせ

DAE とクリーン DNN-HMM の組み合わせによる単語誤り率を表 1 の第 5 行に示す。この組み合わせにより、DAE による強調を行わないクリーン DNN-HMM 単体の精度（第 4 行）に比べて、精度が大幅に向上した。このことから、DAE により効果的に残響除去が行われていることがわかる。興味深いことに、DAE とクリーン DNN-HMM を組み合わせたシステムの認識精度は、DAE による雑音除去を行わないマルチコンディション DNN-HMM 単体の精度（第 7 行）とほぼ同等であった。

5.3.2 DAE とマルチコンディション DNN-HMM の組み合わせ

DAE とマルチコンディション DNN-HMM の組み合わせによる単語誤り率を、表 1 の第 8 行に示す。比較的残響の小さい条件（SimData の Room1）では、DAE により精度が低下したが、より残響の大きい条件（SimData における Room2 および Room3 の Far、および RealData）では、DAE とマルチコンディション DNN-HMM の組み合わせにより顕著に認識精度が向上した。これは、異なるターゲットにより学習した 2 つのネットワークが相補的な効果をもたらしたためと考えられる [*1] 最も困難な RealData において、MLLR 適応した GMM-HMM より 12.94 ポイント単語誤り率が軽減した。

5.3.3 DNN 音素識別情報を用いた特徴量強調

4.2 節で述べたように、DNN の音素識別の出力を用いた DAE を学習・評価した。学習アルゴリズムは、FBANK 特徴

[*1] 予備的に 5 層より多い層を持つマルチコンディション DNN-HMM を学習し、評価を行ったが、精度の向上は見られなかった。すなわち、本タスクでは単に層の数を増やすだけでは精度に寄与しない。

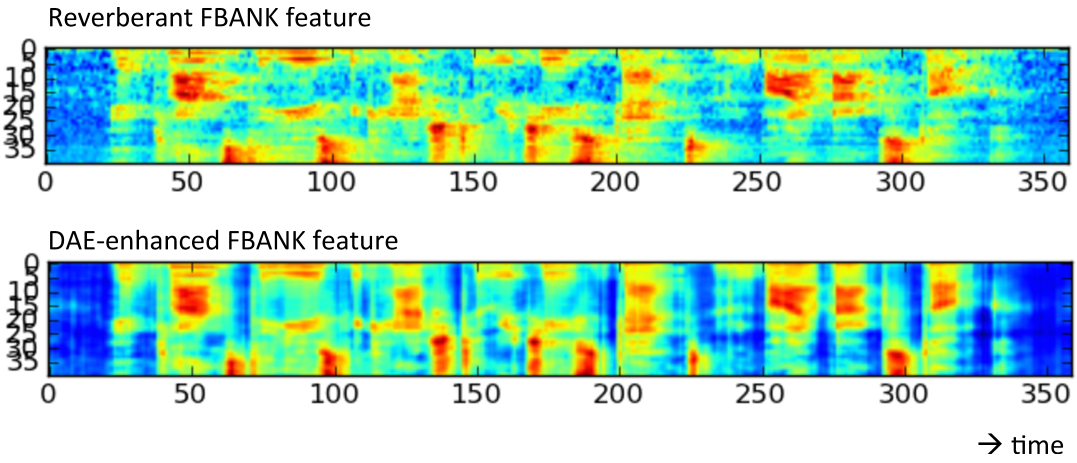

図 2 DAE により特徴量強調された発話の例

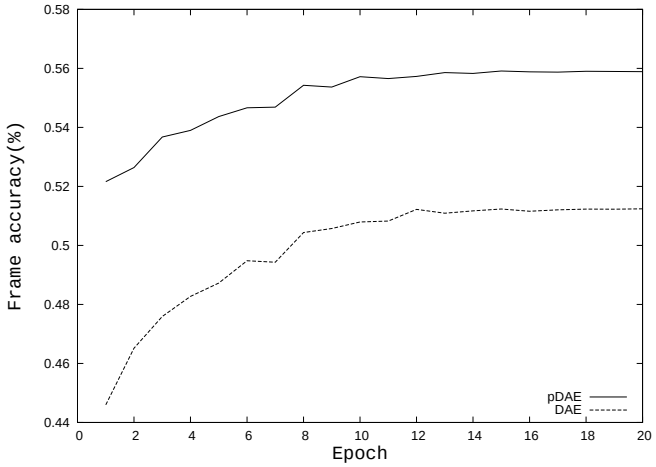

図 3 DAE および pDAE 学習時の開発セットのフレーム識別精度

量 1320 次元に中心フレームの DNN 出力（状態事後確率）を付加する以外は、通常の DAE と同様である。ただし、本実験では、次元数を抑えるためにトライフォンでなくモノフォン状態の事後確率 (145 次元) を用いた。モノフォンをターゲットとする DNN は、別途マルチコンディションデータを用いて作成した。

モノフォン DNN により得られる音素状態事後確率を付加した DAE（pDAE）と通常の DAE の学習時におけるエポック毎の開発セットフレーム識別率を図 3 に示す。フレーム識別率は、各エポック終了時に得られた DAE による強調特徴量をクリーン DNN に入力して計算したものである。この図から、pDAE の与えるフレーム識別率は明らかに通常の DAE より高い領域に存在しており、より効率的に学習が進行していることがわかる。

この pDAE とクリーン DNN-HMM の組み合わせによる認識精度を表 1 の第 6 行に示す。RealData において、通常の DAE より平均で 3.26 ポイント誤り率の改善があった。

pDAE とマルチコンディション DNN-HMM の組み合わせによる認識精度を表 1 の第 9 行に示す。この場合も、RealData において 1.09 ポイントの誤り率の改善があり、DNN の音素識別情報を DAE の入力に追加することの効果が確認できた。

5.3.4 強調特徴量を用いた DNN-HMM の教師なし適応

DAE による強調特徴量と DNN-HMM とのミスマッチを軽減するために、DNN-HMM の教師なし適応を行った。ただし、DNN-HMM の適応については、GMM-HMM における MLLR や MAP のような統計的手法が確立されていない。しかし、評価データを用いて追加の誤差逆伝播学習を行うことで適応に似た効果が得られることが経験的に知られている [27]。本研究では、pDAE により特徴量強調した評価データと初期認識結果を用いて 10 エポックの追加の誤差逆伝播学習を行うことで DNN-HMM の教師なし適応を行った。

MLLR 適応した GMM-HMM と公平な比較を行うために、同一のデータを用いて適応を行った。すなわち、各条件（部屋とマイク距離の組み合わせ・表 1 の各列に対応）の全発話を用いて DNN-HMM の適応を行った（full batch 条件 [25]）。ラベルは、pDAE とマルチコンディション DNN-HMM の組み合わせによる認識結果から作成した。学習係数は 0.001 と比較的小さい値を用いた。

適応実験の結果を表 1 の第 9 行に示す。DNN-HMM の教師なし適応により、すべての条件で認識精度が向上した。RealData に対し、平均で誤り率が 31.95 %となり、MLLR 適応した GMM-HMM に比べて 17.24 ポイント高い精度となった。

6. おわりに

本研究では、DAE による特徴量の強調と DNN-HMM による音響モデルの組み合わせによる残響下音声認識について述べた。

マルチコンディションデータを用いて学習した DNN-HMM は、DAE による残響除去を行わない条件でも、MLLR 適応したベースライン GMM-HMM より顕著に高い認識精度を示した。また、DAE による残響除去により、クリーンデータで学習した DNN-HMM の認識精度が大幅に向上し、マルチコンディション DNN-HMM 単独と同等の性能となった。これにより、

DAE により効果的に残響除去が行えることが確認された。さらに、DAE とマルチコンディション DNN-HMM の組み合わせにより、実残響データなど特に困難な条件で、DNN-HMM 単独より顕著に誤り率が改善した。

これに加えて、DNN の音素識別情報を DAE に追加することにより、通常の音響特徴量のみを用いた DAE より残響除去の性能を改善することができた。

さらに、DAE による強調特徴量に対する DNN-HMM の教師なし適応により、すべての条件で認識精度が向上し、最終的に RealData に対し平均で 31.95 %の誤り率を実現した。

提案手法は学習時・評価時における残響時間の推定等の処理を含まないため、簡易に実装可能であり、どのような環境にでもすぐに適用できる。また、学習データを増やすことにより精度の向上が期待できる。

今回、DAE の入力としては中心フレームおよび前後 5 フレームの計 11 フレーム（110msec に対応）を用いた。これは、学習・評価データにおける T60 より明らかに短い値である。残響除去を目的とした DAE の入力としてどれくらい長いコンテクストを与えるのがよいか、今後調査の必要がある。また、入力のマルチチャネル化も今後の課題である。

参考文献

[1] A.E.Rosenberg, C.H.Lee and F.K.Soong: Cepstral channel normalization techniques for HMM-based speaker verification, *ICSLP*, pp. 1835–1838 (1994).

[2] Krueger, A. and Haeb-Umbach, R.: Model-based feature enhancement for reververant speech recognition, *IEEE Trans. Audio, Speech & Language Process.*, Vol. 18, No. 7, pp. 1692–1707 (2010).

[3] M.Gurelli and C.Nikias: Evam: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals, *IEEE Trans. Audio, Speech & Language Process.*, Vol. 43, No. 1, pp. 134–149 (1995).

[4] M.Delcroix, T.Hikichi and M.Miyoshi: On the use of lime dereverberation algorithm in an acoustic environment with a noise source, *ICASSP*, Vol. 1 (2006).

[5] S.Gannot and M.Moonen: Subspace methods for multimicrophone speech dereverberation, *EURASIP J.Appl.Signal Process.*, Vol. 11, pp. 1074–1090 (2003).

[6] M.Wu and D.Wang: A two-stage algorithm for one-microphone reverberant speech enhancement, *IEEE Trans. Audio, Speech & Language Process.*, Vol. 14, No. 3, pp. 774–784 (2006).

[7] K.Kinoshita, M.Delcroix, T.Nakatani and M.Miyoshi: Suppression of late reverberation effect on speech signal using long-term multiplestep linear prediction, *IEEE Trans. Audio, Speech & Language Process.*, Vol. 17, No. 4, pp. 534–545 (2009).

[8] R.Gomez and T.Kawahara: Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood, *IEEE Trans. Audio, Speech & Language Process.*, Vol. 18, No. 7, pp. 1708–1716 (2010).

[9] G.E.Dahl, D.Yu, L.Deng and A.Acero: Context-dependent pre-trained deep neural networks for large vocabulary speech recognition, *IEEE Trans. Audio, Speech, & Language Proc.*, Vol. 20, No. 1, pp. 30–42 (2012).

[10] T.Ishii, H.Komiyama, T.Shinozaki, Y.Horiuchi and S.Kuroiwa: Reverberant Speech Recognition Based on Denoising Autoencoder, *INTERSPEECH*, pp. 3512–3516 (2013).

[11] X.Lu, Y.Tsao, S.Matsuda and C.Hori: Speech Enhancement Based on Deep Denoising Autoencoder, *INTERSPEECH*, pp. 436–440 (2013).

[12] Bishop, C. M.: *Neural Networks for Pattern Recognition*, Oxford University Press (1995).

[13] G.E.Hinton, S.Osindero and Y.Teh: A fast learning algorithm for deep belief nets, *Neural Computation*, Vol. 18, pp. 1527–1554 (2006).

[14] G.E.Hinton, L.Deng, D.Yu, G.Dahl, A.Mohamed, N.Jaitly, A.Senior, V.Vanhoucke, P.Nguyen, T.Sainath and B.Kingsbury: Deep neural networks for acoustic modeling in speech recognition, *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82–97 (2012).

[15] A.Mohamed, G.Dahl and G.Hinton: Acoustic modelling using deep belief networks, *IEEE Trans. Audio, Speech, & Language Proc.*, Vol. 20, No. 1, pp. 14–22 (2012).

[16] F.Seide, G.Li and D.Yu: Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, *INTERSPEECH*, pp. 437–440 (2011).

[17] N.Morgan: Deep and wide: Multiple layers in automatic speech recognition, *IEEE Trans. Audio, Speech, & Language Proc.*, Vol. 20, No. 1, pp. 7–13 (2012).

[18] G.S.V.S.Sivaram and H.Hermansky: Sparse multilayer perceptron for phoneme recognition, *IEEE Trans. Audio, Speech, & Language Proc.*, Vol. 20, No. 1, pp. 23–29 (2012).

[19] P.J.Bell, M.J.F.Gales, P.Lanchantin, X.Liu, Y.Long, S.Renals, P.Swietojanski and P.C.Woodland: Transcriptions of multigenre media archives using out-of-domain data, *Proc. SLT*, pp. 324–329 (2012).

[20] G.E.Hinton and R.R.Salakhutdinov: Reducing the Dimensionality of Data with Neural Networks, *Science*, Vol. 313, pp. 504–507 (2006).

[21] P.Vincent, H.Larochelle, Y.Bengio and P.A.Manzagol: Extracting and Composing Robust Features with Denoising Autoencoders, *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, pp. 1096–1103 (2008).

[22] Y.Bengio, P.Lamblin, D.Popovici and H.Larochelle: Greedy Layer-Wise Training of Deep Networks, *in Advances in Neural Information Processing Systems 19 (NIPS06)*, pp. 153–160 (2007).

[23] G.Saon, H.Soltau, D.Nahamoo and M.Picheny: Speaker Adaptation of Neural Network Acoustic Models Using I-Vevtors, *Proc. ASRU*, pp. 55–59 (2013).

[24] M.Seltzer, D.Yu and Y.Wang: An Investigation of deep neural networks for noise robust speech recognition, *Proc. ICASSP*, pp. 7398–7402 (2013).

[25] K.Kinoshita, M.Delcroix, T.Yoshioka, T.Nakatani, E.Habets, R.Haeb-Umbach, V.Leutnant, A.Sehr, W.Kellermann, R.Maas, S.Gannot and B.Raj: The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)* (2013).

[26] T.Robinson, J.Fransen, D.Pye, J.Foote and S.Renals: WSJ-CAM0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition, *Proc. ICASSP*, pp. 81–84 (1995).

[27] Y.Xiao, Z.Zhang, S.Cai, J.Pan and Y.Yan: An initial attempt on task-specific adaptation for deep neural network besed large vocabulary continuous speech recognition, *Proc. INTERSPEECH* (2012).