Vol.2013-SLP-97 No.8
2013/7/26

Deep Neural Networkに基づく日本語音声認識の基礎評価

神田 直之[1,a)]　武田 龍[1]　大淵 康成[1]

概要：本稿では Deep Neural Network (DNN) を用いた日本語音声認識に関する検討結果を述べる．DNN とは多数の階層を持った人工ニューラルネットワークモデルである．近年，多層のネットワークでも効率的に最適化できる手法が発表され，各種の認識タスクで従来法を上回る性能を示したことから，再び大きな注目を集めている．音声認識分野においても DNN に基づく音響モデルに関して既に多数の研究が行われ改善が進む一方で，日本語のテストデータを用いた検討結果は限られた学習データを用いた小規模な実験に限られていた．本稿では日本語話し言葉コーパス（CSJ）をテストセットとし DNN に基づく音響モデルに関する各種の評価を行った結果について述べる．特に 270 時間の学習データを用いた評価において，音素誤り最小化 (MPE) 学習された Gaussian Mixture Model に基づく音響モデルと比較して最大 28.2%の認識誤りが削減され，DNN の認識性能の高さを日本語においても確認した．また DNN に基づく音響モデルにおいて，学習用の言語リソースが限られた状況でデータを擬似的に増加させる手法について新たに検討を行い，認識精度がさらに向上することを確認した．

1. はじめに

近年 Deep Learning と呼ばれる深い階層構造を持った一連の手法が様々な評価タスクにおいて従来法を大きく上回る性能を示し注目を集めている．音声認識においてもこれまで Gaussian Mixture Model (GMM) によって表現していた特徴量の出力確率を Deep Neural Network (DNN) に置き換えた手法 (DNN-HMM) が提案され [1]，大幅な認識率の向上が見られた．近年は Recurrent Neural Network[2] や Convolutional Neural Network[3] などを用いた拡張が研究されるとともに，基礎的な性質を調べる多数の研究が行われている [4], [5].

DNN に基づく音声認識はそのほとんどが英語のコーパスに基づいて研究が進められている．多言語での研究も進められており，そのいずれにおいても高い認識性能を示しているものの，日本語における知見は限られている．筆者らが知る限り，日本語を用いて DNN に基づく音声認識の性能を示したものは fujii らの研究 [6]，matsuda らの研究 [7] および西野らの研究 [8] である．fujii ら [6] は Deep-hidden conditional neural fields を用いた音素認識実験において ASJ および JNAS コーパスを用いた精度評価を報告している．また matsuda ら [7] は言語非依存の層を抽出する研究において，フレーム単位の音素認識率によって評価を行っている．これらの研究では評価は音素単位で行われており，大語彙連続音声認識での評価はなされていない．西野らの研究 [8] は文献 [1] で提案された DNN に基づく音響モデル (DNN-HMM) を実装し日本語話し言葉コーパス (CSJ) を用いた大語彙連続音声認識実験によって性能向上を確認した．しかしこの研究では学習データが高々 10 時間の音声にとどまっており，得られた認識性能は CSJ テストセット 2 において 77%弱と，より大規模な学習コーパスを用いた従来のモデルによって得られる結果 (文献 [9] など) を超えるものではなかった．英語で示された DNN の高い性能を鑑み，日本語のテストデータにおいて，より大量の学習データの下で DNN に基づく大語彙連続音声認識システムを評価することは重要であると考える．

本稿では DNN-HMM[1] に基づく大語彙連続音声認識を実装し，日本語話し言葉コーパスを用いて各種の評価を実施した結果について述べる．特に 270 時間の学習データを用いた評価において，音素誤り最小化 (MPE) 学習された GMM-HMM の CSJ テストセット 1 とテストセット 2 における単語認識率が 79.9%と 82.3%であったのに対し，DNN に基づく音響モデルはそれぞれ 85.4%と 87.3% の認識率を示し，DNN の認識性能の高さを日本語においても確認した．

別の観点として，少量の音声リソースしか利用できない場合に如何にして高い認識性能を出すのかという研究が進められている [10], [11]．本稿では少量（約 10 時間）の音声データを用いた場合の DNN-HMM の性能も評価

[1] (株) 日立製作所中央研究所
Central Research Laboratory, Hitachi Ltd., 1-280, Higashi-koigakubo, Kokubunji, Tokyo 185-8601, Japan
[a)] naoyuki.kanda.kn@hitachi.com

し，この中で特徴量の比較や Dropout 法 [12] などの評価も併せて行った．最終的に約 10 時間の音声データから得られた DNN-HMM が 270 時間の音声データに基づいて学習された GMM-HMM に匹敵する性能を示したことから，DNN-HMM がリソースの限られた音声に対する音響モデルとしても極めて有望であることを確認した．

本稿では，この DNN-HMM の高い音声認識性能をさらに向上させるため，音声データを意図的に歪ませることにより擬似的に音声データ量を増加させる手法についての検討を新たに行った．ここでは声道長の変化に相当する歪みと話速変化に相当する歪み，及びランダムに周波数方向に変動する歪みを音声データに付加することで擬似的に音声データ量を増加させた．評価の結果，提案するスペクトル伸縮歪みにより単語認識率が最大で 10.1%向上することを確認した．

以下ではまず 2 章において Deep Neural Network に基づく音響モデルの概要を述べる．続いて 3 章において日本語話し言葉コーパスを用いて行った一連の評価実験について述べる．

2. Deep Neural Network に基づく音響モデル

2.1 Deep Neural Network の概要

Deep Neural Network は多数の階層を持った人工ニューラルネットワーク (Artificial Neural Network; ANN) モデルである．ANN はもともと多層構造のモデルで提案されており，DNN と ANN に本質的な違いはない．しかしながら，ANN の最適化に用いられる Back-Propagation 法 [13] には層数を増やすごとに最適化が困難になる問題 (初期値依存問題および勾配の消失問題) が存在し，従来はネットワークの層数を増やすことが識別性能の向上に寄与しないと考えられていた．これに対し 2006 年に Hinton らが，多層のネットワークでも効率良くネットワークを最適化できるようにするための DNN の逐次的な初期化方法を提案 [14] し，これにより従来法を凌ぐ識別性能を示したことで，近年再び注目を集めている．

ここでは Neural Network の第 l 層のノード数を N_l，ノードの値を $\mathbf{z}^l = (z_1^l, ..., z_{N_l}^l)^T$ と表す．この時 Nural Network で l 層の値 $\mathbf{z}^l$ は以下の式に従って $l+1$ 層へ伝播される．

$$a_j^{l+1} = \mathbf{w}_j^l \cdot \mathbf{z}^l + b_j^l \tag{1}$$

$$z_j^{l+1} = h(a_j^{l+1}) \tag{2}$$

ここで $\mathbf{w}_j^l$ 及び b_j^l は l 層の Neural Network の重みパラメータとバイアスパラメータを表す．また h は各ノードの活性化関数を表す．本研究では中間層における活性化関数として，下記の sigmoid 関数を用いた．

$$h(a_j^{l+1}) = \frac{1}{1+\exp(-a_j^{l+1})} \tag{3}$$

また最終層では活性化関数として多クラス識別に適した softmax 関数を用いた．

$$h(a_j^{L+1}) = \frac{\exp(a_j^{L+1})}{\sum_{k=1}^{N^{L+1}} \exp(a_k^{L+1})} \tag{4}$$

第 1 層のノードの値 $\mathbf{z}^1$ に D 次元の入力特徴量 $\mathbf{x} = (x_1, ..., x_D)^T$ を，また出力層に K 次元ベクトル $\mathbf{y} = (y_1, ..., y_K)^T$ をそれぞれ割り当てた場合，ネットワークの出力 z_k^{L+1} は入力に対するクラス k の事後確率として $p(y_k = 1|\mathbf{x})$ と解釈される[*1]．

Neural Network の最適化は学習データが与えられたとき，下記の事後確率最大化問題として表現される[*2]．

$$\mathcal{L} = \sum_i \log p(y_{(k_i)} = 1|\mathbf{x}_i) \tag{5}$$

ここで k_i は i 番目の学習データの正解クラスである．このとき，パラメータ $\mathbf{w}_j^l$ 及び b_j^l は確率的勾配降下法などの最適化手法によって求めることができる．

$$(\mathbf{w}_j^l, b_j^l) \leftarrow (\mathbf{w}_j^l, b_j^l) + \eta \frac{\mathcal{L}}{\partial(\mathbf{w}_j^l, b_j^l)} \tag{6}$$

2.2 Deep Neural Network の初期化

Hinton らは Restricted Boltzmann Machine (RMB) [15] と呼ばれるモデルをネットワークの各層に対して段階的に適用することで Neural Network の階層が多層に渡る場合でもパラメータを良い状態に初期化できることを示した [14]．RBM によって重みが初期化された DNN は Back-Propagation 法によって最適化される (fine-tuning と呼ばれる)．この手法によって生成された DNN は従来から用いられた 1 層や 2 層の場合と比較して大幅に高精度であることが示されている．一方で近年，識別的 pre-training と呼ばれる手法が提案され，少なくとも英語音声認識においては RBM を若干上回る性能が示されている [4]．識別的 pre-training では以下のプロセスに従って L 層の DNN の初期化を行う．

(1) 2 層の Neural Network を構成しパラメータを学習する．
(2) $l = 3$ として，l を 1 ずつ増加させながら $l = L$ となるまで (3)-(6) を行う
(3) 出力層である l 層目のノードを取り除き，新たに l 層目のノードを追加する．
(4) 出力層として $l+1$ 層目のノードを追加する．
(5) l 層目と $l+1$ 層目のノードをリンクで結び，リンクを小さな乱数で初期化する．これにより l 層の Neural Network が構築される．
(6) l 層の Neural Network のパラメータを最適化する．

文献 [4] では上記の過程において，(1) パラメータの更新係数として大きめの値を用いることと，(2) パラメータの更

[*1] 本稿では DNN の層の数は学習される重みに従って数える．この場合，L 層ネットワークの最終層の値は $\mathbf{z}^{L+1}$ に現れる．
[*2] これはクロスエントロピー誤差最小化問題と等価である．

新を学習データ 1 サンプルあたり 1 回のみ行うこと (early stopping)，の 2 つを行うことにより良い性能が得られると述べている．本研究では識別的 pre-training による初期化を用いた．

2.3 Deep Neural Network に基づく音響モデル

従来から音響モデルは隠れマルコフモデル (Hidden Markov Model; HMM) によってモデル化されており，このうちある状態 s から，入力特徴量 $\mathbf{x}$ が出力される確率 $p(\mathbf{x}|s)$ は GMM によって表現されていた．DNN に基づく音響モデルではこの出力確率を Neural Network に置き換える．Neural Network を用いた場合，状態 s を識別すべきクラスとして割り当てると，2.1 節の議論によって事後確率 $p(s|\mathbf{x})$ が求まる．ここでベイズ則に従うと $p(\mathbf{x}|s)$ は下記のように求められる．

$$p(\mathbf{x}|s) = \frac{p(s|\mathbf{x})p(\mathbf{x})}{p(s)} \tag{7}$$

このうち p(s) は状態 s の生起確率であり，学習データから求めることができる．一方，p($\mathbf{x}$) は入力特徴量の生起確率であり，認識結果には影響を与えないため無視することができる．このため，最終的には下式に従って出力確率の計算を行う．

$$p(\mathbf{x}|s) \propto \frac{p(s|\mathbf{x})}{p(s)} \tag{8}$$

2.4 Dropout 法

DNN はその表現力の高さから，学習データに対して過学習を起こすことが懸念される．この問題に対し Dropout 法と呼ばれる手法が提案されている [12]．Dropout 法とは，DNN の学習時の各反復において隠れ層の各ノードの出力を γ%の確率で 0 にする (dropout する) 手法である．反復ごとに 0 にするノードを変えることにより，全体として弱識別器を多数統合したような効果が生じ，画像認識 [16]，音声認識 [12] のいずれにおいても認識性能が大きく向上することが知られている．一方で学習に要する時間が大幅に増加する欠点が指摘されている．本稿では少量のデータセットにおいて Dropout 法の効果を検証した．

2.5 スペクトル伸縮歪みによる擬似学習データ生成の効果

少量の書き起こし付き音声コーパスしか利用できない状況を想定し，学習データを意図的に歪ませることによって擬似的に学習データを増加させる手法についての検討を行った．擬似的に学習データを歪ませる手法については文字認識の分野では一般的に利用され，認識性能を大幅に向上させることが知られている．例えば文献 [17] では，画像に回転および伸縮を施すことにより文字認識性能が大きく向上することを示した．一方で音声は画像と異なり時間方向と周波数方向で明確に異なる特性を持ち，画像認識で有効な回転処理などは適さないと思われる．そこで本研究では以下の 3 種類の音声歪みについて検討を行った．

2.5.1 声道長歪み (Vocal Tract Length Distortion)

学習データに意図的に声道長を変動させる方向の歪みを施すことにより疑似的に学習データ量を増加させる．これは，通常は認識対象の音声に対して施す声道長正規化処理 [18] を，学習音声に施すことにより実現できる．学習時には各学習サンプルに対してランダムな声道長正規化係数による声道長正規化を施し，そこで得られた特徴量を元にして DNN の重みの更新を繰り返すことで学習を行う．

DNN は学習データに存在しないような音声データの識別において性能が大きく劣化することが指摘されている [19]．一方で文献 [4], [19] などの研究から，多層のネットワークを用いた場合には DNN に基づく音響モデルは声道長の変動に頑健になり，声道長正規化がほとんど効かなくなることも指摘されている．このことから少量の学習データしか用いることができない状況では DNN は声道長の変動を十分に学習できないが，サンプルに声道長歪みを施すことにより声道長に起因する変動を吸収できるネットワークが生成されことが期待される．

2.5.2 話速歪み (Speech Rate Distortion)

上記と同様の議論により，十分な学習データがない状況では話速の違いによるスペクトルの変動を DNN が学習できないことが想定される．このため学習データに話速を変動させる方向の歪みを施すことで擬似的に学習データを増加させることを考える．

このため学習音声にランダムに話速変化を施すことにより擬似的に学習データを増加させる．ここでは学習の各反復ごとにランダムに話速変化を施し，得られた特徴量からネットワークの重みを更新する．周波数特性は変動させず話速のみを変化させるために本研究では Praat[20] を用いた音声長伸縮を行った．

2.5.3 ランダム歪み (Random Distortion)

スペクトル領域において周波数方向にランダムな変動を施す．声道長歪みでは全周波数帯域に渡って同じ方向に歪ませるのに対し，本手法では局所的に異なる方向に周波数パワーを変動させる点が異なる．具体的にはまず全ての時間周波数 bin に対して値域 $[-1, 1]$ の一様乱数を発生させる．

$$r(f,t) \sim U(-1,1) \tag{9}$$

その後，得られた乱数を時間周波数の一定領域内で平均をとることにより，変動係数 $\delta(f,t)$ を求める．

$$\delta(f,t) = \frac{\lambda}{(2p+1)(2q+1)} \sum_{f'=f-p}^{f+p} \sum_{t'=t-q}^{t+q} r(f',t') \tag{10}$$

ここで λ は変動の大きさを制御するパラメータである．また p や q は変動を周波数方向もしくは時間方向に平滑化す

表 1 10 時間の学習データにおける単語認識精度

モデル	特徴量	Word Acc.	
		Set1 (%)	Set2 (%)
GMM-ML	MFCC (39)	72.0	71.7
DNN	MFCC (39)	78.0	80.1
	logMFB (75)	78.8	81.2

表 2 270 時間の学習データにおける単語認識精度

モデル	特徴量	Word Acc.	
		Set1 (%)	Set2 (%)
GMM-ML	MFCC (39)	78.8	80.2
GMM-MPE	MFCC (39)	79.9	82.3
DNN	MFCC (39)	84.6	86.6
	logMFB (75)	85.4	87.3

る効果があり，p について大きな値をとるほど周波数方向に見て変動が大局的（多くの周波数で同じ方向に歪む）になる．同様に q について大きな値をとるほど隣接フレーム間で同じ変動が生じるようになる．変動後のスペクトルパワー $\tilde{S}(f,t)$ を変動前のスペクトルパワー $S(f,t)$ から以下のように求める．

$$\tilde{S}(f,t) = S(f + \delta(f,t), t) \tag{11}$$

ただし実際には $\delta(f,t)$ は整数にならないため，隣接する周波数 bin におけるパワーから内挿により値を求めた．

3. 評価実験

3.1 評価データ

この章では DNN に基づく音響モデルの性能を評価する．評価データとして日本語話し言葉コーパス [21] のテストセット 1(2.3 時間, 男性 10 名) 及びテストセット 2(2.4 時間, 男性 5 名，女性 5 名) を用い，単語認識精度 (word accuracy) を計測した．学習データとしては評価データを除く学会講演 967 講演 (270 時間) を全て用いたものと，その中から男女 5 時間ずつの計 10 時間のみを抜粋して用いたものの 2 種類を用いた．言語モデルは学会講演及び模擬講演のうち評価データを含まない計 2671 講演から学習した 3-gram モデルを用いた．単語辞書には頻度上位の 65,000 単語を登録した．音声認識デコーダは我々が開発した WFST に基づく音声認識デコーダを用いた．

3.2 ベースライン: GMM に基づく音響モデルの評価

まず従来から用いられている GMM に基づく音響モデル (GMM-HMM) の性能を調べる．特徴量として 0 次を含む 13 次元の MFCC(Mel-Frequency Cepstrum Coefficients) 特徴量とその差分，2 次差分を合わせた計 39 次元の特徴量に、平均分散正規化を施したものを利用した．なお MFCC を求める際のフィルタバンク数は 24 とした．

10 時間学習データ及び 270 時間学習データのいずれの設定においても，2,734 状態の状態共有トライフォンに基づき対角共分散の GMM-HMM モデルを学習した．10 時間学習データでは混合数を 8[*3]，270 学習データにおいては混合数を 32 とした．いずれのモデルも最尤推定によりパラメータを推定した．270 時間学習データにおいては音素誤り最小化学習 (MPE 学習) により学習された GMM モデルの評価結果についても評価した．

10 時間学習データにおける結果を表 1 の上段に，また 270 時間学習データにおける結果を表 2 の上段に示す．表において GMM-ML と記載されたものが最尤推定により学習された GMM，GMM-MPE と記載されたものが MPE 学習によって学習された GMM による結果を表す．いずれのテストセットにおいても 270 時間の学習データを用いて MPE 学習を施した場合がもっとも性能が高く，テストセット 1 で 79.9%，テストセット 2 で 82.3%であった．

3.3 DNN に基づく音響モデルの評価

3.3.1 特徴量の比較

続いて DNN に基づく音響モデルの性能を調べる．ここでは中間層に 2048 のノードを持つ 7 層のネットワークを用いた．10 時間の学習データを用いた場合には，ミニバッチサイズ [15] は 128 とし，初期化時のパラメータ更新係数の初期値は 0.05，fine-tuning 時の更新係数初期値は 0.01 とした．270 時間の音声データを用いた場合には，ミニバッチサイズは 1024 とし，初期化時のパラメータ更新係数の初期値は 0.05，fine-tuning 時の更新係数初期値は 0.05 とした．いずれの場合も，更新係数のスケジューリングには AdaGrad 法 [22] を用いた[*4]．また各層の初期化は識別的 pre-training（各層ごとに更新 1 回）によって行った．

DNN の評価にあたり，特徴量として前節と同じ 39 次元の MFCC 特徴量を用いた場合と，log Mel Filter Bank (logMFB) 特徴量を用いた場合の 2 種類の特徴量を比較評価した．logMFB 特徴量は 1 フレームあたり 24 次元の log Mel Filter Bank (logMFB) 特徴量と log パワーをあわせた計 25 次元の特徴量をベースとし，それに差分と 2 次差分を加えた計 75 次元の特徴量を，平均分散正規化して用いた．MFCC および logMFB 特徴量のいずれにおいても，前後 5 フレームの特徴量をあわせた計 11 フレームの特徴量を用いた．

10 時間学習データ及び 270 時間学習データにおける結果をそれぞれ表 1 と表 2 の下段に示す．表から DNN によってテストセット 1，テストセット 2 のいずれにおいても単語認識率が大幅に向上していることが分かる．MFCC 特徴量と logMFB 特徴量を比較した場合にはいずれの条件でも logMFB 特徴量が 0.7 ∼ 1.1 ポイント高い性能を示した．

[*3] 16 混合では性能劣化が見られた．過学習したものと思われる．
[*4] AdaGrad は収束が早くなる反面，悪い値に収束する可能性が指摘されている．ただし本研究のデータ量で評価した範囲では十分良いところに収束した．

表 3 10 時間学習データにおける Dropout 法の効果

モデル	Word Acc.	
	Set1 (%)	Set2 (%)
DNN	78.8	81.2
DNN + Dropout	79.4	82.3

表 4 10 時間学習データにおける声道長歪みの効果

モデル	Distortion Ratio $[\alpha_{min}, \alpha_{max}]$	Frame Acc. (%)	Word Acc. (%)
DNN	-	46.6	81.2
DNN	[0.9, 1.1]	48.2	82.7
	[0.85, 1.15]	48.2	82.9
	[0.8, 1.2]	48.0	82.7

表 5 10 時間学習データにおける話速歪みの効果

モデル	Distortion Ratio $[\beta_{min}, \beta_{max}]$	Frame Acc. (%)	Word Acc. (%)
DNN	-	46.6	81.2
DNN	[0.8, 1.2]	47.4	81.3
	[0.7, 1.3]	47.5	81.4
	[0.6, 1.4]	47.6	81.4

表 6 10 時間学習データにおけるランダム歪みの効果

モデル	Distortion Ratio λ	Frame Acc. (%)	Word Acc. (%)
DNN	-	46.6	81.2
DNN	100	46.7	81.3
	200	46.9	81.4
	400	47.0	81.6
	800	46.9	81.2

DNN においては logMFB 特徴量の方が MFCC 特徴量よりも良い性能を示すことが文献 [23] で指摘されており，この結果を日本語でも確認することとなった．270 時間の学習データを用いた場合にはテストセット 1 で 85.4%，テストセット 2 で 87.3%と MPE 学習によって学習された GMM による結果 (GMM-MPE) から，それぞれ 27.4%，28.2%のエラー削減が行われた．また，10 時間の学習データを用いた場合の DNN の認識性能が 270 時間の学習データを用いた場合の GMM-ML の認識性能を上回る性能を示していることも注目に価する．これらの結果から，DNN の高い認識性能が日本語の大語彙連続音声認識による実験においても確認された．

3.3.2 Dropout 法の効果

10 時間の学習データにおいて Dropout 法を適用した場合の効果について検証した．ネットワークの構成は前節までと同様であるが，各層の初期化時のパラメータ更新係数の初期値は 0.1 とし，pre-training において各層ごとに 20 回の反復を行った．fine-tuning 時のパラメータ更新係数の初期値は 0.1 とした．また，Dropout 法において各ノードを 0 にする確率は 50%とした．特徴量は 75 次元の log Mel-Filter Bank 特徴量を用いた．

結果を表 3 に示す．表よりいずれのテストセットにおいても認識性能が向上した．特にテストセット 2 においては 5.9%のエラー削減率 (81.2% → 82.3%) を示し，Dropout 法が認識性能の向上に有効であることが確認された．

3.4 スペクトル伸縮歪みによる擬似学習データ生成の効果

ここでは 2.5 節で提案したスペクトル伸縮歪みの各手法の効果について検証する．いずれのデータにおいても 10 時間の学習データを用いて評価を行った．特徴量は 75 次元の log Mel-Filter Bank 特徴量を用いた．ここでは単語認識率に加えて，各フレームにおける音素状態の識別率（Frame Accuracy）についても示した．テストセット 2 を用いて各種のパラメータを調整した結果を示す．

まず声道長歪みの効果を示す．学習時には各反復ごとに声道長正規化係数を $[\alpha_{min}, \alpha_{max}]$ の範囲において 0.05 刻みでランダムに変動させることにより，擬似的に声道長を歪ませた音声を生成した．変動幅を表す係数 $\{\alpha_{min}, \alpha_{max}\}$ の組として $\{0.8, 1.2\}$，$\{0.85, 1.15\}$，$\{0.9, 1.1\}$ の 3 パターンを比較した．結果を表 4 に示す．表より，声道長歪みの変動幅はどの設定でも性能が向上したが，変動幅を $\{0.8, 1.2\}$ と大きくし過ぎた場合には精度劣化が見られた．変動幅が $\{0.85, 1.15\}$ の時に性能向上幅が最大になり，フレーム認識率で 1.6 ポイント，単語認識率は 1.7 ポイント精度が向上した．単語認識率においては 9.0%のエラー削減に相当する．

続いて話速歪みの効果を示す．学習時には各反復ごとに話速を $[\beta_{min}, \beta_{max}]$ の範囲において 0.1 刻みでランダムに変動させることにより，擬似的に声道長を歪ませた音声を生成した．変動幅を表す係数 $\{\beta_{min}, \beta_{max}\}$ の組として $\{0.6, 1.4\}$，$\{0.7, 1.3\}$，$\{0.8, 1.2\}$，の 3 パターンを比較した．この評価結果を 5 に示す．ここでもフレーム認識率，単語認識率ともに精度向上を確認した．ただし精度向上幅は声道長歪みと比較した場合には小さく，フレーム認識率で 1 ポイント，単語認識率ではたかだか 0.2 ポイントの改善に留まった．

ランダム歪みの効果を示す．ここではパラメータとして $p = 128, q = 100$ と固定し，λ の値を 100, 200, 400, 800 と変化させた．学習時には学習データに毎回異なるランダム歪みを加えた．この評価結果を 6 に示す．$\lambda = 400$ の時にフレーム認識率，単語認識率ともに最大値を示し，いずれも 0.4 ポイントの性能向上が得られた．

最後にこれらのスペクトル伸縮歪みを全て組み合わせた場合の精度を表に示す．ここでは $\{\alpha_{min}, \alpha_{max}\} = \{0.85, 1.15\}, \{\beta_{min}, \beta_{max}\} = \{0.6, 1.4\}, \lambda = 400$ とした．またテストセット 1 における単語認識率も示す．最終的に，全てのスペクトル伸縮歪みを組み合わせることでテストセット 1 において 79.7%，テストセット 2 において 83.1%の単語認識率（それぞれ 4.2%と 10.1%の誤り削減に

表 7　スペクトル伸縮歪みの効果

モデル	特徴量	学習データ (hour)	声道長歪み	話速歪み	ランダム歪み	Word Acc.	
						Set1 (%)	Set2 (%)
DNN	logMFB (75)	10				78.8	81.2
DNN	logMFB (75)	10	√			79.4	82.9
	logMFB (75)	10		√		78.9	81.4
	logMFB (75)	10			√	79.2	81.6
	logMFB (75)	10	√	√	√	79.7	83.1

相当する）が得られ，スペクトル伸縮歪みにより少量のデータセットにおいて性能が大きく向上することが確認された．

3.5 まとめ

本研究では日本語のテストセットを用いた大語彙連続音声認識において，Deep Neural Network に基づく音響モデルの性能を評価した結果について述べた．270 時間の学習データを用いた評価においては，音素誤り最小化 (MPE) 学習された GMM-HMM の CSJ テストセット 1 とテストセット 2 における単語認識率が 79.9%と 82.3%であったのに対し，DNN に基づく音響モデルは 85.4%と 87.3%の認識率を示し，DNN の認識性能の高さを日本語においても確認した．また，少量の学習データしか利用できない状況で擬似的に音声データを増加させる手法について新たに検討を行い，10 時間の学習データを用いた評価において最大 10.1%の性能向上を確認した．

参考文献

[1] Seide, F., Li, G. and Yu, D.: Conversational speech transcription using context-dependent deep neural networks, *In Proc. Interspeech*, pp. 437–440 (2011).
[2] Graves, A., rahman Mohamed, A. and Hinton, G.: Speech Recognition with Deep Recurrent Neural Networks, *In Proc. ICASSP*, IEEE, pp. 6645–6649 (2013).
[3] Deng, L., Abdel-Hamid, O. and Yu, D.: A Deep Convolutional Neural Network using Heterogeneous Pooling for Trading Acoustic Invariance with Phonetic Confusion, *In Proc. ICASSP*, IEEE, pp. 6669–6673 (2013).
[4] Seide, F., Li, G., Chen, X. and Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription, *In Proc. ASRU*, IEEE, pp. 24–29 (2011).
[5] Seltzer, M., Yu, D. and Wang, Y.: An Investigation of Deep Neural Networks for Noise Robust Speech Recognition, *In Proc. ICASSP*, IEEE, pp. 7398–7402 (2013).
[6] Fujii, Y., Yamamoto, K. and Nakagawa, S.: Deep-hidden Conditional Neural Fields for Continuous Phoneme Speech Recognition, *In Proc. IWSML* (2012).
[7] Matsuda, S., Lu, X. and Kashioka, H.: Automatic Localization of a Language-independent Sub-network on Deep Neural Networks Trained by Multi-Lingual Speech, *In Proc. ICASSP*, IEEE, pp. 7359–7362 (2013).
[8] 西野大輔，篠田浩一，古井貞熙：ディープラーニングを用いた日本語大語彙話し言葉音声認識，日本音響学会 2012 年秋季研究発表会講演論文集 (2012).
[9] 中村篤，大庭隆伸，渡部晋治他：音声認識システム SOLON の日本語話し言葉コーパスによる評価 (2006 年版)，情報処理学会研究報告，Vol. 2006, No. 136, pp. 251–256 (2006).
[10] Swietojanski, P., Ghoshal, A. and Renals, S.: Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR, *In Proc. SLT*, IEEE, pp. 246–251 (2012).
[11] Thomas, S., Seltzer, M. L., Church, K. and Hermansky, H.: Deep Neural Network eatures and Semi-supervised Training for Low Resource Speech Recognition, *In Proc. ICASSP*, pp. 6704–6708 (2013).
[12] Dahl, G. E., Sainath, T. N. and Hinton, G. E.: Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout, *In Proc. ICASSP*, IEEE, pp. 8609–8613 (2013).
[13] Rumelhart, D. E., Hintont, G. E. and Williams, R. J.: Learning representations by back-propagating errors, *Nature*, Vol. 323, No. 6088, pp. 533–536 (1986).
[14] Hinton, G. E., Osindero, S. and Teh, Y.-W.: A fast learning algorithm for deep belief nets, *Neural computation*, Vol. 18, No. 7, pp. 1527–1554 (2006).
[15] Hinton, G.: A practical guide to training restricted Boltzmann machines, *Momentum*, Vol. 9, p. 1 (2010).
[16] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems 25*, pp. 1106–1114 (2012).
[17] Simard, P. Y., Steinkraus, D. and Platt, J. C.: Best practices for convolutional neural networks applied to visual document analysis, *Proceedings of the seventh international conference on document analysis and recognition*, Vol. 2, pp. 958–962 (2003).
[18] Zhan, P. et al.: Vocal Tract Length Normalization for LVCSR, Technical report, Technical Report CMU-LTI-97-150, Carnegie Mellon Univ (1997).
[19] Yu, D., Seltzer, M. L., Li, J., Huang, J.-T. and Seide, F.: Feature Learning in Deep Neural Networks-A Study on Speech Recognition Tasks, *arXiv preprint arXiv:1301.3605* (2013).
[20] Boersma, P.: Praat, a system for doing phonetics by computer, *Glot international*, Vol. 5, No. 9/10, pp. 341–345 (2002).
[21] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *In Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12 (2003).
[22] Duchi, J., Hazan, E. and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159 (2010).
[23] Mohamed, A., Hinton, G. and Penn, G.: Understanding how deep belief networks perform acoustic modelling, *In Proc. ICASSP*, IEEE, pp. 4273–4276 (2012).