Vol.2013-SLP-99 No.4
2013/12/19

音節単位DNN-HMMによる音声認識の検討

関　博史[1,a)]　中川 聖一[1,b)]

概要：近年，音声認識にディープニューラルネットワークを用いることで，従来手法である GMM-HMM と比較し精度が向上するという結果が多数報告されている．本研究では，3 つの年齢層（成人・子供・老人）と性別（男性・女性）に依存した計 6 クラスの学習データベースを使用し，それぞれ音節単位 DNN-HMM を学習した．その結果，従来手法である GMM-HMM と比較して 4 クラスで精度の向上が見られた．そして 6 つのクラスを 1 つのネットワークで学習することにより，5 クラスで精度の向上が見られた．また，直前の音素を考慮した左コンテキスト依存の音節単位 DNN-HMM についても検討した．左コンテキスト依存の音節単位 DNN-HMM は学習すべきパラメータ数が多いため，学習には多くの時間が必要となる．そこで，状態を「結び」にして学習する方法と学習を高速化するために Rectified Linear Unit を導入した結果も報告する．

キーワード：ディープニューラルネットワーク，音節単位，HMM，不特定話者音声認識

Consideration on Syllable-Unit based Deep Neural Network for Speech Recognition

Hiroshi Seki[1,a)]　Seiichi Nakagawa[1,b)]

Abstract: Recently, Deep Neural Networks have been applied to speech recognition and outperformed the conventional GMM based methods. In this paper, we provide 6 class training sets which depend on gender(male, female) and age(elder, adult, child). We trained each syllable-unit based DNN and it outperformed the baseline GMM-HMM for 4 classes. We also trained one DNN using all 6 class training sets and it outperformed the baseline GMM-HMM for 5 classes. In addition, we considered a left context dependent syllable-unit based DNN-HMM. Modeling context dependent phonemes increases parameters to learn, and needs a lot of time. So we also report results about tied state syllable modeling and use of rectified linear unit to train parameters quickly.

Keywords: Deep Neural Network, syllable unit, HMM, speaker independent speech recognition

1. はじめに

ディープニューラルネットワーク (Deep Neural Network: DNN) を音声認識に用いる研究が活発に行われ [1][2], DNN-HMM を用いた日本語音声認識でも従来の GMM-HMM を超える精度が得られたという研究が多数報告されている [3][4]．ニュラルネットワークの特徴は入力フレーム数を 11 フレーム長程度にしても頑健に事後確率が学習される点であり，十分なコンテキストを入力できることから，コンテキストに独立なモデルで高精度な認識率を達成できる．これらのほとんどの研究では，GMM-HMM と DNN-HMM の比較を行っているが，DNN-HMM の出力ラベル単位としてモノフォンとトライフォンの比較を行った研究は少ない．我々は従来から音節単位 GMM-HMM の研究を行っており [5][6]，今回，コンテキスト独立音節単位の GMM-HMM と DNN-HMM，左コンテキスト依存音節単位 DNN-HMM との比較結果を報告する．関連研究として文献 [7] や [8] が

1 豊橋技術科学大学
Toyohashi University of Technology
a) seki@slp.cs.tut.ac.jp
b) nakagawa@slp.cs.tut.ac.jp

ある．文献 [7] では中国語を音節単位 DNN-HMM で学習してコンテキスト独立とコンテキスト依存モデルを比較し，後者のほうが良い性能を得ている．文献 [8] では中国語をトライフォン単位で学習し，fMPE 基準で学習されたトライフォン GMM-HMM よりよい性能を得ている．また，層数や学習データ量と精度の関係について報告している．

本稿では 3 つの年齢層（成人・子供・老人）と性別（男性・女性）ごとに計 6 つのデータベースを用意し，それぞれに対してコンテキスト独立の音節単位 DNN-HMM を学習する．そして音節と直前の音素を考慮した左コンテキスト依存 DNN-HMM も学習し，比較検討を行う．

まず 2 節では DNN-HMM の学習方法について述べる．今回は，制限付きボルツマンマシンによる事前学習を行う方法と，事前学習を行わないで Rectified Linear Unit で学習する方法を比較した．3 節では GMM-HMM による不特定話者認識手法について述べる．4 節では GMM-HMM と DNN-HMM による実験とその結果，5 節ではまとめと今後の課題を述べる．

2. DNN-HMM による音声認識

2.1 ディープニューラルネットワークの概要

ディープニューラルネットワークは多数の隠れ層を持つニューラルネットワークである．従来のニューラルネットワークは，層数が増えるとバックプロパゲーションでパラメータを学習する際入力に近い層まで勾配が伝わらず，うまく学習することが困難であった．しかし，Hinton らによって提案された事前学習アルゴリズムにより適切な初期値を得ることができ [9][10]，このネットワークに対してバックプロパゲーションを行うことで，多数の隠れ層をもつニューラルネットワークを学習することができる [11].

2.2 事前学習

事前学習ではまず第 1 層と第 2 層を Restricted Boltzmann Machine(RBM) とみなし，接続されたノードの重みとバイアスを学習する．第 1 層では Gaussian-Bernoulli RBM，それ以降では RBM が使用される．$\mathbf{v}$ を可視ノード，$\mathbf{h}$ を隠れノードとすると，エネルギー関数はそれぞれ次式で与えられる．

$$E(\mathbf{v},\mathbf{h}) = \sum_{i\in V} \frac{(v_i - a_i)^2}{2} - \sum_{j\in H} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \quad (1)$$

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i\in V} a_i v_i - \sum_{j\in H} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \quad (2)$$

Gaussian-Bernoulli RBM の場合，可視ノードおよび隠れノードの条件付き確率は次式で与えられる．

$$p(v_i = v|\mathbf{h}) = N(v|a_i + \sum_j h_j w_{ij}, 1) \quad (3)$$

$$p(h_j = 1|\mathbf{v}) = sigmoid(b_j + \sum_i v_i w_{ij}) \quad (4)$$

RBM の場合，可視ノードおよび隠れノードの条件付き確率は次式で与えられる．

$$p(v_i = 1|\mathbf{h}) = sigmoid(a_i + \sum_j h_j w_{ij}) \quad (5)$$

$$p(h_j = 1|\mathbf{v}) = sigmoid(b_j + \sum_i v_i w_{ij}) \quad (6)$$

対数尤度は

$$\ln p(\mathbf{v}|\theta) = \ln \frac{1}{Z} \sum_{j\in H} \mathrm{e}^{-E(\mathbf{v},\mathbf{h})} \quad (7)$$

$$= \ln \sum_{j\in H} \mathrm{e}^{-E(\mathbf{v},\mathbf{h})} - \ln \sum_{i\in V, j\in H} \mathrm{e}^{-E(\mathbf{v},\mathbf{h})} \quad (8)$$

となり，パラメータ θ で偏微分すると次式が得られる．

$$-\frac{\partial p(\mathbf{v}|\theta)}{\partial \theta} = < \frac{\partial E}{\partial \theta} >_{data} - < \frac{\partial E}{\partial \theta} >_{model} \quad (9)$$

重み w_{ij} で偏微分すると次式が得られる．

$$-\frac{\partial p(\mathbf{v}|\theta)}{\partial w_{ij}} = < v_i h_j >_{data} - < v_i h_j >_{model} \quad (10)$$

$< v_i h_j >_{data}$ は入力データ及び $p(h_j = 1|\mathbf{v})$ を用いて容易に計算することができる．$< v_i h_j >_{model}$ は直接計算するのが困難なため，ギブスサンプリングを用いて計算する [12]．他のパラメータに関しても同様の計算を行う．

パラメータの学習が終わると，第 2 層と第 3 層を RBM とみなし同様に学習を行う．この時，入力には直前に求めた $p(h = 1|\mathbf{v})$ を用いる．このようにネットワークを下位層から順に RBM とみなし，教師なし学習を行っていく．

2.3 教師有り学習

DNN の教師あり学習としてバックプロパゲーションを用いる．与えられた学習データ $\mathbf{o}$ と正解ラベル $\mathbf{t}$ に対して前向きにスコアを計算し出力層での損失を求め，各層のパラメータの勾配を後ろ向きに計算していく．l 層のノード j の値 o_j は次式のように計算される．

$$u_j = \sum_i w_{ji} o_i \ , \ o_j = \frac{1}{1 + \mathrm{e}^{-u_j}} \quad (11)$$

出力層は softmax 関数を用いて次式のように計算される．

$$o_j = \frac{\exp(u_j)}{\sum_i \exp(u_i)} \quad (12)$$

損失関数にはクロスエントロピーを用いている．

$$E = \sum_j \log o_j \quad (13)$$

ここで，

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial u_j} \frac{\partial u_j}{\partial w_{ji}}, \frac{\partial u_j}{\partial w_{ji}} = o_i \quad (14)$$

$$\Delta w_{ji} = \eta r_j o_j \tag{15}$$

r_j は，出力層の場合と隠れ層の場合でそれぞれ次式で求められる．δ はクロネッカーのデルタである．

$$r_j = \delta_{t_j,j} - o_j \tag{16}$$
$$r_j = o_j(1-o_j)\sum_k r_k w_{kj} \tag{17}$$

2.4 Rectified Linear Unit

Restricted Boltzmann machines(RBM) による事前学習に Rectified Linear Unit を使用することで精度が向上することや [13]、事前学習を行わなくても同等の精度が得られることが報告されている [14][15]．左コンテキストを出力ラベルとして用いた場合，出力層のユニット数は 3712 と多く，隠れ層も増やす必要があるため，多くの計算時間を要する．そこで，学習時間の短縮のため活性化関数として Rectified Linear Unit($o_j = \max(0, u_j)$) を使用し，事前学習は行わず教師有り学習 (fine-tuning) のみを行った．このとき，式 (16)(17) はそれぞれ以下のように変更される．

$$r_j = \delta_{t_j,j} - o_j \tag{18}$$
$$r_j = \max(0, \sum_k r_k w_{kj}) \tag{19}$$

2.5 コンテキスト依存音節単位 HMM のための状態の「結び」

出力ラベルとしてコンテキスト独立音節と左コンテキスト依存音節を用いる．HMM は**図 1** のように 4 状態出力分布を持ち，出力ラベル数はコンテキスト独立の場合，音節 116 種× 4 状態の 464 種，左コンテキスト依存の場合，左コンテキスト 8 種（a,i,u,e,o,N,qs,SIL）×音節 116 種× 4 状態の 3712 種である．また，左コンテキスト依存のうち後半 2 状態をコンテキスト独立にした場合と後半 3 状態をコンテキスト独立にした場合も試した．後半 2 状態を結びにした場合，左コンテキスト依存の音節「a-ka」は a-ka[1],a-ka[2],TC_ka[3],TC_ka[4] の 4 状態で構成され，後半 2 状態はコンテキスト独立音節「ka」である．後半 2 状態を結びにした場合と 3 状態を結びにした場合で，出力ラベル数はそれぞれ 2088，1276 である．

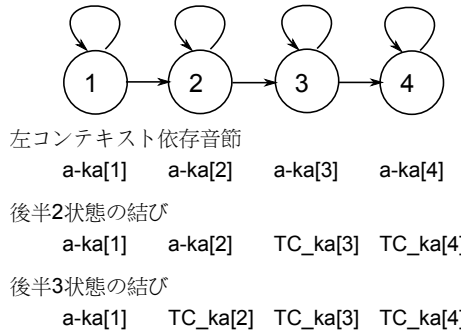

図 1 音節単位 HMM の構造
Fig. 1 Structure of syllable-unit based HMM

3. 不特定話者音声認識

3.1 データベース

GMM-HMM で音響モデルを構築する際に，全学習話者の音声データを用いるのではなく，認識対象話者に類似した学習話者の音声データのみを用いると，話者性の問題に対処でき，認識精度が向上することが知られている [16]．そこで，DNN-HMM による不特定話者音声認識についてもこれを検討をする．年齢・性別非依存の不特定話者音声認識システムを評価するため，3 つの年齢層（成人・子供・老人）と性別（男性・女性）ごとにデータベースを用意し，それぞれ 6 クラスおよびこれらを一つにまとめた 1 クラスに対して認識実験を行った．学習およびテストデータは性別（男女），年齢層（成人，子供，老人）の 6 クラスに分類されてある。各クラスのデータ数を**表 1** に示す．ここで，A-M は成人男性，A-F は成人女性，C-M は子供男性，C-F は子供女性，E-M は老人男性，E-F は老人女性を表している。成人用のデータには ASJ+JNAS コーパス [17] を用いる．各話者の新聞記事読み上げ文 100 文と音素バランス文 50 文から構成されており，話者は 18 歳から 59 歳までの男性 184 名，女性 187 名である．子供用のデータには，CIAIR-VCV コーパスを用いる [18]．大きく 3 つのコンテンツから構成されており，カタカナで表現された 40 単語と 21 種類の数字，童話”マッチ売りの少女”の読み上げ 30 文である．話者は 6 歳から 12 歳の男性 145 名，女性 143 名である．老人用のデータには日本語新聞読み上げコーパス JNAS の老人用である S-JNAS コーパスを用いる [19]．新聞記事読み上げ文 200 文と音素バランス文 50 文から構成されており，話者は 60 歳から 90 歳までの男性 151 名，女性 150 である．テストデータは各クラスとも 100 文である．子供用コーパスは主に童話の読み上げ文から構成されているが，実験で用いている言語モデルは新聞記事から学習している．そのため，子供クラスのテストデータの未知語率は 14%である．成人クラス，老人クラスの未知語率はそれぞれ 0.5%，2.1%である．

3.2 特徴パラメータ

GMM-HMM の学習に用いた特徴量は 12 次元 MFCC，Δ，Δ Δ，および Δ パワー，Δ Δ パワーで計 38 次元である．ここで，各 MFCC は発話ごとに CMN を行っている．GMM-HMM は，音節単位の left-to-right 型で，各 HMM は 4 状態出力分布を持ち，各出力分布は 32 混合の対角共分散正規分布からなる．また，1 クラスモデルは 128 混合を用いた．無音とショートポースを合わせて左コンテキスト依存 928 種類の音節単位 HMM を用いた．コンテキスト独立 HMM を学習した後，コンテキスト依存 HMM を MAP

表 1 各クラスの学習，テストデータ量 (#sentence/#word)
Table 1 Training Data and Test Data

	A-M	A-F	C-M	C-F	E-M	E-F
学習データ量	20337/0	25056/0	3393/7538	3910/7744	24081/0	24061/0
テストデータ量	100/0	100/0	100/0	100/0	100/0	100/0

推定により学習した．子供用コーパスの音声データには，すべての音節が含まれていないため，モデル学習の初期パラメータの推定に成人女性用コーパスを用いて対応した．DNN-HMM の学習には，特徴量としてフレーム周期 10ms ごと 12 次元 MFCC，Δ，ΔΔ，およびパワー，Δパワー，ΔΔパワー計 39 次元を用いた．入力フレーム数は 11 を標準としている．学習データのアライメントはベースラインとなる GMM-HMM でアライメントをとった．

3.3 言語モデル

言語モデルの学習には，毎日新聞の記事のうち 1991 年 1 月から 1994 年 9 月までの 45 ヶ月分および 1995 年 1 月から 1997 年 6 月までの 30 ヶ月分計 75 ヶ月分を使用した．語彙として学習データの中で出現頻度が高い上位 20,000 語使用し，tri-gram 言語モデルを学習した．

3.4 デコーダ

GMM-HMM による大語彙連続音声認識のデコーダには，日本語連続音声認識システム SPOJUS++ (SPOken Japanese Understanding System)[20] を，DNN-HMM において認識実験を行う際には，WFST 版 SPOJUS を用いた．

4. 評価実験

4.1 GMM-HMM の評価

左コンテキスト依存音節単位 GMM-HMM による認識結果を**表 2** のモデル GMM(CD) の欄に示す [5][6]．6 クラスそれぞれで学習した場合と全てのデータをまとめた 1 クラスで学習した場合を比較すると，老人クラスの単語認識精度はほぼ同等の結果だが，それ以外の 4 クラスは個別の音響モデルを使用したほうが 1[%]〜4.7[%] ほど単語認識精度が良い．平均単語認識精度もクラス別（6 クラス）の音響モデルの方が全クラス（1 クラス）の音響モデルより精度は良い．

4.2 コンテキスト独立音節 DNN-HMM の評価

コンテキスト独立音節単位 DNN-HMM による認識結果を表 2 の 6 クラス DNN(CI) の欄に示す．認識に用いる音響モデルは，クラス既知としてそれぞれのクラスに対応するモデルを用いた．ネットワークは 8 層，隠れ層のユニット数は 1024 とした．コンテキスト独立 DNN-HMM(CI) で認識を行った場合，左コンテキスト依存 GMM-HMM と比べて成人クラスと子供クラスは最大 0.7[%]，最小では

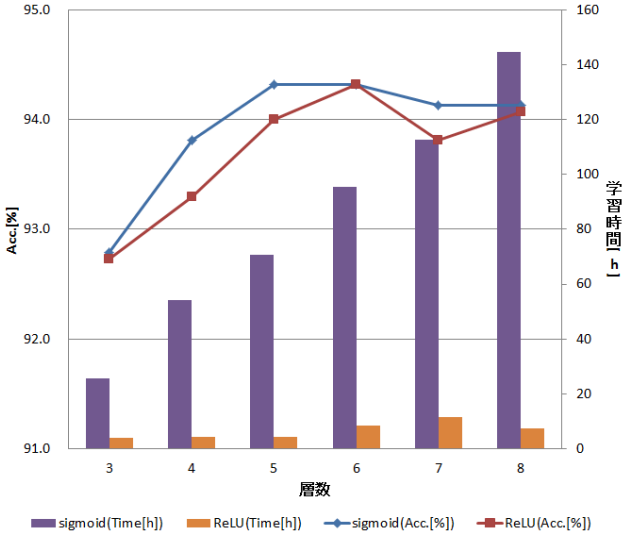

図 2 シグモイド関数と Rectified linear unit による学習時間と Acc.[%] の違い (成人男性クラス，隠れ層 : 1024 ユニット，入力フレーム数 : 11，出力ラベル : コンテキスト独立)

Fig. 2 The training time and accuracy using activation function as sigmoid unit and rectified linear unit(class:adult male, hidden unit:1024, input frame:11, output label:context independent)

0.2[%] 単語認識精度が向上したが，老人クラスだけは単語認識精度が低下した．特に老人女性は GMM-HMM の結果を 2[%] 下回った．このため，6 クラスを平均した単語認識率は GMM-HMM とほぼ同じ結果となった．

4.3 Rectified Linear Unit による高速化

左コンテキスト依存 DNN-HMM を学習するため隠れ層のユニット数を増やす必要があるが，事前学習に時間がかかり大規模なネットワークを学習するのは難しい．そこで，活性化関数として Rectified Linear Unit を使用し，事前学習を行わずバックプロパゲーションによりネットワークを学習した．学習には成人男性クラスのデータを用いた．シグモイド関数を使用した事前学習ありの DNN と Rectified Linear Unit を使用した事前学習なしの DNN の単語認識精度と学習時間を**表 3** と**図 2** に示す．出力ラベルはコンテキスト独立としている．シグモイド関数を使用した時と Rectified Linear Unit を使用した時では動作環境が違っているため単純な比較はできないが（CPU : Core i7-960, GPU : Tesla C2075 × 2，メモリ : 前者 12GB 後者 28GB），それでも Rectified Linear Unit を使用すると，従来の事前学習ありの DNN と比べ大幅に学習時間が削減され，ほぼ同等の単語認識精度を出すことができた．また，隠れ層を

表 2 クラス既知での単語認識精度 (隠れ層：1024 ユニット，入力フレーム数：11)
Table 2 Word recognition accuracy[%] with corresponding to GMM/DNN (hidden unit:1024, input frame:11)

クラス	モデル	A-M	A-F	C-M	C-F	E-M	E-F	ave.
6 クラス	GMM(CD)	93.5	94.6	74.7	78.2	89.4	93.4	87.3
	DNN(CI,layer=8)	94.1	95.0	75.4	78.4	88.8	91.4	87.2
	DNN(CD,layer=8)	89.7	89.9	76.6	79.9	87.0	90.0	85.5
1 クラス	GMM(CD)	91.3	93.6	68.0	74.7	89.5	93.7	85.1
	DNN(CI,layer=5)	93.1	94.5	75.2	77.3	90.3	92.2	87.1
	DNN(CI,layer=6)	93.9	94.9	76.1	78.1	90.3	92.2	87.6
	DNN(CI,layer=7)	94.0	94.9	77.1	78.7	90.4	91.9	87.8

表 3 シグモイド関数を用いた事前学習あり DNN と ReLU を用いた事前学習なし DNN の比較 (成人男性クラス，隠れ層：1024 ユニット，入力フレーム数：11，出力ラベル：コンテキスト独立)
Table 3 Comparison of pre-training DBN and rectifier network (class:adult male, hidden unit:1024, input frame:11, output label:context independent)

層数	3	4	5	6	7	8
シグモイド関数	92.8	93.8	94.3	94.3	94.1	94.1
ReLU	92.7	93.3	94.0	94.3	93.8	93.9

表 4 Rectified Linear Unit を使用した時の単語認識精度 (成人男性クラス，隠れ層：2048 ユニット，入力フレーム数：11，出力ラベル：左コンテキスト依存)
Table 4 Word recognition accuracy using Rectified Linear Unit(class:adult male, hidden unit:2048, input frame:11, outputlabel:left context dependent

層数	4	5	6	7
Acc.[%]	91.1	90.5	91.1	90.0

4096 にまで増やしても Rectified Linear Unit はシグモイド関数を使用した場合と同等もしくはそれ以上の認識率を得ることができた．従って，隠れ層のユニット数を多く必要とする左コンテキスト依存 DNN の学習には Rectified Linear Unit を使用する．

4.4 左コンテキスト依存音節 DNN-HMM の導入

左コンテキスト依存音節 DNN-HMM による認識結果を表 2 の 6 クラス DNN(CD,layer=8) に示す．コンテキスト独立で事前学習したネットワークに出力層を付け加え，バックプロパゲーションを行った．出力層のユニット数が 3712 なのに対し隠れ層のユニット数は 1024 とバランスが悪く，単語認識精度も GMM-HMM より下回った．隠れ層を 2048 として Rectified Linear Unit を用いて学習した時の認識結果を**表 4** に示す．層数を増やしても著しい改善は見られず，表 3 や表 4 から層数はラベルの種類には依存せず，学習量と大きく関わりのあることがわかる．4.5 節で 6 クラスのデータをまとめて学習を行うが，そこでは層数を増やすことで単語認識精度も向上する．

左コンテキスト依存音節単位では，直前の数フレームをネットワークに入力しなくても，左コンテキストに依存したネットワークが学習できると考えた．その時の単語認識精度を**表 5** に示す．入力フレーム数が 5，7，11 のときほぼ同じ精度を示したが，入力フレーム数を 1 とすると悪くなった．この入力条件は左コンテキスト依存 GMM-HMM と同じであるにもかかわらず（厳密に言えば，GMM への入力にはパワーを含んでいない），単語認識精度は GMM-HMM より精度が悪いが，原因は不明である．次に，出力層のユニット数を減らし，学習すべき状態数やパラメータ数を削減する．そのために，左コンテキスト依存音節モデルで後ろ 2 状態を共有する場合 (TC_2state) と後ろ 3 状態を共有する場合 (TC_3state) で実験を行った．この時の単語認識精度を**表 6** に示す．TC_2state は出力ユニット数が 2088，隠れ層のユニット数が 4096 であり，TC_3state は出力ユニット数が 1276，隠れ層のユニット数は 2048 である．後ろ 3 状態を共有した場合，単語認識精度は 92.7[%] となり，左コンテキスト依存 DNN-HMM に関して行った実験の中で最もよい精度となったが，コンテキスト独立 DNN-HMM を上回ることはできなかった．これは，学習データ量とネットワークのパラメータ数の関係に起因していると考えられる．

4.5 1 クラス DNN-HMM

6 クラスすべてのデータをまとめ，ひとつのネットワークでも学習を行った．その結果を表 2 の 1 クラスの欄に示す．特徴パラメータの正規化はクラスごとに行った．GMM-HMM の場合と異なり，全学習データを用いてネットワークを学習することで，クラスごとに学習した DNN-HMM(CI) と同等以上の単語認識精度が得られた．特に，子供男性クラスをみると子供男性クラスのみを用いて学習した 6 クラス DNN-HMM(CI) と比べ単語認識精度は 1.7[%] 改善しており，特定話者集団に依存したデータの質を学習データ量が上回る結果となった．コンテキスト依存 GMM-HMM では，クラス別 (6 クラス) モデルと比べて，全クラス (1 クラ

表 5 入力フレームの変化と単語認識精度の比較 (成人男性クラス，CI : コンテキスト独立 (隠れ層 : 1024 ユニット，層数 : 5) CD : 左コンテキスト依存 (隠れ層 : 4096 ユニット，層数 : 5))

Table 5 Comparison of different input frames and word recognition accuracy(class:adult male, CI:context independent(hidden unit:1024,layer=5), CD:left context dependent(hidden unit:4096,layer=5))

モデル	入力フレーム	Acc.[%]
CI	1	90.7
	11	94.0
CD	1	88.5
	3	89.8
	5	90.6
	7	90.8
	11	90.7

表 6 音節の状態共有に対する単語認識精度の比較 (成人男性クラス，層数 : 5，入力フレーム数 : 11)

Table 6 Comparison of different tied-state model(class:adult male, layer=5, input frame:11)

モデル	隠れ層 ユニット数	出力ユニット	Acc.[%]
TC_2state	4096	2088	91.3
TC_3state	2048	1276	92.7

ス) モデルにすると大幅に認識精度が低下するが，コンテキスト独立 DNN-HMM では，逆に全クラス（1 クラス）モデルの方が認識精度が良くなっている．これよりパターンの変動が大きい音声ほど大量のデータがあれば DNN-HMM の威力が発揮できると考えられる．

5. まとめ

本稿では，性別や年齢に依存したクラス別の 6 つのクラスとそれらをひとつにまとめた 1 クラス DNN-HMM を学習し，従来手法である GMM-HMM との比較を行った．クラス依存で DNN-HMM を学習した場合，平均単語認識率は 87.2[%] となり GMM-HMM（87.3[%]) と比べ同精度にとどまった．1 クラス DNN-HMM を学習した場合，平均単語認識精度は 87.8[%] となり，6 クラス DNN-HMM と比較して 0.5[%] の改善を得ることができ，クラス別（6 クラス）GMM-HMM を上回った．老人クラスはクラス依存で DNN-HMM の学習を行った場合，男性・女性ともに GMM-HMM の単語認識精度を下回った．しかし，1 クラス DNN-HMM を学習することで老人女性クラス以外の全クラスでクラス別（6 クラス）の GMM-HMM を上回った．なお，認識精度が向上すると言われている Dropout 法 [21] を実装し，成人男性クラスに適用したが，精度の改善は見られなかった．

今後は，今回の比較実験についての詳しい調査や 1 クラス DNN-HMM により学習される普遍的な情報と適応学習の組み合わせ，パターン変動の大きい CSJ での評価などを検討したい．

参考文献

[1] 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*, pp. 82–97 (2012).

[2] 中川聖一：再訪：ニューラルネットワークによる音声処理，電子情報通信学会信学技報 SP2013-59，pp. 37–44 (2013).

[3] 神田直之， 武田龍，大淵康成：Deep neural network に基づく日本語音声認識の基礎評価，研究報告音声言語情報処理（SLP)，Vol. 2013-SLP-97, No. 8, pp. 1–6 (2013).

[4] 三村正人，河原達也：CSJ を用いた日本語講演音声認識への DNN-HMM の適用と話者適応の検討，Vol. 2013-SLP-97, No. 9, pp. 1–6 (2013).

[5] 榎並大介：学習データのソフトクラスタリング手法に基づく複数音響モデルによる不特定話者音声認識，修士論文，豊橋技術科学大学 (2012).

[6] 榎並大介，山本一公，中川聖一：性別・年齢非依存の音声認識における話者のソフトクラスタリング手法の検討，日本音響学会春季講演論文集，No. 1-P-27 (2012).

[7] X.Li, C.Hong, Y.Yang and X.Wu: Deep neural networks for syllable based acoustic modeling in Chinese speech recognition, *APSIPA* (2013).

[8] J.Niu, L.Xie, L.Jia and N.Hu: Context-dependent deep neural networks for commercial mandarin speech recognition application, *APSIPA* (2013).

[9] G.E.Hinton: Training products of experts by minimizing contrastive divergence, *Neural Computation*, Vol. 14, pp. 1771–1800 (2002).

[10] G.E.Hinton: A practical guide to training restricted boltzmann machines, Technical Report Technical Report UTML TR 2010-003, Univ. of Toronto (2010).

[11] D.E.Rumelhart, G.E.Hinton and R.J.Williams: Learning representations by back-propagating errors, *Nature*, Vol. 323, No. 6088, pp. 533–536 (1986).

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

[13] V.Nair and G.E.Hinton: Rectified linear unit improve restricted boltzmann machines, *ICML*, pp. 807–814 (2010).

[14] Lãszlo and T.Grosz: A comparison of deep neural network training method for large vocabulary speech recognition, *TSD*, No. LNAI8082, pp. 36–43 (2013).

[15] L.Toth: Phone recognition with deep sparse rectifier neural networks, *Proc.ICASSP*, pp. 6985–6989 (2013).

[16] M.Padmanabhan, L.R.Bahl, D.Nahamoo and M.A.Picheny: Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems, *Speech and audio processing and IEEE Trans.*, Vol. 27, pp. 71–77 (1998).

[17] K.Itou, M.Yamamoto, K.Takeda, T.Takezawa, T.Matsuoka, T.Kobayashi, K.Shikano and S.Itahasi: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, *The Journal of the acoustical society of Japan(E)*, Vol. 20, pp. 199–206 (1999).

[18] : CIAIR 子供の声データベース (CIAIR-VCV), http://research.nii.ac.jp/src/CIAIR-VCV.html.

[19] : 新聞記事読み上げ高齢者音声コーパス (S-JNAS), http://research.nii.ac.jp/src/S-JNAS.html.

[20] Y.Fujii, K.Yamamoto and S.Nakagawa: Large vocabulary speech recognition system:SPOJUS++, *MUSP*, pp. 110–128 (2011).

[21] G.E.Hinton, N.Srivastava, A.Krizhevsky, I.Sutskever and R.Salakhoutdinov: Improving neural networks by preventing co-adaptation of feature detectors, *The Computing Research Repository*, Vol. abs/1207.0580 (2012).

正誤表

音節単位 DNN-HMM による音声認識の検討

平成 25 年 11 月 29 日　　関　博史

ページ等	誤	正
	(削除)	謝辞　WFST版SPOJUSおよびDNN-HMMを開発された藤井康寿氏（現在Google）に感謝します.