Vol.2014-SLP-104 No.29
2014/12/16

年齢・性別に依存しないDNN-HMMによる音声認識法の検討

関 博史[1,a)]　山本 一公[1,b)]　中川 聖一[1,c)]

概要：我々は，従来から音節単位音響モデリングについて研究を行っている．そこで本研究では，まず音節単位および音素単位 DNN-HMM を構築し，これらの認識精度について調査を行った．その結果，triphone，コンテキスト独立音節いずれもほぼ等しい認識精度を示した．次に，3 つの年齢層 (成人・老人・子供) と性別 (男性・女性) ごとに計 6 つのクラスの学習データを用意し，年齢・性別に依存しない DNN-HMM の検討を行った．一般的に，不特定話者を対象とした音声認識システムは，話者特定システムに比べ，認識性能が低下してしまう．しかしクラス毎に特徴量を正規化することで，すべてのクラスを用いて一つのモデルを学習した場合でも，クラスごとに DNN-HMM を学習したモデルを上回る認識精度を得ることが出来た．最後に，クラス情報のネットワークへの組み込みを検討した．

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

Consideration on Age- and Gender-independent Speech Recognition using DNN-HMM

Seki Hiroshi[1,a)]　Yamamoto Kazumasa[1,b)]　Nakagawa Seiichi[1,c)]

Abstract: We have studied syllable-based acoustic modeling for Japanese speech recognition. In this paper, we first investigate the performance of recognition accuracy using phoneme/syllable-based DNN-HMM. The results show that there's no significant difference between phoneme/syllable-based DNN-HMM. Second, we investigate the age- and gender-independent speech recognition using DNN-HMM. We use three types of corpora(adult, elder, child), and each corpus contains male and female speech data. In general, speaker-independent system cannot handle the specific information of speakers, and the recognition performance of speaker independent model is lower that of speaker dependent model. Our experimental results show that one DNN-HMM trained by all corpora with a class-dependent feature normalization method achieves better performance compared to class-dependent DNN-HMMs. Finally, we investigate the incorporation of information on corpora into DNN.

Keywords: Deep Neural Network, HMM, DNN-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]，多くの話者適応手法や学習方法が提案されている [2], [3]．我々は音節単位の音響モデルについて研究を行っており，音節単位 DNN-HMM を用いた音声認識について報告している [4]．文献 [4] では，3 つの年齢層 (成人・老人・子供) と性別 (男性・女性) ごとに計 6 つのデータベースを用意し，モデル化単位および正規化手法の違いによる認識精度の変化について報告した。一般に音声認識システムは，年齢や性別に制限を設けず多くのユーザの使用を仮定しているこ

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

とが望ましい．しかし一方で，不特定話者を対象とした音声認識システムは，話者特定システムに比べ，性能が低下してしまう．また，一般的に学習データが多いほど認識精度は向上する．そこで，認識対象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．DNN は GMM と比べ高い表現能力を持ち，性能の低下は GMM-HMM ほどの変動は見られないものの，話者性の問題は解決されていない [5]．関連研究として文献 [2] があり，ボトルネック特徴量を用いることで年齢の違いによる音響特徴量の変化を抑えている．

本稿では，音響モデルのモデル化単位として用いられる音素および音節単位 DNN-HMM の比較，文献 [4] に引き続き不特定話者音声認識の検討を行う．

本稿の構成を以下に示す．2 節で提案手法である特徴量の正規化，音節単位 DNN-HMM を学習する際に用いた状態の結びについて述べる．3 節で GMM による話者クラスタリングについて述べる．4 節で実験条件及び実験結果を述べ，最後に 5 節で結論を述べる．

2. DNN を用いた音響モデル

2.1 特徴量の正規化

(1) 発話毎の正規化

音声認識システムの動作環境の違いにより，トレーニングデータとテストデータの間の音響特徴量にミスマッチが生じることは多くあり，実環境での動作時に認識精度の低下を招く一因となっている．このミスマッチを抑える手法として，発話毎の Cepstral Mean Normalization (CMN) や Cepstral Variance Normalization(CVN) がある [6]．i 次元目の音響特徴量に対する CMN および CVN は，

$$CMN: \quad \hat{c}_i(t) = c_i(t) - \mu_i \tag{1}$$

$$CVN: \quad \hat{c}_i(t) = \frac{c_i(t)}{\sqrt{(\sigma_i^2)}} \tag{2}$$

$$\mu_i = \frac{1}{T}\sum_{i=1}^{T} c_i(t), \quad \sigma_i^2 = \frac{1}{T}\sum_{t=1}^{T}(c_i(t) - \mu_i)^2 \tag{3}$$

となる．両者を組み合わせると

$$\hat{c}_i(t) = \frac{c_i(t) - \mu_i}{\sqrt{\sigma_i^2}} \tag{4}$$

を得る。この手法を用いて各データの平均，分散を正規化・統一することにより，雑環境下音声認識でもノイズに頑健なシステムが構築される．

(2) クラスごとの正規化

本研究では，年齢・性別ごとの計 6 つの学習データ (クラス) を用いる．各学習データの分散はばらつきを見せており，特徴量の正規化を行うことにより話者間の音響特徴量のばらつきが抑えられると考えられる．また，DNN のプリトレーニングで用いられる制限付きボルツマンマシンでは，学習データを平均 0，分散 1 に正規化しておくことが望ましい．そこで，音響特徴量抽出後クラス毎に平均 0，分散 1 に正規化を行うことにより生じる認識精度の変化について調査を行い，プリトレーニングなしの場合でもこの正規化が有効かどうか検討する．この場合，認識時発話がどのクラスに属するか決定する必要がある．本研究では，GMM を用いてテストデータのクラスタリングを行う (3 節参照)．

2.2 クラス情報の DNN への組み込み

DNN は高い表現力を持ち，クラス情報もネットワーク内で学習されることが期待できるため，以下の二通りの学習を行った．

(1) ユニットの追加: ネットワークの入力層にクラス情報を入力するための 6 ユニット (6 クラス) を追加し，GMM を用いて計算したクラス情報を入力する．2.1 節と対応を取るため，学習時は各発話に対応するクラス情報を既知とし，正解ユニットの値を 1，それ以外のユニットを 0 とする．認識時は，クラス分類用 GMM を用いて各クラスに対する尤度を求め，①尤度の高いクラスのユニットを 1，それ以外のユニットの値を 0 とした場合，②尤度の高いユニットを 1，第 2 候補を 0.5，それ以外のユニットを 0 とした場合，③尤度の高いユニットを 0.7，第 2 候補を 0.3，それ以外のユニットを 0 とした場合の 3 通りの学習方法を検討した．

(2) `Pre-Training` の追加: Pre-Training[7] では，層毎に階層的な特徴を教師なし学習により学習する．そのため，Pre-Training がクラス情報を用いた教師なしクラスタリングの役割を担うか調査を行う．

2.3 学習の高速化

隠れ層のユニットの活性化関数として Rectified Linear Unit($f(x) = \max(0, x)$) を使用することにより，Pre-training を行わない場合でも Pre-training を行った DNN-HMM と同等の精度が得られることが報告されている [8], [9], [10]．学習時間の短縮のため，本研究でも特に断らない限り，Pre-Training なしで活性化関数として Rectified Linear Unit を用いる．

2.4 状態の結び

出力ラベルとしてコンテキスト独立音節および左コンテキスト依存音節，monophone，triphone を用いる．

(1) 音節の場合

音節を出力ラベルとして用いる場合，HMM は**図 1** のように 4 状態出力分布を持つ．DNN の出力ユニット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)

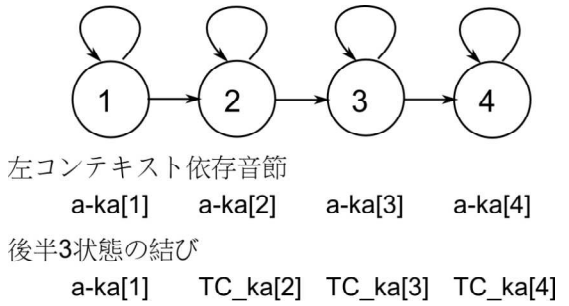

図 1　音節単位 HMM の状態の結び

Fig. 1　Structure of left-context dependent syllable and tied 3 state syllable.

×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 3 状態をコンテキスト独立とした場合 (後半 3 状態の結び; TC3) についても検討を行った．後半 3 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]，TC_ka[2]，TC_ka[3]，TC_ka[4] の 4 状態で構成され，後半 3 状態はコンテキスト独立音節「ka」の各状態と同一である．後半 3 状態をそれぞれ結びにした場合，出力ラベル数は 1276($8 \times 116 \times 1 + 116 \times 3$) である．

(2)monophone と triphone の場合

音素を出力ラベルとして用いる場合，HMM は 3 状態出力分布とする．monophone は 43 種の音素で構成されており，状態数は 129 である．triphone GMM-HMM を構築する際，HMM の状態数は成人男女それぞれ 2000，2500 とした．

3. GMM による話者クラスタリング

性別・年齢別の話者クラスの識別には，GMM を用いる．クラス i の GMM を用いた尤度の計算は，

$$L(X|\lambda_i) = \log p(X|\lambda_i) = \sum_{t=1}^{T} \log p(x_t|\lambda_i) \qquad (5)$$

により行われる．ここで，$X = x_1, x_2, ..., x_T$ は入力系列，λ_i はクラス i の GMM である．また，リアルタイムな音声認識システムを考えた時，すべてのフレームを用いたクラスタリングは音声入力後に音声認識処理開始となるため好ましくない．そこで，発話の先頭 50 フレームのみを用いた話者クラスタリングについても検討する．

4. 評価実験

4.1 実験条件

4.1.1 データベース

GMM-HMM で音響モデルを構築する際，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 [5], [11], [12], [13], [14]．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 1 つにまとめた 1 クラスに対して学習と認識実験を行った．

表 1　各クラスで使用されるトレーニングデータ
Table 1　Training data for Adult, Elder, and Child

ASJ+JNAS		
性別	男性	女性
年齢	18-59	18-59
話者数	184	187
発話数	20,337 (≃33h)	25056 (≃44h)
S-JNAS		
性別	男性	女性
年齢	60-90	60-90
話者数	145	143
発話数	24,081 (≃53h)	24,061 (≃53h)
CIAIR-VCV		
性別	男性	女性
年齢	6-12	6-12
話者数	151	150
発話数	7538(+3393,≃11h)	7744(+3910,≃11h)

実験に用いたコーパスを，**表 1** に示す．学習データ時間は，成人クラスは男性が約 33 時間，女性が 44 時間，子供クラスは男性女性とも約 11 時間，老人クラスは男性女性ともに約 53 時間である．これ以降では成人男性は A-M，成人女性は A-F，子供男性は C-M，子供女性は C-F，老人男性は E-M，老人女性は E-F と表す．

成人用のデータには ASJ+JNAS コーパス [15] 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 [16] を用いる．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”の読み上げ文 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 [17] を用いる．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 名である．

テストデータは，各クラスとも 100 文である．発話者は，それぞれ AM: 23 人，AF: 23 人，EM: 10 人，EF: 10 人，CM: 7 人，CF:8 人である．子供用コーパスはおも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は未知語率は 14%である．成人，老人クラスの未知語率はそれぞれ 0.5%，2.1%である．

表 3　トレーニングデータの違いに対する成人男女クラスの認識精度の変化 (# 32 mixtures, layer=5)

Table 3　Word recognition accuracy using various training data

DNN-HMM	全体学習 (A+E)						全体学習 (A+E+C)					
	AM	AF	EM	EF	Ave.(A)	Ave.(A+E)	AM	AF	EM	EF	Ave.(A)	Ave.(A+E)
CI	95.0	95.7	93.2	94.3	95.4	94.6	94.3	95.4	92.7	95.2	94.9	94.4
TC3	94.2	94.8	92.7	93.8	94.5	94.0	94.9	95.9	92.6	93.2	95.4	94.2

表 2　ASJ+JNAS に対する単語認識精度 (# 32 mixtures, layer=5)

Table 2　Word recognition accuracy on ASJ+JNAS

DNN-HMM		クラス別学習			全体学習		
		AM	AF	Ave.	AM	AF	Ave.
音素	CI	94.8	94.7	94.8	94.2	95.5	94.8
	CD	95.8	95.6	95.7	95.5	95.9	95.7
音節	CI	95.3	95.9	95.6	95.7	96.2	95.9
	TC3	94.1	94.8	94.5	94.1	95.2	94.7
	CD	92.7	93.3	93.0	93.5	93.4	93.4

4.1.2　特徴パラメータと音響モデル

(a) GMM-HMM

GMM-HMM の学習に用いた特徴量は，12 次元 MFCC，ΔMFCC，ΔΔMFCC および Δ パワー，ΔΔ 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．

クラス毎の音節 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 混合としている．116 音節のコンテキスト独立 HMM を学習した後，コンテキスト依存 HMM($116 \times 8 = 928$ 種類) を MAP 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．

クラス毎の音素 GMM-HMM は，音素単位の left-to-right 型で，各 HMM は 3 状態出力分布を持つ．混合数や学習方法は音節 GMM-HMM と同じである．その後，決定木に基づく状態共有を行い triphone GMM-HMM を作成した．

なお，モデルの学習には HTK Toolkit[18] を使用した．

(b) DNN-HMM

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，ΔMFCC，ΔΔMFCC およびパワー，Δ パワー，ΔΔ パワー，計 39 次元を用いた (パワーを用いているため，GMM-HMM より 1 次元多い)．入力フレーム数は 11 とし，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．

4.1.3　クラス分類 GMM

クラス分類のために用いる GMM は，4.1.2(a) と同じく 12 次元 MFCC，Δ MFCC，ΔΔ MFCC，Δ パワー，ΔΔ パワーを用いて学習した．また，混合数 8 の初期モデルの作成には各クラスとも 10000 発話のみを使用し，最終的な GMM の混合数は 128 混合とした．

表 4　50 もしくはすべてのフレームを用いた、年生性別依存 6 クラスへのクラス分類精度

Table 4　Accuracy of classification into six-classes depending on age and gender using 50/all frames

input	output (50 / all frames)					
	AM	AF	EM	EF	CM	CF
AM	84/90	2/1	12/9	0/0	2/0	0/0
AF	0/0	83/96	0/0	13/4	1/0	2/0
EM	19/22	32/0	45/78	4/0	0/0	0/0
EF	30/0	30/17	7/5	29/78	1/0	3/0
CM	2/0	8/2	0/0	1/1	72/92	17/5
CF	0/0	8/1	1/0	2/0	31/14	58/85

4.1.4　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 語を使用し，tri-gram 言語モデルを学習した．カットオフはすべて 1 であり，バックオフの計算にはウィッテンベル法を用いて学習した．

4.1.5　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)[19] を，DNN-HMM のデコーダには，WFST 版 SPOJUS を用いた．

4.2　音素および音節モデルによる認識精度の変化

音素及び音節単位で DNN-HMM を作成し，成人男女の音声認識実験を行った．単語認識精度を**表 2** に示す．音素 (CI, CD) はそれぞれ monophone (出力ユニット数:$43 \times 3 = 129$), triphone(出力ユニット数は AM:2000, AF:2500)，音節 (CI, TC3, CD) はそれぞれコンテキスト独立音節，後半 3 状態の結び，左コンテキスト依存とする．

音節単位でモデル化を行った場合，コンテキスト独立でモデル化を行った DNN-HMM の認識精度が左コンテキスト依存を上回る結果となった．これは，入力が 11 フレームの特徴量であり，コンテキスト情報を十分捉えることができたことと、1 出力ユニット当たりの学習データの増加

表 5　各モデル化手法に対する単語認識精度
Table 5　Word recognition accuracy using different normalization method

class	model	AM	AF	EM	EF	CM	CF	Ave.
6 class_each	GMM(CD)	93.5	94.6	89.4	93.4	74.7	78.2	87.3
6 class_each	DNN(TC3, layer=5)	94.1	94.8	92.8	94.4	78.7	80.7	89.3
1 class_all	DNN(TC3, layer=5)	94.9	95.9	92.6	93.2	77.2	78.8	88.8
1 class_all (正規化無し)	DNN(TC3, layer=5)	94.4	95.8	92.8	93.1	75.8	78.3	88.4
1 class_each ($^+$:正規化時クラス既知)								
1 class_each$^+$	DNN(TC3, layer=5)	95.2	95.6	93.0	95.0	78.4	79.9	89.5
1 class_each (50 frames)	DNN(TC3, layer=5)	95.1	95.4	92.9	93.6	78.6	79.1	89.1
1 class_each (all frames)	DNN(TC3, layer=5)	95.2	95.9	92.3	94.6	79.5	80.2	89.6
Pre-Training 有り DNN-HMM								
1 class_all	DNN(TC3, layer=5)	95.1	95.7	93.4	94.4	75.7	78.9	88.8

表 6　クラス情報を入力した際の単語認識精度
Table 6　Word recognition accuracy on additional class unit

class	model	第 1 候補	第 2 候補	その他	AM	AF	EM	EF	CM	CF	Ave.
1 class_all (all frames)	DNN (TC3, layer=5)	1	0	0	94.2	95.9	91.8	92.4	79.6	80.3	89.0
1 class_all (all frames)	DNN (TC3, layer=5)	1	0.5	0	93.1	95.9	89.6	88.7	55.8	39.7	77.1
1 class_all (all frames)	DNN (TC3, layer=5)	0.7	0.3	0	95.2	95.9	92.0	93.7	79.6	80.2	89.4

のためと考えられる．また，使用したデータが成人男女の場合，クラス別にモデルを学習した場合 (クラス毎に平均 0，分散 1 に正規化，クラス既知) と 2 つのクラスをまとめて 1 つのモデルを学習した場合では，平均単語認識精度に大きな違いは見られず，音節単位でモデル化を行った場合，クラス別学習と比べ全体学習は絶対値で約 0.3%の改善を得た．また，triphone モデルとコンテキスト独立音節モデルはほぼ等しい精度を示した．

次に，S-JNAS および CIAIR-VCV コーパスを用いて 6 クラスを対象にトレーニングデータを増やし，音節単位のモデリング手法についてさらに実験を行った．ASJ+JNAS に S-JNAS を加え学習した場合 (A+E)，さらに CIAIR-VCV を加え学習した場合 (A+E+C) の認識精度を**表 3** に示す．話者性の問題により，コーパスの増加に伴い成人男女の平均単語認識精度は低下したが，CI と TC3 の間で生じていた認識精度の差は小さくなった．

4.3　特徴量の正規化

以下に示す二通りの特徴量の正規化手法を行った．

(1) `_all`: すべてのコーパスをまとめ，トレーニングデータ全体の音響特徴量を平均 0，分散 1 に正規化．

(2) `_each`: 各コーパスのトレーニングデータ (6 クラス) を平均 0，分散 1 に正規化．なお，比較のために正規化しない場合も行った (但し，発話毎の CMN は施した)．

正規化手法_each を用いて評価実験を行うにあたり，テストデータがどのクラスに属するか 128 混合の GMM を用いてクラスタリングを行った．この結果を**表 4** に示す．成人男女のクラス分類は比較的良く，老人男女は成人男女への混同，子供男女は子供男女間への混同が多い．この結果を用いて認識実験を行った結果を**表 5** に示す．クラスタリング精度は悪いものの，各発話の先頭の 50 フレームおよび 1 発話すべてのフレームを用いた場合，いずれも従来手法である 1 class_all を上回る精度を得ることが出来た．また，クラス毎に正規化したモデルを用いた 6 class_each と同等の精度を得た．なお，1 class_all の正規化無しの場合 (発話毎の CMN は採用) は，1 class_all の正規化ありと比べ，若干認識精度が低下した．

次に，1 class_each の各クラスの MFCC12 次元の分散と 1 class_all の MFCC12 次元の分散のユークリッド距離を求めた．その結果，AM: 2.16，AF: 3.23，EM: 2.54，EF: 3.33，CM: 1.99，CF: 2.19 であるように，子供クラスの分散が比較的小さい傾向が得られ，認識の困難さを示している．

4.4　クラス情報のネットワークへの組み込み

(1) クラス情報のネットワークへの入力

クラス情報を入力するユニットを追加した DNN-HMM の単語認識精度を**表 6** に示す．クラス情報を組み込んでいない 1 class_all に比べ精度の向上が見られ，クラス毎の正規化機能を完全には実現できていないが，ほぼ同等の性能を得た．

(2) `Pre-Training` の効果

Pre-Training あり DNN(TC3, layer=5) の単語認識精度を表 5 の Pre-Training 有り DNN-HMM の欄に示す．クラス毎に認識精度の変化が見られるが，6 クラスの平均単語認識精度は Pre-Training 無し 1class_all

と同じく 88.8%であった．Pre-Training ありの DNN-HMM でもクラス毎の正規化機能は実現できていないと言える．

5. まとめ

本研究では，まず音節単位および音素単位 DNN-HMM の比較実験を行った．Triphone およびコンテキスト独立音節 DNN-HMM はほぼ同等の精度を示し，音節単位 DNN-HMM に着目すると学習データの増加に伴い CI および TC3 の認識精度の差が小さくなった．音節単位 DNN-HMM において，学習データを増加させた際 (CSJ コーパス [20] の使用) のモデル化手法 (CI，TC3，CD) と認識精度の関係の調査は今後の課題である．

次に，入力音響特徴量に対するクラス毎の分散正規化手法を提案した．クラス別に特徴量の正規化を行った DNN が今回報告した認識精度の中で最も良いものであり，分散の正規化により話者間の音響特徴量のばらつきが抑えられ，認識精度が改善したと考えられる．この場合，GMM による自動クラスタリング法の組み込みにより，クラス既知と同等の性能が得られた．

最後に，これらのクラス情報もネットワーク内で学習されることが望ましいと考え，クラス情報のネットワークへの入力と Pre-Training の追加，計二通りの学習を行った．いずれもクラスごとに音響特徴量の分散を正規化した DNN-HMM の認識精度の結果を下回っており，今回実験を行ったアーキテクチャではクラス情報をネットワーク内で扱うことは出来なかった．ネットワーク内でのクラス情報の学習は今後の課題である．

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] 林　知樹，北岡教英，武田一哉：深層学習を用いた音声特徴量の年齢の変動に対する頑健性の調査，音講輪 (秋)，pp. 77–80 (2014).
[3] 柏木陽佑，齋藤大輔，峯松信明，広瀬啓吉：制約付き話者コードの同時推定によるニューラルネット音響モデルの話者正規化学習，音講論 (秋)，pp. 7–10 (2014).
[4] 関　博史，中川聖一：音節単位 DNN-HMM の音声認識の評価，音講輪 (春)，pp. 179–182 (2014).
[5] 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).
[6] Olli, V., Bye, D. and Laurila, K.: A recursive feature vector normalization approach for robust speech recognition in noise, *Proc. ICASSP*, pp. 733–736 (1998).
[7] G.E.Hinton, S.Osindero and Y.Teh: A fast learning algorithm for deep belief nets, *Neural Computation*, Vol. 18, pp. 1527–1554 (2006).
[8] L.Toth and T.Grosz: A comparison of deep neural network training methods for large vocabulary speech recognition, *Text, Speech, and Dialogue*, No. LNAI8082, pp. 36–43 (2013).
[9] L.Toth: Phone recognition with deep sparse rectifier neural networks, *Proc. ICASSP*, pp. 6985–6989 (2013).
[10] 関　博史，中川聖一：音節単位 DNN-HMM による音声認識の検討，音声言語情報処理 (SLP)，Vol. 2013-SLP-99, No. 4, pp. 1–6 (2013).
[11] T.Kosaka, S.Matsunaga and S.Sagayama: Tree-structured speaker clustering for speaker-independent continuous speech recognition, *in Proc. ICSLP*, pp. 1375–1378 (1994).
[12] T.Kosaka and S.Matsunaga, S.: Speaker-independent speech recognition based on tree-structured speaker clustering, *Computer speech and language*, pp. 55–74 (1996).
[13] 芳澤伸一，馬場　朗，松浪加奈子，米良祐一郎，山田実一，李　晃伸，鹿野清宏：十分統計量と話者距離を用いた音韻モデルの教師なし学習法，電気情報通信学会誌，Vol. J85-D-, No. 3, pp. 25–30 (2002).
[14] 朱　発強，山本一公，中川聖一：トレーニングデータのソフトクラスタリングに基づく不特定話者の音声認識，音講輪 (春)，No. 1-Q-5, pp. 159–160 (2010).
[15] 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).
[16] : CIAIR 子供の声データベース (CIAIR-VCV)，`http://research.nii.ac.jp/src/CIAIR-VCV.html`.
[17] : 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`.
[18] : HTK Toolkit, `http://htk.eng.cam.ac.uk/`.
[19] Y.Fujii, K.Yamamoto and S.Nakagawa: Large vocabulary speech recognition system:SPOJUS++, *MUSP*, pp. 110–128 (2011).
[20] S.Furui, K.Maekawa and H.Ishihara: A Japanese national project on spontaneous speech corpus and processing technology, *Proc. ASR2000*, pp. 244–248 (2000).