Vol.2016-SLP-113 No.1
2016/10/6

日本語話し言葉音声における半教師ありDNN学習の検討

加藤 拓[1]　篠崎 隆宏[1,a)]

概要：音声認識システム開発における書き起こしコスト削減のために，Deep Neural Network（DNN）音響モデルのための効果的な半教師あり学習アルゴリズムの実現は重要である．少量のラベル付き音声データと大量のラベルなし音声データを用いた半教師あり学習では，まずラベル付きデータを用いて認識システムを学習し，それを用いてラベルなし音声を認識し，その結果をラベル情報の代わりとして用いるのが，典型的な枠組みである．その具体的な実現方法としては，学習時と認識時で同じネットワークを用いる手法の他，学習時には出力層を分岐させたネットワークを用いる手法や，或いはネットワークの出力層を学習の途中で置換する手法などが提案されている．また学習のスケジューリングについても，種々の方法が考えられる．本研究ではこれらの手法を日本語話し言葉音声をタスクとして体系的に比較すると共に，教師なし学習により得られたクラスタ情報と組み合わせた手法について検討を行う．

キーワード：大語彙音声認識，DNN，半教師あり学習

1. はじめに

音声認識システムの開発において，大量の音声データとその書き起こし文書が必要となるが，書き起こし文書作成には多大なコストがかかるという問題がある．このコストを削減するためのアプローチとして少量のラベル付きデータと大量のラベルなしデータを用いて認識システムを構築する半教師あり学習が提案されており [1]，Deep Neural Network（DNN）[2] における半教師あり学習の研究も行われている [3][4].

半教師あり学習ではラベル付きデータを用いて認識システムを構築し，それを用いてラベルなしデータを認識することでラベル情報を取得する．ラベル情報としては隠れマルコフモデル（Hidden Markov Model : HMM）[5] 状態が用いられる．ラベル付きデータとラベルなしデータの学習方法は様々考えられ，通常の DNN を用いた学習の他に，出力層を 2 つ持つ DNN を用いた半教師あり学習が提案されている [6]．また半教師あり学習の後に新たな出力層を用いてラベル付きデータを再学習する手法も提案されており，その他にも両データの学習のスケジューリングについて，様々な順番を考えることが可能である．

しかし，各学習順を変更することによる比較は行われておらず，どの学習順が DNN の学習において有効であるかは示されていない．また学習する言語によって異った傾向があるとも考えられる．そこで本研究では日本語話し言葉音声をタスクとして，出力層を 1 つ持つ DNN と 2 つ持つ DNN において上記の各手法を体系的に比較することによって，半教師あり学習において有効な学習法について検討した．

また半教師あり学習では多くの場合，ラベル付きデータによって学習された認識器を用いて，ラベルなしデータにラベルを付与する．よってラベル付きデータの量が少ない場合は認識器の性能が低くなるので，得られるラベルは誤りを多く含み，かつラベル付きデータの HMM 状態の情報のみを含むと考えられる．そこで本研究では，ラベルなしデータのクラスタリング結果を用いることで，半教師あり学習の性能を更に向上させる 2 つの手法を提案した．教師なし学習によってラベルなしデータをクラスタリングすることで，HMM 状態とは異なる情報を取得することができると考えられる．1 つ目の手法は，クラスタ ID をターゲットとして DNN を学習し，次に新たな出力層を用いてラベルなしデータの HMM 状態とラベル付きデータの HMM 状態を順に学習する手法である．2 つ目の手法は，ラベルなしデータの HMM 状態とクラスタ ID を比較し，同じ HMM 状態とクラスタ ID を持つフレームのみを用いて DNN を学習する手法である．認識結果とクラスタ ID は異なる誤り傾向を持つと考えられるので，共通の HMM 状態とクラスタ ID を持つフレームはより信頼度の高いフレームであると考えることができる．DNN において全ラベルなし

[1] 東京工業大学
Tokyo Institute of Technology
[a)] www.ts.ip.titech.ac.jp

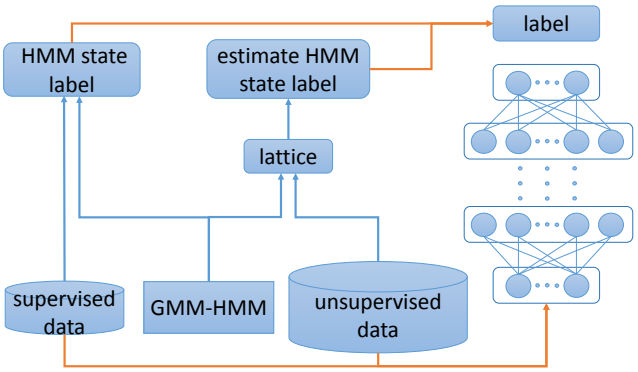

図 1 DNN の半教師あり学習手順.

データの HMM 状態を学習した後に，信頼度の高いフレームによる学習，ラベル付きデータの学習を順に行うことで，より高い性能を持つ DNN を構築できると考えられる．

本論文の構成は以下に示す通りである．第 2 章では DNN における既存の半教師あり学習法について説明し，提案法であるクラスタ ID を用いた半教師あり学習法についての説明を行う．第 3 章では各手法を比較した結果を示し，その内容について考察を行う．第 4 章で本論文のまとめと今後の課題について述べ，結論とする．

2. 半教師あり学習

少量の書き起こし文書と大量の音声データを用いた半教師あり学習法について述べる．ラベルなしデータに対するラベル作成のために，まずラベル付きデータを用いて認識器を構築する．認識器の種類としては GMM-HMM と DNN-HMM の 2 種類が考えられるが，ラベルなしデータのラベルを取得する際に DNN-HMM を用いると学習のパイプラインが長くなる一方，半教師あり学習においてごく小さな影響しかないことがわかっているため [6]，本稿では GMM-HMM を用いてラベルなしデータのラベルを取得することを考える．

半教師あり学習の手順は以下の通りである．図 1 に学習の手順を示す．ラベル付きデータを用いて GMM-HMM を学習し，書き起こし文書を用いてラベル付きデータに対する正解 HMM 状態を得る．次に，学習済みの GMM-HMM を用いてラベルなしデータを認識する．この時に得られるラティスから推定 HMM 状態を取得し，これをラベルなしデータに対するラベルとする．次にラベル付きデータとラベルなしデータの両方を用いて DNN をプレトレーニングする．最後にラベル付きデータに対する正解 HMM 状態と，ラベルなしデータに対する推定 HMM 状態を用いて，DNN を誤差逆伝播法（backpropagation）により学習する．

2.1 Multi-softmax DNN による半教師あり学習

出力層を 2 つ持つ DNN である Multi-sotmax DNN における半教師あり学習が [6] で提案されている．その構造を図 2 に示す．出力層を 1 つ持つ通常の DNN（Single-softmax DNN）を用いた半教師あり学習では，ラベル付きデータとラベルなしデータをシャッフルして同じネットワーク上で学習し，認識する．これに対し Multi-softmax DNN ではラベル付きデータ用の出力層とラベルなしデータ用の出力層を個々に準備し，入力層と隠れ層を全て共有した形となっている．学習の際にはシャッフルされた両データを各々の出力層を用いて学習し，認識実験の際にはラベル付きデータ用の出力層を用いて認識する．またラベルなしデータに含まれる誤りの影響を小さくするために，学習済みの出力層を全て取り除き，新しく乱数で初期化された出力層を用いてラベル付きデータを再学習することで更に認識性能が向上することが示されている．

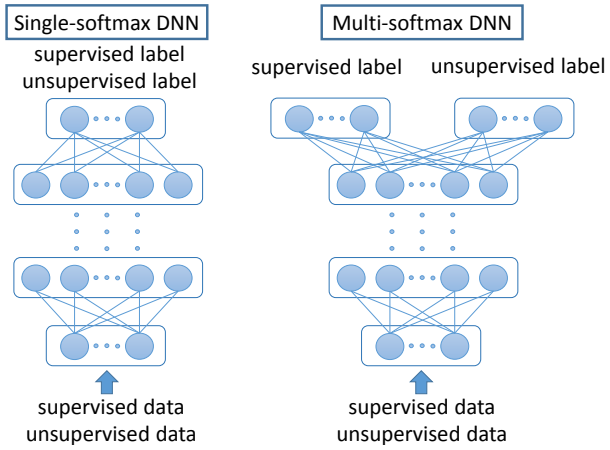

図 2 Single-softmax DNN と Multi-softmax DNN.

2.2 学習順序の比較

半教師あり学習において，ラベルなしデータ中の誤りの影響を小さくするために，新たな出力層を用いて DNN を再学習する手法の有効性が報告されていることを前述した．一方，新たな出力層を用いずに同じ出力層を用いて再学習することや，ラベルなしデータとラベル付きデータをシャッフルせずに順に学習する、という手法も考えられる．そこで本研究では Single-softmax DNN と Multi-softmax DNN の両モデルについて，

(1) 両データをシャッフルして学習
(2) 両データをシャッフルして学習した後，新たな出力層を用いてラベル付きデータの学習
(3) 両データをシャッフルして学習した後，学習済みの出力層を用いてラベル付きデータの学習
(4) ラベルなしデータ，ラベル付きデータを順に学習

の 4 つの学習法について認識実験を行い結果を比較することで，半教師あり学習においてどの学習法が有効であるかを検討する．また，ラベル付きデータとラベルなしデータのデータ量の比によって傾向が変わることも考えられるため，各データ量を変更した実験も行う．

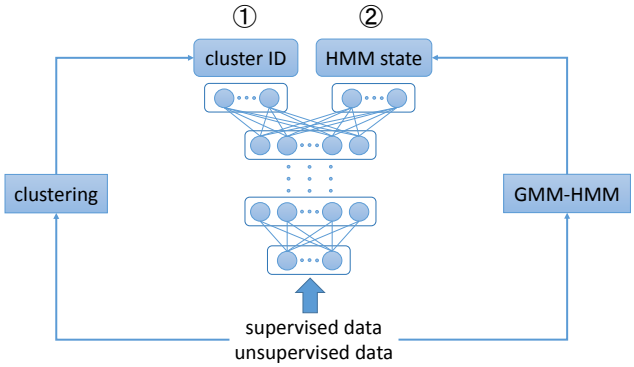

図 3 クラスタ ID の学習による半教師あり学習

2.3 クラスタ ID を用いた DNN 半教師あり学習

ラベルなしデータを教師なし学習によってクラスタリングした結果を，DNN の半教師あり学習に用いる手法を 2 つ提案する．音響特徴量をクラスタリングすることで，音素毎に近い状態でクラスタを作ることができると考えられるが，音声認識において一般的に用いられている MFCC (mel-frequency cepstral coefficients) 特徴量 [7] をクラスタリングすると，音素毎だけでなく話者毎のクラスタが作られる可能性がある．そこで MFCC 特徴量を fMLLR (feature-space maximum likelihood linear regression) 法 [8] によって特徴量空間において話者適応した fMLLR 特徴量をクラスタリングすることで，話者毎ではなく音素毎に近いクラスタを作ることができると考えられる．また話者適応モデルはラベル付きデータにより作成され，ラベルなしデータに対しては作成したモデルを用いて教師なし適応により fMLLR 特徴量を推定している．クラスタ ID を用いた 2 つの提案半教師あり学習法について述べる．

2.3.1 提案法 1：DNN におけるクラスタ ID の学習

クラスタ ID をターゲットとしてラベルなしデータの学習をする．適切なタイミングでクラスタ ID を学習しなければ，DNN による HMM 状態推定の性能を低下させることが考えられる．そこで本実験では以下の手順によって DNN を学習した．学習手順を図 3 に示す．まずプレトレーニング済みの DNN においてクラスタ ID の学習をする．次に，新たな出力層を用いてラベルなしデータの推定 HMM 状態の学習をし，最後に同じ出力層を用いてラベル付きデータの HMM 状態の学習をする．この学習順により，クラスタ ID の情報を学習した後に，徐々に DNN のパラメタを HMM 状態の推定に最適化できると考えられる．

2.3.2 提案法 2：クラスタ ID によって選別したフレームによる学習

ラベルなしデータに対する推定 HMM 状態とクラスタ ID の両方を用いて，より信頼度の高いと判断されたフレームによって DNN を学習する．学習の手順を以下に示す．

(1) 各フレームの HMM 状態（s_i）を求める．
(2) 状態 s_i に割り当てられたフレームに対し，クラスタ ID を求める．
(3) 各クラスタ ID のフレーム数を求め，フレーム数の多い順にクラスタ ID をソートする．
(4) 状態 s_i であるフレームが元のフレーム数の最低 20%以上になるまで，クラスタ ID をソートされた順に選ぶ．
(5) 状態 s_i に対し，選ばれたクラスタ ID を持つフレームを残し，他のフレームを取り除く．
(6) 全状態に対して（2）–（5）を繰り返す．

上記手順により，より信頼度の高いフレームが選択できたと考えられる．各状態毎に，元の最低 20%のフレーム数を使用することにより，状態毎のデータ量の比を大きく変えないようにした．プレトレーニング済みの DNN において，全ラベルなしデータ，ラベルなしデータ中の信頼度の高いデータ，ラベル付きデータ，の順に同じネットワーク上で学習することにより，徐々に高精度なデータで DNN を学習することができると考えられる．

3. 実験

3.1 実験条件

認識システムには Kaldi ツールキット[*1] を用いた．学習・評価セットには日本語話し言葉コーパス（Corpus of Spontaneous Japanese : CSJ）[9] を用いた．学習データとしては 240 時間の学会講演音声を用い，評価セットには CSJ 標準評価セット 1（10 講演）を用いた．本実験では半教師あり学習の実現のため，少量のラベル付きデータと大量のラベルなしデータを学習データ 240 時間のサブセットとして準備した．ラベル付きデータ 1 時間とラベルなしデータ 239 時間の場合と，ラベル付きデータ 10 時間とラベルなしデータ 230 時間の場合とを比較した．それぞれ学習セットとしてはラベル付きデータ中の 90%を使用し，残りの 10%を開発セットに用いた．なお，言語モデルの学習には 240 時間全データを用いている．

特徴量は 13 次元の MFCC を Linear Discriminant Analysis（LDA）[10]+Maximum Likelihood Linear Transform（MLLT）[11] によって変換した後に，特徴量空間における話者適応によって得られる 40 次元の fMLLR 特徴量を用いた．ラベルなしデータ及び評価セットについてはこのモデルを用いて，fMLLR 特徴量を推定した．

Splice は ±17 とし，fMLLR 特徴量の前後 17 フレームの計 35 フレームを DNN の入力として用いた．DNN の入力層は 1900 次元，隠れ層は 6 層でそれぞれ 1905 次元，出力層はラベル付きデータが 1 時間の場合は 816 次元，10 時間の場合は 820 次元である．

DNN の学習では，RBM によるプレトレーニングとクロスエントロピーを誤差関数とするファインチューニングを行った．プレトレーニングにはラベル付きデータとラベルなしデータを合わせた全データ（240 時間）を用いた．

[*1] http://kaldi.sourceforge.net/index.html

ファインチューニングにはミニバッチ毎の確率的勾配降下法による誤差逆伝播法を用い，学習率の初期値は 0.004 とした．

クラスタリング手法としては機械学習ライブラリ scikit-learn[*2]のミニバッチ k-means++法 [12][13] を用いた．クラスタ数は HMM の状態数（1h の場合は 816，10h の場合は 820）を基準にし，k-means++法におけるミニバッチサイズは 10,000 とした．DNN においてクラスタ ID を学習する際は，ラベルなしデータ中の 90%を学習セットに，残りの 10%を開発セットに用いている．

本実験では以下の 12 種の学習方法を比較する．(2)–(12) ではプレトレーニング済みの DNN を用いて学習する．

(1) ランダムに初期化された DNN においてラベル付きデータの学習をする．
(2) プレトレーニングによって初期化された DNN においてラベル付きデータの学習をする．
(3) Singe-softmax DNN において，ラベルなしデータとラベル付きデータをシャッフルして学習する．
(4) (3) で学習済みの DNN の出力層を取り除き，新たにランダムに初期化された出力層を連結し，ラベル付きデータの学習をする．
(5) (3) で学習済みの DNN を用いてラベル付きデータを改めて学習する．
(6) Singe-softmax DNN において，ラベルなしデータ → ラベル付きデータの順に学習する．
(7) Multi-softmax DNN において，ラベル付きデータとラベルありデータをシャッフルして学習する．
(8) (7) で学習済みの DNN の出力層を取り除き，新たにランダムに初期化された出力層を連結し，ラベル付きデータの学習をする．
(9) (7) で学習済みの DNN を用いてラベル付きデータを改めて学習する．
(10) Multi-softmax DNN において，片方の出力層を用いてラベルなしデータの学習をした後に，他方の出力層でラベル付きデータの学習する．
(11) Multi-softmax DNN において，片方の出力層を用いてクラスタ ID の学習をした後に，他方の出力層を用いてラベルなしデータの HMM 状態 → ラベル付きデータの HMM 状態の順に学習する．
(12) Single-softmax DNN において，全ラベルなしデータ → クラスタ ID により選別されたラベルなしデータ → ラベル付きデータの順に学習する．

3.2 実験結果

表 1 にラベル付きデータが 1h と 10h の場合の (1) – (11) の学習法による単語誤り率（word error rate : WER）を示す．1h，10h どちらの場合もプレトレーニングを行うことで認識性能が向上し，ラベルなしデータの HMM 状態の学習をすることで更に性能が向上していることがわかる．

Single-softmax DNN による各半教師あり学習を比較する．(3) – (6) より Single-softmax DNN については，1h の場合は (6) が，10h の場合は (5) と (6) が最も良い結果となった．このことより最後にラベル付きデータの学習をすることで，DNN のパラメタがより適切に調整されたことがわかる．(4) において，新しい出力層を用いてラベル付きデータを学習することで，10h の場合は (3) よりも WER が 1.4%減少しているのに対し，1h の場合は (3) よりも 0.44%増加している．これは 1h の場合はラベル付きデータの量が少ないため，十分に出力層のパラメタの学習が行われなかったためだと考えられる．またどちらも (4) に比べて (5) (6) の WER が良いことから，ラベルなしデータの学習により，出力層におけるパラメタをより適切に設定できると判断できる．

次に Multi-softmax DNN における (7) – (10) の学習について比較する．最も性能が良かった学習法は 1h の場合は (10)，10h の場合は (8) である．どちらも最後にラベル付きデータの学習をすることは共通だが，1h の場合はデータをシャッフルせずに順に学習した方が良い結果となり，10h の場合はデータをシャッフルした後に新たな出力層を用いてラベル付きデータの学習をすることで性能が向上している．これは，1h の場合はラベル付きデータの量が少ないため、ラベルなしデータと共に学習すると，ラベルなしデータの影響が大きくなってしまうことが原因だと考えられる．そのため，それぞれ順番に学習することでラベルなしデータの誤りの影響を小さくすることができる．一方，10h の場合はデータ量が増えたことで，ラベル付きデータとラベルなしデータの量の偏りが小さくなったため，ラベルなしデータの誤りの影響を大きく受けずに学習できたと考えられる．そして新たな出力層を用いて再学習することで，ラベルなしデータ学習の際の誤りを除くことができ，更に性能を向上させることができた．

Single-softmax DNN と Multi-softmax DNN の結果を比較する．両者を比較すると，1h においては (6) が，10h においては (5) と (6) が最も良い WER を示している．出力層を分割するよりも同じ出力層を用いる方が良い理由は，ラベルなしデータによって出力層のパラメタをより適切に学習できるからだと考えられる．また [6] では学習法 (3) (4) (7) (8) を比較した結果，Multi-softmax DNN を用いた (8) の学習法が最も良い結果であったと述べている．日本語話し言葉をタスクとした本実験においても上記 4 つの学習法では (8) が最も良い結果を示しているが，同じ出力層においてラベル付きデータとラベルなしデータを学習した (5) と (6) が更に良い WER を示すことがわかった．

図 4 に学習方法 (6) における評価セットによるエポッ

[*2] http://scikit-learn.org/stable/

表 1　各半教師あり学習法による WER.

DNN の構造	Pre-training	学習方法	1h WER [%]	10h WER [%]
Single-softmax	no	(1) supervised data only	31.17	18.49
	yes	(2) supervised data only	28.28	18.18
		(3) shuffle	22.74	17.77
		(4) shuffle → supervised data (new softmax)	23.18	16.37
		(5) shuffle → supervised data (same softmax)	21.55	**15.87**
		(6) unsupervised data → supervised data	21.48	**15.87**
Multi-softmax		(7) shuffle	23.62	16.64
		(8) shuffle → supervised data (new softmax)	23.13	16.26
		(9) shuffle → supervised data (same softmax)	23.62	16.64
		(10) unsupervised data → supervised data	23.11	16.47
		(11) cluster ID → unsupervised data → supervised data	**21.23**	16.1

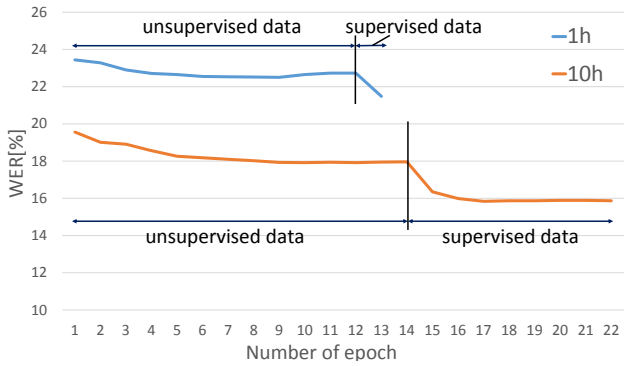

図 4　エポック毎の WER の推移.

ク毎の WER の推移を示す．ラベルなしデータによる学習とラベル付きデータによる学習を連続的にプロットし，ラベル付きデータが 1h の場合は 12 エポック，10h の場合は 14 エポックでラベルなしデータの学習が終了した．図 4 よりまずラベルなしデータの学習によって WER が徐々に小さくなり，続けてラベル付きデータを学習することで更に WER が減少している．このことから，最後にラベル付きデータによって学習する手法が効果的であることがわかる．

次に (11) のクラスタ ID の学習について考察する．クラスタ ID を用いることで，1h の場合は最も性能の良かった (6) よりも WER が 0.25%減少した．しかし 10h の場合は (5) (6) よりも WER が 0.23%増加した．これは 1h の場合はデータ量が少なかったため，クラスタ ID の情報が DNN の学習において有用であったが，10h の場合はデータ量が増加したため，より正しい HMM 状態ラベルをラベルなしデータに付与することができ，クラスタ ID の情報が HMM 状態の推定に不要になってしまったことが原因だと考えられる．そのため，クラスタ ID を最初に学習することはラベル付きデータの量が極端に少ない場合に有効な手法であるといえる．

また，表 2 にクラスタリングのクラスタ数を変更し比較した結果を示す．HMM 状態の数を基準に 0.5 倍したものと，2 倍したものとを比較した．1h の場合はどのクラスタ数も単にラベルなしデータ → ラベル付きデータの学習をするよりも高い性能を示しており，クラスタ数が HMM 状態の数と同数の時に最も良い WER を示している．一方 10h の場合はどのクラスタ数も性能が低下し，クラスタ ID の学習が有効ではないことがわかった．

次に，2 つ目の提案手法であるクラスタ ID を用いて選別したデータによって DNN を学習した結果を表 3 に示す．元のラベルなしデータの中で正しいラベルを持つデータの割合は 76.57%であったのに対し，クラスタ ID でデータを選別することにより，正解ラベルを持つデータの割合は 83.2%まで上昇したので，より高精度なデータが取得できた．このデータを用いて DNN を学習することで性能が向上すると考えられる．しかし全ラベルなしデータで学習した後に選別されたデータで学習し，評価セットに対する WER を算出した結果，選別されたデータで学習する前と比べて WER が 3.72%増加してしまった．

正解データの割合が増加したにもかかわらず性能が低下した原因を探るために，ランダムにフレームを抽出したデータと，正解ラベルをクラスタ ID によって選別したデータについても実験を行った．どちらの場合も状態毎のデータ量の比を変えないために，全状態について元のフレームの最低 20%以上のフレームを抽出した．また各データによる学習性能を正確に比較するために，予めラベルなしデータで学習された DNN を用いるのではなく，プレトレーニング済みの DNN を直接初期モデルとして用いた．表 3 より，ランダムに抽出した方がクラスタ ID を用いるより WER が 6.26%増加した．また，正解ラベルに対してクラスタ ID を用いて選別した結果を見ると，全て正しいラベルで学習しているのにもかかわらず，ランダムに選択するよりも 1.83%増加した．これらの結果と，ラベルの精度が向上している点，状態毎のデータ量の比を変更しないように設定している点を加味すると，性能が低下している原因としてはクラスタ ID によって選択しているフレームに問題があると考えられる．学習においてそれ程有用ではないフレームや同じ話者の連続したフレームばかりを選択した場合，各状態について同じようなフレームが残り，各状態の話者によって異なる情報や頻出しない情報が消失してしまうと考えられる．また特徴量を簡易なクラスタリン

表 2 クラスタ ID の学習による WER.

ラベル付きデータ数	学習方法	クラスタ数	WER [%]
1hour	unsupervised data → supervised data	none	21.48
	Cluster ID → unsupervised data → supervised data	408	21.4
		816	21.23
		1632	21.45
10hour	unsupervised data → supervised data	none	15.87
	Cluster ID → unsupervised data → supervised data	410	16.2
		820	16.1
		1640	16.06

表 3 クラスタ ID により選別したデータを用いた場合の WER. data は全データに対する学習データの割合，purity は正解ラベルの割合を表す.

学習方法	data [%]	purity [%]	WER [%]
all unsupervised	100	76.57	22.8
all→high confidence	100/27.53	76.57/83.2	26.52
high confidence	27.53	83.2	30.25
random select	20	76.58	23.99
true & cluster ID	26.87	100	25.82

グ手法によって分類している点も，同じような情報しか抽出できなかった原因の 1 つだと考えられる．このデータを用いて DNN を学習したため，性能が低下してしまったのだと推測できる．上記問題点の解決策としては，音素毎に高精度に分類できるクラスタリング手法を用いることや，異なる選別方法を用いることなどが挙げられる．

4. まとめ

半教師あり DNN 学習において，ラベル付きデータとラベルなしデータの学習方法や学習順番についての比較を行った．既存研究では，出力層を 2 つ持つ DNN において両データをシャッフルして学習した後に，新たな出力層でラベル付きデータを再学習する手法が良いとされていたが，比較実験の結果から，出力層を 1 つ持つ DNN においてラベルなしデータ → ラベル付きデータの順に学習することで，更に高い性能を示すことがわかった．ラベル付きデータ量が増加した場合は，両データをシャッフルして出力層を 1 つ持つ DNN を学習した後に，同じネットワークでラベル付きデータを学習した場合も同様の WER を示した．また教師なし学習により得たラベルなしデータのクラスタ ID を用いて，DNN の学習を行う手法を 2 つ検討した．出力層を 2 つ持つ DNN においてクラスタ ID→ ラベルなしデータ → ラベル付きデータの順に学習することで，最良の WER を示した上記半教師あり学習に比べて，ラベル付きデータの量が少ない場合は WER を 0.25%削減することができたが，データ量を増やした場合は 0.23%増加し有効ではなくなった．またクラスタ ID によってラベルなしデータから高精度なデータを抽出し，DNN の学習に用いる手法を提案したが，性能を向上させることができなかった．原因としては，選択されたフレームに同じような情報しか残らなかったことなどが考えられる．今後の課題としては，より高精度にクラスタを作る手法の利用や，クラスタ ID をより適切に用いた手法の実現などが考えられる．

謝辞 本研究は JSPS 科研費 26280055 の助成を受けた．

参考文献

[1] X. Cui, J. Huang and J. T. Chien, “Multi-view and multi-objective semi-supervised learning for large vocabulary continuous speech recognition,” Proc. ICASSP, pp.4668-4671, 2011.
[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. 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.
[3] K. Veselý, M. Hannemann, and L. Burget, “Semi-supervised training of deep neural networks,” Proc. ASRU, pp.267-272, 2013.
[4] S. Thomas, M. L. Seltzer, K. Church and H. Hermansky, “Deep neural network features and semi-supervised training for low resource speech recognition,” Proc. ICASSP, pp.6704-6708, 2013.
[5] M. Gales and S. Young, ”The application of hidden Marcov models in speech recognition,” Signal Processing, vol.1, no.3, pp.195-304, 2007.
[6] H. Su and H. Xu, “Multi-softmax deep neural network for semi-superviesd training,” Proc. Interspeech, 2015.
[7] S. B. Davis and P. Mermelstein, ”Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transaction on Acoustic Speech and Signal Processing, vol.28, no.4, pp.357-366, 1980.
[8] A. Ghoshal, D. Povey, M. Agarwal, P. Akyazi, L. Burget, K. Feng, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, ”A novel estimation of feature-space MLLR for full-covariance models,” Proc. ICASSP, pp.4310-4313, 2010.
[9] S. Furui, K. Maekawa, and H. Isahara, “A Japanese national project on spontaneous speech corpus and processing technology,” Proc. ASR’00, pp.244-248, 2000.
[10] R. O. Duda, P. E. Hart, and D. G. Stork, ”Pattern classification,” in Wiley, Nov. 2000.
[11] R. Gopinath, ”Maximum likihood modeling with Gaussian distributions for classification,” in Proc. IEEE ICASSP, vol. 2, pp.661-664, 1998.
[12] D. Arthur, S. Vassilvitskii, “K-means++: the advantages of careful seeding,” Proc. ACM-SIAM Symp. Discrete Algorithms, 2007.
[13] D. Sculley, ”Web-scale k-means clustering,” Proc. ACM WWW, pp.1177-1178, 2010.