Vol.2018-MUS-118 No.15
Vol.2018-SLP-120 No.15
2018/2/21

広帯域用ニューラルネットワーク音響モデル群から狭帯域用音響モデルへの知識蒸留

福田 隆[1]　鈴木 雅之[1]　倉田 岳人[1]　Samuel Thomas[2]　Bhuvana Ramabhadran[2]

概要：本報告では，複数の教師ニューラルネットワーク音響モデルから，高速処理を目的としたコンパクトな生徒音響モデルに知識を蒸留する方法を提案する．教師モデルには，LSTM や VGG といったレイテンシーは低いが認識精度の高いネットワークを用い，そこから効果的に知識を蒸留することを目的とする．提案法は，(1) ミニバッチ単位での教師ソフトラベルの切り替え，および (2) 各学習データサンプルに対して複数の教師ラベルを割り当てるデータ増強戦略の二つから成る．実験では，知識蒸留によって学習されたコンパクトかつ高速な CNN 音響モデルが，VGG や BLSTM などの教師モデルに迫る認識性能を示すことを述べる．次に本報告では，狭帯域用音響モデルの構築を目的として，情報量がより豊富な広帯域用の教師音響モデルからの知識蒸留を試みる．Aurora 4 を用いた雑音環境下認識実験において，提案手法は狭帯域教師モデルのみを用いた知識蒸留法と比較して，相対的に 5.8%の改善が得られた．

1. はじめに

近年の音声認識では，音響モデル構築における様々な段階で，効果的な統合処理を行うことによって高い性能を引き出すアプローチが検討されている．例えば，性質の異なる複数の音響特徴量の結合や，様々なトポロジーを持つニューラルネットワーク音響モデルの上位の層での統合 [1], [2]，複数の音響モデルから予測される音響スコアに基づくシステムコンビネーションなどが挙げられる．複雑かつ大規模な構成のネットワークを高精度に学習するために，様々な種類の雑音を幅広い SNR レンジで学習データに付加したり，それと同時に利用環境を模擬した音響伝達特性を畳み込むことによって，学習データを増大させる処理（以後，データ増強処理と呼ぶ）が併用されることが多い [3]．しかしながら，複雑な構成のシステムは，リアルタイム性や低レイテンシー，低メモリ状態などが要求される商用システムにおいて，現実的に利用不可能なケースがほとんどである．この問題に対して，近年では知識蒸留（knowledge distillation）処理を介したコンパクトかつ高速処理が可能な音響モデルの構築が検討されるようになってきた．

標準的なニューラルネットワーク音響モデルは，発話の書き起こしデータに由来するハードターゲットを用いて学習が進められる一方，知識蒸留に基づく音響モデルは，教師となるニューラルネットワーク音響モデル（以後，教師モデルと呼ぶ）の学習の後，システム要件を満たすコンパクトな音響モデル（以後，生徒モデルと呼ぶ）構築の 2 段階処理によって構成される [4], [5], [6], [7], [8], [9], [10]．教師モデルの学習段階では，上述のように様々なレベルで情報結合を行うことによって，複雑かつ低速ではあるものの，認識率の観点から高精度な音響モデルを構築することを主目的とする．その後，学習データに対するラベルを教師モデルから生成し，生徒モデル用のソフトターゲットとすることによって，最終的にシステムで用いる生徒モデルの学習を行う．ニューラルネットワーク音響モデルの知識蒸留は様々な条件において効果が報告されており，例えば教師あり学習 [6]，半教師あり学習 [9]，マルチリンガル学習を対象としたもの [11]，シーケンストレーニング [12] などにおいて有効性が示されている．

生徒モデルのさらなる性能改善のため，近年では複数の教師モデルを積極的に活用する研究が増えてきた．Hinton らは教師モデル群のバランスを取るために温度パラメータを利用し [7]，Chebotar らは教師モデルが出力する事後確率の重み付きアンサンブル処理を事前に行い，複数の教師モデルから最終的に単一のソフトターゲットを生成して生徒モデルの学習に活用した．一方，Markov らはマルチタスク学習の枠組みを利用して，生徒モデルの学習にハードターゲットと教師ラベル（ソフトターゲット）の併用を試みた [13]．これらのアプローチでは，様々な条件で学習された教師モデルが何らかの形で事前に統合され，入力サン

[1] 日本 IBM 東京基礎研究所
[2] IBM T. J. Watson Research Center

プル毎に単一のソフトターゲットを生成して，生徒モデルの学習を行っている．複数教師モデルの利用は単一教師モデルの利用と比較して高い改善が得られるが，これらの従来手法は教師モデル側での事前統合が基本となっているため，個々の教師モデルが持つ相補的な識別能力を陽に利用しているとは言えない．

本報告では，様々な教師分布ストリームがより直接的に生徒モデルに与えられた時，生徒モデルが入力データに対して様々な解釈（すなわち，より一般化した解釈）を持つことができるようになると仮定し，知識蒸留とデータ増強処理を結合した簡便なモデル構築方法を提案する [14]．提案手法はそれぞれの教師モデルが持つ相補的な情報を兼ね備えた生徒モデルの構築を可能とする．一般に，データ増強処理は入力音声に様々な雑音や歪みを重畳する形で行われるが，提案法では入力音声に対して複数の教師モデルに基づくソフトターゲットを独立に割り当てることでデータ増強を行う．本報告では，コンパクトなサイズの CNN を生徒モデルとした認識実験において，BLSTM や VGG といった強力な教師モデルと同等の性能が得られるようになることを示す．

次に本報告では，このアプローチをより一般化した方法に拡張し，学習時のみに利用可能な「特権」情報の利用を提案する [5]．本報告では，スペクトルの質が劣化した入力データに対して，劣化が起こっていない理想的なデータから得られる情報（ソフトターゲット）を特権情報と呼ぶ．特権情報の効果の検証のため，狭帯域（電話帯域）用音響モデルの構築を目的として，狭帯域専用の教師モデルだけでなく，広帯域用の教師モデルからの知識蒸留も試みる．すなわち，入力音声に対するラベルとして，狭帯域専用教師モデルに基づくソフトターゲット，広帯域用教師モデルに基づくソフトターゲット，そしてハードターゲットを用いる．Aurora 4 を用いた認識実験によって，提案手法は狭帯域教師モデルのみによる知識蒸留と比較して大きな改善が得られることを示す．

2. 複数教師モデルの活用

前述のように，音響的知識を教師モデルから生徒モデルに伝達するための様々な方法が提案されている．一般的なニューラルネットワーク音響モデルの構築では，（正確な）書き起こしに基づくハードターゲットが用られる一方，知識蒸留処理では，ハードターゲットの代わりに次の損失関数が定義される．

$$\mathcal{L}(\theta) = -\sum_i q_i \log p_i, \tag{1}$$

ここで q_i は教師モデルから生成されるソフトラベルであり，擬似ラベルとして用られるものである．p_i は生徒モデルの音素コンテキストクラス i の出力確率である．各学習サンプルのソフトラベル q_i は，値は小さいものの競合するクラスにおいて 0 ではない確率値を持つ．本報告では，次に示す 3 つの戦略を比較検討する．

Interpolated-training 戦略

複数教師モデルを用いる代表的な従来手法は，各モデルからの事後確率について重み付け平均をとることによって，サンプル毎に単一のソフトラベルを生成する [15]．これを定式化すると，次のようになる．

$$q_i = \sum_k w_k q_{ik}, \tag{2}$$

ここで $w_k \in [0, 1]$ は結合重みである．q_{ik} は k 番目の教師モデルからのソフトラベルである．この手法を Algorithm 1 に示し，以後 Intepolated-training 戦略と呼ぶ．

Algorithm 1 intepolated-training

```
for all minibatches in training data do
  pick minibatch i;
  for all teachers in pool of teachers do
    use teacher j to provide soft-targers for minibatch i;
  end for
  combine soft-targets from all teachers with preassinged
  weights w_j for each teacher;
  update neural network model with minibatch i;
end for
```

複数モデルを扱う上でこれは有効なアプローチの一つであるが，重み付け加算処理によって各教師モデルが持つ相補的な情報が弱まってしまう．生徒モデルに様々な解釈を持たせるという意味で，教師音響モデル間の非類似性は，さらに直接的に活用されるべきである．本報告では，より良い生徒モデルの構築を目指して，2 つの戦略を提案する．

Switched-training 戦略

この方法では，各入力音声サンプルに対して教師モデル（ソフトターゲット）をランダムに切り替える．すなわち，Algorithm 2 に示すように教師ラベル q_{ik} をミニバッチの単位で変更し，最急降下法によって生徒モデルの学習を進める．この方法は，各学習サンプルについてどれか一つの教師を選択することになるので，Intepolated-training 戦略と比較して学習データの総量は変わらない．Intepolated-training 戦略とは対照的に，この戦略の利点は，教師モデルのソフトターゲットを結合すための最適な重みを事前に決定する必要がないことである．

Algorithm 2 switched-training

```
for all minibatches in training data do
  pick minibatch i;
  randomly select a teacher j from the pool of teachers to
  provide to provide soft-targets for minibatch i;
  update neural network model with minibatch i;
end for
```

Augmented-training 戦略

ここでは，前述の Switched-training 戦略に，データ増強処理の概念を追加する．すなわち，各学習サンプルに対して教師モデルの数に相当するコピーを生成し，各コピーにそれぞれの教師モデルからのソフトターゲットを割り当てる戦略をとる．入力音声について雑音付加などによりバリエーションを増やすデータ増強処理（その際，ターゲットは同一のものを用いる）とは対照的に，入力サンプルは同一で，ターゲット側のバリエーションを増やすという試みはこれまでに検討されていない．当然のことながら学習データのサイズは，教師モデルの数に応じて増える．この方法では，生徒モデルが入力サンプルに対して複数の音響的解釈を持ち，入力音声に重畳される個人性や外的要因によるスペクトルの揺らぎなどに対して，より頑健に動作することを期待している．この戦略を Algorithm 3 として示す．

Algorithm 3 augmented-training

```
for all minibatches in training data do
  pick minibatch i;
  for all teachers in pool of teachers do
    use teacher j to provide soft-targets for minibatch i;
    update neural network model with minibatch i;
  end for
end for
```

3. 評価実験

種々の音声コーパスを用いて提案手法の有効性を検証する．まず，学習データとして中規模サイズの音声コーパスを用い，様々な実験条件で生徒モデルを比較する (3 章)．次に，サンプリング周波数の異なるニューラルネットワーク音響モデル（特権情報）を教師モデルとして活用し，特権情報が生徒モデルの性能改善にいかに寄与するかを示す (4 章)．

3.1 ベースライン

中規模コーパスに基づく実験では，合計 500 時間の学習データセットを用いる．学習データの半分に相当する 250 時間分は，BN コーパスから 100 時間、Mixer 6 [16] から 100 時間，AMI コーパス [17] から 20 時間，そして独自に収集した読み上げ音声 30 時間を組み合わせて構成する．そして，その 250 時間分について電子協騒音データベース [18] に収録の雑音データと，RWCP 実環境音声・音響データベース [19] に収録の環境インパルス応答を用いて 500 時間に増強する．雑音重畳の際の SNR は 5〜20dB の間で変化させた．

CNN に基づくベースライン音響モデルを上述のデータで学習する．音響特徴量は 40 次元の対数メル周波数スペクトル係数に 1 次と 2 次の動的特徴量を付加した 120 次元のベクトルとした [20]．対数メル周波数スペクトルは，

表 1　標準サイズのベースライン CNN と教師モデルから知識蒸留された生徒モデルの比較

Model	Target	AVG WER
CNN	hard	13.3
VGG	hard	10.5
BLSTM	hard	11.7
CNN: VGG	soft	11.6
CNN: BLSTM	soft	12.8
CNN: VGG+BLSTM	soft/interpolation	11.4
CNN: VGG+BLSTM	soft/switching	11.2

表 2　コンパクトなベースライン CNN と教師モデルから知識蒸留された生徒モデルの比較

Model	Target	Aurora 4
Compact CNN	hard	15.1
Compact CNN: VGG	soft	13.6
Compact CNN: VGG+BLSTM	soft/switching	13.2

フレーム窓長 25ms でシフト幅 10ms 毎に抽出する．そして，対数メル周波数スペクトルについて平均・分散正規化を行った後，前後 5 フレームを連結して計 11 フレームからなる特徴量に拡張する．畳み込み層は隠れノード数 128 と 256 を持つ 2 層からなり，畳み込み層の後にノード数 2048 の全結合層を 4 層追加する．出力層は前後 2 音素のコンテキスト依存決定木に対応する 9300 ノードを持つ．畳み込み層の第一層は，前述の特徴量を入力とするサイズ 9×9 の畳み込みフィルタから構成される．第 2 層はサイズ 3×4 のフィルタを持ち，第 1 層と第 2 層の両方で，マックスプーリングを行う．畳み込み層 2 層目の非線形出力は，全結合層へと接続される．本実験では，全ての隠れ層において活性化関数にシグモイド関数を採用する．以後，本報告では，このネットワーク構成を標準 CNN モデルと呼ぶ．

また本報告では，処理速度をさらに重視したコンパクトな CNN についても認識性能を比較する．コンパクトな CNN モデルは，標準 CNN と同じく 2 層の畳み込み層で構成されるが，ノード数はそれぞれ 64 と 128 とした．畳み込み層の後に続く全結合層も 4 層から 2 層に減らし，またノード数も 768 に削減した．パラメータ数の異なる上記 2 種類の CNN について，ハードターゲットのみで学習されたものをベースラインとして，提案法であるソフトターゲットを用いた知識蒸留に基づく生徒モデルとの比較を行う．

テストデータには Aurora 4 を用い，言語モデルとして WSJ コーパスを用いて構築したタスク標準の bigram モデルを利用する．ここで，学習データには Aurora 4 提供の音声を含んでいないことに注意されたい．Aurora 4 のテストデータは 4 種類のサブセットから構成されており，Set A = クリーン音声/チャネル一致，Set B = 雑音音声/チャネル一致，Set C = クリーン音声/チャネル不一致，Set D = 雑音音声/チャネル不一致として参照される．3 章の実

験では，それらの認識誤り率の平均で評価を行う．

3.2 教師モデルとその性能

本報告では，VGG と BLSTM を教師モデルとして構築し，2 章で述べた提案方法を用いて知識を生徒モデルに蒸留する．教師 VGG モデルは 10 層の畳み込み層から構成され，マックスプーリング層が 3 層おきに挿入されている．そして，ベースライン CNN と同様に，最後の畳み込み層の後にノード数 2048 を持つ 4 層の全結合層を追加している．ここでは，活性化関数として全隠れ層で ReLU 非線形関数を利用している．また，バッチ正規化処理を全ての全結合層で行っている．一方，教師 BLSTM モデルは，512 個のノードを持つ 4 層の bidirectional LSTM であり，出力層の直前に 256 ノードからなるボトルネック層を持つ．これら 2 つの教師モデルはベースライン CNN モデルと同じ構成の出力層を持ち，交差エントロピー基準によって学習した後，さらにシーケンストレーニングによる追加学習を行っている．学習データはベースライン CNN モデルと同じである．ハードターゲットで学習されたベースライン CNN と教師モデルの性能を表 1 の上部に示す．

3.3 生徒モデルとその性能

教師モデルの出力について事後確率の高い上位 50 個のみをソフトラベルとして用い，まず Interpolated-training 戦略と Switched-training 戦略を比較する．本節の実験では，オリジナルのラベル（ハードターゲット）の併用は行わず，ソフトラベルのみで学習を進める．標準サイズとコンパクト CNN の生徒モデルはともに，ランダムに初期化された重みから学習を開始する．

表 1 の下部に標準サイズの生徒 CNN モデルの実験結果を示す．表に示すとおり，VGG と BLSTM 教師モデルからのソフトラベルによって，高精度な生徒モデルが構築できていることがわかる．標準サイズの生徒 CNN とベースライン CNN を比較すると，教師 VGG モデル単独の利用で認識誤りが 13.3%から 11.6%に減少している．さらに教師 LSTM モデルからのソフトラベルも併用して Switched-training 戦略で学習することによって，単独教師の利用と比較して 3.4%の相対的な改善が得られた．表 2 に示すように，コンパクトな CNN の比較においても同様の傾向が見てとれる．このモデルは標準サイズの CNN と比較して RTF を約 23%削減している．また，表 1, 2 において従来手法の Interpolated-training 戦略と提案手法の Switched-training 戦略を比較しており，提案法が教師 VGG，LSTM モデルからの知識を，より効果的に生徒モデルに転移できていると言える．なお，Switched-training 戦略においては，教師モデルをミニバッチの単位でランダムに選択しているが，収束が遅くなるといった現象は特に見られなかった．

表 3 大語彙連続音声認識タスクによる標準サイズのベースライン CNN と生徒モデルの比較

Model	Target	ASpIRE	BN-dev04f
CNN	hard	41.3	18.4
VGG	hard	35.1	14.3
BLSTM	hard	38.7	16.3
CNN: VGG	soft	37.9	15.4
CNN: BLSTM	soft	40.4	17.1
CNN: VGG+BLSTM	soft/interpolation	37.1	15.0
CNN: VGG+BLSTM	soft/switching	36.9	15.1

表 4 大語彙連続音声認識タスクによるコンパクトなベースライン CNN と生徒モデルの比較

Model	Target	ASpIRE	BN-dev04f
CNN	hard	44.2	20.5
CNN: VGG	soft	41.7	17.8
CNN: VGG+BLSTM	soft/switching	41.3	17.7

3.4 LVCSR タスクでの評価

Aurora 4 は主に雑音環境下での評価を対象とした中規模語彙の認識タスクであるため，本節では提案手法のスケーラビリティの検証を目的として，大語彙連続音声認識（LVCSR）タスクである ASpIRE[21] と Broadcast News コーパスを用いて評価実験を行う．両コーパスでは豊富な量の学習データが利用可能であるが，前節との比較という観点から，ここでは学習データは変更せず，3.2，3.3 章で用いた教師・生徒モデルで評価を行う．

表 3 に標準サイズの CNN での比較結果を，教師モデルの性能と共に示す．教師 VGG モデルは両タスクについてベースライン CNN よりも極めて高い性能を示しており，単独教師モデルによる知識蒸留処理においても，この識別能力がうまく生徒モデルに転移できていることがわかる．結果として，生徒 CNN モデルは教師 LSTM モデルの性能を上回った．また，二つの教師モデルを用いた場合に，さらに性能を改善できていることがわかる．表 4 は同様の実験をコンパクト CNN で行った結果である．ベースラインと比較すると，ここでも同様に単独教師モデルの利用と比べて改善が得られた．

4. 広帯域スペクトルからの特権情報の活用

この章では，広帯域モデルからの特権情報の活用によって，狭帯域 CNN 生徒モデルの性能改善を試みる．特権情報は単なる教師モデルアンサンブルだけでなく，Augmented-training 戦略によるデータ増強処理も行って，幅広い知識を生徒モデルに埋め込むことを検討する．したがって，このタスクで利用する学習データは，狭帯域教師モデル群によるソフトターゲット，広帯域教師モデル群によるソフトターゲット，そしてハートダーゲットからなる．生徒モデルへの入力音声は狭帯域に相当する 8kHz のサンプリング周波数であるが，狭帯域教師モデルには生徒モデルと同じ

表 5　ベースライン狭帯域・広帯域モデルおよび教師モデルの性能

Models	A	B	C	D	AVG
Matched CNN-8k	4.6	11.5	6.4	17.7	12.1
Matched CNN-16k	3.9	7.8	6.0	17.5	11.6
VGG-8k	6.3	12.5	7.5	16.2	13.3
VGG-16k	4.8	8.4	6.2	14.3	10.5
BLSTM-8k	6.3	11.8	7.3	15.1	12.5
BLSTM-16k	4.8	9.3	7.4	16.0	11.7

8kHz のデータを入力し，広帯域教師モデルには対となる 16kHz のデータを入力してソフトラベルを生成する．

4.1　ベースライン広帯域・狭帯域モデル

本章の実験では，Aurora 4 で提供される約 15 時間の学習データを用いて，生徒モデルの構築を行う．Aurora 4 では，学習データとしてクリーン音声のみと雑音重畳音声（マルチコンディション学習セット）の 2 種類が定義されており，ここではマルチコンディション学習セットを用いる．ベースライン CNN モデルと教師モデルの性能を表 5 に示す．

表の Matched CNN-8k と Matched CNN-16k はハードラベルのみで学習された結果である．Aurora 4 ではサンプリング周波数 8kHz の学習データは提供されていないので，Matched CNN-8k（ベースライン狭帯域モデル）は，Aurora 4 にもともと存在する 16kHz の学習データを 8kHz にダウンサンプリングして用いた．ここでは，広帯域モデルを含めた様々な教師モデルの活用によって，Matched CNN-8k の性能（12.1%）がどのように改善されるかを比較する．広帯域教師モデルは 3 章で用いた VGG と BLSTM と同様である．ここでは，VGG-16k，BLSTM-16k として参照する．狭帯域教師モデルは 250 時間の Switchboard コーパスについて，3 章で利用した電子協騒音データベースと RWCP 音響環境データベースに収録されているインパルス応答を利用して，合計 500 時間にデータ増強したものを利用する．狭帯域教師モデルは，それぞれ VGG-8k，BLSTM-8k と称することにする．ここで，広帯域・狭帯域共に教師モデルの学習データには，Aurora 4 由来のデータが含まれないことに注目されたい [*1]．

表 6 に様々な組み合わせの狭帯域生徒モデルの性能を示す．本章の実験では，教師モデルからのソフトラベルだけでなくハードラベルも併用する．表に示すとおり，ソフトラベルのみの利用の場合，認識性能は表 5 に示すベースライン（“Matched CNN-8k”）の値よりも低い．ここでの狭帯域教師モデルは，Aurora 4 の音響ドメインに対してオープンなモデルになっているので，Aurora 4 データのみで学習された（Aurora 4 のみに特化した）ベースライン CNN モデルと比較すると性能が劣る結果となった．しかし，その一方で，教師モデルは Aurora 4 に存在しない相補的な音響情報を持つ．その根拠の一つとして，ハードラベルを併用すると性能改善が非常に大きくなり，教師モデルからの相補的な情報が効果的に働いていることを見てとれる．また，狭帯域教師モデルだけではなく，提案手法である広帯域モデルで生成されるソフトラベルを併用した場合，ハードラベルと狭帯域教師モデル併用のケース（“Hard + VGG-8k”）と比べて明らかな差分があり，相対的に 5.8%のさらなる改善 (AVG = 10.3% → 9.7%) が得られた．これらの結果は学習時にのみ利用可能な特権情報の有効性を示唆している．また，提案法である Switched-training 戦略と Augmented-training 戦略の比較において，Switched-training 戦略はハードターゲットを含めた 5 種類の教師モデルを使った場合に性能が飽和しているが，Augmented-training 戦略により学習サンプルあたりのソフトラベルを増加させた場合には，性能をさらに改善させることがわかった．

さらなる検証のため，テストデータに対して教師モデルの音響スコアによるシステムコンビネーションを行い，人手のチューニングによるオラクルのケースでの性能を比較する．表 7 に教師モデルのシステムコンビネーションの結果を示す．表 6 と表 7 からわかるように，提案手法によって構築された生徒モデルは，人手によるチューニングを実施した教師モデルのシステムコンビネーションと同等の性能を示している．教師モデルでデコードする際にかかる実行時間を考慮すると，提案手法の方がリアルタイム性，レイテンシー，メモリ要件の観点から優れていることがわかる．

5.　おわりに

本報告では，複数の教師モデルを用いた知識蒸留に関する効果的な方法（Switched-training，Augmented-training 戦略）を提案した．提案法は，小規模および中規模学習データの両方の条件において好適に動作し，特に小規模データでのモデル構築において大きな効果を発揮する．入力データについてバリエーションを持たせる通常のデータ増強処理とは対照的に，筆者らの知る限り，同一入力データに対して様々な解釈を持たせるべくターゲット側の増強処理を行う試みはこれまでになく，この新しい戦略によって性能を大幅に改善させることに成功した．Aurora 4 を用いた実験では，異なるドメインで学習された汎用教師モデルが，特定ドメイン専用の生徒モデルの改善に大きく寄与することを示した．また，シンプルかつコンパクトな生徒モデルが，VGG や BLSTM のような複雑で大規模な教師モデルに匹敵する性能を達成できることを示した．

[*1] 本章での “教師” モデルは正確には “汎用” モデルと称するべきであるが，本報告全体の用語統一の観点からそのまま教師モデルと称することにする．

表 6　狭帯域生徒モデルの性能

Teachers	Target	A	B	C	D	AVG
VGG8k	soft	6.4	12.4	8.3	17.0	13.6
VGG16k	soft	6.0	11.4	7.8	16.3	12.9
VGG8k + VGG16k	soft/augment	5.7	11.1	7.2	15.6	12.4
Hard + VGG8k	soft/augment	3.9	8.9	5.5	13.6	10.3
Hard + VGG16k	soft/augment	4.0	8.6	5.5	13.7	10.2
Hard + VGG8k + VGG16k	soft/switching	4.3	8.6	5.6	13.2	10.1
Hard + VGG8k + VGG16k	soft/augment	3.8	8.2	5.5	12.9	9.7
Hard + VGG8k + VGG16k + BLSTM8k + BLSTM16k	soft/switching	4.5	8.8	5.9	13.3	10.2
Hard + VGG8k + VGG16k + BLSTM8k + BLSTM16k	soft/augment	3.8	8.3	5.3	12.8	9.7

表 7　教師モデルによるシステムコンビネーションの結果

Teachers	A	B	C	D	AVG
VGG8k+BLSTM8k	5.8	10.7	13.6	6.7	11.3
VGG16k+BLSTM16k	4.7	8.2	6.2	13.9	10.3
VGG8k+VGG16k + BLSTM8k+BLSTM16k	4.8	8.0	5.7	13.0	9.7

参考文献

[1] Soltau, H., Saon, G. and Sainath, T. N.: Joint Training of Convolutional and Non-convolutional Neural Networks, *Proc. IEEE ICASSP*, pp. 5609–5613 (2014).
[2] Fukuda, T., Ichikawa, O., Kurata, G., Tachibana, R., Thomas, S. and Ramabhadran, B.: Efficient Knowledge Distillation from an Ensemble of Teachers, *Proc. Interspeech*, pp. 3697–3701 (2017).
[3] Hsiao et al., R.: Robust Speech Recognition in Unknown Reverberant and Noisy Conditions, *Proc. IEEE ASRU* (2015).
[4] Ba, J. and Caruana, R.: Do Deep Nets Really Need to be Deep?, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc, pp. 2654–2662 (2014).
[5] Vapnik, V. and Izmailov, R.: Learning Using Privileged Information: Similarity Control and Knowledge Transfer, *Machine Learning Research*, Vol. 16, pp. 2023–2049 (2015).
[6] Geras, K. J., Mohamed, A.-R., Caruana, R., Urban, G., Wang, S., Aslan, O., Philipose, M., Richardson, M. and Sutton, C.: Blending LSTMs into CNNs, *ICLR Workshop* (2016).
[7] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *arXiv:1503.02531v1* (2015).
[8] Chan, W., Ke, N. R. and Lane, I.: Transferring Knowledge from a RNN to a DNN, *Proc. Interspeech*, pp. 3264–3268 (2015).
[9] Li, J., Zhao, R., Huang, J.-T. and Gong, Y.: Learning Small-Size DNN with Output-Distribution-Based Criteria, *Proc. Interspeech*, pp. 1910–1914 (2014).
[10] Tang, Z., Wang, D. and Zhang, Z.: Recurrent Neural Network Training with Dark Knowledge Transfer, *Proc. IEEE ICASSP*, pp. 5900–5904 (2016).
[11] Cui, J., Kingsbury, B., Ramabhadran, B., Saon, G., Sercu, T., Audhkhasi, K., Sethy, A., Nussbaum-Thom, M. and Rosenberg, A.: Knowledge Distillation Across Ensembles of Multilingual Models for Low-Resource Languages, *Proc. IEEE ICASSP*, pp. 4825–4829 (2017).
[12] Wong, J. H. M. and Gales, M. J. F.: Sequence Student-Teacher Training of Deep Neural Networks, *Proc. Interspeech*, pp. 2761–2765 (2016).
[13] Markov, K. and Matsui, T.: Robust Speech Recognition using Generalized Distillation Framework, *Proc. Interspeech*, pp. 2364–2368 (2016).
[14] Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J. and Ramabhadran, B.: Effective Joint Training of Denoising Featuer Space Transforms and Neural Network Based Acoustic Models, *Proc. IEEE ICASSP*, pp. 5190–5194 (2017).
[15] Chebotar, Y. and Waters, A.: Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition, *Proc. Interspeech*, pp. 3439–3443 (2016).
[16] Brandchain, L.: The Mixer 6 Corpus: Resource for Cross-Channel and Text Independent Speaker Recognition, *LREC* (2010).
[17] Carletta, J.: Unleashing the Killer Corpus: Experiences in Creating the Multi-everything AMI Meeting Corpus, *Language Resources and Evaluation*, Vol. 41, No. 1, pp. 181–190 (2007).
[18] Itahashi, S.: Recent Speech Database Projects in Japan, *Proc. ICSLP* (1990).
[19] Nakamura, S., Hiyane, K., Asano, F., Nishiura, T. and Yamada, T.: Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-free Speech Recognition, *LREC* (2000).
[20] Fukuda, T., Ichikawa, O. and Nishimura, M.: Long-term Spectro-temporal and Static Harmonic Features for Voice Activity Detection, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 5, pp. 834–844 (2010).
[21] Harper, M.: The Automatic Speech Recognition in Reverberant Environments (ASpIRE) Challenge, *Proc. IEEE ASRU*, pp. 547–554 (2015).