Vol.2018-SLP-124 No.1
2018/10/10

CTC音響モデルのためのシーケンスレベル知識蒸留法の検討

高島 遼一[1,†1,a)]　李 勝[1]　河井 恒[1]

概要：本報告では，CTC 音響モデルのための知識蒸留法について検討する．従来の知識蒸留法では音声フレームごと独立に，教師モデルとのクロスエントロピーを最小化することで生徒モデルを学習しているが，我々の先行研究により，従来のフレームレベルでの知識蒸留法は CTC 音響モデルには有効に働かず，発話シーケンス単位でクロスエントロピーを最小化するシーケンスレベル知識蒸留法であれば有効に働くことが明らかとなった．本報告では，シーケンスレベル知識蒸留法の実装方法として，N-best ベースの手法とラティスベースの手法について検討し，比較を行う．またモデル圧縮実験および耐雑音音響モデル学習の実験において，シーケンスレベル知識蒸留法の性能を評価する．

キーワード：音声認識，音響モデル，connectionist temporal classification, 知識蒸留

Sequence-level Knowledge Distillation for CTC Acoustic models

Takashima Ryoichi[1,†1,a)]　Li Sheng[1]　Kawai Hisashi[1]

1. はじめに

音声認識の分野において，従来の deep neural network-hidden Markov model (DNN-HMM) ハイブリッドモデルに代わる音響モデルとして，connectionist temporal classification (CTC) [1], [2] を用いた音響モデルが研究されている [3], [4], [5], [6], [7], [8], [9], [10], [11]．DNN-HMM では，音声の各フレームに何らかのラベルが付与されることを前提としているため，学習時にはフレームごとのラベル情報が，音声認識時には HMM を用いてフレームごとのラベル推定値から最終的な出力系列への変換が必要である．一方 CTC の枠組みでは，フレームごとにラベルを付与することは前提としておらず，音声系列から直接出力系列を推定する recurrent neural network (RNN) を学習し，従って認識時は HMM によるネットワーク出力の変換が不要である．また DNN-HMM ではラベルが，例えばトライフォンの HMM 状態のような時間的に細かい単位で定義されているのに対して，CTC ではフレーム単位でラベルを定義する必要がないため，モノフォンや文字，単語といった時間的に粗い単位でラベルを定義することが可能である．以上の特徴から，CTC は辞書や言語モデルを用いない End-to-end 音声認識 [10], [11] や，音響イベント検出 [12], [13] など様々な用途に用いられている．

知識蒸留 (knowledge distillation; KD)[14], [15] は teacher-student 学習とも呼ばれる，ニューラルネットワークの学習方法の一つである．KD は，あるモデル (生徒モデルと呼ぶ) を学習する際に，正解ラベルを教師信号にする代わりに，生徒モデルよりも高性能なモデル (教師モデル) の出力を教師信号とすることで，生徒モデルを教師モデルの性能に近くなるように学習する手法である．KD は主にニューラルネットワークのモデル圧縮に使われており，そこでは高性能・大規模なモデルを教師モデル，小規模なモデルを生徒モデルとしている [16], [17], [18], [19]．また，

[1] 情報通信研究機構
National Institute of Information and Communications Technology (NICT), 3–5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0289 Japan
[†1] 現在，日立製作所
Presently with Hitachi Ltd.
[a)] ryoichi.takashima.dh@hitachi.com

耐雑音音響モデルの学習にも使われている [20], [21], [22]．そこではクリーン音声と雑音音声のパラレルデータを用意し，クリーン音声を用いて学習させたモデルを教師モデルとして，雑音音声を用いて生徒モデルを学習させることで，従来のマルチコンディション学習よりも良い性能を示すことが報告されている．

KD は DNN-HMM を用いた音声認識において効果があることが報告されているが，CTC 音響モデル対してはむしろ性能が劣化してしまった例が報告されている [23]．CTC に KD が適用できれば，例えば，認識率は高いがリアルタイム処理が困難な bidirectional-RNN ベースの CTC を教師モデルとして，unidirectional-RNN ベースの CTC を学習することで，高性能かつリアルタイム処理可能な End-to-end 音声認識の実現などの応用が期待される．そこで我々の先行研究 [24] では，CTC 音響モデルのための KD 手法として，フレームレベルの KD 方式とシーケンスレベルの KD 方式を検討した．フレームレベル KD は従来の KD 方式と同様に，フレームごとに教師・生徒モデルのクロスエントロピーを最小化することで学習を行う．一方シーケンスレベル KD は発話シーケンス単位でクロスエントロピーを最小化する．比較実験の結果，従来のフレームレベル KD は CTC に適用すると性能が劣化し，シーケンスレベル KD であれば CTC の性能を改善可能であることを明らかにした．

我々の先行研究では，シーケンスレベル KD 方式として N-best ベースの手法を提案したが，この手法は通常の CTC の学習に対して処理量が N 倍 (N は N-best 仮説の数) に増えるため，学習に要する時間の問題により，仮説数と性能の関係が評価できていなかった．本報告では，N-best ベースの手法を GPU で効率的に計算させることで，先行研究時よりも多くの仮説数で評価を行った．さらに計算効率化のため，ラティスベースの手法を提案し，N-best ベースの手法と認識率・処理速度の観点で比較を行う．また，モデル圧縮の実験と耐雑音音響モデル学習の実験を行い，提案手法の有効性を確認する．

2. 関連研究

2.1 Connectionist temporal classification

一般に，音声認識ではフレームごとに得られるラベル（パス π と呼ぶ）を，フレーム数以下の長さのラベル系列 $\mathbf{l}$ に変換する必要がある．CTC[1] では，同一ラベルの繰り返しの削除と blank ラベル (=ラベル無し) を導入することで，パスからラベル系列への変換を行っている．ここで，この変換関数を $\mathcal{B}$ (ただし $\mathbf{l} = \mathcal{B}(\pi)$) と定義する．同一のラベル系列を出力するパスは複数存在するため，音声系列 $\mathbf{x}$ を入力したときのラベル系列 $\mathbf{l}$ の出力確率は，そのラベル系列に変換される全パスの出力確率の総和で表される．

$$p(\mathbf{l}|\mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi|\mathbf{x}) \tag{1}$$

パス π の出力確率は以下のように計算される．

$$p(\pi|\mathbf{x}) = \prod_{t}^{T} y_{\pi_t}^t \tag{2}$$

ここで $y_{\pi_t}^t$ はフレーム t における RNN の π_t に関するノードの出力値を表す．T はフレーム数である．

CTC 音響モデルは，ラベル系列に対する尤度最大基準から導出される，以下の損失関数を最小化するように学習する．

$$\mathcal{L}_{\mathrm{CTC}} = -\sum_{(\mathbf{x},\mathbf{l})\in Z} \ln p(\mathbf{l}|\mathbf{x}) = \sum_{(\mathbf{x},\mathbf{l})\in Z} \mathcal{F}_{\mathrm{CTC}}(\mathbf{l}|\mathbf{x}) \tag{3}$$

Z は学習データのセットを表す．$\mathcal{F}_{\mathrm{CTC}}(\mathbf{l}|\mathbf{x}) = -\ln p(\mathbf{l}|\mathbf{x})$ は発話ごとの損失関数で，3 節での説明のため定義している．ラベルを k としたとき，back propagation に必要となる，出力ノード y_k^t に関する $\mathcal{L}_{\mathrm{CTC}}$ の勾配は以下のように計算される．

$$\frac{\partial \mathcal{L}_{\mathrm{CTC}}}{\partial y_k^t} = -\frac{1}{p(\mathbf{l}|\mathbf{x})}\frac{1}{y_k^t}\sum_{\pi:(\mathcal{B}(\pi)=\mathbf{l},\pi_t=k)} p(\pi|\mathbf{x}) \tag{4}$$

$\sum_{\pi:(\mathcal{B}(\pi)=\mathbf{l},\pi_t=k)} p(\pi|\mathbf{x})$ は，ラベル系列 $\mathbf{l}$ を出力し，かつフレーム t においてラベル k を通る全てのパスの確率総和を意味する．この確率は forward-backward アルゴリズムによって次式のように計算される．

$$\sum_{\pi:(\mathcal{B}(\pi)=\mathbf{l},\pi_t=k)} p(\pi|\mathbf{x}) = \sum_{s:(l_s=k)} \frac{\alpha_t(s)\beta_t(s)}{y_s^t} \tag{5}$$

$\alpha_t(s)$，$\beta_t(s)$ はそれぞれ，t 番目のフレームにおいて，ラベル系列 $\mathbf{l}$ 中の s 番目のラベルを通るパスの前向き，後ろ向き確率である．$s:(l_s = k)$ は $\mathbf{l}$ 内におけるラベル k の位置を表す．forward-backward アルゴリズムの詳細は文献 [1] を参照されたい．

2.2 知識蒸留

知識蒸留 (KD)[15] はニューラルネットワークの学習方法の一つで，あるモデル (生徒モデル) を学習する際に，正解ラベルを教師信号にする代わりに，生徒モデルよりも高性能なモデル (教師モデル) の出力を教師信号とすることで，教師モデルと似た識別能力を持つ生徒モデルを学習させるという手法である．KD ではまず正解ラベルを用いて教師モデルを学習させる．そして学習データに対する教師モデルの出力 (確率分布) を用いて，生徒モデルをクロスエントロピー基準により学習させる．

$$\mathcal{L}_{\mathrm{KD}} = -\sum_{l} p_{\mathrm{tea}}(l|x) \ln p_{\mathrm{stu}}(l|x) \tag{6}$$

$p_{\mathrm{tea}}(l|x)$ および $p_{\mathrm{stu}}(l|x)$ は，それぞれ教師モデルおよび生徒モデルによって推定された，入力 x に対するラベル l の出力確率を表す．

音声認識における主な従来研究では，式 (6) をフレーム

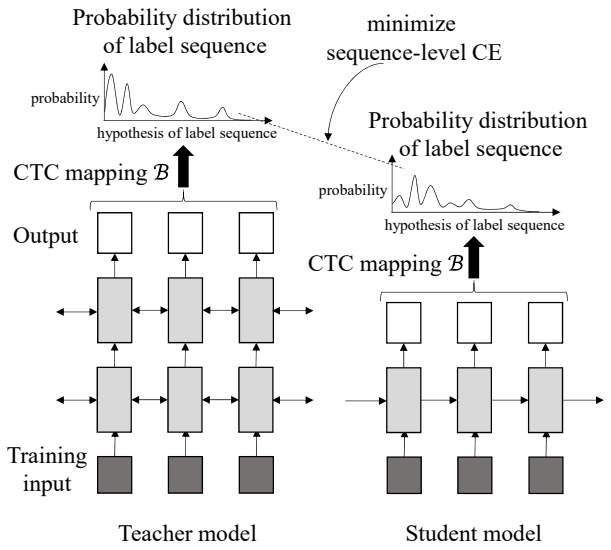

図 1 CTC モデルのためのシーケンスレベル KD 方式

単位で定義し，クロスエントロピー学習による DNN-HMM に対して適用している．一方，MMI 学習への適用 [25] や，機械翻訳における attention モデルへの適用 [26] を目的としたシーケンスレベルでの KD 方式も提案されており，そこではラベル単体ではなく，ラベル系列の出力確率分布を用いて，クロスエントロピーを計算している．

3. CTC 音響モデルのためのシーケンスレベル KD

3.1 概要

我々の先行研究 [24] では，CTC 音響モデルのための KD 手法を検討し，従来のフレームレベルの KD 方式をそのまま CTC モデルに適用しても有効に働かないが，一方シーケンスレベルの KD 方式であれば有効に働くことを明らかにした．本報告ではシーケンスレベル KD の実装方法についてさらに検討する．図 1 に CTC モデルのためのシーケンスレベル KD 方式の概要を示す．関連研究 [25], [26] に倣って，まず学習済みの教師 CTC モデルを用いて，学習データに対するラベル系列の仮説とその出力確率を求める．そして得られた仮説と出力確率を用いて，生徒 CTC モデルを，次式で定義されるシーケンス単位のクロスエントロピーを最小化することで学習する．

$$\mathcal{L}_{\mathrm{CTC-KD_{seq}}} = -\sum_{\mathbf{x}\in Z}\sum_{\mathbf{h}\in\mathcal{H}} p_{\mathrm{tea}}(\mathbf{h}|\mathbf{x}) \ln p_{\mathrm{stu}}(\mathbf{h}|\mathbf{x}) \quad (7)$$

$\mathbf{h}$ は入力系列 $\mathbf{x}$ に対するラベル系列の全仮説 $\mathcal{H}$ 中のある仮説を表す．$p_{\mathrm{tea}}(\mathbf{h}|\mathbf{x})$ および $p_{\mathrm{stu}}(\mathbf{h}|\mathbf{x})$ はそれぞれ教師モデルおよび生徒モデルによって推定された，仮説 $\mathbf{h}$ の出力確率を表す．式 (7) は，次式のようにも表現できる．

$$\mathcal{L}_{\mathrm{CTC-KD_{seq}}} = \sum_{\mathbf{x}\in Z}\sum_{\mathbf{h}\in\mathcal{H}} p_{\mathrm{tea}}(\mathbf{h}|\mathbf{x}) \mathcal{F}_{\mathrm{CTC}}(\mathbf{h}|\mathbf{x}) \quad (8)$$

よって，CTC のためのシーケンスレベル KD の損失関数は，各仮説を正解ラベル系列と見なしたときの CTC の損失関数を，その仮説の出力確率を重みとして加重平均したものと言える．ここで，全ての仮説に対する出力確率 $p_{\mathrm{tea}}(\mathbf{h}|\mathbf{x})$ を展開することは現実的に困難なため，限定された個数の仮説を用いて近似的にシーケンスレベル KD を実現することが求められる．本研究では，N-best 仮説を用いた実装方法と，ラティスを用いた実装方法を検討する．

3.2 N-best ベースのシーケンスレベル KD

N-best ベースのシーケンスレベル KD では，式 (8) を N-best 仮説，すなわち出力確率の最も高い N 個の仮説のみを用いて近似する．

$$\tilde{\mathcal{L}}_{\mathrm{CTC-KD_{Nbest}}} = \sum_{\mathbf{x}\in Z}\sum_{n=1}^{N} \tilde{p}_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x}) \mathcal{F}_{\mathrm{CTC}}(\mathbf{h}_n|\mathbf{x}) \quad (9)$$

$\mathbf{h}_n$ は，N-best 仮説中の n 番目の仮説を表し，$\tilde{p}_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x})$ は次式を用いて確率の総和が 1 となるよう正規化した出力確率を表す．

$$\tilde{p}_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x}) = \frac{p_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x})}{\sum_{n=1}^{N} p_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x})} \quad (10)$$

back propagation に必要となる，y_k^t に関する勾配は以下のように計算される．

$$\frac{\partial \mathcal{L}_{\mathrm{CTC-KD_{Nbest}}}}{\partial y_k^t} = -\sum_{n=1}^{N} \tilde{p}_{\mathrm{tea}}(\mathbf{h}_n|\mathbf{x}) \frac{1}{p_{\mathrm{stu}}(\mathbf{h}_n|\mathbf{x})} \frac{1}{y_k^t} \sum_{\substack{\pi:(\mathcal{B}(\pi)=\mathbf{h}_n,\\ \pi_t=k)}} p(\pi|\mathbf{x}) \quad (11)$$

式 (3) と式 (9)，および式 (4) と式 (11) より，N-best ベースのシーケンスレベル KD は，N-best 仮説の分だけ $\mathcal{F}_{\mathrm{CTC}}$ あるいは $\sum_{\pi:(\mathcal{B}(\pi)=\mathbf{h}_n,\pi_t=k)} p(\pi|\mathbf{x})$ を従来の CTC の学習方法で用いられている forward-backward アルゴリズムを用いて計算し，各仮説の確率で加重平均を取ることで実装できる．そのため，N-best ベースのシーケンスレベル KD は実装が比較的簡単である反面，従来の CTC 学習と比べて計算量が N 倍に増えるという欠点がある．各仮説に対する $\mathcal{F}_{\mathrm{CTC}}$ は仮説ごと独立に計算可能のため，GPU 上で並列計算させることで，一定数の仮説を用いる限りでは学習時間の増大を抑えることが可能である．しかし用いる N-best 仮説の数が増えるにつれて GPU コア数の圧迫による学習速度低下や，メモリ不足に陥るため，現実的に使用可能な仮説数は限られる．

3.3 ラティスベースのシーケンスレベル KD

N-best 仮説はそれぞれ類似したラベル系列になりがちなため，仮説ごと独立に forward-backward 計算をすると，重複した計算がしばしば含まれる．そこでラティスベース KD は複数の仮説をグラフ化し，グラフ上で forward-backward 計算を行うことで，同一計算の重複を防いで計算を効率化する．通常の CTC 学習，N-best ベース KD，ラティス

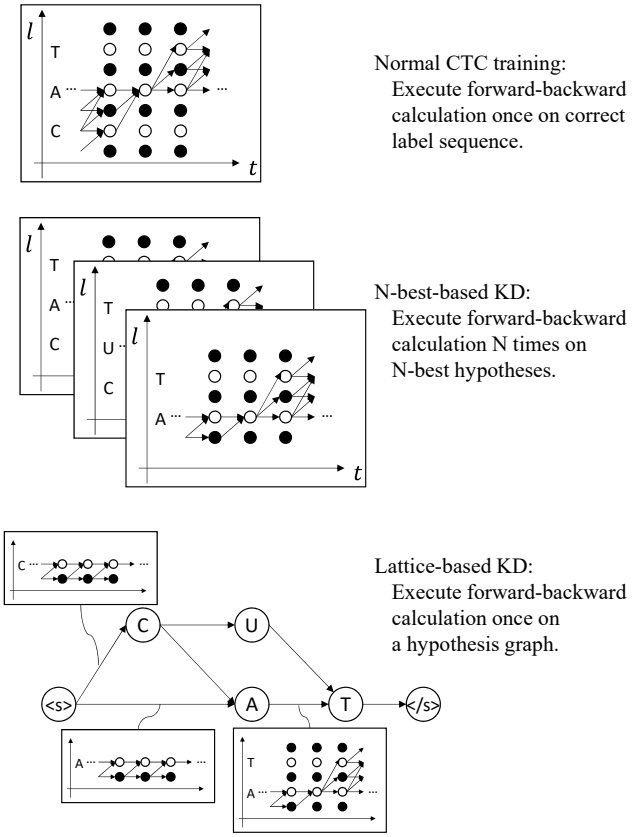

図 2 通常の CTC 学習，N-best ベース KD，ラティスベース KD それぞれにおける forward-backward 計算の概念図

ベース KD それぞれにおける forward-backward 計算の概念図を図 2 に示す．図の例では “CAT” と発声された音声に対して，“CAT”，“CUT”，“AT” の仮説が教師モデルから推定されている．N-best ベース KD では各仮説独立に forward-backward 計算を行うのに対して，ラティスベース KD では，初期状態から ‘C’ への遷移や ‘T’ から終了状態への遷移など，各仮説で共通する部分の計算が 1 回で済む分，仮説が増えても計算量の増大が抑えられると期待される．

ラティスベースでの forward-backward アルゴリズムを説明する．まず計算効率化のため，ラティスデータは事前に「遷移先ステート ID > 遷移元ステート ID」となるようにソートしておく．従来の CTC の forward-backward において，ラベル系列の各ラベルの間に blank を挿入していたのと同様に，ラティスベース forward-backward においても，ラティスの各ステートの間に blank を挿入したグラフ (拡張ラティスと呼ぶことにする) を考える．ラティスのステート ID を $n = 0, \ldots, N$ としたとき，拡張ラティスのステート ID は $2n$ であれば非 blank ラベル，$2n+1$ なら blank ラベルが割り当てられる．

前向き確率計算

前向き確率 $\alpha_t(s)$ の計算について説明する．まず初期化として $t = 0$ の前向き確率を以下のように定義する．

$$\begin{aligned} &\alpha_t(0) = 0, && \alpha_t(1) = y_t^{blk} \\ &\alpha_t(2n) = y_t^{lab(n)} p_{tea}(n|0), && \alpha_t(2n+1) = 0 \end{aligned} \tag{12}$$

y_t^{blk} は blank に関するネットワーク出力，$y_t^{lab(n)}$ は n ステートに割り当てられたラベルに関するネットワーク出力を表す．0 ステートは初期状態でラベルが存在しないため，$\alpha_t(0) = 0$ とする．$p_{tea}(n|m)$ は m ステートから n ステートへの遷移確率で，$m = 0$ のとき，つまり初期状態からの遷移確率と，$n = N$ のとき，つまり終了状態へ遷移確率のみ，1 か 0 の 2 値をとる．

次に $t > 0$ の場合を以下のように定義する．

$$\alpha_t(0) = 0, \qquad \alpha_t(1) = y_t^{blk} \alpha_{t-1}(1)$$

$$\begin{aligned} \alpha_t(2n) \;=\; & y_t^{lab(n)} \Bigg(\alpha_{t-1}(2n) \\ & + \sum_{\substack{m \in 1:n-1 \\ lab(n) \neq lab(m)}} p_{tea}(n|m)(\alpha_{t-1}(2m) + \alpha_{t-1}(2m+1)) \\ & + \sum_{\substack{m \in 1:n-1 \\ lab(n) = lab(m)}} p_{tea}(n|m)\alpha_{t-1}(2m+1) \Bigg) \end{aligned}$$

$$\alpha_t(2n+1) \;=\; y_t^{blk}(\alpha_{t-1}(2n) + \alpha_{t-1}(2n+1)) \tag{13}$$

従来の CTC 学習と異なり，ラティスベース方式では n ステートへ遷移するステートは複数存在し得るため，遷移元 $m \in 1 : n-1$ で総和を取っている点が，従来 CTC 学習における前向き確率計算との違いである．

後向き確率計算

後向き確率 $\beta_t(s)$ の計算について説明する．入力信号の最終フレーム $t = T-1$ における後ろ向き確率を以下のように定義する．

$$\begin{aligned} \beta_t(2n) \;&=\; y_t^{lab(n)} p_{tea}(N|n) \\ \beta_t(2n+1) \;&=\; y_t^{blk} p_{tea}(N|n) \end{aligned} \tag{14}$$

$p_{tea}(N|n)$ は終了状態への遷移確率で，1 か 0 の 2 値をとる．また，$p_{tea}(N|N) = 0$ である．

$t < T-1$ の場合は以下のように定義される．

$$\begin{aligned} \beta_t(2n) \;=\; & y_t^{lab(n)} \Bigg(\beta_{t+1}(2n) + \beta_{t+1}(2n+1) \\ & + \sum_{\substack{m \in n+1:N-1 \\ lab(n) \neq lab(m)}} p_{tea}(m|n)\beta_{t+1}(2m) \Bigg) \\ \beta_t(2n+1) \;=\; & y_t^{blk} \Bigg(\beta_{t+1}(2n+1) \\ & + \sum_{m \in n+1:N-1} p_{tea}(m|n)\beta_{t+1}(2m) \Bigg) \end{aligned} \tag{15}$$

4. 実験

4.1 音響モデル圧縮の実験

提案する KD 方式の有効性を評価するため，まず音響モデル圧縮の実験を行った．KD によるモデル圧縮は，大規模

なモデルを教師モデル，小規模なモデルを生徒モデルとして KD を行うことで，より性能の高い生徒モデルを学習することを目的とする．本実験では，教師 CTC モデルを中間層数 5，メモリセル数 320 の bidirectional long short-taerm memory (LSTM) [27] で定義し，生徒 CTC モデルを中間層数 3，メモリセル数 320 の unidirectional LSTM で定義した．

データセットは WSJ コーパス [28] を用いた．学習データとして “WSJ0” (train_si84，15 時間) を使用し，評価データとして “dev93” および “eval92” を用いた．入力特徴量として，40 次元のメルフィルタバンク特徴とその 1 次，2 次デルタ特徴（計 120 次元）を用い，出力は 72 種類のラベル（69 モノフォン+ノイズ 2 種 (SPN, NSN)+blank）で定義した．学習率は初期値を 0.0004，最終値を 0.000004 として，エポックごとに指数関数に従って減少させた．エポック数は 15 とした．学習において，GPU は NVIDIA Tesla P100 PCIe 16GB を使用した．

CTC の学習・評価ツールとして EESEN ツールキット [5] を使用した．また，フレームレベル KD，シーケンスレベル KD(N-best ベース，ラティスベース) は全て EESEN ツール上で実装した．シーケンスレベル KD においては，EESEN を使用して WFST ベースのビームサーチ [29] により，仮説とその出力確率を求めた．ただしこの処理においては辞書と言語モデルは使用せず，トークン WFST と呼ばれる，フレームごとの出力をラベル系列へ変換する WFST(2.1 節の $\mathcal{B}$ に相当) のみを用いた．

実験結果を表 1 に示す．word error rate (WER [%]) 算出時は辞書および言語モデル (lm_tgpr) を使用しており，phone error rate (PER [%]) 算出時は辞書も言語モデルも使用していない．frame per second (fps) は，学習時において 1 秒間に処理したフレーム数を表し，高いほど学習が高速であることを意味する．表より，フレームレベル KD を用いた場合，KD を用いない場合に比べて生徒モデルの認識率が悪化することが分かる．N-best ベースのシーケンスレベル KD を用いた場合，$N = 10$ 以上において認識率の改善が見られ，また N が大きいほど性能が改善された．ラティスベースのシーケンスレベル KD を用いた場合，辞書・言語モデルを使用せずに PER を評価した場合では最も良い結果となったが，辞書・言語モデルを使用して WER を評価した場合は，KD を用いない場合とほとんど同等の性能となった．また，N-best ベースの方式では $N = 50$ において学習速度が半分程度に低下したが，ラティスベースの方式では学習速度の低下は比較的抑えられていた．

4.2 耐雑音音響モデル学習の実験

KD による耐雑音音響モデル学習実験の概要を図 3 に示す．この実験は文献 [22] に倣い，学習データとして，クリーン音声と，それに雑音を重畳させた雑音音声からなる，パラレルデータを使用する．まずクリーン音声と正解ラベル系列を使用して教師モデルを学習する．次に学習時に用いたクリーン音声を教師モデルに入力し，クリーン学習データに対する仮説および出力確率を計算する．生徒モデルを学習する際は，クリーン音声の代わりに雑音重畳音声を用いて，上記で求めたクリーン学習データに対する仮説・出力確率を教師信号として KD を行う．

表 1　WSJ データセットを用いたモデル圧縮の実験結果

Acoustic Model	KD method	WER w/ LM		PER w/o LM		fps. in training
		dev93	eval92	dev93	eval92	
teacher CTC	none	17.03	10.79	21.19	15.38	1306.3
student CTC	none	20.31	13.59	29.76	24.16	2318.7
	1-best	21.86	14.85	29.87	24.01	1982.0
	10-best	20.46	13.47	28.22	22.46	2324.6
	50-best	**19.99**	**12.94**	28.17	22.13	1288.9
	lattice	20.59	13.52	**28.04**	**21.94**	1879.9
	frame	26.60	17.54	37.94	33.06	4350.1

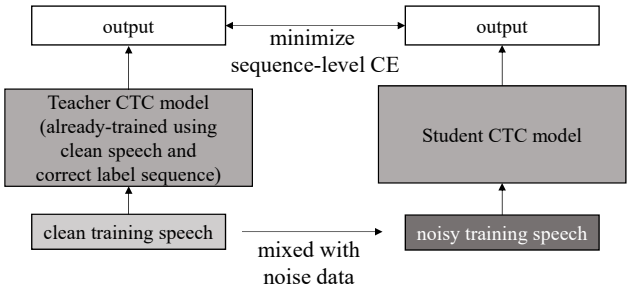

図 3　KD を用いた耐雑音音響モデル学習方法

表 2　CHiME4 データセットを用いた耐雑音音響モデル学習の実験結果

Acoustic Model	KD method	WER w/ LM		PER w/o LM	
		dt05_simu	et05_real	dt05_simu	et05_real
student CTC	none	24.99	54.93	37.65	59.39
	50-best	**22.13**	**52.45**	**33.63**	**55.82**
	lattice	22.99	54.05	33.98	56.27

教師 CTC モデル，および生徒 CTC モデルはどちらも中間層数 5，メモリセル数 160 の bidirectional LSTM を使用した．データセットは CHiME4 コーパス [30] を用いた．CHiME4 コーパスの学習データである “tr05_simu_noisy” セットは WSJ コーパスの WSJ0 セット (train_si84) に雑音を重畳させたデータである．そのため，パラレル学習データとして WSJ0 をクリーン音声，“tr05_simu_noisy” を雑音重畳音声として使用した (それぞれ約 15 時間)．評価セットには “dt05_simu_noisy” と “et05_real_isolated_1ch_track” を使用した．他の実験条件については，4.1 節と同じである．

実験結果を表 2 に示す．表中において KD method が none のモデルでは，従来のマルチコンディション学習と同様に，“tr05_simu_noisy” を学習データとして，正解ラベルを教師信号として学習させている．表より，PER，WER ともにシーケンスレベル KD を用いることによって，通常

のマルチコンディション学習よりも良い性能が得られた．

5. おわりに

本報告では，CTC 音響モデルのための KD 方式としてシーケンスレベル KD について検討し，N-best ベースの方式とラティスベースの方式を提案した．実験の結果，どちらの方式も KD としての効果を確認できた．N-best ベースの方式は用いる仮説数に応じて学習時間も増加する一方，ラティスベースの方式は学習時間の増加は比較的抑えられることを確認した．しかしながら，PER の評価では性能が上がっても，言語モデルと併用して WER を評価した場合では性能が改善されないケースもあった．今後は言語モデルとの統合方法について調査し，さらに temperature などの KD の周辺技術の適用方法についても検討を行う．

参考文献

[1] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *ICML2006*, ACM, pp. 369–376 (2006).
[2] Graves, A. and Jaitly, N.: Towards End-To-End Speech Recognition with Recurrent Neural Networks., *ICML2014*, Vol. 14, pp. 1764–1772 (2014).
[3] Sak, H., Senior, A., Rao, K. and Beaufays, F.: Fast and accurate recurrent neural network acoustic models for speech recognition, *Interspeech*, ISCA, pp. 1468–1472 (2015).
[4] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G. et al.: Deep speech 2: End-to-end speech recognition in english and mandarin, *ICML2016*, pp. 173–182 (2016).
[5] Miao, Y., Gowayyed, M. and Metze, F.: EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding, *ASRU2015*, IEEE, pp. 167–174 (2015).
[6] Kanda, N., Lu, X. and Kawai, H.: Maximum-a-Posteriori-Based Decoding for End-to-End Acoustic Models, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 5, pp. 1023–1034 (2017).
[7] Kanda, N., Lu, X. and Kawai, H.: Mininum bayes risk training of CTC acoustic models in maximum a posteriori based decoding framework, *ICASSP2017*, IEEE, pp. 4855–4859 (2017).
[8] Kim, S., Hori, T. and Watanabe, S.: Joint ctc-attention based end-to-end speech recognition using multi-task learning, *ICASSP2017*, IEEE, pp. 4835–4839 (2017).
[9] Hori, T., Watanabe, S., Zhang, Y. and Chan, W.: Advances in Joint CTC-Attention based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM, *Interspeech*, ISCA, pp. 949–953 (2017).
[10] Audhkhasi, K., Kingsbury, B., Ramabhadran, B., Saon, G. and Picheny, M.: Building competitive direct acoustics-to-word models for English conversational speech recognition, *ICASSP2018*, IEEE, pp. 4759–4763 (2018).
[11] Das, A., Li, J., Zhao, R. and Gong, Y.: Advancing connectionist temporal classification with attention modeling, *ICASSP2018*, IEEE, pp. 4769–4773 (2018).
[12] Wang, Y. and Metze, F.: A first attempt at polyphonic sound event detection using connectionist temporal classification, *ICASSP2017*, pp. 2986–2990 (2017).
[13] Fujimura, H., Nagao, M. and Masuko, T.: Simultaneous speech recognition and acoustic event detection using an LSTM-CTC acoustic model and a WFST decoder, *ICASSP2018*, IEEE, pp. 5834–5838 (2018).
[14] Li, J., Zhao, R., Huang, J.-T. and Gong, Y.: Learning Small-Size DNN with Output-Distribution-Based Criteria, *Interspeech*, pp. 1911–1914 (2014).
[15] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *NIPS Deep Learning and Representation Learning Workshop* (2014).
[16] Chebotar, Y. and Waters, A.: Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition, *Interspeech*, pp. 3439–3443 (2016).
[17] Lu, L., Guo, M. and Renals, S.: Knowledge distillation for small-footprint highway networks, *ICASSP2017*, pp. 4820–4824 (2017).
[18] Watanabe, S., Hori, T., Roux, J. L. and Hershey, J. R.: Student-teacher network learning with enhanced features, *ICASSP2017*, pp. 5275–5279 (2017).
[19] Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J. and Ramabhadran, B.: Efficient Knowledge Distillation from an Ensemble of Teachers, *Interspeech*, pp. 3697–3701 (2017).
[20] Li, J., Zhao, R., Chen, Z., Liu, C., Xiao, X., Ye, G. and Gong, Y.: Developing far-field speaker system via teacher-student learning, *ICASSP2018*, IEEE, pp. 5699–5703 (2018).
[21] Tan, T., Qian, Y. and Yu, D.: Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition, *ICASSP2018*, IEEE, pp. 5714–5718 (2018).
[22] Kim, J., El-Khamy, M. and Lee, J.: Bridgenets: student-teacher transfer learning based on recursive neural networks and its application to distant speech recognition, *ICASSP2018*, IEEE, pp. 5719–5723 (2018).
[23] Senior, A., Sak, H., Quitry, F. d. C., Sainath, T. N. and Rao, K.: Acoustic Modelling with CD-CTC-SMBR LSTM RNNS, *ASRU2015*, IEEE, pp. 604–609 (2015).
[24] Takashima, R., Li, S. and Kawai, H.: An investigation of a knowledge distillation method for CTC acoustic models, *ICASSP2018*, IEEE, pp. 5809–5813 (2018).
[25] Wong, J. H. M. and Gales, M. J. F.: Sequence Student-Teacher Training of Deep Neural Networks, *Interspeech* (2016).
[26] Kim, Y. and Rush, A. M.: Sequence-Level Knowledge Distillation, *EMNLP2016* (2016).
[27] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).
[28] et al., J. G.: CSR-I (WSJ0) Complete LDC93S6A, DVD. Philadelphia: Linguistic Data Consortium (1993).
[29] Mohri, M., Pereira, F. and Riley, M.: Weighted finite-state transducers in speech recognition, *Computer Speech & Language*, Vol. 16, No. 1, pp. 69–88 (2002).
[30] Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J. and Marxer, R.: An analysis of environment, microphone and data simulation mismatches in robust speech recognition, *Computer Speech & Language* (2016).