Vol.2018-MUS-118 No.16
Vol.2018-SLP-120 No.16
2018/2/21

CTCによる文字単位のモデルを併用したAttentionによる単語単位のEnd-to-End音声認識

上乃 聖[1,a)]　稲熊 寛文[1]　三村 正人[1]　河原 達也[1]

概要：End-to-End 音声認識が従来の DNN-HMM ハイブリッド音声認識よりも高速で簡潔であることから注目されている．特に入力の音響特徴量から出力の単語列に直接変換する単語単位 End-to-End 音声認識は外部の言語モデルが必要なく，更なる簡潔性が期待される．しかし，出現頻度の低い単語に関する学習データのスパース性が問題となる．そこで本稿では文字を出力単位としたモデルを併用する単語単位モデルを提案する．文字単位モデルを併せて学習することで単語単位モデルのオーバーフィットを軽減することが期待できる．また，単語単位モデルが未知語を出力した際に文字単位モデルで対応する文字列を参照することで未知語の推定を行う．提案手法を「日本語話し言葉コーパス」(CSJ) で評価を行なった結果，従来のハイブリッド音声認識よりも非常に速い処理時間で同等以上の認識精度を実現し，さらに種々の改善手法により高い性能が得られた．

キーワード：End-to-End 音声認識，Connectionist Temporal Classification (CTC)，
注意機構 (attention mechanism)，マルチタスク学習

1. はじめに

従来の DNN-HMM ハイブリッド音声認識は双方向 LSTM や ResNet を用いることで特定のタスクでは人間レベルの認識精度を達成することが報告されている ?．しかしハイブリッド音声認識は，非常に複雑なデコーダ，大規模な言語モデルや綿密に設計された発音辞書などを必要とする．それに対して，簡単な構造で実現できる End-to-End 音声認識に関する研究が近年行われている．End-to-End 音声認識は LSTM のような RNN を用いて直接音響特徴量から目的の記号 (音素や文字など) に変換するモデルで，高速な認識が実現できる．End-to-End 音声認識に用いられるアプローチは Connectionist Temporal Classification (CTC) ？？？？？？と注意機構 (attention mechanisim) ？？？？？の 2 つに分類される．従来の一般的な End-to-End 音声認識は出力単位を音素や音節、文字などのサブワードとしており，単語系列を出力する際には依然として発音辞書や言語モデルなどの外部機構を必要とする ?．これに対して，音響特徴量から単語を直接出力するモデルについても研究されている？？？？？？．RNN を用いることで言語モデルも包含できるので，言語モデルは必要でなくなり，非常に簡潔で高速な認識が実現できる．本稿では「日本語話し言葉コーパス」(CSJ) を用いて CTC と注意機構モデルの比較を行う．

しかし単語単位モデルでは学習データのスパース性と，未知語が認識不可能なことが問題として挙げられる．コーパス内の単語の種類数はサブワードに比べて非常に多くなり，出現回数の偏りも大きい．これにより学習に十分な量の学習データが得られない単語が多くなる．この問題を解決するためには，出現頻度の低い単語を学習から除外する必要があり，初めから学習データに含まれていない未知語に加えて，出現頻度の低い単語についても未知語の扱いとなり，これらを認識することができなくなる．

そこで本稿では，文字を出力系列とするモデルを併用した単語単位 End-to-End 音声認識を提案する．文字を出力単位とするモデルを併用することで学習データのスパース性の問題を緩和することができる．また文字単位モデルには未知語はなく，何かの認識結果を出力できる．この利点を利用して，単語単位モデルが未知語を出力した際の未知語の推定に文字単位モデルを利用する．図 1 に提案手法の概要を示す．

2. End-to-End 音声認識

End-to-End 音声認識には主に 2 つのアプローチが用いられている．1 つは CTC (Connectionist Temporal Classification) を使う手法であり，もう 1 つは注意機構 (attention

[1] 京都大学 大学院情報学研究科
[a)] ueno@sap.ist.i.kyoto-u.ac.jp

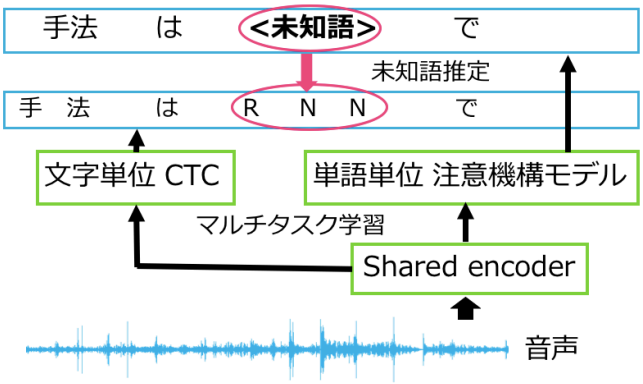

図 1 提案手法の概念図．単語単位モデルと文字単位モデルのマルチタスク学習を行い，また単語単位のデコード時に未知語が推定された際に文字単位モデルを利用する．

mechanism) を用いたエンコーダデコーダモデルである．

本節では $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_T)$ を長さ T の入力音響特徴量の系列とし，. $\boldsymbol{y} = (y_1, ..., y_L)$ を長さ L の出力記号系列とする．ただし，$y_l \in \{1, ..., K\}$ で，K は想定する出力記号の種類数である．

2.1 Connectionist Temporal Classification (CTC)

CTC では，記号の間に”blank”(−) を挿入，あるいは各記号が連続して出力されることを許すことにより記号系列を入力の長さ T に拡張する．各 CTC のパス $\boldsymbol{\pi} = (\pi_1, ..., \pi_T)$ は全てのラベルの繰り返しと”−” を消去することにより元の系列 $\boldsymbol{y}$ が復元される．また，$\pi_t \in \{1, ..., K\} \cup \{-\}$ である．CTC の損失関数は $\boldsymbol{\pi} \in \Omega(\boldsymbol{y})$ 内の全てのパスの確率の和として定義される．

$$p(\boldsymbol{y}|\boldsymbol{X}) = \sum_{\pi \in \Omega(\boldsymbol{y})} p(\pi|\boldsymbol{x}) = \sum_{\pi \in \Omega(\boldsymbol{y})} \prod_{t=1}^{T} p(\pi_t|\boldsymbol{x}_t) \quad (1)$$

ここで事後確率 $p(\pi_t|\boldsymbol{x}_t)$ は双方向 RNN を用いて計算される．CTC の損失関数と勾配は前向き後向きアルゴリズムにより計算される．ただし，各ラベルの出力確率は独立であるという仮定があるため記号間の関係性について学習できない．

2.2 注意機構モデル (attention mechanism)

注意機構を用いたモデルはエンコーダとデコーダの 2 つのサブネットワークから構成される．エンコーダでは LSTM のような RNN を用いて音響特徴量系列を長さ T の分散表現にする．このエンコードされた情報を基にデコーダは長さ L の記号系列を予測する．デコーダではエンコードされた系列表現の関連する度合いを注意機構を使用して計算することで記号系列を順次予測する．本稿ではエンコーダでは複数層の双方向 LSTM を用い，デコーダでは 1 層の単方向 LSTM を使用する．

注意機構の定式化は以下のようになる．エンコーダでは $\boldsymbol{X}$ を中間表現である $\boldsymbol{H} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_T)$ に変換する．デコーダでは l 番目のタイムステップの隠れ状態は以下のように計算される．

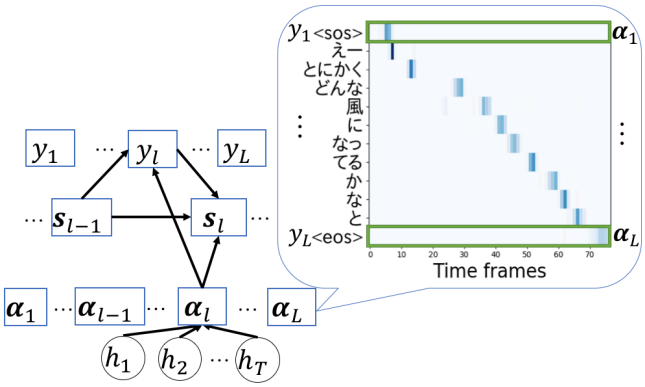

図 2 注意機構のモデル図．h_t はエンコーダの出力で，α_l は l 番目の記号 $\boldsymbol{s}_l$ がどの時間フレームに対応付けれるかを示す．

$$\boldsymbol{s}_l = Recurrency\left(\boldsymbol{s}_{l-1}, \boldsymbol{g}_l, \boldsymbol{y}_{l-1}\right) \quad (2)$$

ただし，$\boldsymbol{y}_{l-1}$ は前のステップの出力であり，$\boldsymbol{g}_l$ はエンコーダの出力の重み付き和で表される．

$$\boldsymbol{g}_l = \sum_t \alpha_{l,t} \boldsymbol{h}_t \quad (3)$$

ここで $\alpha_{l,t}$ は $\boldsymbol{h}_t$ の注意重みと呼ばれる．これは入力系列に対して出力記号系列のどこに対応付けられるかを示す確率であり，次のように計算される．

$$e_{l,t} = Score(\boldsymbol{s}_{l-1}, \boldsymbol{h}_t, \boldsymbol{\alpha}_{l-1}) \quad (4)$$

$$\alpha_{l,t} = \exp(e_{l,t}) / \sum_{t'=1}^{T} \exp(e_{l,t'}) \quad (5)$$

式 (4) にある Score 関数には様々な選択肢がある．本稿では次式で示す畳み込みを用いた注意機構を適用する．

$$e_{l,t} = \boldsymbol{w}^T tanh(\boldsymbol{W}\boldsymbol{s}_{l-1} + \boldsymbol{V}\boldsymbol{h}_t + \boldsymbol{U}f_{l,t} + \boldsymbol{b}) \quad (6)$$

$$\boldsymbol{f}_l = \boldsymbol{F} * \boldsymbol{\alpha}_{l-1} \quad (7)$$

ここで $*$ は 1 次元の畳み込みを示す．$\boldsymbol{g}_l$，$\boldsymbol{y}_{l-1}$ を使って次の記号 $\boldsymbol{y}_l$ を予測する．

$$\boldsymbol{y}_l \sim Generate\left(\boldsymbol{s}_{l-1}, \boldsymbol{g}_l\right) \quad (8)$$

本稿では Generate 関数は以下のように実装される．

$$\boldsymbol{R}\tanh\left(\boldsymbol{P}\boldsymbol{s}_{l-1} + \boldsymbol{Q}\boldsymbol{g}_l\right) \quad (9)$$

図 2 に注意機構のモデル図を示す．注意機構モデルの目的関数は予測記号系列と正解記号系列間の交差エントロピーとなる．注意機構モデルを用いた End-to-End 音声認識では start-of-sentence(sos) と end-of-sentence(eos) の特殊な記号を用いる．デコーダでは eos が出力された時点で処理を完了する．

2.3 単語単位モデル

近年，双方向 LSTM を用いて単語を出力単位とする End-to-End 音声認識モデルの研究も行われている？？？．この単語単位モデルの利点はデコード時間が短くなる点に加えて非常に簡潔な構造で実現できる点である．外部デコーダや言語モデル，発音辞書は必要なく，ニューラルネットワークの出力を抽出するだけで音声認識の結果が得られる．

この単語単位モデルについて，CTC と注意機構モデルで実装することが考えられる．Lu ?らはエンコーダデコーダモデルで，Soltau ら？?や Audhkhasi ら ?は CTC による実装を報告している．注意機構モデルはデコーダで LSTM を用いているため，明示的に前の記号系列の文脈を組み込むことができる．一方，CTC ではエンコーダの LSTM によるフレーム単位での文脈しか学習できない．一単語の音声の長さはサブワードよりも通常長いため，フレーム単位の文脈では不十分な可能性がある．そのため注意機構モデルの方が高い性能が得られることが期待できる．本稿ではまずはじめに単語を出力とした CTC と注意機構モデルを比較する．

簡潔性の反面，単語単位モデルには従来のサブワード単位のモデルと比較して，出力のノード数が非常に多くなり，記号の出現頻度の分布が偏るという大きな問題がある．そのため単語単位モデルは学習データのスパース性によりオーバーフィットを引き起こす可能性がある．Audhkhasi ら ?は学習データ量が大きくない時，乱数により初期化した単語単位モデルは学習が収束しないと報告している．そのため，単語単位モデルには大量の学習データが必要とされる．もう一つの問題は訓練時に登録されている単語しか認識できない点である．サブワード単位のモデルのように辞書中に新たな単語を加えることができない．また前述の通り，出現頻度の低い多くの単語は学習データの不足により除去する必要があり，これらの単語も認識できなくなる．これらの問題を解決するため，本稿では単語単位モデルと文字単位モデルを併用するモデルを提案する．

3. 文字単位モデルを併用した単語単位モデル

3.1 CTC を用いた文字単位モデルとのマルチタスク学習

Audhkhasi ら ?は単語よりも低レベルの出力をする別の CTC を用いて単語単位モデルの事前学習を行ない，Shumbham ら?は音素単位の学習結果を用いてより高レベルのサブワード単位の学習を行なっている．また，Kim ら ?は共通のエンコーダから CTC と注意機構モデルを用いて，文字を出力としてマルチタスク学習を行うことで頑健性が向上し，各モデルを単一で学習するよりも改善が見られることを報告している．Watanabe ?らは文字を出力とした CTC と注意機構モデルとのマルチタスク学習において，精度の改善を実現している．

これらの先行研究をふまえて，CTC による文字単位モデルを併用し，注意機構モデルによる単語単位モデルの学習を行うマルチタスク学習を提案する．文字単位モデルを併用することで学習データのスパース性を緩和しモデルの一般性を改善することが期待される．また，CTC を用いることで学習の際に適切な制約を加えて性能の改善と学習の早期収束を期待する．

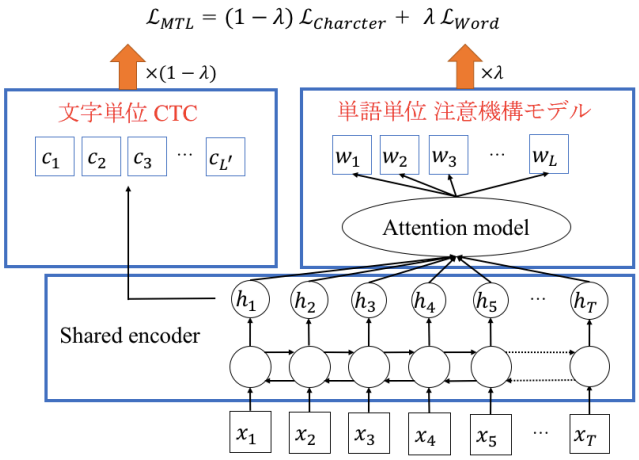

図 3 提案するマルチタスク学習モデルの概要図．マルチタスク学習の損失関数は各モデルの損失の重み付き線形和で表される．

エンコーダでは共通のネットワークを用い、デコーダでは二つのネットワークに分かれる．1 つは注意機構を用いた単語単位出力のモデルであり、出力のノード数は語彙サイズと同じになる．もう 1 つは CTC を用いた文字単位出力のモデルである．このマルチタスク学習の目的関数は両方のモデルの重み付き線形和を用いて定義される．図 3 に提案するマルチタスク学習のネットワーク図を示す．

3.2 未知語推定

単語単位モデルのもう一つの問題は語彙以外の単語を認識できない点である．この問題を解決するため，文字単位のモデルの参照を行う．単語単位モデルが未知語 (UNK) を推定した時，注意機構の重みベクトルから一番大きい値を持つフレーム情報を抽出することで未知語の音声のセグメントを大まかに決定する．そのフレーム情報を文字単位 CTC の出力の記号に対応づける．文字単位モデルでは単語間の境界 (wb) を意味する特殊な記号を導入する．これにより，未知語と文字系列の対応づけが行える．図 4 に未知語推定の概要を示す．<UNK>に関する注意重みから最大の値を持つフレームに対応する CTC の出力は”blank”であり，”blank”を含む <wb>間の系列，つまり「RNN」を未知語の推定結果とする．この機構によりいくつかの未知語を認識することが期待できる．また，未知語の一部しか推定できなかったとしても，<UNK>の記号より有用な情報になる可能性が高い．

4. Sequence-to-Sequence 学習の改善

注意機構モデルのような可変長の入出力系列の学習を行

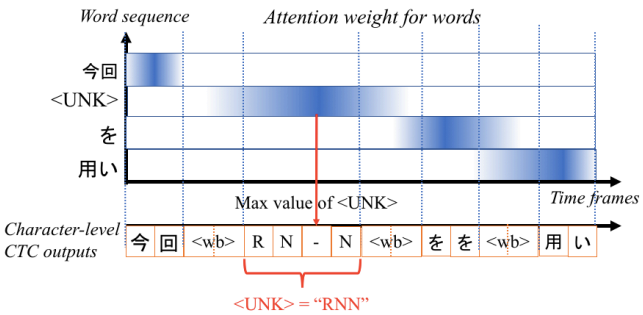

図 4 CTC による文字単位モデルを用いた未知語の推定手法

う Sequence-to-Sequence 学習には学習時，認識時の問題を緩和する手法が提案されている．本稿では，ラベルスムージング ?とスケジュールドサンプリング ? を用いる．

4.1 ラベルスムージング

ラベルスムージングは過学習を防ぐ正則化手法である．これは K 個の記号を予測する学習において教師データとして与える確率を正解記号のみ 1 として与えるのではなく，正解記号以外にも一様分布の確率を与えることで正則化を行う．

$$P_l = \begin{cases} \lambda & (l = True) \\ (1-\lambda) \times \frac{1}{K} & (otherwise) \end{cases}$$

本稿では $\lambda = 0.9$ としている．

4.2 スケジュールドサンプリング

スケジュールドサンプリングとは訓練時には見られない認識時の推論の誤りとの整合性をとる手法である．Sequence-to-Sequence モデルは前の系列記号を用いて次の記号の予測を行う．学習時，モデルが学習に用いる前の系列は正解データから与えられるが，認識時にはモデルが推定した記号を用いる．これには誤りを含むため学習時に想定されていない誤りパターンが生じる．この不整合を解決するため，学習時にも一定の確率でモデルの推定記号を用いて次の記号の予測を行うことで誤りも考慮した学習が行える．学習が初期の段階では正解記号を用い，学習が進むにつれて推定記号を用いる確率を上げていく．本稿では 5 エポック学習が終わるまでは全て正解記号を用い，5 エポックから 15 エポックまで最終的に推定記号を用いる確率が 0.2 になるように線形的に確率を上げ，それ以降は 0.2 のままになるように設定している．

5. 評価実験

5.1 実験設定

手法の評価には『日本語話し言葉コーパス』(CSJ) を用いた．CSJ の学会講演 (CSJ-APS) と模擬講演 (CSJ-SPS) 用いて，学習，評価を行う．訓練時間は CSJ-APS では 224 時間，CSJ-SPS では 251 時間で，評価には CSJ の標準的な評価セットで行う．CSJ-APS では評価セット 1(男性 10 人，計 10 講演) を用い，CSJ-SPS では評価セット 3(男性 5 人，女性 5 人，計 10 講演) を用いた．また各コーパスにおいて，出現回数が 3 回以上の単語のみ語彙に含める．

音響特徴量は 40 次元の Log Mel-scale filterbank とした．エンコーダは 3 層の 320 次元の隠れ層を持つ双方向 LSTM で構成し，注意機構を用いたデコーダは 1 層の 320 次元の隠れ層を持つ単方向 LSTM で構成し，その後出力単語数分のノードを持つ softmax の出力層となる．CTC と注意機構モデルの目的関数の重みは 0.2 およびと 0.8 とした．最適化アルゴリズムは Adam ?を用い，Gradient Clipping の閾値を 5.0 とした．全ての全結合層，畳み込み層の重みは (-0.1, 0.1) の一様分布で初期化し，双方向 LSTM の重みは He ら ?によって提案された初期化手法を用いる．学習初期に長いフレームのデータを与えた場合，収束が遅くなるので，学習時の入力データはフレームによりソートし短いものから学習を行なっている．注意機構による単語単位モデルによる認識にはビーム幅 4 のビームサーチを行なっている．また特定の実験では L ラベルスムージングとスケジュールドサンプリングを用いる．以上のネットワークの実装は Chainer v2.1.0?を用いている．

ベースラインには従来法である DNN-HMM と言語モデルを使用したハイブリッドシステムと，End-to-End 音声認識の一種である音素を出力単位とする CTC と言語モデルを使用した．DNN-HMM システムは 2048 のノード数を持つ中間層を 7 層にし，出力層は 4829 のノード数を持つネットワークである．音素を出力とする CTC システムは 3 層の双方向 LSTM と出力層を持つ．認識時は DNN-HMM ハイブリッドモデルは Julius を，音素単位 CTC は EESEN WFST デコーダを用いた ?.

5.2 実験結果

CSJ-APS 評価セット 1 に対する WER と実時間係数 (RTF) を測定した結果を表 1 に示す．ベースラインの音響モデルは CSJ-APS で学習されており，言語モデルについては CSJ-APS と CSJ-SPS の両方から学習している．単語単位モデルはベースラインシステムに比べて非常に高速に認識が行えていることがわかる．また，単語単位の CTC よりも注意機構モデルを用いた方が 2.3 の WER が改善することが示された．これは注意機構のモデルにより単語単位の言語モデルの学習ができているためと考えられる．提案手法であるマルチタスク学習により 0.83 ポイントの改善が見られた．

マルチタスク学習を実現するためのモデルとして提案手法以外にも組み合わせが考えられる．種々の補助タスクの組み合わせによる結果を表 2 に示す．ただし，他の実験ではミニバッチサイズが 30 であるが，単語 CTC の入力が処理できなかったため表 2 ではミニバッチサイズを 10 とし

表 1 CSJ-APS における手法の比較．単語誤り率 (WER) と実時間係数 (RTF) を計測．(・) 内は補助タスクとタスクを行うモデルを示す．

モデル	WER(%)	RTF
DNN-HMM + largeLM	13.62	0.925
音素 CTC + largeLM	14.15	0.581
単語 CTC	16.97	0.010
単語 attention	14.67	0.035
単語 attention (文字 CTC)	13.84	0.035

表 2 CSJ-APS における補助タスクの組み合わせの比較．メインタスクは単語 attention で固定．ただし他の実験よりミニバッチサイズは小さい．

補助タスクとモデル	WER(%)
補助タスクなし	17.25
文字 CTC (Proposed)	**16.11**
文字 attention	17.55
単語 CTC	17.03

た．単語 CTC と組み合わせることで 0.22 ポイントの性能の改善が見られたが，提案手法の組み合わせが最も有効であることが確認された．

表 3 に提案手法に対してラベルスムージング，スケジュールドサンプリング，未知語推定を CSJ-APS と CSJ-SPS の評価セットに適用した結果を示す．上記 3 つの手法により性能が改善され，これらを用いることで従来の DNN-HMM ハイブリッドモデルより高い性能になった．

表 4 に CSJ-APS で単語単位の注意機構モデルで未知語と推定された単語を提案手法で推定した例を示す．「組み合わせ」のように辞書に含まれているが，言い淀みやフィラーなどの要因により未知語と推定される場合もある．一部しか認識できなかった単語でも人間が推定できるものが多いことがわかる．

6. おわりに

本稿では，注意機構モデルを用いた単語単位モデルと CTC を用いた文字単位モデルとのマルチタスク学習モデルを提案した．CSJ の 2 つのデータセットに対して実験を行い，従来のハイブリッドモデルに匹敵する性能が得られ，非常に高速な音声認識が達成された．また両コーパスともラベルスムージング，スケジュールドサンプリング，未知語推定により，さらなる性能の向上が見られた．

表 3 CSJ-APS と CSJ-SPS における改善手法の効果．数値は WER(%)

モデル	APS	SPS
DNN-HMM+largeLM	13.62	12.80
単語 attention (文字 CTC)	13.84	11.94
+label smoothing	13.33	10.97
+scheduled sampling	13.68	11.84
+未知語推定	13.65	11.80
+上記全て	12.91	10.52

表 4 CSJ-APS における単語単位注意機構が未知語を出力した際に，文字単位 CTC を用いて行われた未知語推定例．

推定結果	**正解**
組み合わせ PSJ	組み合わせ PSJ
法置音 ノイズっぽイ コミュンティーメンバー プ R ロ 0 グ	報知音 ノイズっぽい コミュニティーメンバー プロローグ

参考文献

[1] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-li Lim, Bergul Roomi, and Phil Hall, “English Conversational Telephone Speech Recognition by Humans and Machines,” in Interspeech 2017, 2017.

[2] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber, “Connectionist temporal classification : Labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd international conference on Machine Learning, pp. 369-376, 2006.

[3] Alex Graves and Navdeep Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in Proceedings of the 31st International Conference on Machine Learning, pp. 1764-1772, 2014.

[4] Haşim Sak, Andrew Senior, Kanishka Rao, and Francçoise Beaufays, “Fast and accurate recurrent neural network acoustic models for speech recognition,” arXiv preprint arXiv:1507.06947, 2015.

[5] Haşim Sak, Félix de Chaumont Quitry, Tara Sainath, Kanishka Rao, “Acoustic modelling with CD-CTC-SMBR LSTM RNNs,” in Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE, pp. 604-609, 2015.

[6] 伊藤均, 萩原愛子, 一木麻乃, 三島剛, 佐藤庄衛, 小林彰夫. ”漢字の読みを考慮した End-to-end 音声認識,” 日本音響学会春季研究発表会講演論文集, 2017.

[7] 増田嵩志, 齋藤大輔, 峯松信明. ”敵対的学習を適用した End-to-end 音声認識,” 情報処理学会 音声言語情報処理研究会, SLP119, 2017.

[8] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “End-to-end continuous speech recognition using attention-based recurrent NN: first results,” in NIPS: Workshop Deep Learning and Representation Learning Workshop, 2014.

[9] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio , “Attention-based models for speech recognition,” in Advances in Neural Information Processing Systems (NIPS), pp. 577-585, 2015.

[10] Rohit Prabhavalkar, Tara N Sainath, Bo Li, Kanishka Rao, and Navdeep Jaitly, “An analysis of “attention” in sequence-tosequence models,” in Interspeech 2017, pp. 3702-3706, 2017.

[11] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pp. 4960-4964, 2016.

[12] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, “End-to-end attention-based large vocabulary speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4945-4949, 2016.

[13] Yajie Miao, Mohammad Gowayyed, and Florian Metze, “EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding,” in Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on , pp. 167-174, 2015.

[14] Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, David Nahamoo , “Direct acoustics-to-word models for english conversational speech recognition,” in Interspeech2017, pp. 959-963, 2017.

[15] Hagen Soltau, Hank Liao, and Hasim Sak, “Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition” , in Interspeech2017, pp. 3707-3711, 2017.

[16] Liang Lu, Xingxing Zhang, and Steve Renais, “On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition,” in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pp. 5060-5064, 2016.

[17] Jinyu Li, Guoli Ye, Rui Zhao, Jasha Droppo, and Yifan Gong, ”Acoustic-To-Word Model Without OOV,” 2017 IEEE Automatic Speech Recognition and Understanding Workshop , pp. 111–117,2017.

[18] Hagen Soltau, Hank Liao, and Hasim Sak, ”Reducing the computational complexity for whole word models,” 2017 IEEE Automatic Speech Recognition and Understanding Workshop , pp. 63–68, 2017.

[19] Shubham Toshniwal, Hao Tang, Liang Lu, and Karen Livescu, “Multitask learning with low-level auxiliary tasks for encoderdecoder based speech recognition,” in Interspeech2017, pp. 3532-3536, 2017.

[20] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, ”Joint CTC-attention based end-to-end speech recognition using multi-task learning,” Proc. ICASSP’17, pp. 4835–4839 , 2017.

[21] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey and Tomoki Hayashi, ”Hybrid CTC/Attention Architecture for End-to-End Speech Recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240-1253, 2017.

[22] Christian Szegedy,Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, ”Rethinking the inception architecture for computer vision,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826, 2016.

[23] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer, ”Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks”, NIPS, 1171–1179, 2015.

[24] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization” , arXiv preprint arXiv:1412.6980, pp. 1-15, 2014.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, ”Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”, In Proceedings of the IEEE international conference on computer vision, pp. 1026–1034, 2015.

[26] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton, “Chainer: a next-generation open source framework for deep learning,” in Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), 2015.

正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

訂正箇所	誤	正
1 ページ 3 行目	特定のタスクでは人間レベルの認識精度を達成することが報告されている?	特定のタスクでは人間レベルの認識精度を達成することが報告されている[1]
1 ページ 11-12 行目	Connectionist Temporal Classification (CTC) ? ? ? ? ? ? と注意機構 (attention mechanisim) ? ? ? ? ?	Connectionist Temporal Classification (CTC) [2] [3] [4] [5] [6] [7] と注意機構 (attention mechanism) [8] [9] [10] [11] [12]
1 ページ 15-18 行目	単語系列を出力する際には依然として発音辞書や言語モデルなどの外部機構を必要とする?．これに対して，音響特徴量から単語を直接出力するモデルについても研究されている? ? ? ? ? ?.	単語系列を出力する際には依然として発音辞書や言語モデルなどの外部機構を必要とする[13]．これに対して，音響特徴量から単語を直接出力するモデルについても研究されている[4] [14] [15] [16] [17] [18].
3 ページ 9-11 行目	Lu ?らはエンコーダデコーダモデルで，Soltau ら? ?やAudhkhasi ら?はCTC による実装を報告している．	Lu [16] らはエンコーダデコーダモデルで，Soltau ら[15] [18] やAudhkhasi ら[14]は CTC による実装を報告している．
3 ページ 24-26 行目	Audhkhasiら?は学習データ量が大きくない時, 乱数により初期化した単語単位モデルは学習が収束しないと報告している．	Audhkhasiら[14] は学習データ量が大きくない時, 乱数により初期化した単語単位モデルは学習が収束しないと報告している．
3 ページ 37-41 行目	Audhkhasi ら?は単語よりも低レベルの出力をする別のCTC を用いて単語単位モデルの事前学習を行ない，Shumbham ら?は音素単位の学習結果を用いてより高レベルのサブワード単位の学習を行なっている．また，Kimら?は共通のエンコーダからCTC と注意機構モデルを用いて, 文字を出力としてマルチタスク学習を行うことで頑健性が向上し, 各モデルを単一で学習するよりも改善が見られることを報告している．Watanabe ?らは文字を出力としたCTC と注意機構モデルとのマルチタスク学習において, 精度の改善を実現している．	Audhkhasi ら[14] は単語よりも低レベルの出力をする別のCTC を用いて単語単位モデルの事前学習を行ない，Shumbham ら[19] は音素単位の学習結果を用いてより高レベルのサブワード単位の学習を行なっている．また，Kimら[20] は共通のエンコーダからCTC と注意機構モデルを用いて，文字を出力としてマルチタスク学習を行うことで頑健性が向上し, 各モデルを単一で学習するよりも改善が見られることを報告している．Watanabe [21] らは文字を出力としたCTC と注意機構モデルとのマルチタスク学習において，精度の改善を実現している．

4 ページ 2-3 行目	本稿では，ラベルスムージング?とスケジュールドサンプリング? を用いる．	本稿では，ラベルスムージング[22] とスケジュールドサンプリング[23] を用いる．
4 ページ 2 列目 11-14 行目	最適化アルゴリズムはAdam ?を用い，Gradient Clippingの閾値を5.0 とした．全ての全結合層，畳み込み層の重みは(-0.1, 0.1)の一様分布で初期化し，双方向LSTM の重みはHe ら?によって提案された初期化手法を用いる	最適化アルゴリズムはAdam [24] を用い，Gradient Clippingの閾値を5.0 とした．全ての全結合層，畳み込み層の重みは(-0.1, 0.1)の一様分布で初期化し，双方向LSTM の重みはHe ら[25] によって提案された初期化手法を用いる．
4 ページ 2 列目 21 行目	以上のネットワークの実装はChainer v2.1.0?を用いている	以上のネットワークの実装はChainer v2.1.0[26] を用いている．
4 ページ 2 列目 29-30 行目	音素単位CTC はEESEN WFST デコーダを用いた?.	音素単位CTC はEESEN WFST デコーダを用いた[13]