Vol.2021-SLP-138 No.17
2021/10/19

Triggered attention型ストリーミング音声認識における Mask-CTCを用いた事前学習

趙 懐博[1]　樋口 陽祐[1]　小川 哲司[1]　小林 哲則[1]

概要：Mask-CTC と Triggered attention 機構を組み合わせ，低遅延で高性能なストリーミング End-to-End 音声認識システムの構築を試みた．Triggered attention 機構とは，Connectionist Temporal Classification (CTC) による記号の出力タイミングに基づいて推論を行う手法であり，ストリーミング音声認識における有効性が示されている．しかし，CTC 出力に基づくアライメント推定を高精度に行うためには，多少の遅延を許しても未来の情報を入力して推論を行うことが望ましい．一方で，ストリーミング音声認識では，遅延を低く抑えつつ，高い認識精度を達成することが望ましい．そこで，本研究では，未来の情報を含む長期的な文脈を考慮して特徴表現を学習する枠組みである Mask-CTC をエンコーダの事前学習に導入することで，低遅延かつ高精度なストリーミング音声認識を実現することを目指す．WSJ データを用いたストリーミング音声認識実験により，従来の Triggered attention 型ストリーミング音声認識モデルと比較して，提案モデルは遅延を低く抑えながら高い認識性能を与えることが明らかになった．

キーワード：Triggered attention 機構, Mask-CTC, End-to-end モデル, ストリーミング音声認識

An Investigation of Enhancing CTC Model for Triggered Attention-based Streaming ASR

Huaibo Zhao[1]　Yosuke Higuchi[1]　Tetsuji Ogawa[1]　Tetsunori Kobayashi[1]

1. はじめに

音声認識の問題を音声からテキストへの単一の系列変換モデルとして表現する End-to-End 音声認識モデリングの進展 [1–3] により，音響モデル，単語辞書，言語モデルといった複数の要素の組み合わせにより実現されていた音声認識システムの実装は著しく容易になっている．特に，注意機構に基づくエンコーダ・デコーダモデル [2, 3] は，End-to-End 音声認識モデルの標準的な構造と言える．このモデルは，特徴量を抽出するエンコーダモジュール，注意機構，および注意機構を用いてエンコーダ出力からテキストを予測するデコーダモジュールから構成される．このような構造は，Long Short-Term Memory (LSTM) ネットワーク [3–5] や Transformer [6–8] など，様々な深層ニューラルネットワークに対して実現されている．また，Connectionist Temporal Classification (CTC) も End-to-End 音声認識における重要技術であり，マルコフ仮定と動的計画法 [1] を用いて，入力系列と出力系列のアラインメントを陽に求めながら損失関数を計算するのが特徴である．さらに，CTC と注意機構に基づくエンコーダ・デコーダモデルを統合することで，モデルの学習と推論をさらに改善できることが知られている [8–10].

音声認識をユーザインタフェースとして利活用する場合，ストリーミング機能が必要となる．しかし，大域的な注意機構に基づく End-to-End 音声認識モデルと，音声入力に対して低遅延で逐次出力を行うストリーミングとの相性は良いとは言えない．この問題を解消するために，ストリーミング End-to-End 音声認識に関する方式検討も進んでいる．例えば，Monotonic Chunk-wise Attention (MoChA) [11]

[1] 早稲田大学 基幹理工学部 情報通信学科
Department of Communications and Computer Engineering, Waseda University

は，適応的に区切られた部分系列に対する soft attention を利用し，注意機構に基づくモデルを用いたストリーミングを実現している．また，ストリーミング音声認識の性能低下を防ぐために，注意機構を用いたエンコーダ・デコーダに基づく neural transducer の提案もなされている [12]. Triggered attention 型ストリーミング音声認識 [13] では，CTC に基づくエンコーダネットワークからアラインメントを生成し，CTC 出力のスパイクをトリガーとしてデコーダを駆動する．この Triggered attention 型モデルは，フレーム同期でのデコーディングを可能にするものの，その性能は CTC ネットワークが推定するアラインメントの精度に大きく依存する．一方，正確なアライメントを得るためには，エンコーダネットワークへの入力は長時間であることが望ましく，その結果遅延時間は長くなる．これは，ストリーミング音声認識の要件 (低遅延でかつ高精度) とは相反する．

それに対し本研究では，Triggered attention 型モデルに焦点を当て，高精度かつ低遅延のストリーミング音声認識を実現するための特徴表現を抽出する方式について検討を行う．そのために，Mask-CTC [14,15] の枠組みをエンコーダネットワークの事前学習に用いることを提案する．Mask-CTC は，CTC と Conditional Masked Language Model (CMLM) [16,17] のマルチタスク学習により，長期文脈情報を考慮した特徴表現抽出を可能とするエンコーダを得ようとするものである．Triggered attention 型ストリーミング音声認識システムにおけるエンコーダを Mask-CTC の枠組みで事前学習することで，長期文脈を考慮した特徴表現が抽出され，例え入力系列が短くても CTC におけるアライメント精度の向上が期待できる．また，長期文脈を考慮することで将来観測される情報の先読み能力も向上し，低遅延で高い認識精度を維持できることも期待される．

本稿の構成は以下の通りである．まず，2 章で提案手法に必要な要素技術を概観する．続いて，3 章で提案手法である Triggered attention 型ストリーミング音声認識モデルの学習方法について説明する．4 章では，WSJ コーパスを用いた音声認識実験について述べ，提案方式の有効性を明らかにする．最後に，5 章において本稿のまとめと今後の課題について述べる．

2. 関連技術

End-to-End 音声認識は，入力系列 $X = (\mathbf{x}_t \in \mathbb{R}^D | t = 1, ..., T)$ から出力系列 $Y = (y_l \in \mathcal{V} | l = 1, ..., L)$ を生成する問題である．ここで，X は長さ T の音響特徴量系列であり，$\mathbf{x}_t$ は時刻 t における D 次元の特徴ベクトルである．Y は語彙 $\mathcal{V}$ に含まれる記号の長さ L の系列であり，y_l は l 番目の記号を表す．

2.1 Attention span を用いたストリーミングエンコーダ

Transformer を用いた End-to-End 音声認識モデル [7,8] のエンコーダに attention span [18,19] を導入することで，過去と未来の文脈情報を制御できる．注意重みは，固定サイズの span マスク W を用いて以下のように書ける．

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}W\right)V \quad (1)$$

ここで，$Q \in \mathbb{R}^{n_q \times d_q}$ はクエリ，$K \in \mathbb{R}^{n_k \times d_k}$ はキー，$V \in \mathbb{R}^{n_v \times d_v}$ はバリューの行列を各々表し，n_* と d_* は，各々系列長と特徴ベクトルの次元数を表す．本研究では，先読みの範囲にのみ制限を設け，過去のデータには常にアクセス可能とする．

2.2 Triggered attention 型ストリーミングデコーダ

Triggered attention 型ストリーミングデコーダ [13,20] は，CTC モジュールが出力したスパイク状の事後確率値をトリガーとして，注意機構とフレーム同期のデコーダを駆動する [20].

Triggered attention 型デコーダは，エンコードされた特徴系列 H と出力ラベル系列 Y のアライメントを必要とする．これは，CTC の出力に基づいて計算され，デコーダの学習に利用される．具体的には，アライメント情報は，CTC に基づくトリガーネットワークを用いた Viterbi デコーディングによって与えられ，注意機構に基づくデコードが過去のエンコードされたフレームと過去の出力のみに条件づけられるよう，以下のように使用される．

$$P_{\text{ta}}(Y|H) = \prod_{l=1}^{L} P(y_l | y_{1:l-1}, \mathbf{h}_{1:n_l}) \quad (2)$$

ここで，n_l は CTC アライメントにおいて記号 y_l が最初に出現する位置を表す．未来の特徴抽出を必要としない式 (2) は，ストリーミング音声認識に適している．学習時は，全てのモジュールが事前学習された非ストリーミングの CTC-attention 型 モデル [21] で初期化され，ストリーミングに適する形で fine-tuning される．エンコーダは，2.1 にて説明した attention span を用いて学習され，推論時は CTC と デコーダによる joint decoding が行われる．

2.3 Mask-CTC

Mask-CTC [14] は，高精度かつ高速な End-to-End 音声認識を実現するための学習フレームワークである．Mask-CTC では，CTC と CMLM のマルチタスク学習によって，注意機構に基づくエンコーダ・デコーダモデルを学習する．エンコーダの出力から CTC による損失を計算し，デコーダでは CMLM によるマスク推定 [16] に基づいた損失を計算する．マスク推定では，正解系列中の記号をランダムにマスク記号に置き換え，このマスクされた記号を文脈情報に基づいて予測する．入力系列 X と観測記号 Y_{obs}

に対して，マスクされた記号 Y_{mask} の事後確率分布は次のようにモデル化される．

$$P_{\text{cmlm}}(Y_{\text{mask}}|Y_{\text{obs}}, X) = \prod_{y \in Y_{\text{mask}}} P_{\text{cmlm}}(y|Y_{\text{obs}}, X). \quad (3)$$

ここで，$Y_{\text{obs}} = Y \setminus Y_{\text{mask}}$ である．CTC と CMLM のマルチタスク学習により，エンコーダは長期文脈情報を考慮した特徴抽出が可能となる．また，将来観測される記号の先読み能力も向上し，CTC におけるアラインメント精度の向上が期待できる．

3. 低遅延・高精度なストリーミング音声認識のための特徴表現学習

低遅延・高精度な Triggered attention 型ストリーミング音声認識モデルの学習法について述べる．提案の学習法は，以下の 3 段階から成る (図 1).

- **Stage 0 (先読みに適した特徴表現学習)：**Mask-CTC (CTC と CMLM によるマルチタスク学習) により，長期文脈情報を考慮可能なエンコーダを学習し，高精度な CTC モジュールを獲得する．
- **Stage 1 (ストリーミング音声認識モデルの事前学習)：**Stage 0 で得られたエンコーダと CTC モジュールを用いて，CTC-attention 型の非ストリーミングモデル [21] を学習する．これは，後続の Triggered attention 型ストリーミングデコーダの学習に向けて，エンコーダを自己回帰的なデコーダに適応することを目的としている．また，Mask-CTC で得られた CTC のアライメント精度が維持されることを期待する．
- **Stage 2 (ストリーミング音声認識モデルの学習):** Stage 1 で学習された CTC-attention 型モデルのパラメータを用いて，Attention span を導入したエンコーダと Triggered attention 型デコーダを初期化し，ストリーミング音声認識モデルを学習する．

以下では，Stage 0 の Mask-CTC による特徴表現学習に期待する効果と，Stage 1 および 2 のストリーミング音声認識モデルの学習の詳細について述べる．

3.1 Mask-CTC による特徴表現学習の効果

提案するストリーミングに適した特徴表現の学習法は，Mask-CTC [14] の枠組みで学習したエンコーダが高精度な CTC 出力を与えることから着想を得た．CTC 出力の精度において，Mask-CTC が CTC-attention 型モデルよりも優れていることから，CTC と CMLM のマルチタスク学習により，エンコーダでの特徴抽出処理の精度が向上し，CTC モジュールでのアライメントの精度が向上することが示された．CMLM は，過去から未来までの長期的な文脈を考慮して予測を行うモデルであるため，CMLM の枠組みでエンコーダを学習することで，先読みを可能にするような特徴表現が学習されることが期待される．このような特性は，ストリーミング音声認識タスクにおいて，特に低遅延のみ許される場合には望ましいものと言える．

Triggered attention 型のストリーミングモデルを学習する際には，アライメントの生成と各モジュールの初期化が必要であり，従来では非ストリーミングの CTC-attention モデルが初期モデルとして用いられている．提案手法では，Stage 0 で Mask-CTC により得られたモジュールで初期化された CTC-attention モデル (以降では，Enhanced CTC-attention モデルと呼び区別する) を，ストリーミングモデルの事前学習において利用する．こうすることで，Mask-CTC により学習される，低遅延・高精度なストリーミングに適した特徴抽出プロセスが，Triggered attention 型ストリーミングモデルのための初期モデルにも導入されることを期待する．

3.2 ストリーミング音声認識モデルの学習過程

Stage 1 では，エンコーダ，CTC モジュール，自己回帰型デコーダから成る CTC-attention モデルを事前学習する．このモデルは，次の Stage 2 において Triggered attention 型ストリーミングモデルを学習するための初期モデルとして用いられる．ただし，提案アプローチで構築する Enhanced CTC-attention モデルでは，エンコーダに Mask-CTC の効果が導入され，従来の CTC-attention モデルのエンコーダよりも高精度な CTC 出力が得られることを期待する．このとき，Mask-CTC は非自己回帰型で，かつ非ストリーミング型のモデルであるため，この Stage 1 では，非ストリーミングの枠組みで，Triggered attention 型ストリーミングモデルの初期化に使われるエンコーダと CTC モジュールを事前学習することを試みる．

Stage 2 では，2.2 で述べた手順に従って，Triggered attention 型ストリーミングモデルを構築する．従来アプローチと異なる点は，Stage 1 で構築された Enhanced CTC-attention モデルをアライメントの生成やストリーミングモデルにおける CTC モジュールおよびエンコーダの初期モデルとして利用している点である．Enhanced CTC-attention モデルには Mask-CTC の優れた特徴抽出機能が組み込まれており，これにより，低遅延でも信頼性の高いアライメントを生成し，Triggered attention 機構に優れたトリガを与えることを期待できる．このように，Enhanced CTC-attention モデルを初期モデルとして用いることで，既存のストリーミング音声認識モデルよりも，特に低遅延で高精度な認識が可能になると期待できる．

4. ストリーミング音声認識実験

4.1 実験データ

モデルの学習・評価には，英語による読み上げ音声である Wall Street Journal (WSJ) コーパス [22] を用いた．音

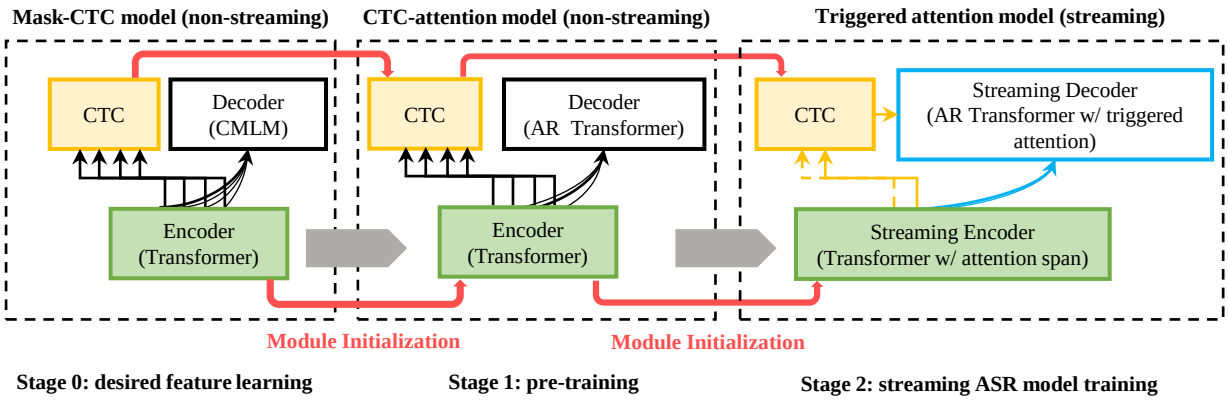

図 1: 高精度・低遅延のストリーミングに適した Triggered attention 型ストリーミング音声認識モデルの構築

響特徴量は 80 次元の対数メルフィルタバンク出力にピッチ情報を加えた 83 次元であり，Kaldi [23] を用いて抽出した．出力系列の単位は文字 (character) とした．

4.2 実験条件

End-to-End 音声認識モデルとして，エンコーダブロック 12 層，デコーダブロック 6 層から成る Transformer を構築した [6, 7]．各 Self-attention 層において，Multi-head Attention のヘッド数は 4，埋め込み次元は 256，全結合層のユニット数は 2048 とした．モデルパラメータの最適化には Adam を利用し，学習率は文献 [6] と同様のスケジューリングを行なった．非ストリーミングモデルの学習では 50 から 100 エポック，ストリーミングモデルの学習では 120 エポックの更新を行った．モデルの学習後，最終 5 エポック分のチェックポイントを用いてモデルのパラメータを平均し，最終的なモデルを得た．推論時，CTC の性能を評価する場合，Best-path decoding [24] による貪欲アルゴリズムを用いた．また，CTC と Triggered attention 機構に基づいた推論には，Frame-synchronous one-pass decoding [13] を用いた．このとき，CTC-attention の重みは 0.5 とし，ビーム幅 10 のビームサーチを行った．

4.3 評価項目

提案手法の有効性を実証するために 2 つの実験を行った．ここでは，各実験において比較したモデルについて説明する．なお，全てのシステムは ESPnet [25] を用いて開発し，推論時には外部言語モデルは使用していない．

4.3.1 実験 1

ストリーミング音声認識モデルの事前学習段階 (Stage 1) に Mask-CTC を導入することの効果を調査した．そのため，以下のモデルから生成された貪欲アルゴリズムによる CTC の出力を比較した．

- **CTC-attention** [21]: 既存の非ストリーミングモデル．自己回帰型 Transformer に基づくデコーダ，Transformer に基づくエンコーダ，CTC モジュールから成る．
- **Mask-CTC** [14]: Mask-CTC の枠組みで構築された非ストリーミングモデル．Transformer に基づく CMLM デコーダ，Transformer に基づくエンコーダ，CTC モジュールから成る (2.3)．本モデルは，提案法における Stage 0 で構築される．
- **Enhanced CTC-attention**: 提案法の Stage 1 で得られる，Mask-CTC の効果が組み込まれた非ストリーミング CTC-attention モデル．エンコーダと CTC モジュールを Mask-CTC モデルで初期化した後，全モジュールを CTC-attention の枠組みで学習する．

このとき，Enhanced CTC-attention モデルが既存の CTC-attention モデルの性能を上回った場合，CTC-attention モデルに代えて Enhanced CTC-attention モデルを初期モデルとして用いることで，Triggered attention 型ストリーミング音声認識の性能改善が期待できる．Enhanced CTC-attention モデルは Mask-CTC が持つ先読みを考慮した特徴抽出機能が導入されており，特に，低遅延で高精度なストリーミングの実現が期待される．

4.3.2 実験 2

Stage 2 において構築された，Triggered attention 型ストリーミング音声認識モデルの有効性を調査した．Mask-CTC を学習に組み込むことで認識精度を維持したまま遅延を低減できるかを明らかにするため，以下のモデルを比較した．

- **Triggered attention-based streaming**: CTC-attention の枠組みで構築された既存の Triggered attention 型ストリーミング音声認識モデル [13]．アライメントの生成とモジュールの初期化には，既存の CTC-attention モデルを用いる．
- **Enhanced triggered attention-based streaming**: Mask-CTC を特徴表現学習に組み込んだ，Triggered attention 型ストリーミング音声認識モデル．Enhanced CTC-attention モデルを用いて，アライメントの生成とモジュールの初期化を行う．

表 1: CTC-attention モデルを Mask-CTC に基づき初期化する効果．各モデルの CTC 出力により評価した．

Model		Encoder	Decoder	eval92	dev93
CTC-attention	(Conventional)	Transformer randomly initialized	AR Transformer	16.6	21.2
Mask-CTC	(Stage 0 in Fig. 1)	Transformer randomly initialized	CMLM Transformer	13.9	17.9
Enhanced CTC-attention	(Stage 1 in Fig. 1)	Transformer pre-trained with Mask-CTC	AR Transformer	**13.3**	**16.8**

表 2: Triggered attention 型ストリーミング音声認識モデルの事前学習に Mask-CTC を用いる効果．

Model	Pre-trained model	Encoder latency [ms]	eval92	dev93
Triggered attention-based streaming ASR	CTC-attention	160	28.2	34.5
		320	22.4	27.5
		480	18.9	24.1
		640	17.0	20.2
Enhanced triggered attention-based streaming ASR	Enhanced CTC-attention	160	21.3	25.9
		320	15.5	19.5
		480	14.3	19.1
		640	14.1	18.1

両システムともに，推論時は joint CTC-attention decoding を行った．モデルのストリーミング特性を分析するために，エンコーダの遅延時間が認識精度に与える影響を調査した．ここで，エンコーダ遅延 (encoder latency) とは，エンコーダが入力フレームごとに特徴抽出を行うための先読みの範囲と定義され，2.1 で述べた attention span 機構によって制御される．提案の狙いは，このエンコーダ遅延を低減しながら，高い認識精度を達成することである．

4.4 実験結果

4.4.1 実験 1

比較した 3 つのモデルの CTC 出力に対して単語誤り率 (WER) を計算し，表 1 に示す．まず，Mask-CTC モデルが CTC-attention モデルの性能を上回っていることから，CTC-CMLM の枠組みで学習された長期的な文脈を考慮した特徴抽出過程 (エンコーダ) の有効性を確認できた．また，Enhanced CTC-attention モデルが最良の性能を与えたことは，CTC-attention モデルの事前学習が Mask-CTC により効果的に行われたことを意味する．この結果は，Triggered attention 型ストリーミング音声認識モデルの初期モデルとして，従来の CTC-attention モデルに代えて Enhanced CTC-attention モデルを用いることの有効性も示唆している．

4.4.2 実験 2

実験 2 の結果を表 2 に示す．提案モデルは，エンコーダ遅延の値に依らず既存モデルの WER を削減した．また，提案モデルは，Triggered attention 型ストリーミング音声認識の精度を維持しつつ，低遅延での処理を可能とすることもわかる．表 2 より，両システムともにエンコーダ遅延が短くなるにつれて認識性能が劣化しているものの，提案モデルを用いた場合は性能劣化の度合いが小さいことがわかる．また，提案モデルは，既存モデルを用いた場合よりも低遅延で高い認識精度を達成した．例えば，320ms の遅延での提案モデルの WER (eval92 で 15.5%，dev93 で 19.5%) は，640 ms の遅延での既存モデルの WER (eval92 で 17.0%，dev93 で 20.2%) を上回った．このような，低遅延でも認識性能を維持できるという性質から，提案の学習法は，利用可能な将来の情報が少なくても高精度な特徴抽出とアライメント生成が可能であることを示唆している．一方，遅延時間を 320ms から 160ms とした際に提案モデルの性能も急激に低下しており，160ms 程度の短いエンコーダ遅延が要求される場合は抜本的な改善が必要であることがわかった．

5. まとめ

本稿では，低遅延かつ高精度な Triggered attention 型ストリーミング音声認識モデルの学習法について述べた．Mask-CTC によって導入される将来の文脈を考慮した特徴表現学習をストリーミング音声認識モデルの事前学習において組み込むことで，既存の Triggered attention 型ストリーミング音声認識と比較して，低遅延で高精度なストリーミングを実現した．今後は，認識性能と遅延のトレードオフに関して更なる調査を行うとともに，様々な言語を対象として提案モデルの有効性を検証する予定である．

参考文献

[1] Graves, A. and Jaitly, N.: Towards End-to-End Speech Recognition with Recurrent Neural Networks, *Proceedings of International Converence on Machine Learning (ICML)*, pp. 1746–1772 (2014).
[2] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. and

Bengio, Y.: Attention-Based Models for Speech Recognition, *Proceedings of the Advances in Neural Information Processing Systems 28 (NeurIPS)* (2015).

[3] Chan, W., Jaitly, N., Le, Q. V. and Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964 (2016).

[4] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, pp. 1735–1780 (1997).

[5] Moritz, N., Hori, T. and Roux, J. L.: Unidirectional Neural Network Architectures for End-to-End Automatic Speech Recognition, *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2019).

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 5998–6008 (2017).

[7] Dong, L., Xu, S. and Xu, B.: Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition, *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888 (2018).

[8] Karita, S., Yalta, N., Watanabe, S., Delcroix, M., Ogawa, A. and Nakatani, T.: Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration, *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2019).

[9] Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Yalta, N., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T. and Zhang, W.: A Comparative Study on Transformer vs RNN in Speech Applications, *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 449–456 (2019).

[10] Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, K., Jaitly, N., Li, B., Chorowski, J. and Bacchiani, M.: State-of-the-Art Speech Recognition with Sequence-to-Sequence Models, *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4774–4778 (2018).

[11] Chiu, C.-C. and Raffel, C.: Monotonic Chunkwise Attention, *ArXiv*, Vol. abs/1712.05382 (2018).

[12] Sainath, T. N., Chiu, C.-C., Prabhavalkar, R., Kannan, A., Wu, Y., Nguyen, P. and Chen, Z.: Improving the Performance of Online Neural Transducer Models, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5864–5868 (2018).

[13] Moritz, N., Hori, T. and Le, J.: Streaming automatic speech recognition with the Transformer model, *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6074–6078 (2020).

[14] Higuchi, Y., Watanabe, S., Chen, N., Ogawa, T. and Kobayashi, T.: Mask CTC: Non-Autoregressive End-to-End ASR with CTC and Mask Predict, *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3655–3659 (2020).

[15] Higuchi, Y., Inaguma, H., Watanabe, S., Ogawa, T. and Kobayashi, T.: Improved Mask-CTC for Non-Autoregressive End-to-End ASR, *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8363–8367 (2021).

[16] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (2019).

[17] Ghazvininejad, M., Levy, O., Liu, Y. and Zettlemoyer, L.: Mask-predict: Parallel decoding of conditional masked language models, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6114–6123 (2019).

[18] Sukhbaatar, S., Grave, E., Bojanowski, P. and Joulin, A.: Adaptive attention span in Transformers, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 331–335 (2019).

[19] Chang, X., Subramanian, A. S., Guo, P., Watanabe, S., Fujita, Y. and Omachi, M.: End-to-End ASR with Adaptive Span Self-Attention, *Proceedings of 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2020).

[20] Moritz, N., Hori, T. and Roux, J. L.: Triggered Attention for End-to-end Speech Recognition, *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5666–5670 (2019).

[21] Watanabe, S., Hori, T., Kim, S., Hershey, J. R. and Hayashi, T.: 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] Paul, D. B. and Baker, J. M.: The design for the wall street journal-based CSR corpus, *ICSLP* (1992).

[23] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al.: The Kaldi speech recognition toolkit, *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2011).

[24] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 369–376 (2006).

[25] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. and Ochiai, T.: ESPnet: End-to-End Speech Processing Toolkit, *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2207–2211 (2018).