Vol.2022-MUS-134 No.61
Vol.2022-SLP-142 No.61
2022/6/18

Tranducer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]　木田 祐介[2]　小川 哲司[1]　小林 哲則[1]

概要：Transducer 型の End-to-End 音声認識モデルを Mask-CTC を用いて事前学習することで，低遅延で高精度なストリーミング音声認識を実現することを試みた．Transducer に基づく End-to-End 音声認識モデルは，入力音声の各フレームに対してトークンの予測を行うフレーム同期型のモデルであり，Acoustic Encoder における先読み範囲を制御することで，ストリーミングによる認識が可能となる．このとき，Acoustic Encoder として Transformer などの大域的な注意機構を用いるネットワークを用いる場合，先読み範囲を長く取る (将来の文脈を考慮する) と高い認識精度が得られる一方で，遅延量は増大する．そこで本研究では，低遅延かつ高精度なストリーミング音声認識の実現を目指し，Transducer 型モデルにおける Acoustic Encoder の事前学習に Mask-CTC を利用することを試みた．Mask-CTC は，将来の情報を含む長期的な文脈を考慮した音響特徴表現を抽出可能にする枠組みであり，Acoustic Encoder の事前学習に利用することで，先読み範囲を短くすることが望ましいストリーミング音声認識に適した音響特徴表現を学習しやくすなることを期待する．提案法の有効性を実証するために，Wall Street Journal と TED-LIUM2 を用いて音声認識実験を行ったところ，提案モデルは遅延を短く抑えながら高い認識精度を与えることがわかった．

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

An Investigation of Enhancing Transducer-based Streaming ASR with Mask-CTC Pre-training

Huaibo Zhao[1]　Yosuke Higuchi[1]　Yusuke Kida[2]　Tetsuji Ogawa[1]　Tetsunori Kobayashi[1]

1. はじめに

深層ニューラルネットワーク (Deep Neural Networks; DNN) は音声認識 (Automatic Speech Recognition; ASR) における基幹技術となっている [1,2]．特に，独立に設計・最適化がなされてきた音響モデルや言語モデルといった機能要素を単一のモデルとして構築できる End-to-end モデリングへの利用は，音声認識システムの開発を容易にすることに大きな貢献を果たしている [2–4] . End-to-End 音声認識モデルとしては，Recurrent Neural Network Transducer (RNN-T) [5] などの Transducer 型モデルや，注意機構に基づく Encoder-Decoder モデル [2,4] などが一般的に用いられており，その有効性が示されている．

一方で，音声会話システムのようにリアルタイムでの音声認識が必要なアプリケーションでは，入力に同期して逐次認識結果を得るストリーミング音声認識の実現が求められる．Transformer [6–8] を用いた Encoder-Decoder モデルによりストリーミングを実現する場合，一般的に Decoder のフレーム同期化が複雑になるため，その対策が進んでいる．Monotonic Chunk-wise Attention (MoChA) [9] は，固定幅の Chunk (部分系列) に対して注意重みを計算しながら逐次認識を行うことで，ストリーミング音声認

[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
[2] LINE 株式会社
LINE Corporation

識を実現する．また，Triggered attention 型のストリーミング音声認識 [10] では，CTC [11] のアライメント情報をトリガーとして用いて Decoder を駆動する．一方，Transducer 型 End-to-End 音声認識モデルは，Acoustic Encoder の出力と Label Encoder の出力のアライメントを取るように学習されるため，デコードはフレーム同期となり，ストリーミング音声認識に適した枠組みと言える [12]．Transducer 型モデルの各モジュールには，これまで Long Short-Term Memory (LSTM) [5,13,14] が用いられてきたが，最近では，Transformer [6,12,15] を用いる Transformer-Transducer (Transformer-T) の検討が進んでいる．このとき，Transformer における注意重みの計算に将来の情報を用いることは，結果を得るまでの遅延に直結する．

ストリーミング音声認識では，一般的に，遅延時間と認識精度はトレードオフの関係にある．例えば，Transformer-T モデルを用いる場合，結果を得るまでの遅延を短くするためには Acoustic Encoder における先読みの範囲を短くすることが望ましいが，それに伴い音声認識精度は劣化する．一方で，認識精度を高く保持するために先読み範囲を長く取れば，結果を得るまでの遅延時間は長くなる．

そこで，本研究では，Transducer 型モデルに焦点を当て，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ための Acoustic Encoder の学習法について検討を行う．具体的には，Mask-CTC [16,17] により Acoustic Encoder を事前学習することで，先読みに適した特徴抽出を可能にすることを目指す．Mask-CTC は，Conditional Masked Language Model (CMLM) [18,19] と CTC のマルチタスク学習により，出力記号の長期依存関係を考慮した音響特徴表現の抽出過程を学習する枠組みであり，先読みを含め文脈を高精度に捉えるような特徴抽出が期待できる．この性質は，先読み範囲を短くしながら (低遅延で) 高い性能を与えることが求められるストリーミング音声認識と相性が良い．実際，Mask-CTC を用いた事前学習によって，Triggered Attention に基づくストリーミング音声認識モデルの性能が改善されることがわかっている [20]．本研究では，Transducer 型音声認識システムにおいても，Mask-CTC を用いて Acoustic Encoder を事前学習することで，認識精度を保持したまま遅延を短くできることを明らかにする．

本稿の構成は以下の通りである．まず，2 章で，本研究の要素技術である Transducer 型ストリーミング音声認識システムと Mask-CTC について概観する．3 章では，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ための Transducer 型モデルの構築法について述べる．続いて，4 章では，音声認識実験を通じて，提案法の有効性を実証する．最後に，5 章で本研究のまとめと今後の課題を述べる．

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 $Y = (y_u \in \mathcal{V} | l = 1, \cdots, U)$ 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 $\mathcal{V}$ に含まれる記号長 U の系列であり，y_u は u 番目の記号を表す．

次節以降，本研究の要素技術である，Transducer 型の ストリーミング End-to-End 音声認識, Chunk-wise Attention Mask，および Mask-CTC について概観する．

2.1 Transducer 型ストリーミング音声認識

Transducer 型モデルは，Acoustic Encoder，Label Encoder，Joint Network から成り，記号の出力確率は以下のように計算される．

$$\mathbf{h}^A_{1:t} = \text{AcousticEncoder}(\mathbf{x}_{1:t}) \tag{1}$$

$$\mathbf{h}^L_{1:u-1} = \text{LabelEncoder}(y_{1:u-1}) \tag{2}$$

$$\mathbf{h}_{A,L} = \text{Tanh}(\text{Linear}(\mathbf{h}^A_{1:t}) + \text{Linear}(\mathbf{h}^L_{1:u-1})) \tag{3}$$

$$P(y_u | y_{1:u-1}, \mathbf{x}_{1:t}) = \text{SoftMax}(\mathbf{h}_{A,L}) \tag{4}$$

Acoustic Encoder は，入力系列 $\mathbf{x}_{1:t}$ をベクトル系列 $\mathbf{h}^A_{1:t}$ に埋め込む (式 (1))．同時に，Label Encoder は過去に出力した記号系列 $y_{1:u-1}$ から $\mathbf{h}^L_{1:u-1}$ を得る (式 (2))．この 2 つの出力を結合ネットワークに入力し，同じ次元数のベクトルに埋め込み，加算する (式 (3))．その計算結果から語彙 $\mathcal{V}$ に対して次に出力される記号の確率分布が計算される (式 (4))．

Transducer の枠組みは，各時間フレームに対して，過去の出力記号列に基づいて次の記号を予測することで，ストリーミングの機能を実現する．このとき，Acoustic Encoder，Label Encoder には様々な構造のニューラルネットワークを利用可能である．双方の Encoder に LSTM を用いるモデルは，RNN-T として知られる [5,14]．本研究では，Label Encoder に LSTM [15] を，Acoustic Encoder には Transformer [6] と Conformer [21] を用いた．

Transformer は様々なタスクにおいて，LSTM よりも大幅に高い性能を与えた [7,22]．Transducer 型 End-to-End 音声認識でも同様である [12,15,23]．Transformer は Self-attention 機構を用いて，入力系列に対して，以下の式 (5) により注意重みを計算する．

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

ここで，クエリ $Q \in \mathbb{R}^{n_q \times d_q}$，キー $K \in \mathbb{R}^{n_k \times d_k}$，バリュー $V \in \mathbb{R}^{n_v \times d_v}$ は同じ入力系列を異なる行列に掛けて生成され，n_* と d_* は，系列長と特徴量の次元数である．Transformer は大域的な特徴を良く捉えるが，局所的な情報の

表現能力に改善の余地がある [21]．それに対し，Transformer に畳み込み機構を導入した Conformer が提案され，Transformer を上回る性能を達成している．本研究では，この 2 種類のネットワークを用いて Transducer 型ストリーミング音声認識の Acoustic Encoder を構築する．完全なストリーミングを実現するため，以下の式 (6) のように，SoftMax 関数の入力に Attention Mask W をかけることで，先読み範囲を制御する．

$$\mathrm{MaskedAttention}(Q,K,V) = \mathrm{SofMax}\left(\frac{QK^T}{\sqrt{d_k}} \cdot W\right)V \tag{6}$$

Conformer の場合，畳み込み層による遅延の増加を避けるため，causal な Depth-wise 畳み込み層を用いた．

2.2 Chunk-wise Attention Mask による先読みの制限

Transformer をストリーミング音声認識モデルとして用いる場合，Self-attention 機構において未来の情報へのアクセスを制限する必要がある [10,12]．本研究では，Acoustic Encoder に Chunk-wise Attention Mask [15] を導入することで，ストリーミング Transformer-Transducer を構築する．Chunk-wise Attention Mask は，図 1(左) のような行列で示され，これを Self-attention 機構の注意重みの計算に用いる．ここで，行列成分の値が 1 のフレームは注意重みの計算に利用し，0 のフレームは利用しないことを示す．Chunk-wise Attention Mask では，入力列を固定幅の Chunk（部分系列）に分け，各 Chunk 内のフレームに対して注意重みを計算する．ここで，2 つのフレームが同じ Chunk 内にある場合お互いに見えるが，2 つのフレームが異なるチャンクにある場合，オフセットに関わらず，左側のフレームは注意の計算で右側のフレームを見ることができない．これより，先読みの範囲が Chunk の幅によって制御される．Transformer 層を重ねる場合，図 1(右) のように，過去フレームの範囲は層数によって増加するが，先読みは常に Chunk サイズに限られる．本研究の遅延時間の計算は Chunk サイズから得る最大の先読み幅を 10ms のフレームレートに掛けた値とした．

2.3 Mask-CTC

Mask-CTC [16] では，CMLM と CTC のマルチタスクにより，Encoder-Decoder モデルを学習する．CMLM のマスク推定 [18] に基づく Decoder を学習することで，出力記号の長期依存関係を考慮した Encoder を学習することが可能であり，これはストリーミング音声認識の事前学習に有効であることがわかっている [20]．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，マスクされた記号を文脈情報に基づいて予測する．入力列 X と観測記号 Y_{obs} に対して，マスクされた記号 Y_{mask} の出力確率分布は次のように計算される．

$$P_{\mathrm{cmlm}}(Y_{\mathrm{mask}}|Y_{\mathrm{obs}},X) = \prod_{y\in Y_{\mathrm{mask}}} P_{\mathrm{cmlm}}(y|Y_{\mathrm{obs}},X). \tag{7}$$

ここで，Y_{obs} は $Y \setminus Y_{\mathrm{mask}}$ である．

3. 低遅延・高精度な Transducer 型ストリーミング音声認識の構築

Mask-CTC を用いて Acoustic Encoder を事前学習することで，低遅延かつ高精度な Transducer 型ストリーミング音声認識モデルを構築する．図 2 に提案モデルの学習法を示す．また，提案モデルは以下の 2 段階で構築される．

- **Stage 1 (先読みに適した特徴表現学習)：**Mask-CTC (CTC と CMLM のマルチタスク) の学習により，出力記号間の長期的な文脈情報を抽出可能な Encoder を獲得する．
- **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 Encoder を用いて，Chunk-wise Attention Mask を導入した Encoder のパラメータを初期化する．これを Acoustic Encoder として，Transducer 型ストリーミング音声認識モデルの学習に用いる．

以上の手法より，Mask-CTC により学習される，低遅延かつ高精度なストリーミング音声認識に適した特徴抽出プロセスが，Transducer 型のストリーミング音声認識モデルの学習を促進することを期待する．

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，英語の読み上げ音声コーパスである Wall Street Journal (WSJ) [24] および自然発話音声コーパスである TED-LIUM2 (TED2) [25] を用いた．モデルの入力として，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 次元の音響特徴量を Kaldi [26] を用いて抽出した．出力系列の単位は文字またはサブワードとした．サブワード語彙は SentencePiece [27] を用いて各訓練データから構築し，WSJ と TED2 に対する語彙サイズはそれぞれ 80 と 300 とした．

4.2 学習・推論条件

全ての実験は ESPnet2 [28, 29] を用いて行なった．Transducer 型の音声認識モデルとして，12 層の Acoustic Encoder と 1 層の Label Encoder から構成される Transformer-Transducer (Transformer-T) または Conformer-Transducer (Conformer-T) を構築した [12,15]．各 Self-attention 層において，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．Conformer の畳み込みモジュールにおけるカーネルサイズは 31 とし

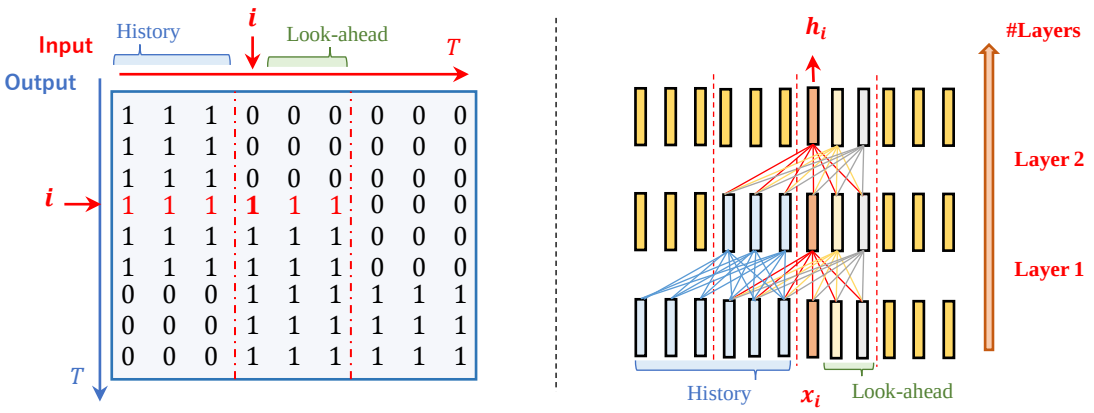

図 1: Chunk-wise Attention Mask と Self-attention 層によるフレームに含まれる過去・未来の情報.

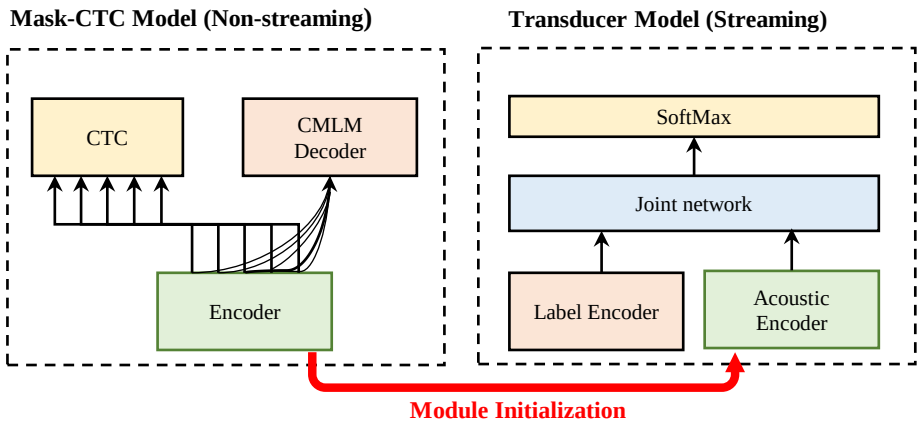

図 2: 高精度・低遅延な Transducer 型ストリーミング音声認識モデルの構築．Mask-CTC により音響エンコーダの事前学習を行う．

表 1: 遅延時間を 120ms としたときの WSJ におけるストリーミング音声認識結果.

Model	Initialization	Output unit	WER [%] eval92	WER [%] dev93
Baseline				
Transformer-T	None	character	21.5	24.1
Transformer-T	None	subword	19.5	23.3
Conformer-T	None	subword	15.5	19.2
Proposal				
Transformer-T	Mask-CTC	character	17.0	21.0
Transformer-T	Mask-CTC	subword	16.6	20.8
Conformer-T	Mask-CTC	subword	**14.9**	**18.5**

表 2: 遅延時間を 120ms としたときの TED-LIUM2 における ストリーミング音声認識結果.

Model	Initialization	Output Unit	Test WER [%]
Baseline			
Transformer-T	None	subword	15.3
Proposal			
Transformer-T	Mask-CTC	subword	**14.1**

た．ストリーミングモデルでは，Self-attention の注意重みに Chunk-wise Attention Mask を適用した．Conformer の場合，畳み込み層に対して Causal なカーネルを設計した [30]．モデルパラメータの最適化は Adam を利用し，文献 [6] と同様の学習率のスケジューリングを行なった．モデルの学習は 120 エポック行なった．推論は [5] の推論アルゴリズムに従い，ビームサーチのビーム幅は 10 とした．

Mask-CTC に基づいた Acoustic Encoder の事前学習は，文献 [16, 17] に従い，200 エポック学習を行なった．

4.3 評価項目

提案手法の有効性を実証するために，以下のモデルを比較した．

- **Baseline** [15]: 既存の Transducer 型 End-to-End 音声認識モデル．Acoustic Encoder を含む全てのモジュールのパラメータをランダムに初期化した．ストリーミングモデルとして学習する場合は，遅延時間を 120ms, 160ms または 200ms とした．非ストリーミン

表 3: WSJ における Transformer-T のストリーミング音声認識結果．出力単位はサブワードとし，異なる遅延時間に対して性能を評価した．

Model	Latency [ms]	Initialization	WER [%] eval92	WER [%] dev93
Baseline				
Transformer-T	120	None	19.5	23.3
	160		16.8	20.9
	200		15.1	18.9
	∞		14.7	17.3
Proposal				
Transformer-T	120	Mask-CTC	16.6	20.8
	160		15.0	19.0
	200		14.8	18.5

グの場合は，遅延時間を ∞ と示す．

- **Proposal**: 提案の Transducer 型 ストリーミング音声認識モデル．モデルの構造は Baseline モデルと同様だが，Acoustic Encoder を Mask-CTC を用いて初期化した．遅延時間は 120ms, 160ms または 200ms とした．

遅延時間は，Chunk-wise Attention Mask の Chunk 幅によって制御される，各入力フレームに対する Acoustic Encoder の先読み時間である (2.2 節参照)．

4.4 実験結果

4.4.1 Mask-CTC による事前学習の性能改善効果

WSJ と TED-LIUM2 における各モデルの単語誤り率 (Word Error Rate; WER) を表 1 と表 2 に示す．遅延時間は 120ms とし，WSJ では出力単位に関する比較も行なった．両コーパスにおける各モデルの結果より，**Proposal** が **Baseline** の性能を上回っていることから，Acoustic Encoder の事前学習に Mask-CTC を用いることの有効性を確認できた．また，提案手法は Acoustic Encoder の構造に依らず効果的であり，Conformer を用いることで最良の性能を与えた．表 1 で異なる出力単位の Transformer-T の結果を比較すると，サブワードを用いることでより高い性能を与えた．また，文字を用いた時の方がサブワードを用いた時よりも大きい改善率を与えたことから，提案手法はより細かい単位の出力記号の双方向依存関係モデリングに役立つこともわかる．

4.4.2 遅延時間に関する性能変化の分析

各遅延時間 (120ms, 160ms, 200ms) における Transformer-T モデルの単語誤り率を表 3 に示す．**Proposal** は遅延時間に依らず **Baseline** の性能を上回った．特に遅延時間を 120ms とした時，提案モデルがより良好な性能を与えたことから，提案手法は先読み幅が短くても高精度な特徴抽出が可能であることを示唆している．また，**Proposal** は **Baseline** よりも少ない遅延時間で高い認識精度を達成した．例えば，遅延時間 160ms の **Proposal** の WER (eval92 で 15.0%，dev93 で 19.0%) は，遅延時間 200ms の **Baseline** の WER (eval92 で 15.1%，dev93 で 18.9%) と同等である．また，遅延時間 200ms の **Proposal** と非ストリーミングの **Baseline** を比べると，eval92 において同等な精度を達成できることを確認した．

5. まとめ

本研究では，Transducer 型モデルを用いて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ために，Mask-CTC により Acoustic Encoder を初期化することを試みた．Mask-CTC は出力記号に関する長期文脈を考慮した音響特徴表現の学習法であり，ストリーミングにおける先読みに貢献することが期待される．WSJ と TED-LIUM2 を用いた実験を通して，提案モデルが既存の Transducer 型ストリーミング音声認識システムと比較して低遅延で高精度な認識を達成できることが明らかになった．

参考文献

[1] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal processing magazine*, Vol. 29, No. 6, pp. 82–97 (2012).

[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] 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).

[4] 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).

[5] Graves, A.: Sequence Transduction with Recurrent Neural Networks, *ArXiv*, Vol. abs/1211.3711 (2012).

[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] Chiu, C.-C. and Raffel, C.: Monotonic Chunkwise Attention, *ArXiv*, Vol. abs/1712.05382 (2018).
[10] 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).
[11] 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).
[12] Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S. and Kumar, S.: Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7829–7833 (2020).
[13] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, pp. 1735–1780 (1997).
[14] He, Y., Sainath, T. N., Prabhavalkar, R., McGraw, I., Álvarez, R., Zhao, D., Rybach, D., Kannan, A., Wu, Y., Pang, R., Liang, Q., Bhatia, D., Shangguan, Y., Li, B., Pundak, G., Sim, K. C., Bagby, T., yiin Chang, S., Rao, K. and Gruenstein, A.: Streaming End-to-end Speech Recognition for Mobile Devices, *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6381–6385 (2019).
[15] Chen, X., Wu, Y., Wang, Z., Liu, S. and Li, J.: Developing Real-Time Streaming Transformer Transducer for Speech Recognition on Large-Scale Dataset, *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5904–5908 (2021).
[16] 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).
[17] 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).
[18] 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).
[19] 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).
[20] Zhao, H., Higuchi, Y., Ogawa, T. and Kobayashi, T.: An Investigation of Enhancing CTC Model for Triggered Attention-based Streaming ASR, *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 477–483 (2021).
[21] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition, *ArXiv*, Vol. abs/2005.08100 (2020).
[22] 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).
[23] Yeh, C.-F., Mahadeokar, J., Kalgaonkar, K., Wang, Y., Le, D., Jain, M., Schubert, K., Fuegen, C. and Seltzer, M. L.: Transformer-Transducer: End-to-End Speech Recognition with Self-Attention, *ArXiv*, Vol. abs/1910.12977 (2019).
[24] Paul, D. B. and Baker, J.: The design for the Wall Street Journal-based CSR corpus, *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York* (1992).
[25] Rousseau, A., Deléglise, P. and Estève, Y.: Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks, *Language Resources and Evaluation Conference (LREC)* (2014).
[26] 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).
[27] Kudo, T. and Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, *arXiv preprint arXiv:1808.06226* (2018).
[28] 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).
[29] Boyer, F., Shinohara, Y., Ishii, T., Inaguma, H. and Watanabe, S.: A Study of Transducer Based End-to-End ASR with ESPnet: Architecture, Auxiliary Loss and Decoding Strategies, *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 16–23 (2021).
[30] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: Wavenet: A generative model for raw audio, *arXiv preprint arXiv:1609.03499* (2016).