Vol.2022-MUS-134 No.14
Vol.2022-SLP-142 No.14
2022/6/17

ゼロ資源言語の認識に向けたMAMLに基づくend-to-end音声認識の検討

周　鋭[1,a)]　伊藤 彰則[1,b)]　能勢 隆[1,c)]

概要：高い精度の音声認識システムを開発するためには，大量のラベル付け音声データが必要である．しかし，世界中にある言語の大部分については，そのように多くの学習データを用意することはできない．このような言語の音声認識システムを開発するため，学習時に目標言語のデータを使用しない，または非常に少ないデータのみを使って開発する音声認識システムがゼロリソース音声認識システムである．本稿では，メタラーニングの学習方法である MAML とハイリソース言語データを用いて，モデルを事前学習する．そして，少数の目標言語のデータを使って，モデルを微調整する．20 分程度の目標言語データを使用して，MAML による学習を行ったところ 40 %程度の CER が得られた．

キーワード：ゼロリソース，メタラーニング，音声認識

MAML-based End-to-End Speech Recognition for Zero-Resource Language Recognition

Zhou Rui[1,a)]　Ito Akinori[1,b)]　Nose Takashi[1,c)]

Abstract: High-precision speech recognition systems require large amounts of labeled speech data. However, most of the languages in the world do not have that much training data. A speech recognition system that uses no speech data, or uses very little data from such training target languages, is called a zero resource speech recognition system. In this paper, we pre-train a model using a training method called MAML and high-resource language data. Then we use a data of a minority language to fine-tune the model. When we use 20 minutes of the target language data, we obtained a character error rate of about 40%.

Keywords: Zero Resource, Meta Learning, Speech Recognition

1. はじめに

自動音声認識（ASR）は，入力された音声信号から，音声とテキストを出力する技術である．音声認識技術は，私たちの日常生活の中で幅広く活用されている．例えば，スマートフォンで調べたいことを口に出して検索したり，チャットの際に文字を打ちたくない場合は直接音声で入力したり，自動車に音声で指示を出したりすることができる．音声認識モデルには，2 つのタイプがある．それは，End-To-End モデルと非 End-To-End モデルである.2010 年以前には，主に非 End-To-End モデルである GMM-HMM[1] が使われていた．しかし，ニューラルネットワークの発展により，GMM から DNN，すなわち DNN-HMM[2] への置き換えが進んだ．非 End-To-End モデルは，HMM と DNN，CNN，RNN のハイブリッドモデルであり，音響モデル，組合せ発音モデル，言語モデルの 3 つのモデルを組み合わせたものである．これに対して，End-To-End モデルは，すべての音声認識ステップを 1 つのモデルで行うため，よりシンプルで，パラメータも少ない．

このような音声認識モデルを学習させる通常の方法で

[1] 東北大学工学研究科
Tohoku University Graduate School of Engineering
[a)] zhou.rui.p1@dc.tohoku.ac.jp
[b)] akinori.ito.a2@tohoku.ac.jp
[c)] nose@tohoku.ac.jp

は，大量のラベル付き音声データが必要である．しかし，世界には 6000 以上の言語が存在し，これらの言語の多くはそれほど多くのデータを有していない．また，これらの言語のデータを記録するには，多くの費用と時間がかかるため，通常の方法では音声認識モデルを学習させることができない．このような目標言語を学習に用いずに，あるいはごくわずかなデータで音声認識システムを構築することを，ゼロリソース音声認識と呼ぶ．ゼロリソース音声認識を行う際には，高リソース言語が多数存在することを利用し，大量の高リソース言語を使ってモデルを事前学習させ，少数のターゲットとなるゼロリソース言語を使ってモデルを微調整すれば，ターゲット言語での音声認識が可能になると考えられる．

本論文では，まず，事前学習法であるメタ学習を紹介する．次に，音声認識のための End-To-End モデルの構成とパラメータを説明する．最後に，メタ学習法の有効性を証明するために 3 つの実験を行い，結論を導き出す．

2. メタラーニング

2.1 メタラーニングとは

通常の機械学習は，機械に特定のタスクを完了する能力を待たせる学習である．例えば，動物の写真を機械に与えると，機械はこれらの動物の分類を学習する．しかし，今まで登場したことのない他の写真を分類することはできない．しかし，メタ学習は学習することを学習する方法であり [3]，機械は決まったタスクを実行する能力を学習するのではなく，機械学習機能を実行するために学習するのである．例えば，写真の分類モデルを学習する場合，まず，猫と犬の写真をモデルに入力すると，モデルは猫と犬を分類する能力を持ち，次に，リンゴとオレンジの写真も同様に，モデルはリンゴとオレンジを分類することができる．このように多くの分類課題を与えると，モデルは決まった方法で分類することを学習するだけでなく，この課題を分類することを学習するようになる．したがって，モデルに車と自転車の絵を与えると，モデルはすぐに車と自転車を分類するタスクを学習することができるようになる．

2.2 メタラーニングと音声認識

先ほどの例は，メタ学習による画像分類の例である．音声認識もこれと同じように実現できる．まず，日本語，中国語，英語など多くの高リソース言語のデータを使ってモデルを事前に学習する．そして，そのモデルに音声を認識する能力を与える．次に，この学習済みのモデルを用いて，対象となるゼロリソース言語のわずかなデータを用いてモデルを微調整し，この対象言語での音声認識を実現する．

2.3 Model-Agnostic Meta Learning

Model-Agnostic Meta Learning (MAML) はメタ学習法の一つであり，Model-Agnostic はモデルに依存せず，全てのモデルがこの方法を利用することができる [4]．深層学習では，多くのメタ学習モデルが提案されており，それらは初期化モデル重みの学習と転移可能なオプティマイザの学習に大別される．その中でも MAML は前者に属する．MAML は，新しいタスクに迅速に適応するために，良好な初期化パラメータを学習する．すなわち，少ない学習データで素早く収束し，微調整を完了することができ，少量データへの過剰適合に悩まされることがない．メタ学習はタスクを最小単位とし，言語の種類ごとにタスクを作成し，各タスクは 1 つの音声とする．各タスクセットにはサポートデータとクエリデータが 1 つずつ含まれ，サポートデータとクエリデータは同じ言語のタスクからランダムに選択される．図 1 は MAML 学習のステップである，最初にモデルをコピーしてコピーモデルを作成する．コピーモデルはサポートデータを使用してロスを計算し，コピーモデルのパラメータを一回更新する．次に，更新したモデルとクエリデータを使ってロスを計算する．このロスを使って，更新モデルの勾配を計算し，さらに，本来のモデルのパラメータを更新する．このような過程を繰り返して，全てのタスクを学習することを一つのエポックとする．

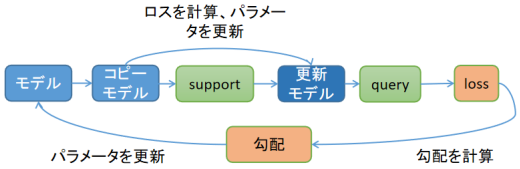


図 1: MAML 学習ステップ
Fig. 1 MAML Training Step

3. 音声認識モデル

3.1 モデルの構造

事前学習や微調整をするためには，音声認識モデルとして End-To-End モデルを用いる必要がある．広く使われている End-To-End モデルにはいくつかの種類がある．それらは，Listen, Attend and Spell(LAS)[5], CTC[6], RNN-Transducer[7], Transformer-Transducer[8], Monotonic Chunkwise Attention[9] などである．微調整に用いるデータ数は非常に少ないため，モデルが複雑であってはならず，音声認識モデルとして LSTM-CTC を採用することにした．図 2 は音声認識モデルの構造である．入力 X は音声信号であり，事前学習で共有する部分は，信号処理の VGG Extractor[10] と Bi LSTM[11] である．言語が異なると出力される文字の種類が異なるため，CTC Loss を計算するために，言語ごとの head layer が必要である．表 1 はモデルのパラメータである．LSTM は 6 層で，セルは 360 である，ドロップアウトはすべて 0.1 とした．

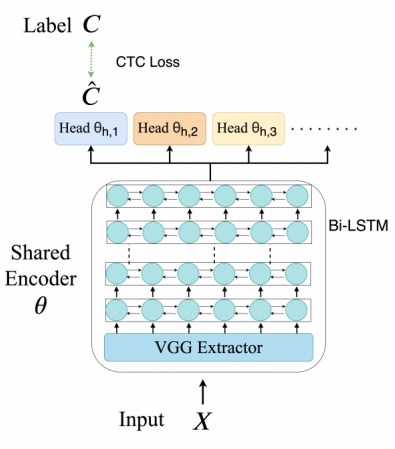


図 2: モデルの構造
Fig. 2 Structure of model

表 1: モデルのパラメータ
Table 1 Parameter of model

モデル	パラメータ
LSTM Layer	6
LSTM Cell	360
LSTM Dropout	0.1
VGG Dropout	0.1

3.2 CTC Loss

音声認識における音声は，一般的にテキストよりも長くなる．そのため，テキストに対応する音声の長さを決めることができないが，CTC はこの長さの問題を解決する．音声信号の長さを L，テキストの文字数を U とし，テキストの全長が L となるように，L-U の数の空白トークンをランダムにテキストトークンに挿入する．このような挿入を行う場合のすべてのアライメント確率の和を使ってロスを計算する．Loss = − log(P(Y|X))，ここで X は認識される音声で，Y は認識したテキスト結果で，P は音声認識結果の確率である．

4. 実装と実験

MAML の効果を証明するために，三つの実験を行った．三つの実験で用いる事前学習モデルは同じなので，まず事前学習のデータを示す．

表 2 は事前学習データの情報である．言語データは common voice コーパス [12] のものを使用した．これは九種類の言語で，キャラクターはアルファベットである．言語ごとに 2000 のタスクセットを作成したので，全部で 18000 のタスクセットを使用した．

4.1 大量の英語データを用いた実験

最初の実験は，異なる事前学習エポックに対応するモデルで，5000 個の英語ファイルを学習した結果である．英語ファイルは，Librispeech コーパス [13] の 360 時間の学習データから選択された．事前学習エポックは 3，微調整エポックは 10 である．評価指標は Train-Loss，Valid-Loss，Character Error Rate，Word Error Rate とした．図 3 は，5000 個の英語ファイルの微調整の結果を示している．epoch0 は事前学習モデルを用いない通常の学習方法，epoch1〜epoch3 は事前学習エポックの数である．その結果，事前学習モデルは，通常の学習方法よりも早く低損失を実現し，文字誤り率，単語誤り率も低くなっていることがわかる．しかし，エポック数を増やしても，効果はあまり上がらない．

表 2: 事前学習データ情報
Table 2 Pre-training data information

コーパス	時間 (h)	音声ファイル数 (個)	キャラクター
Chuvash	1.38	1211	86
Frisian	5.05	3707	77
Greek	2.12	1976	93
HakhaChin	0.64	815	58
Kyrgyz	2.28	1788	76
Maltese	2.44	1977	81
Sakha	2.34	1444	73
Slovenian	1.33	1348	65
Latvian	2.77	3198	80

表 3: 英語データの情報
Table 3 English data information

ファイル数	5000	3000	1000	500	100
時間 (h)	17.3	10.3	3.4	1.74	0.35

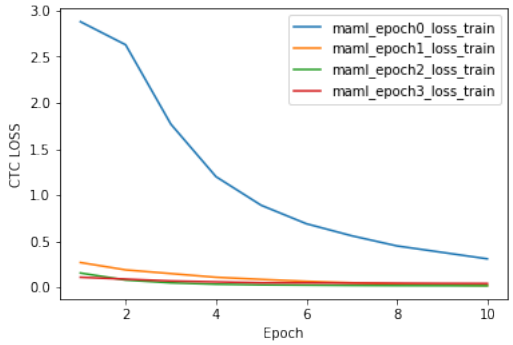
4.2 少量の英語データによる実験

前記の実験は，英語の微調整用ファイルを 5000 個使用した実験である．これに対し，リソースが少ない場合の効果を実証するため，微調整用の英語ファイルの数を減らして，実験を実施した．このとき使用した事前学習モデルは，8 エポックモデルである．英語ファイルの数は表 3 に示す通りである．英語ファイル数は 5000, 3000, 1000, 500, 100 であり，微調整エポック数は 10 である．評価指標は Train-Loss,Valid-Loss,Character Error Rate,Word Error Rate である。

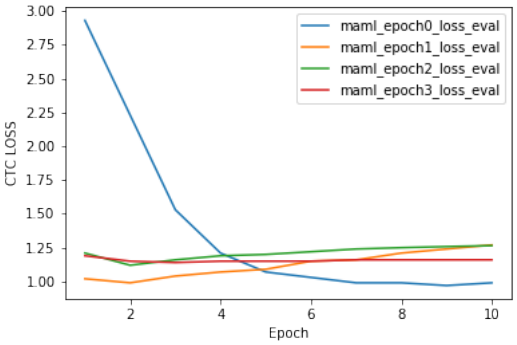
結果を図 4 に示す．結果から見ると，5000 ファイルと 3000 ファイルの場合の最終的な CER と WER はほぼ同じである．ファイルの数が減少するにつれて，CER と WER は増加する．しかし，一番少ない 100 個のファイルでも，CER が 28%，WER が 70%であり，通常の学習方法よりも良い結果となった．

4.3 少数言語の実験

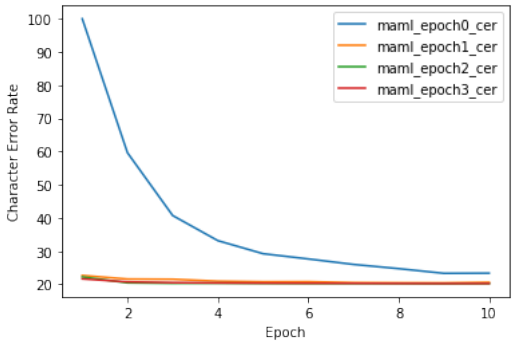
前節の英語実験の結果から，MAML を使うことにより，微調整データが非常に少ない場合でも，ある程度の認識性



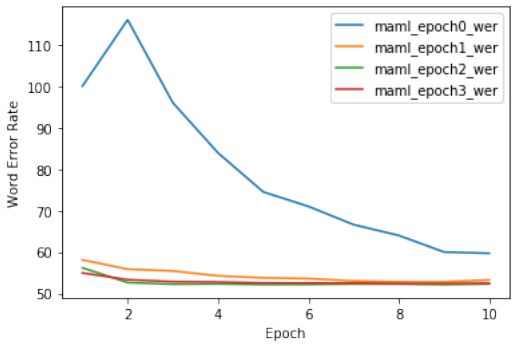
(a) Train-Loss



(b) Valid-Loss



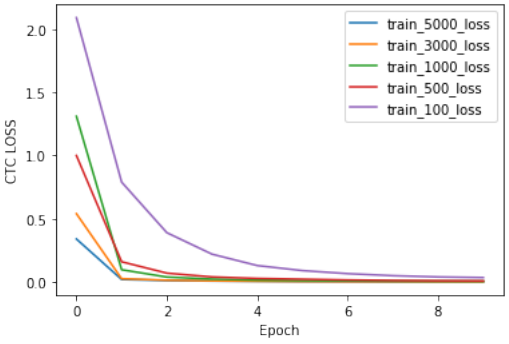
(c) Character Error Rate



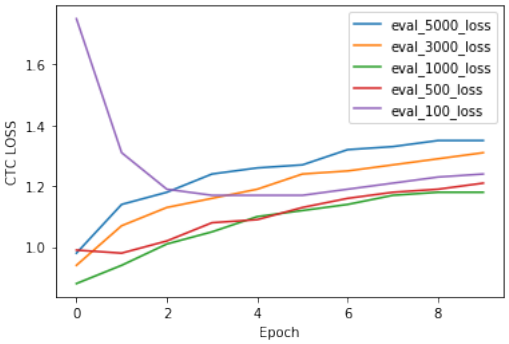
(d) Word Error Rate

図 3: 5000 英語微調整ファイル，異なる epoch の結果

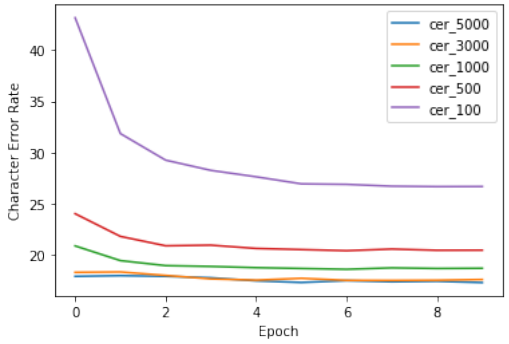
Fig. 3 5000 English fine-tuning files, different epoch results



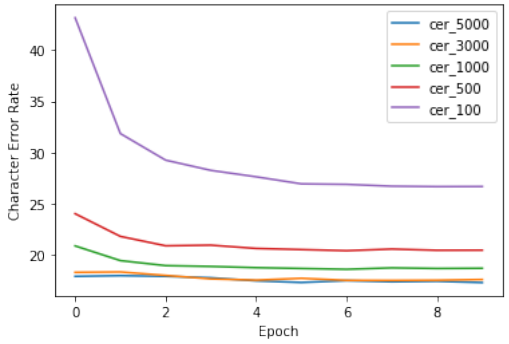
(a) Train-Loss



(b) Valid-Loss



(c) Character Error Rate



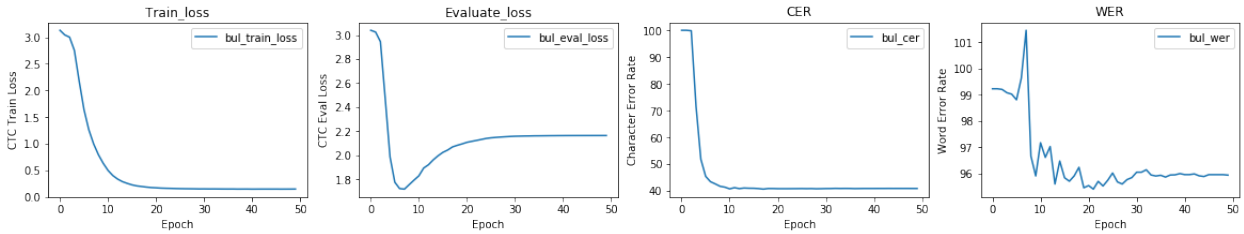
(d) Word Error Rate

図 4: 異なる微調整英語ファイル数の結果

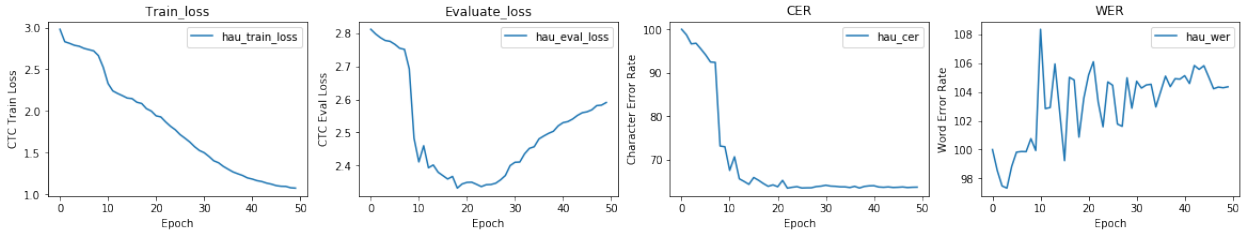
Fig. 4 Results of different number of fine tune English files

能が得られた．この結果を受けて，少数言語の認識実験を行った．表 4 は微調整用の少数言語の情報である．実験に使用したデータの長さは 30 分程度であり，通常の学習方法では全く認識をすることができない．前節の英語実験と同じように 8 epoch の事前学習モデルを使用して，この少数言語で微調整を行う．微調整の epoch 数は 50 とし，評価指標は前節と同じとした．

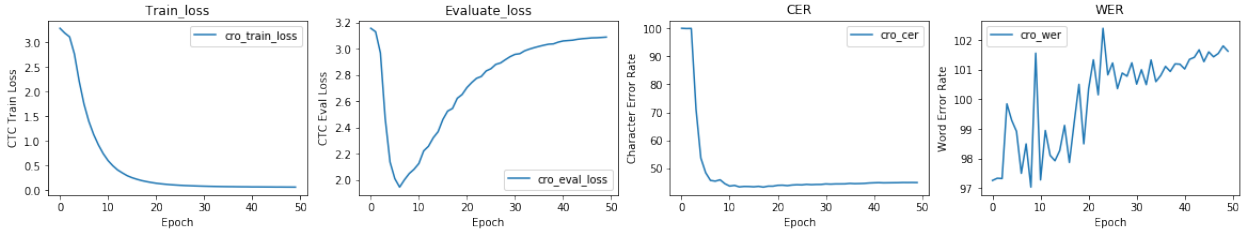
図 5 はそれぞれの言語の結果である．30 分程度の学習



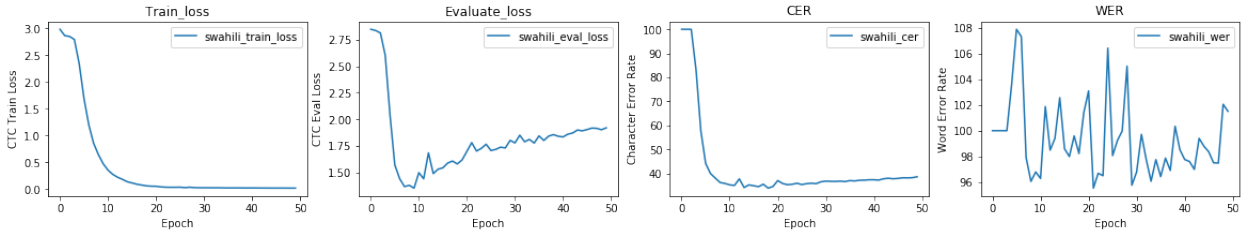
(a) Bulgarian 語の結果



(b) Hausa 語の結果



(c) Croatian 語の結果



(d) Swahili 語の結果

図 5: 少数言語実験の結果

Fig. 5 Results of minority language experiments

表 4: 微調整用少数言語のデータ情報

Table 4 Minority language data information for fine tuning

コーパス	時間 (h)	音声ファイル数 (個)	文字種
Bulgarian	0.44	200	66
Croatian	0.35	150	77
Hausa	0.24	200	55
Swahili	0.25	150	39

表 5: ゼロリソース言語の実験結果

Table 5 Experiment results of zero resource language

言語	CER(%)	WER(%)	AVG_LOSS
English	40.66	84.7	2.93
Bulgarian	93.16	99.9	5.63
Croatian	90.77	99.9	5.53
Swahili	90.2	100	4.94
Hausa	93.6	100	6.28

データにより，50%以下の CER が得られた．通常の学習方法では全く認識できないので， MAML の効果があったことが分かる．

4.4 ゼロリソースの実験

前の実験では目標言語のデータを少量使用して微調整を行った．最後の実験では事前学習モデルだけを使用して，微調整なしに目標言語の音声を認識する．言語は四種類の少数言語と英語を使用した．結果を表 5 に示す．

ゼロリソースの実験結果から，微調整をしない場合，英語の CER だけが 40%であったものの，他の言語はあまり認識されていない．また，LOSS も大きくなっている．

5. おわりに

本論文では，まず，ゼロリソースの音声認識手法であるメタ学習を紹介した．その後，メタ学習を実装し，実験を行った．結論は 3 つある．第一に，メタ学習は，事前に学習したモデルを用いることで，高リソース音声認識の効果を向上させることができる．第二に，事前学習済みモデルを用い，非常に少ないデータで微調整を行うことで，ある程度の認識が可能となる．わずか 30 分程度の音声データで，50%以下の文字誤り率を得ることができた．最後に，対象言語での微調整を行わず，事前学習済みモデルのみを用いた場合，少数言語の認識性能はよくなかった．今後は，さらに性能を向上させるため，言語間の文字表記の統一，より少ない微調整データによる認識方法の検討などを行っていく．

参考文献

[1] Jaitly, N., Nguyen, P., Senior, A., & Vanhoucke, V. (2012). Application of pretrained deep neural networks to large vocabulary speech recognition.
[2] Li, J., Yu, D., Huang, J. T., & Gong, Y. (2012, December). Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM. In 2012 IEEE Spoken Language Technology Workshop (SLT) (pp. 131-136). IEEE.
[3] Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. Artificial intelligence review, 18(2), 77-95.
[4] Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126-1135). PMLR.
[5] Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4960-4964). IEEE.
[6] Paterlini-Brechot, P., & Benali, N. L. (2007). Circulating tumor cells (CTC) detection: clinical impact and future directions. Cancer letters, 253(2), 180-204.
[7] Graves, A. (2012). Sequence transduction with recurrent neural networks. arXiv preprint arXiv:1211.3711.
[8] Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., & Kumar, S. (2020, May). Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7829-7833). IEEE.
[9] Chiu, C. C., & Raffel, C. (2017). Monotonic chunkwise attention. arXiv preprint arXiv:1712.05382.
[10] Beckmann, P., Kegler, M., Saltini, H., & Cernak, M. (2019). Speech-vgg: A deep feature extractor for speech processing. arXiv preprint arXiv:1910.09909.
[11] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 18(5-6), 602-610.
[12] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670.
[13] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5206-5210). IEEE.