End-to-End モデルによる音声認識

河原達也（京都大学）

1. はじめに

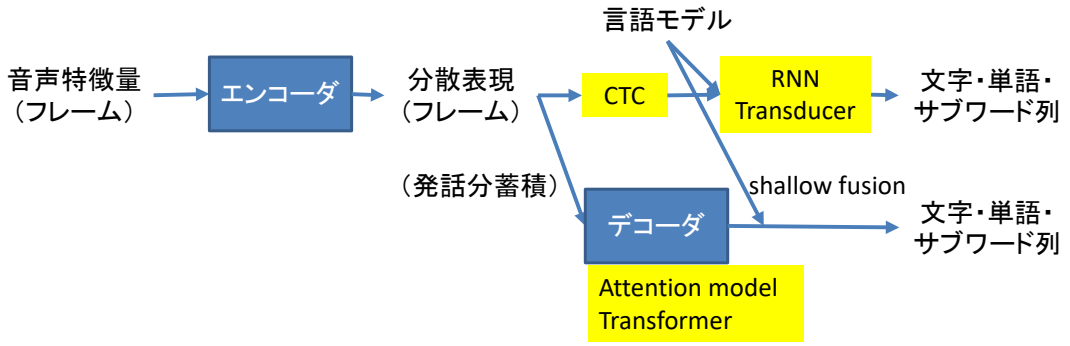
音声認識は、時系列である音声波形あるいは周波数スペクトログラムから、文字列や単語列へ変換を行う処理であり、時間的な変動も扱う必要がある。したがって、単純な深層ニューラルネットワーク(DNN)ではモデル化・実装できない。そのため、局所的に DNN を適用し、その出力を隠れマルコフモデル(HMM)で処理する方式(DNN-HMM)が長く用いられてきた。その学習には、音声を HMM の各状態に対応する局所的なパターンに区分化する前処理が必要で、単語辞書や言語モデルも個別に学習され、全体として最適化がされていなかった。

これに対して、HMM や単語辞書を介することなく、音声から文字列や単語列を直接認識する End-to-End モデル[1][2]及びその学習法が検討されるようになり、現在では主流となっている。本節ではその主なモデル・手法を紹介する。

2. 音声認識のための End-to-End モデルの分類

現在音声認識において主に用いられている End-to-End モデルを分類したものを図 1 に示す。音声の特徴量をフレーム毎に処理する部分をエンコーダ(encoder)と呼ぶ。これについては、すべてのモデルで同様である。この結果をパイプライン的に処理するのが、CTC であり、これに言語モデルを統合したものが RNN トランスデューサ(RNN

Transducer)である。一方、エンコーダで得られる分散表現を発話ブロックで蓄積して、別のデコーダ(decoder)を適用するアテンションモデル(Attention model)やトランスフォーマ(Transformer)などの方式もある。これは、自然言語処理で用いられているモデルと本質的に同じであるが、テキストを入力とする場合に比べて音声を入力する場合は、音声を処理するエンコーダの比重（パラメータ数）が大きくなる。



	エンコーダ	デコーダ	
CTC	LSTM/Transformer/Conformer		（フレーム単位）
RNN Transducer	LSTM/Transformer/Conformer	RNN言語モデル	
Attention model	bi-LSTM	uni-LSTMデコーダ	（発話・ブロック単位）
full Transformer	Transformer/Conformer	Transformerデコーダ	

図 1　音声認識のための End-to-End モデルの分類

次に音声認識に用いられる単位について述べる。従来の DNN-HMM などのモデルでは、音素が認識の単位として一般に用いられていた。音素は、音声に直接写像すると考えられるが、文字と音素の変換（G2P ツールや発音辞書）が必要となり、その誤りや曖昧性もあるため、End-to-End モデルではあまり用いられない。

代わりに、文字を直接用いることができる。多くの言語ではエントリ数は少ないが、言語モデルとしては弱い。ただし、日本語や中国語ではエントリ数が多く、現実的な選択の１つである。

End-to-End モデルでは単語または形態素を直接認識の単位とすることもできる。言語的制約は強くなるが、エントリ数が多くなり、頻度の少ない単語や未知の単語への対応の問題も生じる。

したがって、現在最も一般的に用いられているのは、文字と単語の中間のサブワード単位である。このサブワードを構成する方法にはいくつかある。BPE (Byte-Pair Encoding) は、文字(unicode の Byte)の系列の頻度に基づいて結合するもので、大規模言語モデルでも一般的に採用されている。Word piece は、単純な頻度でなく、頻度比に基づいて結合するものである。これらは通常、単語の境界をまたがないように構成されるが、事前に単語単位に分割されていることが前提である。これに対して、Sentence piece では unigram モデルに基づいて文の尤度を最大化するように単位を構成する。これは確率的に分割を行うもので、単語境界が明確でない日本語・中国語などで有用である。大規模言語モデルでは、数十万以上の BPE トークンが用いられることが一般的であるのに対して、音声認識では数千程度のサブワード単位が一般的である。

3. Connectionist Temporal Classification (CTC)

HMM を用いることなく、ニューラルネットワークのみで時系列パターンを分類しよ

うと定式化されたのが CTC[3]である。CTC では通常、文字やサブワード単位の LSTM が用いられる。これに加えて、どの文字・サブワードでもないブランク記号（_）を導入し、各文字・サブワードの間に挿入する。入力時間フレーム毎にこれらの記号が出力され、これを集積する。この際に、ブランク記号を消去し、時間的に連続した同一の出力記号を 1 つに縮約する操作を行う（図 2）。

これを確率的に定式化すると、各フレームの音響特徴量に対して各記号の事後確率を計算し、同じ記号列 S に縮約されるものの総和を計算することになる。例えば、以下の記号系列はすべて hai を表現するものとしてまとめられる。

_h_____a_____i__

_hh____aa____ii_

_h____aaaa__iii_

モデル学習の際には、正解記号系列 S（上記の例では hai）に縮約されるすべての系列の尤度の総和を求める。この尤度計算及び勾配計算は、HMM の尤度計算と同様に、前向き・後向きアルゴリズムで効率的に実現できる。この対数尤度を元に、LSTM の各パラメータを更新する学習則が導出される。CTC では LSTM により時間フレーム間の依存性はモデル化されるが、記号間の関係は独立に扱われている。CTC は非常に簡易で、双方向のモデルを用いなければリアルタイムに動作するが、言語モデルを用いないため、認識精度には限界がある。

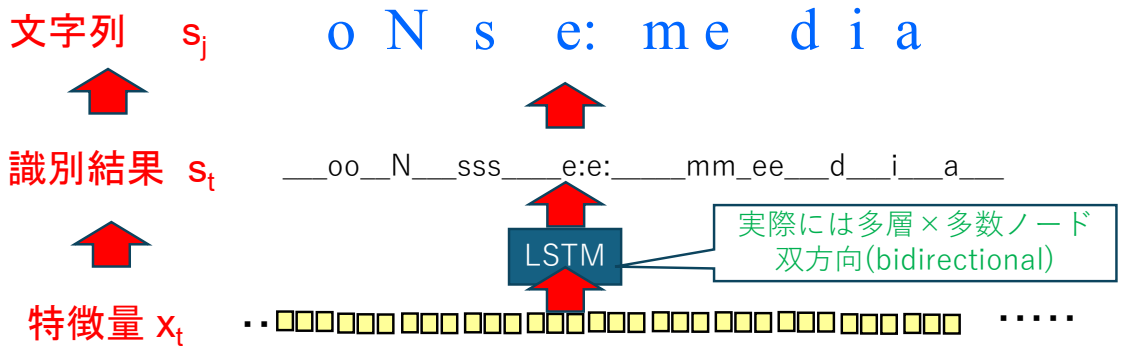


図２　CTC

4.　RNN トランスデューサ（RNN transducer）

CTC のモデルに再帰型ニューラルネットワーク(RNN)に基づく言語モデルを統合したのが RNN トランスデューサ[4]である。フレーム毎の音響エンコーダ（図１のモデルに相当）の出力と各文字・サブワード単位の言語モデルの出力を統合し、次の文字・サブワードの確率を逐次計算していく（図３）。RNN トランスデューサは、実装が大がかりであるが、リアルタイムに動作し、認識精度も高い。

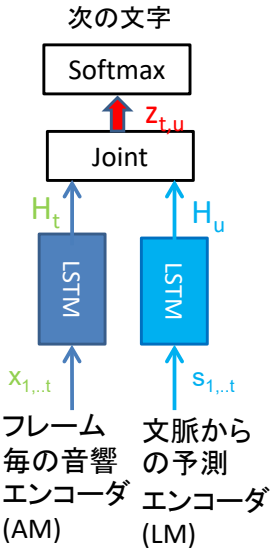


図３　RNN トランスデューサ

5.　アテンションモデル（Attention model）

アテンションモデル[5]は、正確には注意機構付きエンコーダ－デコーダ(encoder-decoder)モデルであり、符号部と復号部から構成される（図４）。

エンコーダでは、入力フレーム毎に音響特徴量を LSTM により別の数値ベクトル（分散表現）に符号化する。デコーダは、この符号化された分散表現の系列を入力として、文字やサブワードなどの記号の系列を順次予測する。その際に、例えば最初の方の音素は音声の最初の方の情報を用いるのが有用であるので、重みを付ける。これが注意機構であり、この重み自体も動的に計算され、重みを計算するパラメータは統合的に学習される。

その上で、出力の文字やサブワードの記号はこの内部状態に基づいて計算される。デコーダは、文頭(sos)記号から予測を開始し、文末(eos)記号が出力されると終了する。デコ

ーダ LSTM は次の記号を予測する際に、直前の状態に加えて、直前の記号を用いており、言語モデルの機能を包含している。エンコーダ・デコーダ及び注意機構の学習は、正解記号系列と予測記号系列のクロスエントロピに基づいて統合的に行われる。以上をまとめると、入力音響特徴量系列 X をいったん分散表現の系列 H に変換した上で、p(S|H)=p(S|X)が最大となる記号系列 S を出力するモデルであるので、エンコーダは音響モデル、デコーダが言語モデルに対応すると考えられる。また、注意機構は音声と記号の対応付け（アライメント）を行うものと捉えられる。このアライメントは原理的には任意の対応付けが可能であるが、音声の場合は時間方向に単調に文字やサブワード等の記号に対応付けられるので、直前の重みに依存させるとともに、CTC とマルチタスク学習を行うことが多い。また、アテンションモデルでは、入力の一部（雑音区間など）をスキップしたり、出力が字幕テキストのように音声に忠実でなくてもある程度対応づけができる。アテンションモデルではすべての入力をいったん処理しないと認識を開始することができないが、高精度の認識を行うことができる。

　デコーダに内包される言語モデルは、学習音声データに対応するテキストのみからしか学習されないので、より大規模なテキストのみのデータから学習した外部言語モデルを統合(Shallow fusion)することも行われる。

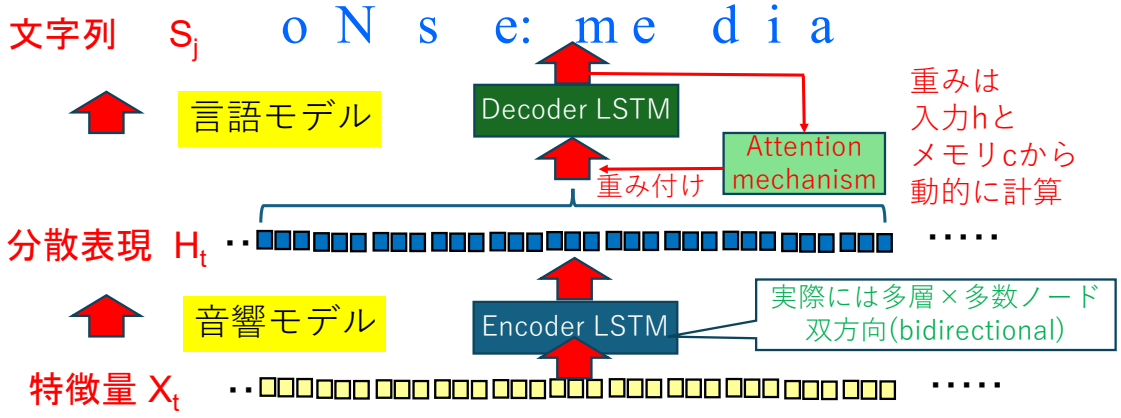


図４　アテンションモデル

6.　トランスフォーマ（Transformer）

トランスフォーマ[6]は前記のアテンションモデルで用いられていたLSTMなどのリカレントなモデルの代わりに自己注意機構(Self-Attention)を導入したものである（図５）。本来、音声やテキストの入力系列の長さは可変であるが、（０パディングなどにより）むりやり固定長にしている。これにより、入力全体及び特徴軸全体における関係や重要性を考慮した特徴抽出が行われる。これを多段に行うことで、抽象化が行われる。また、認識の際には入力に応じた重みづけが行われる。LSTM 等の再帰型ネットワークと異なり、深層化及び並列化の効果が大きい。

自己注意機構を図６に示す。音声の場合、行列の横軸を時間フレーム、縦軸を周波数ビン、つまりスペクトログラムと考えるとわかりやすい。自己注意機構の行列は、通常複数の部分空間（マルチヘッド）を用いて構成され、その処理結果が組み合わされる。また残

差型のネットワークとして構成され、レイヤー正規化が行われる。レイアー正規化は、事前に時間方向に、事後にチャネル方向に行われるのが一般的である。また、デコーダの自己注意機構では、将来の情報をみないようにマスクをかける。

自己注意機構にさらに、コンボリューション層を追加したものをコンフォーマ(Conformer)と呼ぶ。これにより、周波数軸・時間軸の局所的変動を吸収し、自己注意機構の大域的特徴抽出を補完する。コンフォーマは現在、エンコーダとしては最高の性能を示している。

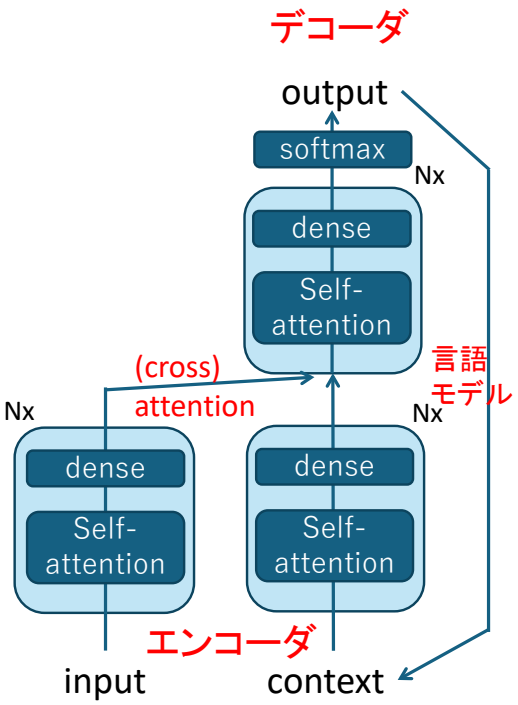


図5　トランスフォーマ

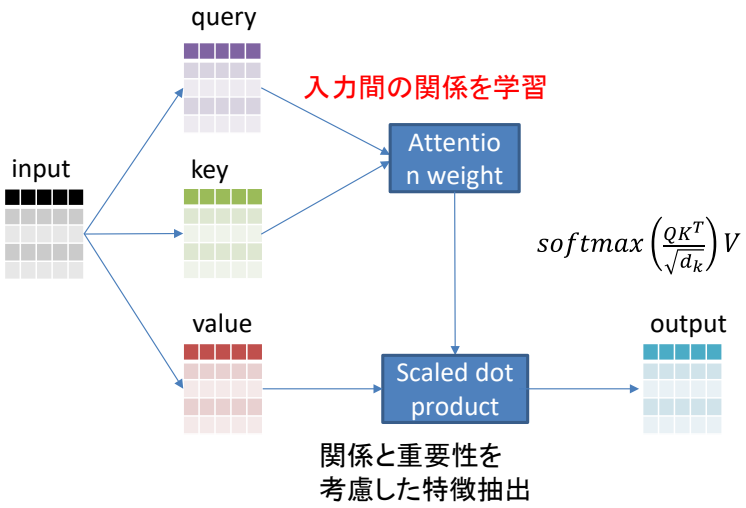


図 6　自己注意機構

7.　自己教師付き学習に基づく大規模事前学習モデル

End-to-End モデルは、認識システムの構成が単純で、音響モデル・言語モデルの統合的最適化が行われるが、学習に膨大なペアデータが必要となる。その反面、テキストのみのデータや音声のみのデータを活用できない。テキストのみのデータでトランスフォーマの自己教師付き学習(SSL: Self-Supervised Learning)を行う大規模言語モデルが大きな成功を収めたのを機に、音声のみのデータで自己教師付き学習を行う方法も研究されている[7]。

現在一般的に用いられている大規模事前学習モデルは、トランスフォーマと量子化器（コードブック）を組み合わせたものであり、概ね図 7 に示す枠組みである。その代表的なものが、wav2vec 2.0[8]と HuBERT[9]である。

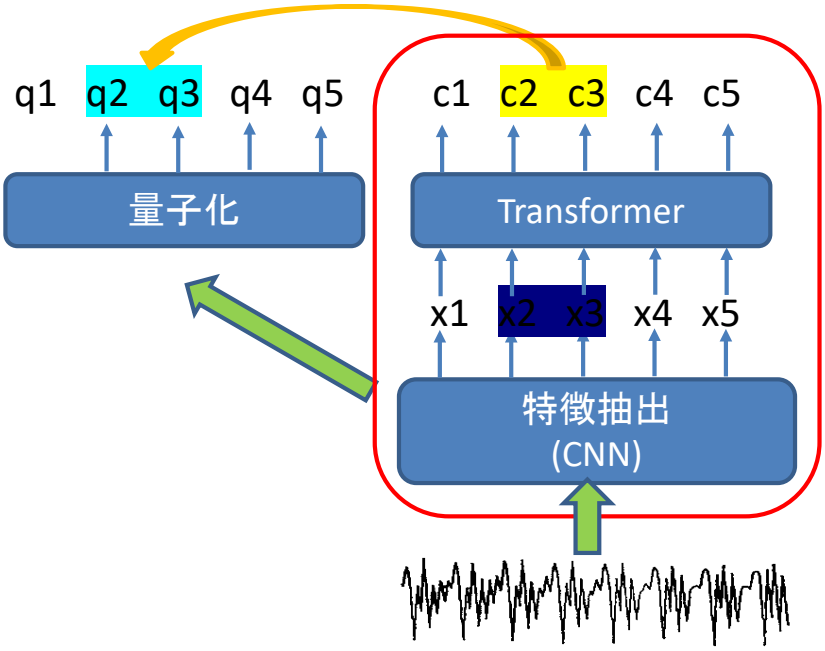


図 7　wav2vec 2.0 / HuBERT

入力音声またはその周波数特徴量は、CNN (Convolutional Neural Network)に入力され、フレーム毎に特徴抽出が行われる。その出力がトランスフォーマに入力される。その際に、一部のフレームがマスクされ、トランスフォーマで前後の文脈を用いてその復元が行われる。ただし、音声では同じ音素が複数フレーム連続するのが一般的であるので、ある程度の長さの連続するフレームをまとめてマスクする。一方、CNN の出力はフレーム毎にベクトル量子化も行われる。トランスフォーマの出力とこの量子化の結果を照合することで損失が定義され、学習が行われる。

HuBERT[9]では、この量子化器は別途あらかじめ k-means などにより構成され、トランスフォーマの出力も量子化符号に変換してクロスエントロピ損失を計算する。これに対して、wav2vec 2.0[8]では、トランスフォーマの出力と符号化ベクトルとの対照学習

(contrastive learning)を行う。また、量子化に Gumbel Softmax 関数を導入して微分可能な形にし、トランスフォーマなどと一体的に End-to-End 学習を行う。その際に、できるだけ多くの符号が一様に用いられるような損失(diversity loss)も追加する。

この学習の結果得られる CNN とトランスフォーマは一体的な DNN（図 8 の枠）で、音声認識などのエンコーダとして用いることができる。この上に CTC やトランスフォーマデコーダを結合することで音声認識が行われる。音声認識の学習にはある程度のラベル付きデータが必要となるが、End-to-End モデルをスクラッチから学習するよりはるかに少量のデータで高い性能が実現される。

大規模事前学習モデルの音声認識への活用には 2 通りの方法がある。1 つは、このモデルに認識単位に相当する出力層を付加して、CTC 損失によりモデルをファインチューニングする方法である。この場合も、モデルのトランスフォーマ層のみをファインチューニングし、CNN 層は固定する場合が多い。この方法は簡潔で、特にデータ量が少ない場合に有効である。もう一つの方法は、自己教師付き学習による大規模事前学習を表現学習とみなし、このモデルを特徴抽出器として用い、さらにトランスフォーマのようなエンコーダ・デコーダモデルに基づく音声認識モデルを構成するものである。この場合は、事前学習モデルのパラメータは固定する。

8. 主な大規模事前学習モデル

現在広く用いられている代表的な大規模事前学習モデルを紹介する。

8.1. XLS-R

XLS-R[10]は wav2vec 2.0 の多言語版であり、128 言語、43.6 万時間のデータで学習されている。モデルのサイズによって複数のモデルがあり、それらの仕様を表 1 に示す。XLS-R のファインチューニングにより、事前学習に含まれない未知の言語も含めて様々な低資源言語の音声認識が効率的に実現できることが示されている。また、英語のみを用いた事前学習と比較して、多言語データの有効性も示されている。

表 1　XLS-R の仕様

	層数	状態数	パラメータ数
BASE	12	256	95M
LARGE	24	768	317M
X-LARGE	48	1024	964M
XLS-R (2B)	48	1920	2162M

8.2. Whisper

Whisper[11]は Open AI が開発・公開している汎用的な音声認識モデルである。これ自体で音声認識が可能で、99 言語をカバーしている。学習データは 68 万時間で、うち日本語は 2.3 万時間である。

基本的な（コンフォーマでない）トランスフォーマに基づいており、モデルの大きさによって複数のモデルがあり、その仕様を表２に示す。動画に付与された字幕のような音声に忠実でないテキストを使用した弱教師付き学習に基づいて構成されている。そのため、フィラーのない整形されたテキストを出力するが、音声にないテキストを生成すること（ハルシネーション）もある。

Whisper のモデルを対象タスクドメイン、例えば、特定の騒音環境やアプリケーションのデータでファインチューニングすることにより、それに適応した音声認識システムを構成することもできる。また、未知の言語にもある程度適応することもできる。

表２　Whisper の仕様

	層数	状態数	パラメータ数
Tiny	4	384	39M
Base	6	512	74M
Small	12	768	244M
Medium	24	1024	769M
Large	32	1280	1550M

9. まとめ

本節で解説したEnd-to-Endモデルは、この１０年ほどの間で研究開発されたものであるが、学習データや計算環境の大規模化と相乗して、音声認識の性能を大きく向上させた。特に、１つのモデルで多くの言語をカバーできるようになったのは大きい。

かなりうるさい環境やマイクとの距離が遠い条件、さらには複数人が発話する状況ではまだ課題があるが、音声認識はかなり成熟したといって過言ではない。

[1] A.Graves and N.Jaitly. Towards End-to-End speech recognition with recurrent neural networks. Proc. ICML, 2014.

[2] 河原達也.音声認識技術の変遷と最先端－深層学習によるEnd-to-Endモデル－. 日本音響学会誌, Vol.74, No.7, pp.381--386, 2018.

[3] A.Graves, S.Fernandez, F.Gomez, and J.Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In ICML, 2006.

[4] A.Graves. Sequence Transduction with Recurrent Neural Networks. Proc. ICML workshop on Representation Learning, 2012.

[5] J.Chorowski, D.Bahdanau, D.Serdyuk, K.Cho, and Y.Bengio. Attention-based models for speech recognition. Proc. NIPS, 2015.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and

I. Polosukhin. Attention Is All You Need. Proc. NIPS 2017.

[7] 河原達也, 三村正人. 大規模事前学習モデルに基づく音声認識. 日本音響学会誌, Vol.79, No.9, pp.455--460, 2023.

[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Proc. NIPS 2020.

[9] W.-N. Hsu, B. Bolte, Y.-H. Hubert Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, IEEE/ACM Trans. Audio, Speech & Language Proc., Vol. 29, pp. 3451–3460, 2021.

[10] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. Proc. Interspeech, 2022.

[11] A. Radford, J-W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, arXiv:2212.04356, 2022.