解説

大規模事前学習モデルに基づく音声認識 *

河原達也，三村正人（京都大学）**

1. はじめに

ちょうど５年前に「AI 時代の音響学」という小特集が組まれ、深層学習による End-to-End モデルについて解説[1]を行った。この５年間に世の中では様々なことがあったが、AI ブームはしぼむどころか、ますます隆盛している。この AI ブームの基盤となっているのが深層学習である。深層学習の進展は急速で、５年以上前の技術はほとんど残っていないといっても過言でない。実際に５年前の解説記事[1]の内容の多くは時代遅れになっているが、End-to-End モデルの概念、及び CTC (Connectionist Temporal Classification) などの基本的な手法は現在も重要である。

この５年間の最大の進展は、トランスフォーマーの出現と大規模事前学習である。昨今連日のように話題となっている GPT が Generative Pretrained Transformer の略であることからも明らかである。トランスフォーマーの大規模事前学習モデルは、自然言語処理だけでなく、画像処理、音声処理など多くの領域でそれまでの常識を塗り替えている。

最近の大規模事前学習は、自己教師付き学習の枠組みで行われるのが一般的である。これは、学習に必要な教師信号を自ら得て、人手ラベルを必要としないもので、学習データを飛躍的に増やすことができ、モデルの大規模化・高精度化につながった。実際に、従来より１～２桁多い数万時間以上の音声データを活用して、音声認識や感情認識などの様々なタスクの基盤となる普遍性の高い表現学習が実現されている。本稿では、音声認識を主に想定して、大規模事前学習モデルについて解説する。

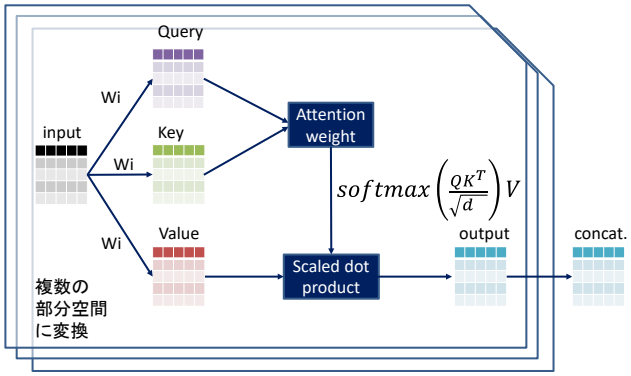

図１　自己注意機構

2. 基本的概念

まず音声に限らず、最近の大規模事前学習に必要かつ重要な概念を簡潔に述べる。

2.1 トランスフォーマー

トランスフォーマー(Transformer)[2]は、自己注意機構(Self-Attention)を多段に接続したネットワークである。自己注意機構を図１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（対数メルフィルタバンク出力など）と考えるとわかりやすい。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 (Long Short-Term Memory) 等の自己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。

* Automatic Speech Recognition based on Large Pretrained Models

** Tatsuya Kawahara and Masato Mimura (Kyoto University)

図２　BERT のマスク言語モデル

2.2　BERT

自然言語処理に大変革をもたらした **BERT** (Bidirectional Encoder Representations from Transformers)[3]は、トランスフォーマーを自己教師付き学習したもので、その主なものはマスク言語モデルである。これは、入力系列の一部のトークン（単語など）を遮蔽して、トランスフォーマーを通じて復元するものである（図２参照）。１つずつ次のトークンをそれまでの文脈履歴から予測する場合が（生成）言語モデルと呼ばれ、その代表例が GPT であるが、BERT では入力中のあらゆる箇所をランダムに遮蔽して、その前だけでなく、その後の文脈も用いて双方向的にモデル化・学習を行うのが特徴である。

2.3　対照学習

BERT を音声に適用する最も単純な方法は、音声の各フレームを符号化して BERT に入力するものであり、**vq-wav2vec**[4]により行われたが、符号化による情報損失は不可避である。

自然言語のようにトークンを扱う場合は、一致／不一致で損失が定義できるが、音声などのように数値ベクトルを扱う場合は類似度を測る。類似度は相対的なものなので、以下の式のように「自身のバリエーション」と他のものとの比をとる。これを**対照学習(contrastive learning)**と呼ぶ。

$$-\frac{\exp\left(sim(c_i, q_i)\right)}{\sum_j \exp\left(sim(c_i, q_j)\right)}$$

ここで、c がトランスフォーマーの出力、q が元の入力、またはその符号に対応する。sim はベクトル間の類似度を定義する関数で、コサイン距離などが用いられる。分母は、分子と異なるものを複数サンプルして計算する。雑音や変形を付加しても他のものと混同しないように対照学習を行うことで、頑健性が増すことが、様々な領域で知られている。

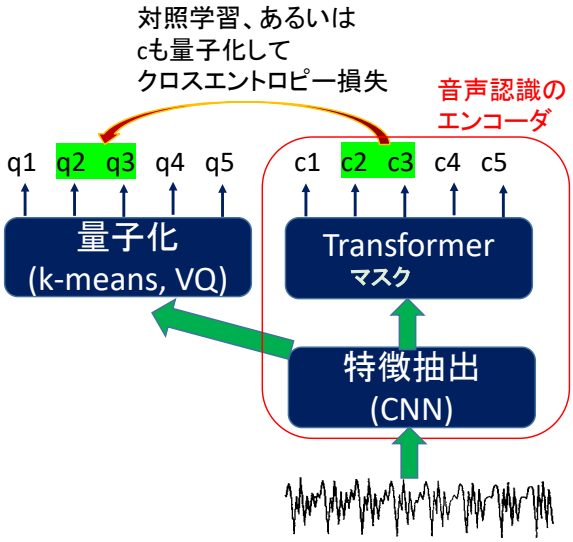

図３　wav2vec 2.0 / HuBERT の枠組み

3.　音声の大規模事前学習の枠組み

音声認識などで現在一般的に用いられている大規模事前学習モデルは、トランスフォーマーと量子化器（コードブック）を組み合わせたものであり、概ね図３に示す枠組みである。その代表的なものが、**wav2vec 2.0**[5]と **HuBERT**[6]である。

入力音声またはその周波数特徴量は、CNN 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ーに入力される。その際に BERT と同様に、一部のフレームがマスクされ、トランスフォーマー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。トランスフォーマー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[6]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、トランスフォーマーの出力も量子化符号に変換してクロスエントロピー損失を計算する。これに対して、**wav2vec 2.0**[5]では、トランスフォーマーの出力と符号化ベクトルとの対照学習を行う。また、量子化に Gumbel Softmax 関数を導入して微分可能な形にし、トランスフォーマーなどと一体的に 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ーは一体的なDNN（図3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ー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。ファインチューニングを行う際にはトランスフォーマー（の上位の層）のみを更新すればよい。

4. 大規模事前学習モデルの展開

本章では、様々な大規模事前学習モデルについて概観する。図4に代表的な手法の展開・関係を示す。

4.1 wav2vec 2.0 以前

音声の表現学習は、かつては制約ボルツマンマシンやオートエンコーダに基づく手法が多く(例えば[8], [9], [10], [11])、入力信号を再構成できるようなコンパクトな表現ベクトルを抽出することに主眼が置かれていた。しかし、音声は連続量であり、音声認識等のタスクと無関係な情報（例えば話者や音環境）も内包するため、入力を忠実に復元できる表現が必ずしも望ましいわけではない。

CPC (Contrastive Predictive Coding)[7]は、観測済みデータの復元でなく、観測されていない未来のフレームを予測することで、局所的な変動の影響を受けにくい、大域的な特徴量（“slow features”）の獲得をめざした手法である。CPCモデルは、音声波形データからフレーム毎の特徴量ベクトルz_tを抽出する特徴量エンコーダと、時刻tまでの特徴量系列$[z_1, z_2, \dots, z_t]$を要約して文脈表現ベクトルc_tを生成するコンテクストモジュールの2つのサブネットワークから構成される。このc_tから、未来の複数フレーム$[z_{t+1}, z_{t+2}, \dots, z_{t+K}]$を予測する。

これらのモジュールでは、文脈表現c_tを単一の線形変換W_kからなる予測器で変換したベクトル$W_k c_t$と予測ターゲットのフレームにおける特徴量z_{t+k}の類似度が、それ以外のフレーム（負例）との類似度より大きくなるように対照学習が行われる（対照的予測基準)。後続タスクの入力には、z_tまたはc_tを用いる。原論文[7]の評価実験はフレームごとの音素識別であるが、CPC 特徴量が単純な周波数特徴量より大幅に高い精度となっている。また、単一フレームではなく複数の、しかもより遠い（K=12 程度）フレームをターゲットにすることや、負例を同一発話中からサンプリングするのが効果的であるなど、以降の研究でも用いられる重要な知見が示されている。

wav2vec[12]は、学習基準やネットワーク構成はCPCとほぼ同一であるが、最新の音声認識モデル（wav2letter++[13]）との組合せにおいて、対照的予測基準に基づく事前学習が教師あり音声認識の性能を有意に改善できることを示した。

vq-wav2vec[4]は、wav2vecとBERT[3]をカスケード的に組み合わせた手法である。つまり、対照的予測基準により獲得された特徴量z_tの系列を入力として、別途マスク言語モデル基準によりBERTを学習し、後続タスクではこのBERTの出力を入力として用いる。ただし、特徴量系列をBERTのような言語モデルに入力するには、各フレームを離散トークンとして扱う必要がある。そこで、VQ-VAE[11]で提案された微分可能なコードブックを用いてベクトル量子化を導入している。この枠組みが、音声の事前学習において大きなエポックの一つとなった。CPCやwav2vecにおける対照学習ではターゲットフレームの特徴量ベクトルz_{t+k}自体を予測するのに対して、vq-wav2vecでは、対象フレームが割り当てられるコードのクラスタ中心$\widehat{z_{t+k}}$を予測するように学習する。これにより、コードブックとその他のモジュールが一体的に学習される。音響モデルとしてwav2letterを用いた音声認識実験において、クラスタ中心$\hat{z}_t$を入力として用いることで、量子化誤差のないwav2vev特徴量z_tよりも性能が低下する反面、クラスタID系列をBERTで変換したものでは大幅な性能向上が見られた。このことから、トランスフォーマーを用いて系列全体の文脈から特徴量抽出する重要性が確認できる。

4.2 wav2vec 2.0 以降

これまでの手法は古典的な表現学習の改良版で、評価実験も小規模なデータセットを用いた概念検証にとどまっていた。そこから大規模データ

を用いた事前学習の有用性を示したのが、**wav2vec 2.0**[5]である。vq-wav2vec で導入されたトランスフォーマー(BERT)とクラスタリングを踏襲するが、主な違いは、トランスフォーマーと特徴抽出モジュール、量子化モジュール（コードブック）が、マスク言語モデル基準に基づいて統合的に End-to-End に学習される点である。つまり、CPC 以来の自己回帰的ネットワークに基づくコンテクストモジュールが削除され、マスクされたフレームのクラスタ中心をトランスフォーマー自体が予測し、文脈表現ベクトルc_tを生成する。マスクは特徴抽出の出力系列に対して行われ、一定の確率(0.065)で開始時刻を選択した上で、そこから連続するフレーム（10 フレーム）を学習可能なマスクベクトルで置換する。トランスフォーマーはもはや言語モデル（BERT）ではないので、vq-wav2vec で導入されたコードブックは不要のように思われるが、正則化のボトルネック[11]として活用されていると捉えられる。原論文[5]の評価実験は、6 万時間のラベルなしデータを用いてモデル学習を行い、10 分の教師ありデータのみで CTC を用いたファインチューニングにより、高い汎化性能を持つ音声認識モデルが実現可能なことを示している。

XLSR-53[14]は wav2vec 2.0 の多言語版であり、CommonVoice などから収集した 53 ヵ国語、5 万 6 千時間のデータを用いて学習された大規模事前学習モデルである。後継の **XLS-R**[21]では、128 言語、約 50 万時間のデータに拡張されている。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かった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。

wav2vec 2.0 は、様々な技術の集合体であるが、従来の教師あり学習に基づく音声認識の手法の適用が難しかった。**HuBERT**[6]では、対照学習と量子化モジュールを削除して、代わりにオフラインの k-means クラスタリング（音響特徴量または学習済みモデルの中間表現の量子化）と、クラスタ ID を用いた（通常の音声認識と同様の）クロスエントロピー損失により学習を行う。評価実験では、特に教師ありデータが少ない条件で HuBERT は一貫して wav2vec 2.0 より高い音声認識性能となっている。ただし、高い性能を得るためには、初期ターゲット（音響特徴量のクラスタ ID）で学習したモデルの中間表現で再び k-means クラスタリングを行い、ターゲットの再構築とモデルの再学習を行う必要がある。また、トランスフォーマーの深い（中間以降の）層の出力を離散化することにより音素に近いトークンが獲得できることが示されている。

wavLM[15]は、HuBERT にデノイジングの機構を導入したものである。クリーンな音声に雑音や他の音声信号を様々な SN 比で付加した混合音を入力として、クリーン音声のクラスタ ID をターゲットとして HuBERT モデルを学習する。通常の英語音声認識以外にも、SUPERB ベンチマークにおける様々な音声処理タスクで HuBERT を凌ぐ性能を達成している。また、雑音下音声認識[16]や音声強調[17]などでも良好な性能を示している。

BEST-RQ[18]も HuBERT の延長と捉えられ、生の音響特徴量にオフラインで離散 ID を割り当て、マスクした区間のターゲットとして用いる。ただし、k-means クラスタリングでなく、乱数で初期化した線形写像とコードブックを用意し、音響特徴量をこの写像で変換したベクトルに最も近いコードの ID をそのフレームのラベルとする。また、これまでのモデルと異なり、マスクを中間表現でなく入力の音響特徴量にかける。この写像とコードブックのパラメータの更新は行わないため、HuBERT と同様オフラインのクラスタリングとなっている。BEST-RQ は非常に単純で柔軟なアルゴリズムであり、標準的な構成の音声認識システム（対数メルフィルタバンク出力、コンフォーマに基づくエンコーダ、RNN-トランスデューサー）をそのまま適用できる。未来の情報を参照できないストリーミング音声認識のための因果的コンフォーマなどでは、既存の手法より高い性能を示すことが報告されている。最近の Google の大規模多言語音声認識システムである Google USM[19]では、BEST-RQ が（複数コードブックに拡張された上で）事前学習に用いられており、他の自己教師付き学習との比較はないが、Whisper で用いられている弱教師あり学習（忠実な書き起こしでない字幕をそのまま用いた教師

あり学習）より有効であると報告されている。

これらを概観すると、特徴抽出モジュールや学習に基づくクラスタリング、対照学習など、当初重要と思われた要素が次々と省略され、結局はトランスフォーマーでマスク言語モデル学習を行うことのみが本質的であると結論づけられる。BEST-RQ の比較実験では、量子化手法の性能は最終的な自己教師付き学習の性能にほぼ影響しないと述べられており、また data2vec[20]のように離散トークンを一切用いない手法もあるため、クラスタリングの重要性も認められない。なお、wav2vec 2.0 や HuBERT でも、複数のコードブックを用いて実質的なクラス数を非常に大きくすることが重要とされている。例えば wavLM の SUPERB ベンチマークに対する分析[15]において、タスクによって注目する層が大きく異なる（話者同定には入力に近い層、意図認識には出力に近い層の出力が重要であるなど）ことから、トランスフォーマーによって大域的かつ階層的な特徴量の抽出が行われていると推察される。

近年の論文は速報性を重んじるため、先行研究との理論的・実験的な比較が不十分なことが多く、ここで挙げた手法間に実際にどのくらいの性能差があるのか明確でない。また、これらの大規模モデルを自ら学習したり比較したりするのは、（特に大学等の機関では）計算資源の面で現実的でない。低資源言語の音声認識には 128 言語版の XLS-R、雑音下音声認識などその他のタスクには wavLM を用いるのが、現時点では合理的な選択といえる。いずれも Hugging Face のリポジトリなどから事前学習済みモデルを取得できる。

5. 大規模事前学習モデルの適用例

大規模事前学習の適用例として、著者らの最近の研究から簡単に紹介する。XLS-R モデル[21]を用いて、クメール語の音声認識とアイヌ語の音声認識を行った。

カンボジア特別法廷におけるクメール語の音声認識[22]では、1 時間、5 時間、10 時間の音声データでファインチューニングを行い、文字誤り率(CER)が各々21.7%、12.1%、11.1%となった。トランスフォーマーに基づく End-to-End 音声認識をスクラッチから学習するにはこの何倍ものデータが必要であり、45 時間のデータで 12%程度であったことから、この枠組みの有効性が確認できる。また、言語とドメイン各々に、逐次的にかつマルチタスク学習の枠組みでファインチューニングする効果も示している。

異なる方言を対象としたアイヌ語の音声認識[23]では、1 時間、4 時間、10 時間、33 時間の音声データでファインチューニングを行い、音節誤り率(CER)が各々19.6%、16.5%、15.5%、14.1%となった。33 時間のデータでスクラッチから学習した Bi-LSTM モデルでは、31.3%であった。

上記の実験結果から、5〜10 時間のデータでもほぼ収束し、良好な性能が得られることがわかる。

また、IEMOCAP コーパスを用いた感情音声認識[24]も行った。960 時間の LibriSpeech コーパスから構築された wav2vec2-base モデルを、12 時間のコーパスでファインチューニングしたところ、重みなし精度(UA)、重みつき精度(WA)が各々72.1%、71.6%となった。スクラッチから学習した CNN-LSTM ベースラインモデルでは各々55.47%、55.81%であり、大きく上回っている。ここでも、音声認識や性別認識とマルチタスク学習を行う効果を確認している。

6. おわりに

第一著者の個人的見解であるが、パターン認識において教師なし学習やコードブックを用いる手法が有効であると考えておらず、様子見をしていたら、急速に高い性能を出すようになっていた。

もちろん、英語や日本語のように大規模のラベル付きデータがあれば、スクラッチから End-to-End モデルを学習すればよい。ただし、そのような音声認識は完成度の高い領域に達している。少数言語の音声認識や感情認識などの学習データ量が少ないタスクが研究課題となっている中で、大規模事前学習は非常に有効である。また、多言語の音声認識や、言語情報と感情の同時認識など複合的な処理を行う上でも効果的である。

ただし、このような大規模事前学習モデルを自ら構築するのは容易でなく、GPT と同様な感想も禁じ得ない。

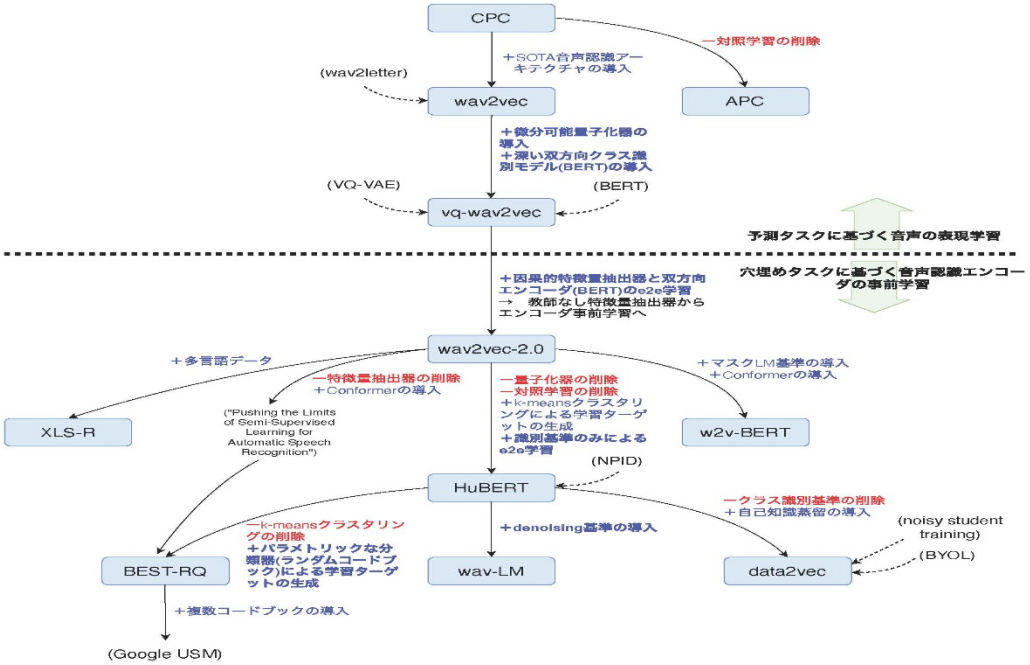

図４　大規模事前学習モデルの展開

文　献

[1] 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.

[2] 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.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proc. NAACL-HLT 2019.

[4] A. Baevski, S. Schneider, M. Auli. VQ-wav2vec: Self-Supervised Learning of Discrete Speech Representations, arXiv:1910.05453, 2019.

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

[6] 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.

[7] A. van den Oord, Y. Li and O. Vinyals. Representation Learning with Contrastive Predictive Coding, arXiv:1807.03748, 2018.

[8] N. Jaitly, G. Hinton, Learning a Better Representation of Speech Sound Waves using Registered Boltzmann Machines, ICASSP, 2011

[9] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, Unsupervised speech representation learning using WaveNet autoencoders, IEEE/ACM Trans. Audio, Speech, and Language Processing, Volume 27, pp.2041-2053, 2019

[10] W. Wang, R. Arora, K. Livescu and J. Bilmes, On Deep Multi-View Representation Learning, Proc. ICML, 2015

[11] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning”, NIPS, 2017

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, Wav2vec: Unsupervised Pre-training for Speech Recognition, Proc. Interspeech, 2018

[13] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert. Wav2letter++: The fastest open-source speech recognition system, arXiv, abs/1812.07625, 2018.

[14] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised Cross-lingual Representation Learning for Speech Recognition, Proc. Interspeech, 2021

[15] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, WavLM: Large-scale self-supervised pre-training for full stack speech processing,”, IEEE J. Selected Topics in Signal Processing, Volume 16, 2022

[16] Y. Masuyama, X. Chang, S. Cornell, S. Watanabe, N. Ono, End-to-End Integration of Speech Recognition, Dereverberation, Beamforming, and Self-Supervised Learning Representation, Proc. IEEE-SLT, 2023

[17] H. Song, S. Chen, Z. Chen, Y. Wu, T. Yoshioka, M. Tang, J. W. Shin, and S. Liu, Exploiting WavLM on Speech Enhancement, Proc. IEEE-SLT, 2023

[18] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, Self-supervised learning with random-projection quantizer for speech recognition, Proc. ICML, 2022

[19] Y. Zhang, et.al, Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, arXiv:2303.01037, 2023

[20] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language, Proc. ICML, 2022

[21] 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

[22] K. Soky, S. Li, C. Chu, and T. Kawahara. Domain and language adaptation using heterogeneous datasets for wav2vec2.0-based speech recognition of low-resource language. Proc. IEEE-ICASSP, 2023.

[23] J. Lee, M. Mimura, and T. Kawahara. Embedding articulatory constraints for low-resource speech recognition based on large pre-trained model. Proc. Interspeech, 2023.

[24] Y. Gao, C. Chu, and T. Kawahara. Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with ASR and gender pretraining. Proc. Interspeech, 2023.