UTMOS:VoiceMOS Challenge2022 に向けた UTokyo-SaruLab チームの自然性 MOS 予測モデル *

☆中田 亘 [1] (東大 工学部), 辛 徳泰 [1]，佐伯 高明 [1]（東大院・情報理工）
郡山 知樹（東大院・情報理工/サイバーエージェント）
高道 慎之介，猿渡 洋（東大院・情報理工）

1 はじめに

合成音声の評価には主観評価が用いられるが，その金銭的，時間的コストが課題である．このコストの低減を目的として，主観評価値を自動予測する機械学習手法 [1,2] が複数提案されているものの，更なる予測精度向上やドメイン外のデータへの適用など課題も多い．こういった中で The VoiceMOS Challenge [3] が開催された．本チャレンジの参加者は，合成音声とその 5 段階自然性 Mean Opinion Score （MOS）値から成る共通データベースを用いて，MOS 値予測モデルを構築し，未知の合成音声テストデータに対して予測した MOS を提出する．

本稿では，我々が構築した MOS 値予測モデルである UTMOS について述べる．UTMOS は，アンサンブル学習に加え，対照学習，聴取者依存モデリング，音素エンコーディングなどの手法により予測性能を改善する．本チャレンジにおける予測結果に加え，UTMOS の各要素に対する ablation study の結果を報告する．なお，UTMOS の実装は https://github.com/sarulab-speech/UTMOS22 で公開されている．

2 VoiceMOS Challenge 2022

VoiceMOS Challenge 2022 では，main track と OOD track がある．各 track におけるデータセットの統計については，論文 [3] を参照されたい．

Main track Main track では，共通データベースとして BVCC が提供された．これは，過去の Blizzard Challenge と Voice Conversion Challenge，ならびに ESPnet-TTS [4] の英語合成音声と，その音声に対して新たに実施された 5 段階自然性 MOS 評価（聴取実験）の結果から成る．

OOD track OOD track のデータベースは，Blizzard Challenge 2019 に提出された音声合成システムによる中国語合成音声と，main track と別に実施された聴取実験の結果に加え，評価スコアの無い合成音声から成る．

各 track において，テストセットに対する予測 MOS 値は，平均平方誤差 (MSE)，相関係数 (LCC)，スピアマンの順位相関係数 (SRCC)，ケンドールの順位相関係数 (KTAU) で評価された．各評価指標は，発話レベル（各音声サンプルでの平均値）とシステムレベル（音声サンプルに対する MOS を各合成音声システムで平均した値）のそれぞれで計算される．

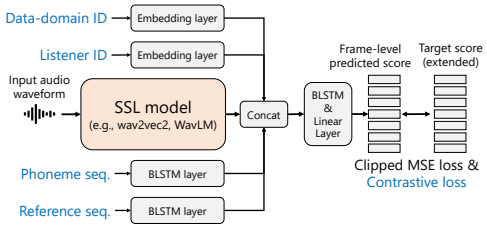

Fig. 1: 強学習器のモデル構造

3 UTMOS

UTMOS では，**強**学習器と**弱**学習器からなる複数のモデルを用いてアンサンブル学習を行う．強学習器は，self-supervised learning (SSL) モデルを基本とするニューラルネットワークから構成され，音声波形を入力とする．一方弱学習器は，古典的なリッジ回帰やサポートベクトルマシンなどの機械学習器からなり，事前学習済み SSL モデルに対して音声を入力することにより得られる特徴量を入力とする．

3.1 強学習器

3.1.1 基本構造

Fig. 1に強学習器の基本構造を示す．先行研究 [5] と同様，強学習器では事前学習済み SSL モデルを用いて音声入力から特徴量を得る．まず，発話に対応する音声波形が SSL モデルに入力され，フレームレベルの特徴量が得られる．先行研究 [5] では，フレームレベルの平均が特徴量として使用されていたが，UTMOS ではフレームレベルの特徴量を Bidirectional long short-term memory (BLSTM) と全結合層に入力し，自然性 MOS をフレームごとに予測する．訓練時には，ラベルとなる MOS をフレーム方向に複製し，フレームレベルの損失関数を定義した．我々は，事前実験により，フレームレベルで損失関数を定義することで，フレーム間で平均してから損失を取る場合よりも予測性能が改善することを確認している．推論時には，フレームレベルで予測したスコアの平均を取ることにより，MOS を推定する．この基本構造に対して 3.1.2 – 3.1.5節に示す改善を行った．

3.1.2 対照学習

対照学習は，SSL の学習において注目されている手法であり，ラベルを用いずに，データ同士を比較することにより学習を行う．近年では音質評価において広く使われている [6]．UTMOS では，SRCC などの順位相関係数の改善を目的として，対照学習を行った．

---

* UTMOS: Team UTokyo-SaruLab MOS prediction model for VoiceMOS Challenge 2022 by NAKATA, Wataru[1], XIN, Detai[1], SAEKI, Takaaki[1] (The University of Tokyo), KORIYAMA, Tomoki (The University of Tokyo / CyberAgent, Inc.), TAKAMICHI, Shinnosuke, SARUWATARI, Hiroshi (The University of Tokyo). ·[1] indicates equal contribution. ·[1] で示された著者の貢献度は同じ

s_1, s_2 をそれぞれ異なる発話に対する主観評価とすると，2つの発話に対する評価の差異は $d_{x_1,x_2} = s_1 - s_2$ と表すことができる．予測された各音声に対する主観評価をそれぞれ $\hat{s_1}$，$\hat{s_2}$ とすると，予測された各音声に対する評価の差異 $\hat{d}_{x1,x2} = \hat{s_1} - \hat{s_2}$ は $d_{x1,x2}$ に近くなることが期待される．よって，対照学習の損失関数を，$\mathcal{L}^{con}_{x_1,x_2} = \max(0, \|d_{x_1,x_2} - \hat{d}_{x_1,x_2}\| - \alpha)$ とした．ここで α はマージンと呼ばれるハイパーパラメータであり，予測結果に対する小さな誤差を無視するために設定される．実装では，ミニバッチ内で対照学習の損失は計算されるため $\mathcal{L}^{con} = \sum_{i \neq j} \mathcal{L}^{con}_{x_i,x_j}$ となる．また，対照学習の損失に加えて，clipped MSE loss [2] を回帰損失として利用する．clipped MSE loss は $\mathcal{L}^{reg}(y, \hat{y}) = \mathbb{1}(\|y - \hat{y}\| > \tau)(y - \hat{y})^2$ と定義した．

学習に用いた最終的な損失は以下のように定義する．ここで β 及び γ はハイパーパラメータである．

$$\mathcal{L} = \beta \mathcal{L}^{\text{reg}} + \gamma \mathcal{L}^{\text{con}} \tag{1}$$

3.1.3 聴取者依存モデリングとデータドメインによる条件付け

先行研究 [2,7] では聴取者依存モデリングを行う事により予測精度が改善することが確認されている．これは，聴取者により評価値の分布が異なることに起因する．これを踏まえて，聴取者依存モデリングを強学習器に導入した．図 1に示す通り，聴取者の分散表現が SSL モデルにより抽出された音声の特徴量に連結され，予測を行うことにより，聴取者依存の評価値を出力する．また，学習時には，データに存在する聴取者に加えて “mean listener” が使用される．これは先行研究 [7] と同様，全ての聴取者による評価値の平均を評価値とするような仮想的な聴取者である．推論時，聴取者のデータは与えられていないため，“mean listener” を用いて発話レベルの MOS を推論する．

加えて，聴取実験ごとのバイアスを考慮する必要がある．そのため，Fig. 1に示したように，聴取者に加え，domain ID を用いたデータドメインによる条件付けをする．これにより，main, OOD, external (3.2で説明）の 3 つの異なる聴取実験から得られたデータを用いて同時に学習することが可能となる．また，“mean listener” の MOS は，各ドメインごとの聴取者の評価の平均を求めることで計算を行う．

3.1.4 音素エンコーディング

UTMOS では，音声認識結果を強学習器の入力として使用する．また，複数の言語を扱うために書記素列ではなく音素列を強学習器の入力とする．加えて，実際の発話内容（発話テキスト）と合成音声から得られる音声認識結果が異なると合成音声の明瞭性が低いと考えられるため，入力されたテキストも強学習器の入力として使用する．音声合成モデルに入力されたテキストは音声認識結果を DBSCAN [8] を用いてクラスタリングし，各クラスタのメジアンとなるテキストを求めることで，推定した．この推定されたテキストを reference sequence とし，音声合成モデル入力されたテキストと仮定する．DBSCAN を行う際の距離関数には，normalized Levenshtein 距離を使用した．音素列と reference sequence は，Fig. 1に示すように，多層 BLSTM に入力され，最初と最後の隠れ状態を連結した後，フレーム数分複製された後，SSL モデルの出力に連結される．

3.1.5 データ拡張

過学習を抑制するため，話速・ピッチ変換に基づくデータ拡張を適用した．訓練時には，f_t と f_p はそれぞれ $[1 - F_t, 1 + F_t]$ と $[-F_p, F_p]$ の範囲からランダムに決定される．ここで，f_t と f_p はそれぞれ話速，ピッチを制御するための係数である．また，F_p は聞いた際の音声が元の音声と大きく変化しないよう調節された．

3.2 独自に行った聴取実験データの利用

OOD track には，ラベル付きのデータが 136 発話存在したが，安定的に MOS 予測モデルを学習するには不十分であった．そこで，540 発話存在するラベルなしデータに対して，独自に聴取実験を行い MOS 評価値を得ることで，ラベルなしデータを活用した．聴取実験では，まず MOS 評価が一番高いシステム（BC2019-A）に対して中国語母語話者の確認の上で，自然音声と仮定した．その後，540 個のラベルなし発話及び 249 のラベル有り発話に対して 5 段階 MOS 評価を行った．聴取者数は 32 人であり，全員中国語母語話者である．各聴取者は 55 発話評価したため，各発話に対する評価数は平均で 2 個である．独自に行った聴取実験と VoiceMOS Challenge2022 で配布されたデータの間の発話レベル SRCC は 0.757 であり，強い相関が確認された．

3.3 強学習器と弱学習器によるアンサンブル学習

予測結果をよりロバストにするため図 2に示すようなアンサンブル学習の一種であるスタッキング [9] を使用する．転移学習された SSL モデルに加え，発話レベルの特徴量から単純な回帰モデルを用いて MOS 予測を行った．前者を強学習器，後者を弱学習器とする．

弱学習器は，特徴量抽出器と回帰モデルからなる．特徴量抽出器には，事前学習済み SSL モデルから得られる特徴量の時間方向に平均をとったものを利用する．これは，先行研究 [5] で提案されている手法と類似している．回帰モデルには，線形回帰や，決定木，カーネル法を利用する．一般に，モデルの多様性がアンサンブル学習において重要であると考えられているため [10]，複数の SSL モデルを特徴量抽出に利用することにより，モデルの数を増やす．加えて，OOD track に関しては，異なる言語や聴取実験からなる複数のデータドメインを用いる事により，弱学習器の多様性を確保する．

スタッキングには，ステージ 0 からステージ 3 の手順が存在する．まず，特徴量を抽出した後，強学習器と弱学習器をそれぞれ学習し，交差検証を用いて予測結果を出力する．その後，予測結果を用いてメタ学習器を学習する．最後に，ステージ 2 のモデルの予測結果を用いてステージ 3 のモデルを学習し，最終的な予測結果を得る．

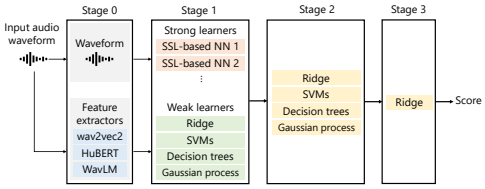

Fig. 2: 強学習器と弱学習器を用いたスタッキング

Table 1: 強学習器に対する各種要素の評価結果

(a) Main

	Utterance-level				System-level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
UTMOS strong	0.276	0.883	0.881	0.708	0.148	0.930	0.925	0.774
w/o contrastive loss	0.241	0.881	0.879	0.706	0.114	0.932	0.930	0.781
w/o listener ID	0.307	0.880	0.878	0.704	0.160	0.935	0.933	0.784
w/o phoneme encoder	0.249	0.881	0.882	0.709	0.119	0.935	**0.936**	**0.790**
w/o data augmentation	0.226	**0.885**	**0.882**	**0.710**	**0.103**	**0.936**	0.933	0.784
w/o MSE loss	**0.219**	0.882	0.880	0.707	0.114	0.932	0.929	0.778
SSL-MOS	0.380	0.869	0.871	0.695	0.223	0.920	0.918	0.758

(b) OOD

	Utterance-level				System-level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
UTMOS strong	0.378	0.891	0.871	0.690	0.248	**0.970**	**0.972**	**0.879**
w/o contrastive loss	0.407	0.870	0.862	0.676	0.272	0.945	0.957	0.841
w/o listener ID	0.636	0.847	0.825	0.638	0.490	0.931	0.944	0.820
w/o phoneme encoder	0.390	**0.893**	**0.881**	**0.702**	0.258	0.966	0.967	0.868
w/o data augmentation	**0.322**	0.887	0.872	0.691	**0.191**	0.960	0.967	0.872
w/o external data	0.412	0.883	0.868	0.684	0.253	0.960	0.961	0.861
SSL-MOS	0.676	0.872	0.842	0.654	0.500	0.957	0.964	0.862

4 実験的評価

4.1 実験条件

強学習器に関しては，3.1.2節から 3.2節にかけて説明した各種法の ablation study を行った．さらに，main track では，MSE loss の使用，OOD track では独自に取得した聴取データの利用についても調査した．提案法に対応するのは “UTMOS-strong” であり，この条件が VoiceMOS Challenge 2022 での提出に使われたモデルである．音声の前処理として，全ての音声は 16kHz にダウンサンプリングし，音量を正規化した．また訓練時には，5 段階の自然性評価値を $[-1, 1]$ に正規化した．また，強学習器に使用した事前学習済み SSL モデルには，LibriSpeech [11] で事前学習した wav2vec 2.0 [12] base モデル[3] を使用した．また，音素エンコーディングに使用する音素列の取得には，Xu らが提案している音声認識モデル [13] を使用した．音素エンコーダには 3 層の BLSTM を使用し，BLSTM の隠れ層は 256 次元とした．聴取者とドメインの埋め込みは 128 次元とした．main track では，main track のデータセットのみを用いて学習し，OOD track では，OOD track のデータセットに加え，独自に行なった聴取実験のデータを用いて学習を行った．ただし，“w/o external” においては，OOD track のデータセットのみを用いて学習した．式 (1) に示した各種ハイパーパラメータは，“w/o contrastive loss” と “w/o MSE loss” 以外に関しては，$\beta = 1, \gamma = 0.5$ とした．“w/o contrastive loss” と “w/o MSE loss” に関しては，それぞれ $\beta = 1, \gamma = 0$ と $\beta = 0, \gamma = 1$ とした．α と τ に関しては，“w/o listener ID” を除き $\alpha = 0.5$, $\tau = 0.25$ とした． “w/o listener ID” に関しては，$\alpha = 0.1$, $\tau = 0.1$ とした．データ拡張に関するハイパーパラメータは，$F_t = 0.1$,$F_p = 300$ とした．また，“w/o data augmentation” に関しては，データ拡張は行わなかった．最適化には，Adam [14] ($\beta_1 = 0.9$, $\beta_2 = 0.99$) を使用し，学習率スケジューリングには，4000 ステップのウォームアップの後，11000 ステップ線形減少させた．バッチサイズは 12 であり，勾配積算を 2 ステップごとに行い，仮想バッチサイズを 24 とした．強学習器の学習はランダムシードによって性能が変化するため，異なるランダムシードを用いて 5 回学習を行い最終的な予測スコアは，それぞれのシードによる予測スコアの平均を取ることで得た．

モデルスタッキングに使用される強学習器に関しては，Optuna [15] によるハイパーパラメータチューニングを用いて development セットに対して最も高いシステムレベルの SRCC を持つモデルが選ばれた．最大で，17 個の強学習器を main track で使用し，6 個の強学習器を OOD tack で使用した．弱学習器で使用する事前学習済み SSL モデルから得る特徴量に関しては，wav2vec 2.0 [12] を 4 つ, HuBERT [16] を 2 つ，WavLM [17] を 2 つ使用した．これらのモデルは全てモデルサイズや学習に使用されたデータセット，学習手法で差異を持つ．メタ学習器に使用された単純な回帰モデルに関しては，2 つの線形回帰モデル（リッジ回帰，線形サポートベクトル回帰 (SVR)），2 つの決定木モデル（ランダムフォレスト及び LightGBM [18]），2 つのカーネル法（kernel SVR および，ガウス過程回帰）が使用された．これらの SSL モデルと回帰モデルの組み合わせにより，全部で 48 個の弱学習器を訓練した．

弱学習器及びメタ学習器の訓練では main track においては main track データセットのみを利用し，OOD

[3] https://github.com/pytorch/fairseq/blob/main/examples/wav2vec

Table 2: スタッキングに対する評価結果．“Strong” と “Weak” はそれぞれ，スタッキングに使用された強/弱学習器の数を示す． “Strong” が 1 のときに限っては弱学習器は使用されなかった．つまり，強学習器単独による推論結果である．OOD track に使用した弱学習器の数は 48, 96, 144 とし, それぞれ，OOD のみ，OOD 及び独自の聴取実験，main，OOD 及び独自の聴取実験で訓練したモデルを使用している．

(a) Main

		Utterance-level				System-level			
Strong	Weak	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
1	0	0.216	0.894	0.890	0.720	0.105	0.937	0.934	0.792
17	0	0.169	0.896	0.893	0.725	**0.088**	**0.939**	**0.936**	0.792
0	48	0.186	0.887	0.885	0.714	0.108	0.928	0.927	0.777
1	48	0.172	0.896	0.894	0.726	0.098	0.935	0.933	0.789
5	48	0.169	0.898	0.895	0.728	0.095	0.938	**0.936**	0.793
12	48	0.169	0.898	0.895	0.728	0.094	0.938	0.935	0.792
17	48	**0.165**	**0.899**	**0.896**	**0.730**	0.090	**0.939**	**0.936**	**0.795**

(b) OOD

		Utterance-level				System-level			
Strong	Weak	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
1	-	0.280	0.905	0.885	0.704	0.160	0.972	0.965	0.858
6	0	**0.155**	**0.920**	**0.896**	**0.720**	0.029	0.988	0.975	0.886
0	48	0.204	0.893	0.858	0.674	0.033	0.985	0.963	0.860
0	96	0.179	0.907	0.877	0.696	0.030	0.988	0.975	0.890
0	144	0.176	0.909	0.882	0.702	0.033	0.987	0.974	0.888
1	144	0.174	0.910	0.883	0.704	0.033	0.986	0.976	0.894
6	144	0.162	0.917	0.892	0.715	**0.028**	**0.989**	**0.977**	**0.900**

track では OOD track データセットに加え，main track データセット，独自に行なった聴取実験で得たデータを使用し，ステージ 2 において，結果を統合した．よって OOD track に使用された弱学習器の数は全体で 144 個である．OOD track のメタ学習器に関しては OOD track のデータのみを用いて学習した．

4.2 VoiceMOS Challenge 2022 おける結果

2節で説明した通り，main track, OOD track の双方において，発話レベル（Utt.）及びシステムレベル（Sys.）の評価指標が計算された [3]．3 つのベースライン手法に加え，21 チームが main track の予測結果を提出し，OOD track には 3 つのベースライン手法に加え 15 チームが予測結果を提出した．我々の Team ID は “T17” である．

main track の結果は，Utt. MSE = 0.165 (1), Utt. SRCC = 0.897 (1), Sys. MSE = 0.090 (1), Sys. SRCC = 0.936 (3), OOD track の結果は,Utt. MSE = 0.162 (1), Utt. SRCC = 0.893 (2), Sys. MSE = 0.030 (1), Sys. SRCC = 0.988 (1) である．カッコの中の数字は全体における順位を示す．

4.3 強学習器に対する ablation study 結果

3.1で説明した各種要素の性能への影響を調査した．全ての要素を使用したモデルを “UTMOS strong” とし，SSL モデルの転移学習によるベースライン手法 [5] を “SSL-MOS” として示す．結果を表 1に示す．最も良い結果は**太字**で示されており，“SSL-MOS” および “UTMOS strong” を除き最も悪い結果は下線で示されている．

結果から，ablation study での比較手法は，ほとんど全ての指標において “SSL-MOS” より改善していることが確認できる．加えて，main track では，UTMOS strong からデータ拡張や音素エンコーディングを使用しないモデルがよりよい結果を示した．これは main track ではデータセットが大きかったことが理由であると考えられる．一方で，OOD track では UTMOS strong が最も良い結果を示した．これは，提案手法が小さいデータにおいては有効であることを示唆している．main track と OOD track 双方において，聴取者の条件付けを行わない場合，多くの場合性能が劣化することが確認された．このことから，聴取者依存モデリングの有効性が確認できる．

4.4 スタッキングに対する評価

強学習器と弱学習器を用いたスタッキングの有効性を調べるために，強学習器，弱学習器の数を変化させた際の予測精度の変化を計算した．表 2に結果を示す．強学習器は development set に対するシステムレベルの SRCC をもとに貪欲的に 1, 5, 12 個選択された．

結果から，単独の強学習器のみを用いた場合，SRCC は高いものの，MSE が大きくなっていることが確認できる．一方で，スタッキングを使用することで，SRCC は高いまま MSE が改善することが確認された．加えて，特徴量抽出手法が比較的単純にも関わらず，弱学習器のみを用いたスタッキングにおいても高い SRCC が確認された．また，強学習器や弱学習器の数を増やす事により，予測精度が改善する傾向が確認された．これは，複数のハイパーパラメータや複数のドメインのデータを用いてモデルを用意することで，予測精度が改善できることを示唆している．

5 まとめ

本稿では，VoiceMOS Challenge2022 に向けて構築した自然性 MOS 予測モデルについて述べた．今後は，汎用的な MOS 予測モデルの構築を目指す．

謝辞：本研究は，JSPS 科研費 21H04900, 21K11955，JST 次世代研究者挑戦的研究プログラム JPMJSP2108, JST ムーンショット型研究開発事業 JPMJMS2011 の支援を受けた．

参考文献

[1] B. Patton et al., “AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech,” *arXiv preprint arXiv:1611.09207*, 2016.
[2] Y. Leng et al., “MBNET: MOS prediction for synthesized speech with mean-bias network,” *Proc. ICASSP*, pp. 391–395, 2021.
[3] W.-C. Huang et al., “The VoiceMOS Challenge 2022,” *arXiv preprint arXiv:2203.11389*, 2022.
[4] T. Hayashi et al., “ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” *Proc. ICASSP*, pp. 7654–7658, 2020.
[5] E. Cooper et al., “Generalization ability of MOS prediction networks,” *arXiv preprint arXiv:2110.02635*, 2021.
[6] P. Manocha et al., “NORESQA: A framework for speech quality assessment using non-matching references,” *Proc. NeurIPS*, vol. 34, 2021.
[7] W.-C. Huang et al., “LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech,” *arXiv preprint arXiv:2110.09103*, 2021.
[8] M. Ester et al., “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996, p. 226–231.
[9] L. Breiman, “Stacked regressions,” *Machine learning*, vol. 24, pp. 49–64, 1996.
[10] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.
[11] V. Panayotov et al., “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. ICASSP*, South Brisbane, Australia, Apr. 2015, pp. 5206–5210.
[12] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv preprint arXiv:2006.11477*, 2020.
[13] Q. Xu et al., “Simple and Effective Zero-shot Cross-lingual Phoneme Recognition,” *arXiv preprint arXiv:2109.11680*, 2021.
[14] D. Kingma, B. Jimmy, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
[15] T. Akiba et al., “Optuna: A next-generation hyperparameter optimization framework,” in *Proc. KDD*, 2019.
[16] W.-N. Hsu et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *arXiv preprint arXiv:2106.07447*, 2021.
[17] S. Chen et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *arXiv preprint arXiv:2110.13900*, 2021.
[18] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in *Proc. NIPS*, vol. 30, 2017.