UTMOSv2: 自然性 MOS 予測におけるスペクトログラム特徴量と SSL 特徴量の統合的利用*

☆馬場 凱渡（東大・工），中田 亘，齋藤 佑樹，猿渡 洋（東大院・情報理工）

1 はじめに

Mean Opinion Score (MOS) などの合成音声品質評価値の自動予測 [1, 2] は，人間を介入した主観評価の実施コストを削減でき，音声合成分野における新興の研究対象である．特に，高品質合成音声に対する MOS の正確な予測は，最先端の Deep Neural Network (DNN) ベース音声合成システムを公平に比較する上で非常に重要である [3].

本稿では，我々が VoiceMOS Challenge (VMC) 2024 Track 1 に向けて構築した MOS 予測システムである “UTMOSv2” を紹介する．この Track は，VMC 2022 で使用された BVCC データセット [4] に含まれる高品質音声合成システムのみを対象として再評価された MOS（即ち，zoomed-in MOS 評価 [5] の結果）の予測を目的としている．UTMOSv2 は，過去の VMC [4,6] で有効であった Self-Supervised Learning (SSL) 音声特徴量の活用 [7] や，複数の音声特徴量の fusion [8] を踏襲しつつ，音声スペクトログラムを画像と解釈して MOS 予測のための特徴量抽出を行う枠組みを新たに探究する．また，SSL 特徴量/スペクトログラムから MOS を予測するモデルを段階的に学習させ，これらの fusion に基づいて最終的な MOS 予測システムを構築する multi-stage learning の有効性も検証する．我々の UTMOSv2 は，VMC 2024 Track 1 において 16 評価指標のうち 9 項目で 1 位，残り 7 項目で 2 位となり，3 位以降に大きな差をつけて上位を独占した．本稿では，UTMOSv2 の構成要素（特徴量 fusion, multi-stage learning, 学習データ）に関する ablation study を実施した結果も報告する．UTMOSv2 は我々の GitHub リポジトリで公開予定である．

2 UTMOSv2

2.1 Basic Architecture

Figure 1に UTMOSv2 の概略図を示す.

スペクトログラム特徴抽出器: 音響/音声処理タスクにおける成功例 [9–11] をもとに，UTMOSv2 は音声スペクトログラムを画像とみなして MOS 予測のための特徴量を抽出する．Figure 2にスペクトログラム特徴抽出器の概略図を示す.

UTMOSv2 では，スペクトログラムの時間・周波数解像度を両立するために，まず，複数の窓長パラメータで短時間フーリエ変換を行う．次に，その結果から複数のメルスペクトログラムを抽出し，サイズを固定させたもとでそれらを画像とみなし，ImageNet [12] で事前学習された Convolutional NN (CNN) に入力する．最後に，各 CNN により抽出された特徴量の重み付き和を取り，その結果に対して attention [13] と

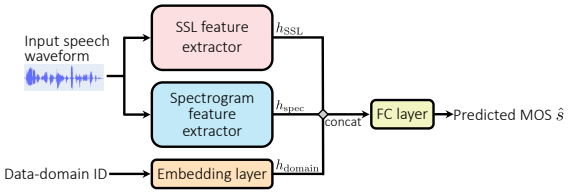

Fig. 1: UTMOSv2 の概略図

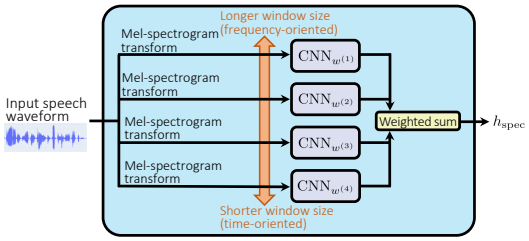

Fig. 2: スペクトログラム特徴抽出器の概略図

max-pooling の組み合わせを適用することでスペクトログラム特徴量 $\boldsymbol{h}_{\mathrm{spec}}$ を得る.

SSL 特徴抽出器: MOS 予測に関する先行研究 [2,7] を参考に，UTMOSv2 は音声波形からの特徴抽出に事前学習済み SSL モデルを活用する．特徴抽出プロセスは SSL モデルの各層から抽出された隠れ状態の重み付き和に対する attention と max-pooling で構成される．以降，この結果として得られる特徴量を $\boldsymbol{h}_{\mathrm{SSL}}$ と表記する.

データドメイン埋め込み: 前身の UTMOS [7] を参考に，UTMOSv2 はデータドメインで条件付けされた MOS 予測を採用する．具体的には，学習に用いられた MOS データセットの ID をドメインとみなし，ID を trainable look-up embedding table に入力した結果として得られるデータドメイン埋め込み $\boldsymbol{h}_{\mathrm{domain}}$ で MOS 予測のための全結合層を条件付けする．推論時における未知データドメインからの音声入力に対する MOS 予測は，既知データドメインの ID を入力するか，学習済み複数データドメインを仮定した予測結果を平均することで得られる [7].

特徴量 fusion: 前述の $\boldsymbol{h}_{\mathrm{spec}}, \boldsymbol{h}_{\mathrm{SSL}}, \boldsymbol{h}_{\mathrm{domain}}$ を結合し，次式のように全結合層に通すことで最終的な MOS 予測値を出力する.

$$\hat{s} = \mathrm{FC}\left(\mathrm{Concat}(\boldsymbol{h}_{\mathrm{spec}}, \boldsymbol{h}_{\mathrm{SSL}}, \boldsymbol{h}_{\mathrm{domain}})\right) \quad (1)$$

ここで，FC(·) と Concat(·) はそれぞれ全結合層と特徴量の結合処理を意味する.

2.2 追加された学習データセット

VMC 2024 Track 1 では公式の学習データセットが存在しないため，いくつかの MOS データセットを混合させて学習データセットとした．収集されたデータセットは BVCC [14]，Blizzard Challenge (BC) 2008 [15],

*UTMOSv2: Integrating Spectrogram and SSL Features for Naturalness MOS Prediction BABA, Kaito, NAKATA, Wataru, SAITO, Yuki, and SARUWATARI, Hiroshi (The University of Tokyo).

2009 [16], 2010 [17], 2011 [18], SOMOS [19], そして 50% zoomed-in BVCC (sarulab-data) である．

BC データセットでは，英語の合成音声サンプルのみを用いた．BC2008 では，“EUC” とマークされた評価者のスコアは 5 段階 MOS ではないため，データセットから除外した．BC2010 では，{ EH1, EH2, ES1, ES3 } タスクの MOS を用いた．ES2 は自然性 MOS 評価ではないため除外した．

2.3 損失関数

MOS 予測モデル学習時の損失関数は，UTMOS [7] で採用された contrastive loss と，Mean Squared Error (MSE) loss の重み付きである．Contrastive loss は，次式で定義される．

$$\mathcal{L}_{\text{con}}(s,\hat{s}) = \sum_{i\neq j} \max(0, |(s_i - s_j) - (\hat{s}_i - \hat{s}_j)| - \alpha) \quad (2)$$

ここで，s と $\hat{s}$ はそれぞれ予測対象の MOS とモデルから予測された MOS である．$\alpha > 0$ はマージンパラメータであり，小さな予測誤差をモデルが考慮しないよう制御する．最終的な損失関数は $\lambda_{\text{con}}\mathcal{L}_{\text{con}}(s,\hat{s}) + \lambda_{\text{mse}}\mathcal{L}_{\text{mse}}(s,\hat{s})$ で定義され，λ_{con} と λ_{mse} はそれぞれ第一項と第二項に対する重みを調整する非負のハイパーパラメータである．

2.4 Multi-stage learning

UTMOSv2 を構成するモデルは大きく，スクラッチでの学習は困難であるため，我々は各モデルを段階的に学習する multi-stage learning を導入する．

Stage 1: スペクトログラム特徴抽出器と SSL 特徴抽出器を個別に学習する．各特徴抽出器に MOS 予測のための FC 層を用意し，データドメイン埋め込み $\boldsymbol{h}_{\text{domain}}$ と各モデルの出力（$\boldsymbol{h}_{\text{spec}}$ もしくは $\boldsymbol{h}_{\text{SSL}}$）から MOS を予測するように特徴抽出器を最適化する．

Stage 2: 2 つの特徴抽出器のモデルパラメータを固定し，特徴量 fusion のための新たな FC 層とデータドメイン埋め込み $\boldsymbol{h}_{\text{domain}}$ を学習する．

Stage 3: 全てのモデルパラメータを小さい学習率で fine-tuning する．

同様の multi-stage learning は SSL 特徴抽出器の事前学習にも適用される．具体的には，まず SSL モデルのパラメータを固定させたもとで MOS 予測の FC 層のみを学習し，次に SSL モデルも含めて特徴抽出器全体を fine-tuning する．一方で，予備実験の結果から，スペクトログラム特徴抽出器に対する multi-stage learning は有意な改善を示さなかった．そのため，この特徴抽出器のモデルパラメータは，ImageNet で事前学習された EfficientNetV2 [20] を MOS 予測タスクで fine-tuning することによって初期化した．

3 実験的評価

3.1 評価指標

評価指標は VMC2024 Track 1 に準拠して設定した．即ち，評価セットは zoom-in rate を 25%, 12%として BVCC データセットに対して MOS 評価を再実施した結果である．これらの評価セットに対して MOS 予測を行い，system-/utterance-level の MSE, Linear Correlation Coefficient (LCC), Spearman’s rank CC (SRCC), Kendall’s RCC (KTAU) を計算した．本稿では紙面の都合上，zoom-in rate が 12%の評価セットに対する MSE と SRCC の評価結果のみを掲載するが，大まかな傾向は他の評価セット・評価指標に対しても同様であった．

3.2 共通の実験条件

UTMOSv2 のスペクトログラム特徴抽出器に含まれる CNN は ImageNet [12] で事前学習された EfficientNetV2 [21] を用いた．SSL 特徴量は LibriSpeech [22] で事前学習された wav2vec2.0 [23] baseで抽出した．データドメイン埋め込みの次元は 1 とした．

学習時の損失関数 (Eq. (2)) に対するマージンパラメータは全ての実験で $\alpha = 0.2$ とした．損失関数の重みパラメータは $\lambda_{\text{con}} = 0.2, \lambda_{\text{mse}} = 0.7$ とした．Optimizer は AdamW [24] を使用し，weight decay 係数を 1×10^{-4} とした．学習率のスケジューリングは cosine annealing [25] によって減衰され，multi-stage learning の各 stage ごとに調整した．MOS 予測に関する先行研究 [26] を参考に，学習時のデータ拡張として mixup [27] を使用した．

モデルのチェックポイント選択は system-level SRCC の平均を基準とした 5-fold 交差検証によって行なった．加えて，推論時には入力音声波形から無作為にフレームを 5 回抽出し，それらに対する予測を平均することにより最終的な MOS を出力した．

3.3 特徴量 fusion に関する評価

UTMOSv2 の特徴量 fusion の有効性を検証した．

実験条件: 本実験では以下の MOS 予測システムを比較した．

- **UTMOSv2**: 特徴量 fusion を用いる提案システム
- **UTMOSv2 w/o SSL**: スペクトログラム特徴量だけを用いる提案システム
- **UTMOSv2 w/o spec.**: SSL 特徴量だけを用いる提案システム
- **B01:** オリジナルの BVCC データセットで学習された SSL-MOS [2] システム．本システムは VMC2024 Track 1 のベースラインである．
- **UTMOS [7]:** 提案システムの前身

“UTMOSv2 w/o SSL” は学習率を 1×10^{-3} から 1×10^{-7} まで減衰させて学習され，バッチサイズとエポック数はそれぞれ 10 と 20 とした．Section 2.4で述べたように，“UTMOSv2 w/o spec.” は 2 段階で学習された．まず，学習率を 1×10^{-3} から 1×10^{-7} まで減衰させ，MOS 予測のための FC 層とデータドメイン埋め込みのみを，バッチサイズ 32 で 20 エポック学習させた．次に，学習率を 3×10^{-5} から 1×10^{-9} まで減衰させ，全てのモデルパラメータをバッチサイズ 32 で 5 エポック学習させた．特徴量 fusion を用いる “UTMOSv2” は，これらの 2 システムに基づいて構築された．具体的には，“UTMOSv2 w/o spec.” と “UTMOSv2 w/o SSL” で学習された特徴抽出器を初期値として，Section 2.4の Stage2–3 によってモデル

Table 1: UTMOSv2 とベースラインモデル（B01 と UTMOS）の比較．**太字**と下線付きスコアはそれぞれ 3 つの UTMOSv2 の ablation study 結果の中で最良，最悪の評価結果である．

	Utterance-level		System-level	
	MSE↓	SRCC↑	MSE↓	SRCC↑
UTMOSv2	0.459	**0.579**	0.288	**0.854**
w/o SSL	**0.357**	0.516	**0.188**	0.770
w/o spec.	0.673	0.529	0.497	0.793
B01	0.741	0.417	0.589	0.609
UTMOS [7]	0.541	0.300	0.378	0.367

Table 2: 提案する multi-stage learning の比較．**太字**と下線付きスコアはそれぞれ各列の中で最良，最悪の評価結果である．

	Utterance-level		System-level	
	MSE↓	SRCC↑	MSE↓	SRCC↑
UTMOSv2	0.459	**0.579**	0.288	**0.854**
w/o Stage 2	0.342	0.505	0.108	0.816
w/o Stage 1&2	**0.293**	0.423	**0.097**	0.672

パラメータを更新した．この際，FC 層とデータドメイン埋め込みはランダム初期化された．Stage 2 では学習率を 1×10^{-3} から 1×10^{-5} まで減衰させ，バッチサイズを 16 として 8 エポック学習させた．Stage 3 では学習率を 5×10^{-5} から 1×10^{-8} まで減衰させ，バッチサイズを 8 として 2 エポック学習させた．この ablation study では，学習データは Section 2.2で述べた全ての MOS データセットを混合させ，UTMOSv2 推論時のデータドメインは “BVCC” を指定した．

結果と考察: Table 1に評価結果を示す．SRCC については，UTMOSv2 が一貫して 2 つのベースラインを上回る性能を達成している．また，“w/o SSL” と “w/o spec.” の間で SRCC に性能差はほぼ観測されないが，特徴量 fusion を用いる “UTMOSv2” はこれらから大きく改善していることがわかる．これらの結果から，我々の UTMOSv2 は特徴量 fusion を活用し，zoomed-in MOS 予測においてベースラインシステムを大きく上回る SRCC を達成したと言える．

興味深い点として，“w/o SSL” は最良の utterance-/system-level MSE を達成したが，SRCC は UTMOSv2 ベースのシステムの中で最低値となった．一方で，“w/o spec.” は最悪の MSE を記録したが，SRCC の観点では “w/o SSL” を上回った．この観点から，UTMOSv2 に搭載されているスペクトログラム特徴抽出器は，スペクトログラム画像の微細構造を高い精度で捉えて絶対的な MOS 予測性能を改善できる一方で，SSL 特徴量はシステム間の相対的な順位関係を捉えるのに適している可能性が示唆されている．総括すると，UTMOSv2 はこれらの 2 つの特徴量を fusion し，良い MSE をキープしつつより最良の SRCC を達成した．

3.4 Multi-stage learning に関する評価

提案する multi-stage learning の有効性を検証した．

実験条件: 本実験では，Section 3.3における “UTMOSv2” を以下の MOS 予測システムと比較した．

- **UTMOSv2 w/o Stage 2**: Stage 2 学習を実施しない場合
- **UTMOSv2 w/o Stage 1&2**: Stage 3 学習のみ実施する場合

これらの 2 システムを構築する際，モデル学習のバッチサイズ，エポック数はそれぞれ 8, 20 とした．学習率のスケジューリングは，“UTMOSv2 w/o Stage 2” と “UTMOSv2 w/o Stage 1&2” に対してそれぞれ学習率を 1×10^{-4} から 1×10^{-7}，1×10^{-3} から 1×10^{-7} まで減衰させた．学習データセットと予測時のデータドメイン ID 設定は Section 3.3の実験と同じである．

結果と考察: Table 2に評価結果を示す．Multi-stage learning の段階を減らし，2 つの特徴抽出器を MOS 予測タスクで事前に学習させなくなるにつれて，構築されたシステムの挙動は “UTMOSv2 w/o SSL” に近づく（即ち，低い MSE と低い SRCC を記録する）ことがわかる．この傾向は合成音声の絶対的な MOS を予測したい場合においては望ましいが，複数の音声合成システム間の性能差を議論する際にはそうとは限らない．総括すると，提案する multi-stage learning は SSL 特徴量の長所である，複数の合成音声サンプル間の相対的な品質差を考慮した MOS 予測を実現する上で必須であることが示唆された．

3.5 データセットに関する調査

Zoomed-in MOS 予測システム構築におけるデータセットの影響を調査するための ablation study を実施した．

実験条件: 本実験では “UTMOSv2” (Section 3.3) に基づいて MOS 予測システムを構築し，学習データセットと推論時のデータドメイン ID のみ変更した．具体的には，“UTMOSv2” のモデルを “All datasets” もしくは “All w/o {BVCC, BC, SOMOS, sarulab-data}” で学習し，データセットを除外した際の zoomed-in MOS 予測の性能差を調査した．加えて，推論時に指定するデータドメイン ID を変えることで，どのデータセットのドメインが VMC2024 Track 1 評価セットに近いかどうかも調査した．

結果と考察: Table 3に評価結果を示す．まず，学習データセットの違いについて，多くの場合で “All datasets” が最良の評価値を達成しているが，“BVCC” や “BC” を除外した学習が最良となっている場合もある．加えて，“SOMOS” や “sarulab-data” を学習データから除外すると，概して評価値は悪化する傾向にあることがわかる．これらの結果から，zoomed-in MOS 予測システムを構築する際には (1) 低品質な音声合成システムを可能な限り除外しつつ，(2) 最先端の DNN 音声合成システムを多く含むようなデータセットで MOS 予測モデルを学習させることが重要であると示唆された．

次に，推論時に指定するデータドメイン ID の違いについて，MSE は “sarulab-data”（即ち 50% zoomed-in BVCC）を指定したときに最良であり，“BVCC” を指定したときに最悪となった．この結果は，先行研究において議論されていた range-equalizing bias [5]

Table 3: 学習時のデータセットと推論時のデータドメイン ID が MOS の予測性能に与える影響に関する ablation study の結果．例えば，第 2–3 列は VMC2024 Track 1 の評価セットの MOS を “BVCC” データドメインで予測したときの結果を示す．**太字**と下線付きスコアはそれぞれ各列の中で最良，最悪の評価結果である．

(a) Utterance-level の評価結果.

	BVCC		BC		SOMOS		sarulab-data	
Training datasets	MSE↓	SRCC↑	MSE↓	SRCC↑	MSE↓	SRCC↑	MSE↓	SRCC↑
All datasets	0.459	**0.579**	**0.262**	0.584	**0.234**	0.579	**0.238**	0.582
w/o BVCC	–	–	0.541	**0.626**	0.324	**0.629**	0.297	**0.629**
w/o BC	**0.393**	0.473	–	–	0.299	0.493	0.360	0.450
w/o SOMOS	0.447	0.376	0.443	0.375	–	–	0.443	0.378
w/o sarulab-data	0.484	0.430	0.312	0.431	0.392	0.428	–	–

(b) System-level の評価結果

	BVCC		BC		SOMOS		sarulab-data	
Training datasets	MSE↓	SRCC↑	MSE↓	SRCC↑	MSE↓	SRCC↑	MSE↓	SRCC↑
All datasets	0.288	**0.854**	**0.088**	**0.851**	**0.056**	0.844	**0.058**	**0.838**
w/o BVCC	–	–	0.343	0.832	0.128	**0.846**	0.101	0.836
w/o BC	**0.145**	0.819	–	–	0.069	0.823	0.122	0.805
w/o SOMOS	0.224	0.696	0.221	0.682	–	–	0.221	0.700
w/o sarulab-data	0.282	0.647	0.102	0.661	0.186	0.690	–	–

を実証している．しかしながら，この傾向は SRCC に関しては観測されない．以上より，range-equalizing bias の悪影響は MOS の絶対的な予測において顕著に現れることが示唆された．

4 おわりに

本稿では，我々が VMC2024 Track 1 に向けて構築した UTMOSv2 システムを紹介し，MOS 予測実験を通じて本システムの構成要素の影響を評価した．今後は，自然性以外の評価指標における MOS 予測システムの構築について探求する．

謝辞: 本研究は JST ムーンショット型研究開発事業 JPMJMS2011 の支援を受けたものです．また，議論にご協力いただいた，東京大学の山内一輝氏，武伯寒氏，濱田誉輝氏に感謝します．

参考文献

[1] C.-C. Lo et al., “MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion,” in *Proc. Interspeech*, 2019, pp. 1541–1545.
[2] E. Cooper et al., “Generalization ability of mos prediction networks,” in *Proc. ICASSP*, 2022, pp. 8442–8446.
[3] X. Chang et al., “The Interspeech 2024 Challenge on Speech Processing Using Discrete Units,” in *Proc. Interspeech*, 2024.
[4] W. C. Huang et al., “The VoiceMOS Challenge 2022,” in *Proc. Interspeech*, 2022, pp. 4536–4540.
[5] E. Cooper, J. Yamagishi, “Investigating range-equalizing bias in mean opinion score ratings of synthesized speech,” in *Proc. Interspeech*, 2023, pp. 1104–1108.
[6] E. Cooper et al., “The voicemos challenge 2023: Zero-shot subjective speech quality prediction for multiple domains,” in *Proc. ASRU*, 2023, pp. 1–7.
[7] T. Saeki et al., “UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022,” in *Proc. Interspeech*, 2022, pp. 4521–4525.
[8] Z. Qi et al., “LE-SSL-MOS: Self-supervised learning MOS prediction with listener enhancement,” in *Proc. ASRU*, 2023.
[9] J. Szep, S. Hariri, “Paralinguistic classification of mask wearing by image classifiers and fusion,” in *Proc. Interspeech*, 2020, pp. 2087–2091.
[10] J. S. Chung et al., “VoxCeleb2: Deep Speaker Recognition,” in *Proc. Interspeech*, 2018, pp. 1086–1090.
[11] S. Amiriparian et al., “Sentiment analysis using image-based deep spectrum features,” in *Proc. ACIIW*, 2017, pp. 26–29.
[12] J. Deng et al., “ImageNet: A large-scale hierarchical image database,” in *Proc. CVPR*, 2009, pp. 248–255.
[13] A. Vaswani et al., “Attention is all you need,” in *Proc. NeurIPS*, 2017.
[14] E. Cooper, J. Yamagishi, “How do voices from past speech synthesis challenges compare today?” in *Proc. SSW 11*, 2021, pp. 183–188.
[15] V. Karaiskos et al., “The Blizzard Challenge 2008,” in *Proc. Blizzard Challenge Workshop*, 2008.
[16] A. W. Black et al., “The Blizzard Challenge 2009,” in *Proc. Blizzard Challenge Workshop*, 2009, pp. 1–24.
[17] ——, “The Blizzard Challenge 2010,” in *Proc. Blizzard Challenge Workshop*, 2010.
[18] S. King, V. Karaiskos, “The Blizzard Challenge 2011,” in *Proc. The Blizzard Challenge Workshop*, 2011, pp. 1–10.
[19] G. Maniati et al., “SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis,” in *Proc. Interspeech*, 2022, pp. 2388–2392.
[20] M. Tan, Q. V. Le, “EfficientNetV2: Smaller models and faster training,” in *Proc. ICML*, 2021.
[21] ——, “EfficientNetV2: Smaller models and faster training,” in *Proc. ICML*, 2021.
[22] V. Panayotov et al., “Librispeech: an asr corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
[23] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.
[24] I. Loshchilov, F. Hutter, “Decoupled weight decay regularization,” in *Proc. ICLR*, 2019.
[25] ——, “SGDR: Stochastic gradient descent with warm restarts,” in *Proc. ICLR*, 2017.
[26] K. Wang et al., “MOSPC: MOS prediction based on pairwise comparison,” in *Proc. ACL*, 2023, pp. 1547–1556.
[27] H. Zhang et al., “mixup: Beyond empirical risk minimization,” in *Proc. ICLR*, 2018.