小特集 「Beyond MOS — 音声評価法の将来展望 —」

最先端の予測性能を持つ合成音声品質の自動評価システムUTMOSについて*

佐伯高明，高道慎之介（東京大学）**

1. はじめに

テキスト音声合成や音声変換などの音声生成タスクの評価には，かねてより人間による主観評価実験が最も重要な評価指標として用いられてきた．例えば，mean opinion score (MOS) による主観評価実験では，人間の聴取者が合成音声のサンプルを聴き，その自然性などを 5 段階の整数値で評価した結果の平均値として得られる MOS 値を比較する．一方で，このような主観評価は，時間的・金銭的コストが高く，さらに異なる聴取実験間の評価スコア同士を比較できないという問題がある．したがって，音声合成タスクの自動評価は，音声合成研究を加速する上で非常に重要な研究課題である．

音声生成の自動評価の推進に向けて，MOS 予測タスクの性能を競うコンペティションである VoiceMOS Challenge 2022 [1] が開催された．チャレンジには合計 22 のチームが参加し，様々な MOS 予測のアプローチが試みられた．著者を含む東京大学のチームが開発した UTMOS [2] は，Challenge の 16 個の評価指標のうち 10 個の指標で 1 位を獲得した手法であり，Challenge で最先端の性能を示した MOS 予測手法である．本稿では，この UTMOS について説明し，また，VoiceMOS Challenge 2022 以後の UTMOS に関連する研究動向について概説する．さらに，UTMOS を含む現状の MOS 予測モデルの課題と，今後の音声合成の自動評価の展望についても述べる．

2. MOS 予測モデルの概況

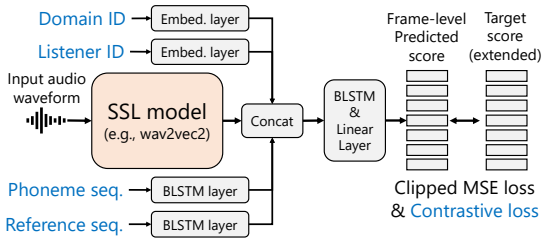
音声合成タスクでは，かねてより主観評価が重要な評価方法とされてきた [3] が，それに加えて様々な客観評価指標が用いられてきた．例えば，音声合成や音声強調において，mel cepstral distortion（MCD）[4] や perceptual evaluation of speech quality (PESQ) [5] のような客観評価指標が参照音声と生成音声に対して計算される．しかし，MCD のような単純な音響特徴に基づく指標は，人間の知覚的な自然性は高いが異なる音響的・韻律的特性を持つ発話において，人間の主観的評価との乖離が生じる．また，生成音声と同一の発話内容・話者性を持つ参照音声が必要であることから，ユースケースが限定されるという問題もある．

そのため，統計モデルを用いて入力音声から主観的評価値を予測するフレームワークが提案されてきた [6–10]．この MOS 予測の枠組みは，従来の主観的評価実験を代替するだけでなく，参照音声が得られない条件でも評価が可能であるという利点がある．例えば，Quality-Net [11] は双方向リカレントネットワークによって，音声強調の出力音声の発話レベル音声品質をフレーム単位で予測する枠組みであり，予測スコアと PESQ との間に高い相関があることを確認している．Voice Conversion Challenge 2018 [12] の聴取実験を用いて学習された MOSNet [13] は，畳み込みネットワークと双方向リカレントネットワークの構造によって，音声変換による合成音声サンプルから MOS を予測するモデルである．また，MOS は聴取者によってばらつきがあると考えられるため，MOS の聴取者への依存性を明示的にモデル化するような手法 [14, 15] も提案されてきた．さらに，合成音声の評価実験では，異なる合成音声対の評価値の順序を適切に評価できることが重要であるため，明示的に合成音声対を比較するような MOS 予測の手法も提案されている [16].

近年では，大規模なラベルなし音声データで学習された自己教師あり学習モデル [17, 18] による

* UTMOS: A state-of-the-art mean opinion score prediction system developed by UTokyo SaruLab.

** Takaaki Saeki and Shinnosuke Takamichi (University of Tokyo)



図–1 UTMOS [2] の強学習器.

音声表現が，音声認識をはじめとする様々な音声処理タスクで有用であることが報告されている．このような自己教師あり学習モデルを活用した MOS 予測の枠組み [19, 20] が多く提案されており，ラベル付きデータのないドメインへの汎化性能を改善することが報告されている [19]．UTMOS もこのような流れの中で開発された手法であり，3 節で述べるように，自己教師あり学習（self-supervised learning; SSL）モデルの転移学習や聴取者依存のモデリング，異なる合成音声対を比較する対照学習などを活用している．VoiceMOS Challenge 2022 [1] 以後，合成音声の自動評価の関心は高まっており，リファレンス音声を用いず，かつ教師なしでの自動評価手法 [21] も提案されている．このような教師なし手法と教師あり手法を組み合わせることで，out-of-domain の MOS 予測の性能を改善するような研究 [22] もある．さらに，リファレンスを用いた自己教師あり学習ベースの自動評価により，out-of-domain でより人間の主観評価値に高く相関するような手法も提案されている [23]．このような音声合成の自動評価手法が今後さらに研究されることで，音声合成研究を加速させることが多いに期待される．

3. UTMOS について

UTMOS では，図 1 に示すように，SSL モデルの fine-tuning に基づく MOS 予測モデルに様々な改良を行うことで，予測性能の高い強学習器を構築する．さらに，図 2 に示すような，複数の強学習器と弱学習器を用いたアンサンブル学習を行うことで，より汎化性能の高い MOS 予測モデルを構築する．この弱学習器は，事前学習済み SSL モデルの特徴量に対してリッジ回帰やサポートベクトルマシンなどの機械学習手法を用いたモデルとなっている．本節では，この UTMOS の概要と，VoiceMOS Challenge 2022 での結果について述べる．より詳細な説明については元論文 [2] を参照されたい．

3.1 UTMOS の強学習器

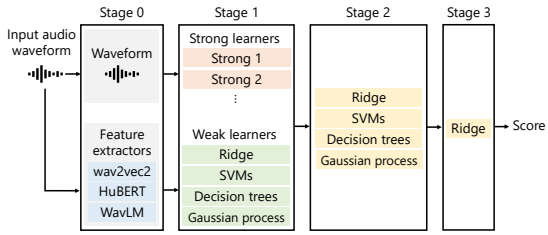
図 1 は強学習器の基本構造を示している．既存研究で，大規模な音声データで学習された SSL モデルを MOS 予測タスクに対して fine-tuning することで，より高い汎用性を実現することを確認している [19]．UTMOS の強学習器は，この SSL モデルの fine-tuning に基づくモデルに対して，複数の手法を導入したモデルである．

先行研究 [14, 15] では，聴取者によって評価値の分布が異なることに着目し，聴取者依存モデリングによる予測精度の改善を行なっている．さらに，評価値の分布は，聴取実験の設定（ドメイン）によっても異なると考えられる．そこで，UTMOS では，聴取者および聴取実験ドメインによるバイアスを考慮した学習を行うために，図 1 に示すように聴取者 ID とドメイン ID による条件付けを行う．その際，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，聴取者依存の評価値が出力される．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 [15] 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．このとき，予測するラベルは各発話に対して平均された MOS ではなく，とびとびの整数値となるため，以下のような clipped MSE loss [14] を回帰損失として用いる．

$$\mathcal{L}^{\mathrm{reg}}(y, \hat{y}) = \mathbb{1}(\|y - \hat{y}\| > \tau)(y - \hat{y})^2 \quad (1)$$

ただし，$\mathbb{1}(\cdot)$ は，括弧内の条件が真の時に 1 を出力し，そうでない時は 0 を出力する指示関数である．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．

対照学習は，ラベルなしでデータ間を比較し学習を行う自己教師あり学習手法である．UTMOS では，合成音声サンプル対のスコアの順位を正しく予測するために，図 1 に示すような contrastive loss を導入している．異なる発話 x_1, x_2 に対する主観評価値の差異を $d_{x_1,x_2} = s_1 - s_2$ で表し，予測された評価の差異を $\hat{d}_{x_1,x_2} = \hat{s_1} - \hat{s_2}$ とするとき，この差異が実際の評価差に近づくことが期



図–2 UTMOS [2] のアンサンブル学習.

待される．対照学習の損失関数は $\mathcal{L}x_1, x_2{}^{\mathrm{con}} = \max(0, |d_{x_1,x_2} - \hat{d}_{x_1,x_2}| - \alpha)$ で定義され，α は予測誤差を許容するためのマージンである．損失はミニバッチ内の全ペアに対して計算され $\mathcal{L}^{\mathrm{con}} = \sum_{i \neq j} \mathcal{L}^{\mathrm{con}}_{x_i,x_j}$ となる．これと式 (1) とを合わせて，最終的な損失関数は以下のように定義される．

$$\mathcal{L} = \beta \mathcal{L}^{\mathrm{reg}} + \gamma \mathcal{L}^{\mathrm{con}} \qquad (2)$$

ここで β 及び γ は損失の重み付けのハイパーパラメータである．

UTMOS では，合成音声の明瞭度を捉えるため，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場合，合成音声の明瞭性が低いと考えられるため，音声合成に入力されたと考えられる“擬似的な入力テキスト”も強学習器の入力として使用する．この擬似的な入力テキストは，音声認識結果を DBSCAN [24] 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．図 1 の reference sequence が，この擬似的な入力テキストに対応している．音素列と reference sequence は，図 1 に示すように，多層双方向リカレントネットワークに入力され，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．

さらに，UTMOS の強学習器では，過学習を抑制するための話速・ピッチ変換に基づくデータ拡張や，out-of-domain (OOD) track のための新たなラベル収集などを導入している．これらの手法により，UTMOS の強学習器は，従来の SSL-MOS [19] と比較して，VoiceMOS Challenge 2022 の全ての指標について高い性能を実現している．さらに，各手法についての ablation study の結果では，Main track と OOD track のほぼ全ての指標において，聴取者・ドメイン依存のモデリングの有効性が確認された．また，データ拡張や音素エンコーディングなどの手法は，データ量がより少なくなる OOD Track での有効性が確認された．

3.2 UTMOS のアンサンブル学習

UTMOS では，アンサンブル学習の一種であるスタッキング [25] を導入することで，さらに汎化性能を改善している．ここでは，3.1 節で述べた強学習器に加え，発話レベルの SSL 特徴量から単純な回帰モデルを用いて MOS 予測を行う弱学習器を導入する．このアンサンブル学習手法を用いた推論プロセスを図 2 に示す．弱学習器では，強学習器のように SSL モデルの重みを更新するのではなく，事前学習済み SSL モデル特徴量を時間方向に平均をとった特徴量を利用する．

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ため [26]，複数の SSL モデルからの特徴量を用いた別々のモデルを構築することで，モデルの種類を増やす．図 2 に示すように，このスタッキング手法による推論プロセスは，ステージ 0 からステージ 3 までの手順からなる．まず，ステージ 0 の特徴量抽出では，強学習器については音声波形をそのまま用い，弱学習器については複数の SSL モデルによる特徴量を得る．その後，ステージ 1 で，それぞれの入力特徴量を用いて強学習器と弱学習器を個別に学習し，交差検証により予測を行う．ステージ 2 では，ステージ 1 の予測結果を用いて学習されたメタ学習器を用いて予測を行う．最後のステージ 3 では，ステージ 2 の予測結果を用いて学習されたモデルによってスコアを予測する．

元論文 [2] では，このスタッキングの有効性を調査するために，強学習器，弱学習器の数を変化させた場合の比較を行っている．結果的に，単独の強学習器を用いた場合よりも，異なるハイパーパラメータから得られる複数の強学習器を用いてスタッキングを行ったほうが性能が高くなることが確認された．さらに，弱学習器を含めたスタッキングを行うことで，さらに性能を改善できることが示されている．このことは，モデルの多様性が，MOS 予測の汎化性能向上に寄与することを示し

ID	Utterance-level				System-level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
B01	17	12	12	12	18	11	12	12
B02	19	20	20	20	19	20	20	20
B03	21	22	22	22	20	22	22	22
T01	9	11	9	9	11	14	14	13
T02	8	9	10	10	10	9	10	10
T04	22	21	21	21	21	21	21	21
T05	5	6	7	7	8	8	8	9
T06	13	15	15	15	17	12	11	11
T07	14	10	11	11	4	4	5	6
T08	15	17	17	17	15	18	16	16
T09	16	19	19	19	14	19	19	19
T10	7	7	6	8	7	7	7	5
T11	3	2	2	3	5	1	1	1
T12	2	3	3	2	3	6	6	8
T13	18	16	16	16	22	17	17	18
T14	23	24	24	24	23	23	23	23
T15	24	23	23	23	24	24	24	24
T16	20	18	18	18	16	16	18	17
T17	1	1	1	1	1	3	3	2
T18	10	8	8	6	12	10	9	6
T19	4	4	5	5	2	2	2	3
T20	6	5	4	4	9	5	4	4
T21	11	14	13	14	6	13	13	14
T22	12	13	14	13	13	15	15	15

(a) Main track

ID	Utterance-level				System-level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
B01	8	6	10	10	9	5	2	2
B02	11	14	15	15	6	8	9	9
B03	13	16	16	16	8	15	12	13
T02	18	18	17	17	18	18	17	17
T03	5	4	4	4	4	3	14	12
T05	10	11	7	7	10	6	2	4
T06	16	12	11	11	16	12	5	5
T07	12	9	9	8	13	9	7	7
T10	6	10	8	9	7	14	13	15
T11	2	1	1	1	2	2	10	8
T12	7	5	5	6	12	7	11	11
T15	14	15	14	14	15	16	15	15
T17	1	2	2	2	1	1	1	1
T18	3	3	3	3	3	4	4	3
T19	4	7	6	5	5	11	6	5
T20	15	13	12	12	14	10	7	9
T21	17	17	18	18	17	17	18	18
T22	9	8	13	13	11	13	16	14

(b) OOD track

図–3 VoiceMOS Challenge 2022 での順位表 [1]．表内の数字は順位を表し，UTMOS は T17 である．

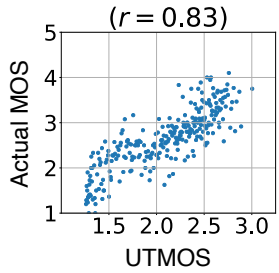
ている．VoiceMOS Challenge 2022 で UTMOS の他に高い性能を示したモデル [27][1]も，UTMOS のようなスタッキングを用いた予測は行わないものの，複数の SSL モデルを組み合わせた予測を行っている．このことから，複数の SSL モデルの特徴を MOS 予測に用いることの有効性が示唆される．

3.3 VoiceMOS Challenge 2022 について

VoiceMOS Challenge 2022 では，Main track と OOD track が設けられた．Main track は，大規模な聴取テストのデータセットを用いて高性能なモデルを構築することを対象としている一方，OOD track では，異なる言語や聴取テスト環境などのドメインに対してモデルを適応することに焦点を当てている．Main track では BVCC dataset [28] が使用された．この BVCC dataset は，過去の Blizzard Challenge と Voice Conversion Challenge, ESPnet-TTS [29] 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 評価の結果から成る．OOD track のデータセットは，Blizzard Challenge 2019 [30] 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 とは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 (mean squared error; MSE)，線形相関係数 (linear correlation coefficient; LCC)，スピアマンの順位相関係数 (Spearman's rank correlation coefficient; SRCC)，ケンドールの順位相関係数 (Kendall rank correlation coefficient; 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．より詳細な Challenge の実施方法やデータセットの統計については，UTMOS の元論文 [1] を参照されたい．

Main track では，3 つのベースライン手法に加えて 21 チームが予測結果を提出した．OOD track では，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．UTMOS の Main track での結果は，Utt. MSE = 0.165 (1), Utt. SRCC = 0.897 (1), Sys. MSE = 0.090 (1), Sys. SRCC = 0.936 (3)，OOD track での結果は,Utt. MSE = 0.162 (1), Utt. SRCC = 0.893 (2), Sys. MSE = 0.030 (1), Sys. SRCC = 0.988 (1) であった．ただし，Utt. と Sys. はそれぞれ発話レベル・システムレベルの結果を表し，括弧の中の数字は全体における順位を示す．図 3 は，main track と OOD track における順位を表しており，T17 が我々のチームの ID である．このように，UTMOS は，合計 16 種類の指標のうち 10 個の指標で 1 位を獲得しており，最も順位を落とした指標でも 3 位以内に入っていることがわかる．このことから，UTMOS は，VoiceMOS Challenge 2022 において最先端の性能を実現したモデルであると言える．

[1] これは 16 個の指標のうち 6 個の指標で 1 位を獲得したモデルである．



図–4 日本語多話者音声合成における，UTMOS による予測 MOS 値と実際の MOS 値の散布図 [35]．r は線形相関係数であり，各点は話者ごとの値を示す．

4. VoiceMOS Challenge 2022 以後の影響

VoiceMOS Challenge 2022 以後，テキスト音声合成 [31] やニューラルボコーダ [32] など，複数の音声合成に関する研究で，UTMOS が評価指標として用いられている．このような研究の中でも，主観評価による自然性 MOS と UTMOS による予測スコアが高く相関することが言及 [31] されている．UTMOS は，公式実装 [2]だけでなく，他の開発者によって fairseq への依存性を持たない使いやすいツールキット [3]としても公開されている．さらに，UTMOS による自動評価は，著者によって End-to-end 音声処理のツールキットである ESPnet [33, 34] にも実装されている．

また，UTMOS についての興味深い分析も行われている．関らの研究 [35] では，日本語多話者音声合成のタスクについて，UTMOS による予測 MOS 値と実際の主観評価 MOS 値についての分析が行われている．図 4 は，この分析結果を示す散布図である．ここで用いられている UTMOS のモデルは，英語と中国語で学習されたモデルであるが，日本語の合成音声の評価についても，UTMOS による予測 MOS 値と実際の MOS 値の話者レベルの相関が 0.8 を上回ることが確認された．この分析結果は，UTMOS は，一定の性能を維持しながら言語横断的な適用が可能であることを示唆している．また，著者による最新の研究 [23] では，リファレンスを用いた自動評価手法との比較も行われている．結果的に，信号処理ベースの評価指標である MCD [4] と比較すると，学習データに含まれないドメインの音声を評価する zero-shot の UTMOS は，多くのケースで人間の主観評価値と同程度もしくはより高く相関することが示されている．さらに，先行研究の実験結果 [31] から，UTMOS 値は合成音声の主観評価値にはよく相関するものの，自然音声に対して低い値を示す傾向が見られる．これは，UTMOS が合成音声のデータセットで学習されていることが主な要因であると考えられる．

さらに，UTMOS を発話合成音声の評価以外に適用した研究も存在する．例えば，歌声変換のチャレンジである Singing Voice Conversion Challenge 2023 [36] では,UTMOS を評価指標の一つに採用している．このサマリー論文 [36] の中で，UTMOS は，MCD や SSL-MOS [19] などの指標より，主観評価 MOS と高く相関することが言及されている．さらに，音声強調の評価に UTMOS を使用した研究 [37, 38] もある．[37] の著者らの事前実験において，UTMOS が PESQ [5] や SI-SDR, DNSMOS [10] などの手法よりも人間の主観評価 MOS に高く相関したことが言及されている．このことから，UTMOS は，学習データセットに含まれていないような，歌声合成や音声強調などのタスクにおいても有効であることが示唆される．

最後に，定量的な評価は公表できていないが，UTMOS を著者の研究グループで使用した際の所感について触れておく．まず，簡易的な実験により，UTMOS には一定の話者依存性があることを確認している．つまり，同じ条件で収録されたスタジオ録音品質の自然音声についての平均的な UTMOS のスコアは，話者によって異なることを確認している．これは学習データに含まれる音声の話者性の偏りによるものであると考えられるが，このような話者依存性は自然性スコアを予測する UTMOS の目的の上では好ましくない特性である．また，怒り音声など，感情音声については，自然音声であっても低い予測スコアを示す傾向が見られている．さらに，UTMOS 値は相対的なスコアとしては比較的優れている，つまり主観評価スコアとの相関は高いものの，絶対的なスコアとしての信頼度は高くないと考えられる．UTMOS による評価は主観評価に伴う金銭的・時間的コストを削減できる一方，上述のような特性に配慮した上で，音声合成タスクの評価に使用する必要がある．

[2] https://github.com/sarulab-speech/UTMOS22
[3] https://github.com/tarepan/SpeechMOS

5. 現状の課題と今後の展望

UTMOS は，3 節に述べた様々な手法により，学習データに含まれるドメインの MOS 予測に関しては非常に高い性能を実現している．一方で，異なる聴取実験設定や異なる言語など，未知のドメインへの zero-shot な予測性能を改善することが，非常に重要な今後の課題である．UTMOS のように，SSL model の転移学習を行うことで，未知のドメインに対してより高い性能を実現できることが示されている [19] が，依然として in-domain の予測性能とは乖離がある．このような問題意識から，VoiceMOS Challenge 2023 [39] は，out-of-domain の予測にフォーカスしている．フランス語合成音声や歌声合成音声，ノイジー音声，音声強調の出力音声などの評価について，ターゲットドメインの MOS ラベル付き学習データが提供されない条件で Challenge が行われた．この Challenge に向けて，教師あり MOS 予測手法と教師なし自動評価手法とのハイブリッドな手法 [22] や，聴取者依存のモデリングを用いた手法 [40] など．様々な手法が提案された．さらに，著者による最新の研究 [23] では，テキスト生成の評価尺度から着想を得た，参照音声を用いる自動評価手法を提案しており，ツールキットも公開している [4]．この研究では，自己教師あり音声表現を用いるものの，UTMOS とは違い下流タスクでの学習を行わないことによって，MOS 予測モデルの out-of-domain 予測の問題点にアプローチしている．このように，実際のユースケースに向けて zero-shot な音声自動評価の性能を改善する方向性は，今後ますます重要になると考えられる．

また，現状の自動評価は，発話文単位で合成された読み上げ音声の自然性評価を対象とするにとどまっている．一方で，近年のニューラル音声合成の研究では，長文の合成 [41] や対話の合成 [42] など，より柔軟な音声合成タスクに向けた研究が行われている．したがって，音声の自動評価もまた，今後，より幅広い音声生成タスクに対応させる必要がある．例えば，長文の読み上げ音声合成を自動評価するため，長文の合成音声と，その自然性の主観評価値を含むデータを収集することが必要である．また，合成音声の自然性以外にも，発話の表現力や対話としての自然性など，音声合成モデルが柔軟なタスクを扱えるようになるのに合わせて，自動評価の評価指標も拡張していく必要がある．

さらに，予測された MOS の結果が，入力音声のどのような特性に基づいたものであるかを推論する枠組みも重要である．このような枠組みは MOS 予測の解釈性を向上させ，当該タスクに MOS 予測モデルを適用できるかどうかを定性的に把握するのに役立つと考えられる．さらに，自然言語で制御できる MOS 予測も興味深いトピックである．現状の UTMOS は自然性に関する主観評価 MOS を予測するにとどまっているが，どのような指標についての評価を行うかが自然言語で制御できるようになれば，より適用範囲を拡張できると考えられる．近年，audio language model の text prompt によって，様々なタスクについて参照音声なしで音声の自動評価を行う手法 [43] も提案されており，自然言語ベースの自動評価は今後の発展が期待されるトピックである．

6. ま と め

本稿では，VoiceMOS Challenge 2022 で最先端の予測性能を実現した合成音声品質の自動評価システム UTMOS について述べた．UTMOS は，SSL モデルの転移学習に基づく MOS 予測モデルに対し，聴取者・ドメインを考慮したモデリング，対照学習，音素エンコーディング，アンサンブル学習など，様々な手法を導入することで性能を改善している．UTMOS は複数のツールキット上に実装されており，実際に近年の合成音声の評価でも用いられているため，今後音声合成研究をより加速させるものとして期待される．一方で，zero-shot での適用性能など，UTMOS には依然として課題も多い．今後，音声合成の自動評価をさらに推進するためのデータセット整備などが特に重要な課題である．

謝辞:本研究は科研費 21H04900, 22H03639, 23H03418, 23K18474，JST 創発的研究支援事業 JP23KJ0828，ムーンショット JPMJPS2011 の助成を受けた．本解説記事の執筆に際し，東京大学大学院の関健太郎氏の助言を受けた．

[4] https://github.com/Takaaki-Saeki/DiscreteSpeechMetrics

文　献

[1] W.-C. Huang et al., "The VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.

[2] T. Saeki et al., "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.

[3] A. W. Black, K. Tokuda, "The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. Interspeech*, 2005, pp. 77–80.

[4] T. Fukada et al., "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.

[5] A. W. Rix et al., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

[6] T. Yoshimura et al., "A hierarchical predictor of synthetic speech naturalness using neural networks." in *Interspeech*, 2016, pp. 342–346.

[7] B. Patton et al., "Automos: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv preprint arXiv:1611.09207*, 2016.

[8] A. R. Avila et al., "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 631–635.

[9] G. Mittag et al., "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.

[10] C. K. A. Reddy et al., "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2021, pp. 6493–6497.

[11] S.-W. Fu et al., "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *arXiv preprint arXiv:1808.05344*, 2018.

[12] J. Lorenzo-Trueba et al., "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *The Speaker and Language Recognition Workshop*, 2018, pp. 195–202.

[13] C.-C. Lo et al., "MOSNet: Deep learning-based objective assessment for voice conversion," *Proc. Interspeech*, pp. 1541–1545, 2019.

[14] Y. Leng et al., "MBNET: MOS prediction for synthesized speech with mean-bias network," *Proc. ICASSP*, pp. 391–395, 2021.

[15] W.-C. Huang et al., "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *Proc. ICASSP*, 2021, pp. 896–900.

[16] K. Wang et al., "MOSPC: MOS prediction based on pairwise comparison," in *Proc. ACL*, 2023, pp. 1547–1556.

[17] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[18] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM TASLP*, vol. 29, pp. 3451–3460, 2021.

[19] E. Cooper et al., "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022, pp. 8442–8446.

[20] T. Sellam et al., "Squid: Measuring speech naturalness in many languages," in *Proc. ICASSP*, 2023.

[21] S. Maiti et al., "SpeechLMScore: Evaluating speech generation using speech language model," in *Proc. ICASSP*, 2023.

[22] Z. Qi et al., "Le-ssl-mos: Self-supervised learning mos prediction with listener enhancement," in *Proc. ASRU*. IEEE, 2023.

[23] T. Saeki et al., "SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics," *arXiv preprint arXiv:2401.16812*, 2024.

[24] 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.

[25] L. Breiman, "Stacked regressions," *Machine learning*, vol. 24, pp. 49–64, 1996.

[26] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[27] Z. Yang et al., "Fusion of self-supervised learned models for mos prediction," *arXiv preprint arXiv:2204.04855*, 2022.

[28] E. Cooper, J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" in *Proc. ISCA SSW*, 2021, pp. 183–188.

[29] 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.

[30] Z. Wu et al., "The blizzard challenge 2019," in *Proc. Blizzard Challenge Workshop*, vol. 2019, 2019.

[31] S.-H. Lee et al., "Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis," *arXiv preprint arXiv:2311.12454*, 2023.

[32] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," in *Proc. ICLR*, 2024. [Online]. Available: https://openreview.net/forum?id=vY9nzQmQBw

[33] S. Watanabe et al., "ESPnet: End-to-end speech processing toolkit," *Proc. Interspeech*, pp. 2207–2211, 2018.

[34] T. Hayashi et al., "ESPnet2-TTS: Extending the edge of tts research," *arXiv preprint arXiv:2110.07840*, 2021.

[35] K. Seki et al., "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," in *Proc. ICASSP*, 2023.

[36] W.-C. Huang et al., "The singing voice conversion challenge 2023," in *Proc. ASRU*. IEEE, 2023.

[37] P. Andreev et al., "Iterative autoregression: a novel trick to improve your low-latency speech enhancement model," *arXiv preprint arXiv:2211.01751*, 2022.

[38] L. Sun et al., "Dual-branch modeling based on state-space model for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[39] E. Cooper et al., "The voicemos challenge 2023: zero-shot subjective speech quality prediction for multiple domains," in *Proc. ASRU*, 2023.

[40] K. Shen et al., "Sqat-ld: Speech quality assessment transformer utilizing listener dependent modeling for zero-shot out-of-domain mos prediction," in

Proc. ASRU. IEEE, 2023.
[41] R. Clark et al., “Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs,” in *Proc. SSW*, 2019.
[42] T. A. Nguyen et al., “Generative spoken dialogue language modeling,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.
[43] S. Deshmukh et al., “PAM: Prompting audio-language models for audio quality assessment,” *arXiv preprint arXiv:2402.00282*, 2024.

佐伯高明

2019 年に東京大学工学部を卒業．2021 年に東京大学大学院情報理工学系研究科博士前期課程を修了．2021 年より東京大学大学院情報理工学系研究科博士後期課程．音声合成，音声表現学習，音声信号処理の研究に従事．

高道慎之介

2011 年に長岡技術科学大学を卒業．2013 年・2016 年それぞれに奈良先端科学技術大学院大学博士前期・後期課程を修了．2023 年より東京大学講師（現職）．博士（工学）．音声合成変換，音声信号処理の研究に従事．