一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2023-133,SIP2023-180,SP2023-115(2024-03)

テキスト生成の自動評価尺度に基づく音声生成の自動評価

佐伯 高明[1,a)]　マイティ ソウミ[2,b)]　高道 慎之介[1,c)]　渡部 晋治[2,d)]　猿渡 洋[1,e)]

概要：音声生成の評価において，主観的評価は長らく最も重要な基準であったが，メルケプストラル歪み（mel cepstral distortion: MCD）や mean opinion score（MOS）予測モデルなどの客観評価尺度も使用されてきた．これらの客観評価指標は，時間的・金銭的コストが低く，異なる評価結果同士を比較できるという利点があり，人間の主観的判断と高い相関を持つ客観評価尺度が求められている．本稿では，テキスト生成の評価指標に基づく，音声生成のための自動評価手法を提案する．提案する *SpeechBERTScore* は，生成された音声と参照音声から得られた自己教師あり音声特徴量系列に対して BERTScore を計算する．また，提案する *SpeechBLEU* と *SpeechTokenDistance* では，自己教師ありの離散音声トークンを用いて評価尺度を定義する．合成音声に関する実験的評価では，提案手法の SpeechBERTScore が，MCD や最先端の MOS 予測モデルよりも人間の主観的評価と高く相関することを示した．さらに，提案手法は劣化音声の評価に対しても効果的であり，言語横断的な適用が可能であることが明らかとなった．

キーワード：音声自動評価，音声生成，自己教師あり音声表現，テキスト生成自動評価

Takaaki Saeki[1,a)]　Soumi Maiti[2,b)]　Shinnosuke Takamichi[1,c)]　Shinji Watanabe[2,d)]
Hiroshi Saruwatari[1,e)]

1. はじめに

主観的な聴取テストは，生成された音声 [1] や劣化した音声 [2] の品質を評価するための最も重要な基準とされてきた．しかし，聴取テストの時間・金銭的コストの大きさや，異なる聴取テスト同士の結果を比較できない問題などから，さまざまな客観的評価指標が使用されてきた．例えば，音声合成や音声強調において，メルケプストラル歪み（MCD）[3] 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，このような単純な音響特徴に基づく指標は，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，人間の主観的評価から逸脱しうる．そのため，統計モデルを用いて入力音声から mean opinion score (MOS) などの主観的評価値を予測するフレームワークにも焦点が当てられている [4–8]．しかし，これらの MOS 予測モデルの性能は，下流タスクの教師あり学習で使用されたドメイン以外のデータに対して低下し，実用性に制限がある [5]．

かねてより，自然言語処理（natural language processing; NLP）の分野で，人間の主観的判断と高い相関を持つ自動評価指標が提案されている [9,10]．BLEU [11] は，参照テキストと生成テキストの間の n-gram の重複を数えることで内容の一致を捉え，BERTScore [9] は，言語モデルによる意味表現を使用して文脈的な合致を捉える．最近では，音声データから潜在的な音声情報を学習する自己教師あり表現学習（self-supervised learning; SSL）フレームワークが提案されている [12–14]．そのような表現は，音声トークン系列をテキストトークン系列のように扱うことを可能にし，音声コーパス上での言語モデルの構築をも可能にしている [15]．よって，これらの SSL 音声特徴表現に対し自動テキスト生成指標を適用することは，人間の主観的スコアと高く相関する音声評価指標の設計において有望である．

我々は，NLP の自動評価指標に基づく，音声生成のための新たな自動評価指標を提案する．提案する *SpeechBERTScore* は，生成された音声と参照音声の両方からの

[1] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan.
[2] カーネギーメロン大学 言語技術研究所
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
a) takaaki_saeki@ipc.i.u-tokyo.ac.jp
b) smaiti@andrew.cmu.edu
c) shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
d) shinjiw@ieee.org
e) hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

表 1 従来の客観評価指標と提案手法の比較.

Method	Need reference	Need labelled data	Use SSL pretraining	Need down stream training
MCD [3], PESQ [16]	Y	N	N	N
SpeechLMScore [17]	N	N	Y	Y
MOSNet [4]	N	Y	N	Y
UTMOS [8]	N	Y	Y	Y
Ours	Y	N	Y	N

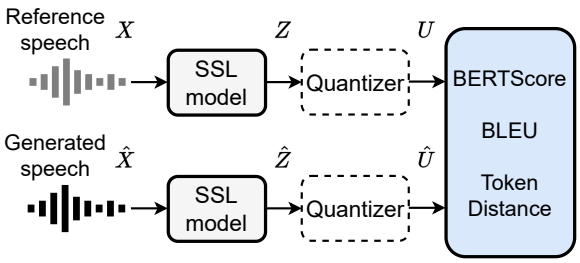

図 1 提案する音声自動評価指標．SpeechBERTScore は SSL 連続特徴量を用いて計算される．SpeechBLEU と SpeechTokenDistance は量子化された離散音声トークンに対して計算される．

SSL 連続値特徴に対して BERTScore を計算し，その意味的一致性を捉える．*SpeechBLEU* と *SpeechTokenDistance* は，それぞれ生成された音声と参照音声の離散トークン系列に対して BLEU と文字レベルの距離を計算する．提案するこれらの評価指標は，参照音声と生成音声が異なる系列長を持つ場合にも適用可能である．表 1は，従来の客観評価指標と我々の提案手法の相違を示している．従来の信号処理に基づく客観的指標 [3, 16] とは異なり，提案手法は，より意味的情報を捉えた評価のために SSL 音声特徴量を使用する．また，従来の自動評価フレームワーク [4, 8, 17] とは異なり，我々の方法は下流タスクでの学習を必要とせず，非常に利用コストが低い．評価実験により，我々の手法が，従来の自動評価指標よりも人間の主観的評価値と高く相関することを示した．また，提案する SpeechBERTScore は，ノイジー音声に対しても有効であり，言語横断的な適用が可能であることを示した．本研究で提案する評価指標は，オープンソースのツールキットとして公開している[*1]．

2. 提案する自動評価指標

本節では，提案する音声生成の客観的指標について述べる．これらの尺度は，NLP タスクの評価に使用されてきた客観的指標に基づいて，音声生成の評価ために設計されている．提案する評価尺度は，いずれも，生成音声と参照音声が異なる系列長を持つ場合にも適用でき，下流タスクでの学習が不要である．

2.1 SpeechBERTScore

BERTScore [9] は，テキスト生成タスクに対し広く用いられている自動評価手法である．BERTScore では，各テキストトークンに対応する BERT [18] 埋め込みを用いて，生成されたテキストと参照テキストとの類似性が計算される．本研究では，BERTScore を音声生成の評価指標に対して適用する *SpeechBERTScore* を提案する．*SpeechBERTScore* は，生成された音声と参照音声の両方からの SSL 連続値特徴量系列に対して BERTScore を計算することで，その意味的一致を捉える．

生成された音声波形を $\hat{X} = (\hat{x}_t \in \mathbb{R} | t = 1, \cdots, T_{\text{gen}})$，参照音声波形を $X = (x_t \in \mathbb{R} | t = 1, \cdots, T_{\text{ref}})$ と表す．ここで，波形の長さ T_{gen} と T_{ref} は一致しなくてもよいとする．$\hat{Z} = (\hat{\boldsymbol{z}}_n \in \mathbb{R}^D | n = 1, \cdots, N_{\text{gen}})$ および $Z = (\boldsymbol{z}_n \in \mathbb{R}^D | n = 1, \cdots, N_{\text{ref}})$ は，それぞれ $\hat{X}$ と X から得られた SSL 連続値特徴量を表し，以下のように書ける．

$$\hat{Z} = \text{Encoder}(\hat{X}; \theta), \qquad Z = \text{Encoder}(X; \theta), \quad (1)$$

ここで θ は事前学習されたエンコーダモデルのモデルパラメータを表す．N_{gen} と N_{ref} は，それぞれ T_{gen} と T_{gen} によって一意に決まり，エンコーダのサブサンプリング率に依存する．事前学習されたエンコーダモデルには，SSL によって学習されたモデル [12, 13] を使用する．

元の BERTScore は，精度，再現率，F1 スコアを定義しているが，ここでは，予備実験で最も良い結果を示した精度を使用した．この時，SpeechBERTScore は次のように定義される．

$$\text{SpeechBERTScore} = \frac{1}{N_{\text{gen}}} \sum_{i=1}^{N_{\text{gen}}} \max_j \cos(\hat{\boldsymbol{z}}i, \boldsymbol{z}j) \quad (2)$$

ここで，$\cos(\cdot)$ は二つの特徴間のコサイン類似度を表す．

2.2 SpeechBLEU

BLEU [11] では，n-gram の一致に基づいてスコアが計算され，機械翻訳されたテキストの品質を人間による翻訳文と比較する用途で使用される．提案する SpeechBLEU では，参照音声と生成された音声の品質を評価するために，離散音声トークンに対して BLEU が計算される．

生成された離散ユニット系列を $\hat{U} = (\hat{u}_n \in \mathbb{R} | n = 1, \cdots, N_{\text{gen}})$，参照離散ユニット系列を $U = (u_n \in \mathbb{R} | n = 1, \cdots, N_{\text{ref}})$ と表す．外部の量子化器を使用することで，離散ユニットは次のように得られる．

$$\hat{U} = \text{Quantizer}(\hat{Z}; \phi), \qquad U = \text{Quantizer}(Z; \phi), \quad (3)$$

ここで ϕ は量子化器のパラメータを表す．本研究では，量子化器には k-means アルゴリズムを使用する．この時，SpeechBLEU は以下のように定義される：

$$\text{SpeechBLEU} = \text{BLEU}\, \boldsymbol{w}(\hat{U}, U), \quad (4)$$

ここで $\boldsymbol{w}$ は各 n-gram の重みを表す．BLEU を計算する

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

ために使用される最大の n-gram 数 を G とするとき，$G \in \mathbb{Z}, G \geq 1$ が満たされる．$\boldsymbol{w}$ は $\boldsymbol{w} = (w_g \in \mathbb{R} | g = 1, \cdots, G)$ として記述される．ここでは，各 n-gram に一様な重み $w_g = 1/G$ を使用する．

2.3 SpeechTokenDistance

NLP タスクの評価では，BLEU のように言語構造を捉える指標だけでなく，文字レベルの類似性を計算する原始的な指標も使用されてきた．本研究では，2 つの音声離散トークン系列に対してそのような文字レベルの距離を計算することで，生成された音声を評価する．これまで，様々な文字レベルの距離尺度が提案されているが，我々は代表的な Levenshtein 距離 [19] と Jaro-Winkler 距離 [20] を調査する．

Levenshtein 距離は，一方のテキストを他方に変化させるために必要な単一文字編集の最小数を計算する．Jaro-Winkler 距離は，2 つのテキスト間の類似性を測る尺度であり，Jaro 距離 [21] は，共通する文字の数と順序に基づいて類似性を計算し，Winkler の拡張 [20] は接頭辞により多くの重みを与える．この時，SpeechTokenDistance は，式 (3) で与えられる離散トークンに対して以下のように計算される．

$$\text{SpeechTokenDistance} = \text{DistanceMeasure}(\hat{U}, U), \quad (5)$$

ここで，DistanceMeasure には Levenshtein 距離または Jaro-Winkler 距離が用いられる．

3. 実験的評価

3.1 実験条件

3.1.1 データセット

本研究では，3 種類の評価用データセットを使用し，サンプリン周波数を 16kHz に設定した．まず，英語の合成音声の評価には，SOMOS データセット [22] を使用した．これには，200 種類の異なるテキスト音声合成 （TTS） の音響モデルと LPCNet [23] ニューラルボコーダーによって合成された LJSpeech [24] の合成音声と，それに対応する主観評価値が含まれている．評価には参照音声が必要なので，SOMOS データセットから LJSpeech ドメインのみを使用し，信頼性が低い評価を除外する SOMOS-clean サブセットを採用した．結果として，1000 個の合成音声発話，それに対応する評価と自然音声発話が用いられた．

提案手法の言語横断的な適用性を調査するために，中国語の合成音声データセットも使用した．BVCC データセット [25] の Blizzard Challenge 2019 （BC2019） [26] のサブセットを使用した．これは 1300 サンプルの合成音声と，それに対応する評価値と自然音声発話を含む．

さらに，劣化音声の評価には，NISQA コーパス [6] の `NISQA_VAL_SIM` サブセットを使用した．このデータセットには，様々な音響的歪みによって生成された 2500 個の劣化音声発話と，それに対応する評価値とクリーンな音声発話が含まれている．

3.1.2 自己教師あり事前学習モデル

SpeechBERTScore の評価では，Wav2vec 2.0 [12], HuBERT [13], WavLM [14], Encodec [27] を含む複数の SSL モデルを利用した．具体的には，fairseq [28] で利用可能なモデルとして，Wav2vec 2.0 Base (`wav2vec2-base`), Wav2vec 2.0 Large (`wav2vec2-large`), HuBERT Base (`hubert-base`), HuBERT Large (`hubert-large`), WavLM Base (`wavlm-base`), WavLM Base+ (`wavlm-base+`), WavLM Large (`wavlm-large`) を使用した．Encodec (`encodec`)[*2] については，residual vector quantizer 層の前の連続値特徴を使用した．§ 3.4以外の評価結果では，最も性能の良かった層の特徴量を使用した．

中国語の合成音声の評価では，公開 SSL モデルを使用した[*3]：Wav2vec 2.0 Base (`wav2vec2-base-cmn`), Wav2vec 2.0 Large (`wav2vec2-large-cmn`), HuBERT Base (`hubert-base-cmn`), HuBERT Large (`hubert-large-cmn`), 53 (`xlsr-53`) および 128 言語 (`xlsr-128`) で訓練された多言語 XLSR モデル [29] を使用した．

SpeechBLEU および SpeechTokenDistance の評価では，LibriSpeech 960h [30] で訓練された k-means モデルを使用し，HuBERT Base 特徴を離散音声トークンに変換し，クラスタサイズとして 50，100，200 を比較した．§ 3.4以外の評価結果では，最も性能の良かった層から得られた音声トークンを使用した．

3.1.3 比較手法

合成音声の評価には，音声合成評価で広く使用される参照音声を用いた客観的指標である MCD [3] と Log F0 root mean squared error （RMSE） を使用した．MCD と Log F0 RMSE には，ESPnet2-TTS [31, 32] 内の評価スクリプトを使用した．また，参照音声が不要な教師なし評価指標として，SpeechLMScore [17] を使用した．この SpeechLMScore には，LibriSpeech 960h [30] でトレーニングされたトークン語彙サイズ 50 の公開モデル[*4] を使用した．参照音声が不要な教師あり評価指標としては，UTMOS [8] を使用した．これは，BVCC データセットでトレーニングされ，VoiceMOS Challenge 2022 [33] で最も高い性能を示した MOS 予測モデルの一つである．

劣化音声の評価では，参照音声を用いる客観評価指標として，PESQ [16], STOI [34], ESTOI [35], SDR を使用した．さらに，合成音声の評価と同一設定の SpeechLMScore [17]

[*2] https://github.com/facebookresearch/encodec
[*3] https://github.com/TencentGameMate/chinese_speech_pretrain
[*4] https://github.com/soumimaiti/speechlmscore_tool

表 2 合成音声に対する主要な評価結果.

	Utterance-level		System-level	
	LCC	SRCC	LCC	SRCC
Traditional reference-aware metrics				
MCD [3]	0.356	0.330	0.541	0.518
Log F0 RMSE	0.050	0.057	0.116	0.123
Unsupervised, without reference				
SpeechLMScore [17]	0.164	0.127	0.268	0.246
Supervised, without reference				
UTMOS [8]	0.363	0.340	0.537	0.575
Proposed (Unsupervised, reference-aware metrics)				
SpeechBERTScore	**0.581**	**0.563**	**0.781**	**0.760**
SpeechBLEU ($G = 2$)	0.427	0.423	0.680	0.659
SpeechTokenDistance (Levenshtein)	0.247	0.210	0.362	0.414
SpeechTokenDistance (Jaro-Winkler)	0.407	0.427	0.663	0.681

表 3 劣化音声に関する主要な評価結果.

	LCC	SRCC
Traditional reference-aware metrics		
PESQ [16]	**0.841**	0.840
STOI [34]	0.741	0.825
ESTOI [35]	0.764	0.826
SDR	0.346	0.741
Unsupervised, without reference		
SpeechLMScore [17]	0.583	0.583
Supervised, without reference		
DNSMOS [7] (BAK)	0.542	0.567
DNSMOS [7] (SIG)	0.595	0.642
DNSMOS [7] (OVRL)	0.674	0.697
Proposed (Unsupervised, reference-aware metrics)		
SpeechBERTScore	0.824	**0.868**
SpeechBLEU ($G = 2$)	0.821	0.827
SpeechTokenDistance (Levenshtein)	0.762	0.800
SpeechTokenDistance (Jaro-Winkler)	0.778	0.777

モデルを，参照不要な教師なし指標として使用した．また，事前訓練された DNSMOS モデル [7] を参照音声不要な教師ありモデルとして使用し，これには signal quality (SIG)，background quality (BAK)，overall quality (OVRL) の 3 つの指標が含まれている．

3.1.4 評価方法

各指標と主観的評価との相関は，線形相関係数 （linear correlation coefficient; LCC） とスピアマンの順位相関係数 （Spearman's rank correlation coefficient; SRCC） を使用して評価した．合成音声の評価では，発話レベルとシステムレベルの指標を定義でき，システムレベルの指標は各 TTS システムに対する発話レベルの指標を平均して計算される．劣化音声については，発話レベルの指標のみを使用した．Log F0 RMSE は低いほど良い結果を示し，SpeechBERTScore は高い方が良い結果を示すため，我々の評価では LCC と SRCC の絶対値を使用した．

3.2 主要な評価結果

まず，§ 3.1.1で述べたように，英語の合成音声に対して評価を行った．SpeechBERTScore には，§ 3.5のモデル別比較で最も良い結果を示した `wavlm-large` モデルを使用した．SpeechBLEU には，§ 3.3の評価結果に基づき，トークンの繰り返しを除いた語彙サイズ 200 のケースを使用した．SpeechTokenDistance には, § 3.3の評価結果に基づき，トークンの繰り返しを除去しない語彙サイズ 200 のケースを使用した．表 2に結果を示す．提案された指標，SpeechBERTScore，SpeechBLEU，および SpeechTokenDistance （Jaro-Winkler） は，すべての基準で従来の参照音声を用いる評価指標を上回った．また，それらの提案手法は，教師なしの SpeechLMScore および教師ありの UTMOS よりも高い相関を示した．これらの結果から，提案された客観的指標が，人間の主観的評価とより高く相関するような合成音声の自動評価が可能であることが確認できる．. 最も高い相関を示した手法は SpeechBERTScore で，発話レベルの SRCC が 0.581，システムレベルの SRCC が 0.760 であった．

また，§ 3.1.1で述べたように，劣化音声に対しても評価を行った．SpeechBERTScore，SpeechBLEU，および SpeechTokenDistance については，表 2の合成音声評価と同じ設定を使用した．表 3に結果を示す．PESQ が最も高い LCC を示した一方，提案手法の SpeechBERTScore は最も高い SRCC を示している．SpeechBERTScore と SpeechBLEU は，PESQ を除いて，すべての比較手法を LCC および SRCC の両方で上回った[*5]．総括すると，提案する評価指標は合成音声の評価において従来の参照音声を使用する指標よりも高い性能を示し，また劣化音声の評価にも有効であった．

表 4 トークンの重複と語彙サイズに関する分析.

Repetition	Vocab.	Utterance-level SpeechBLEU ($G = 2$)		Utterance-level SpeechTokenDistance (Jaro-Winkler)	
		LCC	SRCC	LCC	SRCC
w/ rep	km50	0.341	0.325	0.284	0.275
	km100	0.364	0.346	0.354	0.356
	km200	0.407	0.396	**0.407**	**0.427**
w/o rep	km50	0.357	0.342	0.202	0.202
	km100	0.386	0.369	0.304	0.329
	km200	**0.427**	**0.423**	0.370	0.379

3.3 離散トークンを用いた評価指標に関する分析

提案する音声トークンベースの指標 (§ 2.2と § 2.3) に関する評価を実施した．先行研究 [15, 17] では，ユニット言語モデルの学習時に音声離散トークンの繰り返しを除去する試みがなされている．このような繰り返しの除去はトークンの冗長性や全体の系列長を低減させる一方，音声トークンの継続長情報が無視される．この重複トークン除去の影響について調査するため，音声トークンの繰り返しがある場合 （`w/ rep`） とない場合 （`w/o rep`） を比較した．

[*5] SDR の低い LCC は，その値域の広さによるものである．

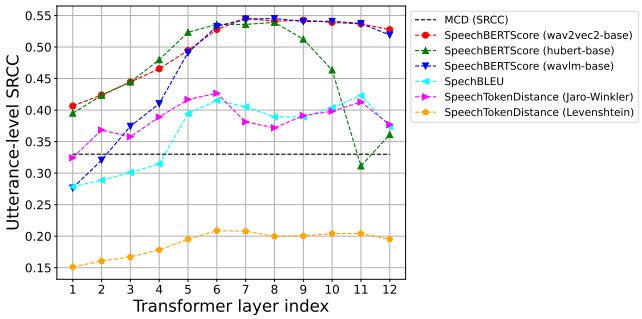

図 2 SSL モデル特徴量の層に関する分析.

また，先行研究 [17] にならい，トークン語彙サイズを 50，100，200 に変化させた．事前実験の結果に基づき，式 (4) の G を 2 に設定した．

表 4に評価結果を示す．SpeechBLEU および SpeechTokenDistance （Jaro-Winkler） は，語彙サイズが大きくなるにつれて相関が高くなることが確認できる．これは，より豊かな音声情報を持つトークンの使用の有効性を示唆している．SpeechBLEU については，`w/o rep` がより良い性能を示した一方で，SpeechTokenDistance には `w/ rep` がより効果的であった．SpeechTokenDistance では文字レベルの類似性を計算するためにトークンの継続長情報が重要である一方，SpeechBLEU ではトークンの繰り返しを除去することが内容の類似性をより良く捉える可能性が示唆される．

3.4 SSL モデル特徴量の層に関する分析

SSL モデル特徴量を取得するためのトランスフォーマー層の選択が，提案された評価指標の性能にどのように影響するかを調査した．図 2は，異なるトランスフォーマー層の特徴を使用した際の，各指標の発話レベル SRCC のプロットを示す．ここで，低い番号は入力に近い層に対応している．結果的に，1 番目または 2 番目の層の特徴量を使用すると，提案された指標の性能が低下することが確認できる．一方で，5 番目以上の層を使用することで，ほとんどのケースで SpeechBERTScore と SpeechBLEU は MCD を上回った．これは，より高い層の特徴が，より多くの意味情報を含んでおり [36]，提案する NLP の評価指標に基づく音声生成の評価指標にとってより効果的であることを示している．さらに，SSL モデル間で異なる傾向が見られ，`wav2vec2-base` と `wavlm-base` は，`hubert-base` と比較して層の選択に関してより頑健であった．

3.5 SSL モデルに関する分析

英語の合成音声について，§ 3.1.2で述べた異なる SSL モデルの比較を行った．表 5に結果を示す．まず，`encodec` は，すべての指標で MCD よりも大幅に悪い結果となった．これは，neural audio codec の表現が音響情報をうまく捉えている一方で，音声における意味情報が SpeechBERTScore への応用において重要であることを示唆している．他のモデルについては，large モデルが base モデルよりも良い結果を示す傾向があったが，Wav2vec 2.0 と HuBERT では差が小さかった．`wavlm-large` はすべての指標で最良の結果を示し，より幅広い音声タスクに向けて設計された WavLM の学習規範の有効性が示唆される．

表 5 SOMOS (英語) についてのモデルごとの比較結果.

	SOMOS (English)			
	Utterance-level		System-level	
	LCC	SRCC	LCC	SRCC
Traditional reference-aware metrics				
MCD	0.356	0.330	0.541	0.518
Proposed reference-aware metrics (SpeechBERTScore)				
`encodec`	0.087	0.074	0.158	0.144
`wav2vec2-base`	0.560	0.539	0.776	0.745
`wav2vec2-large`	0.566	0.547	0.770	0.744
`hubert-base`	0.564	0.545	0.775	0.740
`hubert-large`	0.563	0.548	0.766	0.730
`wavlm-base`	0.559	0.545	0.769	0.739
`wavlm-base+`	0.566	0.551	0.767	0.741
`wavlm-large`	**0.581**	**0.563**	**0.781**	**0.760**

表 6 BC2019 (中国語) についてのモデルごとの比較結果.

	BC2019 (Chinese)			
	Utterance-level		System-level	
	LCC	SRCC	LCC	SRCC
Traditional reference-aware metrics				
MCD	0.156	0.300	0.153	0.362
Proposed reference-aware metrics (SpeechBERTScore)				
`wav2vec2-large`	0.746	0.644	0.834	0.725
`hubert-large`	0.735	0.682	0.867	0.787
`wavlm-large`	0.748	0.654	0.849	0.755
`wav2vec2-large-cmn`	0.753	0.684	0.856	0.785
`hubert-large-cmn`	**0.781**	0.701	0.879	0.819
`xlsr-53`	0.750	**0.706**	**0.904**	0.874
`xlsr-128`	0.742	0.679	0.884	**0.889**

最後に，提案する SpeechBERTScore の言語横断的な適用性について，§ 3.1.1で述べた中国語の BC2019 データセットを用いて調査した．表 6に結果を示す．中国語を含むデータセットで学習されたモデルが，英語音声コーパスのみで学習されたモデルよりも高い性能を示していることが確認できる．しかし，英語コーパスのみで学習されたモデルでも，すべての指標で MCD を上回っており，提案手法の言語横断的な適用性を示している．これは，本研究の提案手法が，当該言語の SSL モデルを持たない言語にも使用できることを示唆している．

4. おわりに

本稿では，客観的なテキスト生成の評価指標に基づく音声生成の自動評価手法を提案した．評価実験により，提案手法が，従来の参照音声を用いた客観評価指標や近年の MOS 予測モデルよりも人間の主観的評価とより高く相関することを示した．また，我々の手法は，言語横断的な適

用性を持ち，高い実用性を持つことを示した．今後の課題としては，MoverScore [10] など，より多くの NLP の評価指標を検討することが挙げられる．

謝辞 本研究の一部は，JSPS 科研費 23H03418, 23K18474, 22H03639, 21H05054, 22KJ0838 ムーンショット研究開発費 JPMJPS2011，および JST FOREST JPMJFR226V によって支援された．

参考文献

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

[2] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," Recommendation, International Telecommunication Union, 1996.

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

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

[5] E. Cooper, W.-C. Huang, T. Toda, et al., "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022, pp. 8442–8446.

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

[7] C. K A Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2021, pp. 6493–6497.

[8] T. Saeki, D. Xin, W. Nakata, et al., "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[9] T. Zhang, V. Kishore, F. Wu, et al., "BERTScore: Evaluating text generation with bert," in *Proc. ICLR*, 2019.

[10] W. Zhao, M. Peyrard, F. Liu, et al., "MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance," in *Proc. EMNLP-IJCNLP*, 2019, pp. 563–578.

[11] K. Papineni, S. Roukos, T. Ward, et al., "BLEU: a method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

[12] A. Baevski, Y. Zhou, A. Mohamed, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12449–12460, 2020.

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

[14] S. Chen, C. Wang, Z. Chen, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[15] K. Lakhotia, E. Kharitonov, W.-N. Hsu, et al., "On generative spoken language modeling from raw audio," *TACL*, vol. 9, pp. 1336–1354, 2021.

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

[17] S. Maiti, Y. Peng, T. Saeki, et al., "SpeechLMScore: Evaluating speech generation using speech language model," *arXiv preprint arXiv:2212.04559*, 2022.

[18] J. Devlin, M.-W. Chang, K. Lee, et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[19] V. I Levenshtein et al., "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, 1966, vol. 10, pp. 707–710.

[20] W. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage," *Proc. Section on Survey Research Methods*, 01 1990.

[21] M. A Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[22] G. Maniati, A. Vioni, N. Ellinas, et al., "SOMOS: The samsung open mos dataset for the evaluation of neural text-to-speech synthesis," *arXiv preprint arXiv:2204.03040*, 2022.

[23] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, 2019, pp. 5891–5895.

[24] K. Ito and L. Johnson, "The LJ Speech Dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

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

[26] Z. Wu, Z. Xie, and S. King, "The blizzard challenge 2019," in *Proc. Blizzard Challenge Workshop*, 2019, vol. 2019.

[27] A. Défossez, J. Copet, G. Synnaeve, et al., "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[28] M. Ott, S. Edunov, A. Baevski, et al., "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. NAACL (Demonstrations)*, 2019, pp. 48–53.

[29] A. Conneau, A. Baevski, R. Collobert, et al., "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[30] V. Panayotov, G. Chen, D. Povey, et al., "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

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

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

[33] W.-C. Huang, E. Cooper, Y. Tsao, et al., "The VoiceMOS Challenge 2022," *arXiv preprint arXiv:2203.11389*, 2022.

[34] C. H Taal, R. C Hendriks, R. Heusdens, et al., "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[35] J. Jensen and C. H Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM TASLP*, vol. 24, no. 11, pp. 2009–2022, 2016.

[36] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ASRU*, 2021, pp. 914–921.