大規模音声自己教師あり学習モデルにもとづく音声好感度予測*

◎須田仁志（産総研），高道慎之介（慶應義塾大／産総研），深山覚（産総研）

1 はじめに

音声の聴感上の好感度は，音声の持つ主観的なパラ言語・非言語情報の 1 つであり，パートナー選択をはじめ人間の社会的活動に大きな影響を与えている [1]．音声とその好感度を対応付けたコーパスの構築とその分析により，種々の音響的特徴が好感度に影響しているほか，聴感上の好感度は聴取者によって大きく異なることが示されている [2]．音声を入力とした好感度予測には，メルスペクトログラムのような音響的特徴のみを表現した特徴量が用いられるが，話者情報や言語的内容を容易に扱える特徴量を活用することで，さらなる予測性能の向上が期待される [3]．

言語内容や話者情報などのラベルを付与せず，大規模な音声データのみから自己教師あり学習（self-supervised learning; SSL）により構築された大規模音声 SSL モデルは，音声認識や話者認識などの種々のダウンストリームタスクに有効な特徴量を出力することが知られている [4]．本稿では，音声 SSL モデルを利用した音声好感度予測手法を提案および評価し，音声好感度の有効な分析手法を明らかにする．

2 音声 SSL モデルにもとづく好感度予測

本節では，音声 SSL モデルから得た特徴量を利用した音声好感度予測手法を提案する．音声 SSL モデルでは，Transformer 中の attention ブロック（Transformer 層）の出力を時系列特徴量として利用できる．提案手法では，各層の出力の重み付き和を特徴量とみなして，時系列方向の self-attention および平均と標準偏差を得る stats pooling 層を介して好感度評点を予測する．聴取者による聴感上の好感度の違いを表現するため，提案法では性別（男性・女性）と年代（30 代以下・40 代以上）にもとづいて聴取者を分類し，各属性による平均好感度評点をそれぞれ単一の数値として予測する．この構成を Fig. 1 に示す．

音声好感度予測は，合成音声の自然性に関する平均オピニオン評点（mean opinion score; MOS）予測に類似した問題である．提案手法では，MOS 予測手法である UTMOS [5] において有効性が示されている対照損失を導入する．提案手法では，次式で示す損失関数 $\mathcal{L}$ を規準に最適化する．

$$\mathcal{L} = \beta \sum_i \mathcal{L}^{\mathrm{reg}}_{x_i} + \gamma \sum_i \mathcal{L}^{\mathrm{con}}_{x_i,x_j} \tag{1}$$

式 (1) におけるクリップ付き二乗誤差 $\mathcal{L}^{\mathrm{reg}}$ [6] および対照損失 $\mathcal{L}^{\mathrm{con}}$ は次式で定義される．

$$\mathcal{L}^{\mathrm{reg}}_{x_i} = \mathbb{1}\left[|y_i - \hat{y}_i| > \tau\right](y_i - \hat{y}_i)^2 \tag{2}$$

$$\mathcal{L}^{\mathrm{con}}_{x_i,x_j} = \max\left(0, |(y_i - y_j) - (\hat{y}_i - \hat{y}_j)| - \alpha\right) \tag{3}$$

ここで，x_i，y_i，$\hat{y}_i$ はそれぞれミニバッチ内で i 番目の音声，正解評点，予測評点を表し，j は $i \neq j$ なる無作為に選ばれたインデックスを表す．また，α，β，γ，τ はハイパーパラメータである．式 (2) において，$\mathbb{1}[\cdot]$ は，条件が満された場合に 1，満たされない場合に 0 となる関数である．提案手法では各聴取者属性について上述の損失関数を計算し，この総和を最終的な損失関数として利用する．なお，式 (1) 第 2 項の対照損失項について，UTMOS においてはミニバッチ内のすべての組み合わせに対する総和（$\sum_{i\neq j} \mathcal{L}^{\mathrm{con}}_{x_i,x_j}$）が採用されているが，提案手法では各音声と無作為に選択した異なる単一音声との対照損失を採用する．

3 実験条件

音声好感度コーパスには CocoNut-Humoresque [2] を採用した．このコーパスは，Coco-Nut [7] 中の 1800 音声に対する，それぞれ 11 人以上の聴取者による主観的な音声好感度評点が含まれる．聴取者を性別（男性・女性）と年代（30 代以下・40 代以上）にもとづき 4 グループに分割し，各聴取者グループ内での平均評点を正解評点とした．評点は $[-1, 1]$ に線形に正規化し，評点が存在しないグループについては 0 を正解評点とした．学習，検証，評価セットは CocoNut-Humoresque の分割をそのまま採用し，それぞれ 1200，300，300 音声からなる．各音声は約 4 秒で，標本化周波数が 44 100 Hz のステレオ音声である．学習時および評価時には，無作為にチャンネルを選択した．

音声 SSL モデルとして，日本語音声で学習された HuBERT モデル `imprt/kushinada-hubert-large`（くしなだ，3 億パラメータ）[8] および Wav2vec 2.0 モデル `imprt/izanami-wav2vec2-large`（いざなみ，3 億パラメータ）[9]，複数言語で学習された Wav2vec 2.0 モデル XLS-R [10] のうち `facebook/wav2vec2-xls-r-300m`（XLS-R 300M，3 億パラメータ）および `facebook/wav2vec2-xls-r-2b`（XLS-R 2B，20 億パラメータ）を利用した．入力信号の標本化周波数は 16 000 Hz，フレーム周期は 20 ms である．特徴量には，各 Transformer 層の出力に加え前段の畳み込み層の出力を用い，XLS-R 2B では 49 層，そのほかのモデルでは 25 層の出力の重み付き和を利用した．これらの音声 SSL モデルをそれぞれ利用した場合に加え，くしなだといざなみから得られた 50 層の特徴量の重み付き和を利用した予測モデルを構築した．

比較手法として，対数メルスペクトログラムを入力とし，時間遅れニューラルネットワーク（time-delay neural network; TDNN）を利用した音声好感度予測モデルを構築した．入力信号の標本化周波数は 22 050 Hz，フレーム周期は 256 サンプル，メル周波数軸上のビン数は 0 Hz から 8000 Hz までの 80 ビンである．損失関数は音声 SSL モデルを利用した場合と同一である．

評価指標には，平均二乗誤差（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's Tau coefficient; KTAU）を用いた．

*Voice likability prediction using large-scale self-supervised speech models. by SUDA, Hitoshi (AIST), TAKAMICHI, Shinnosuke (Keio University/AIST), and FUKAYAMA, Satoru (AIST)

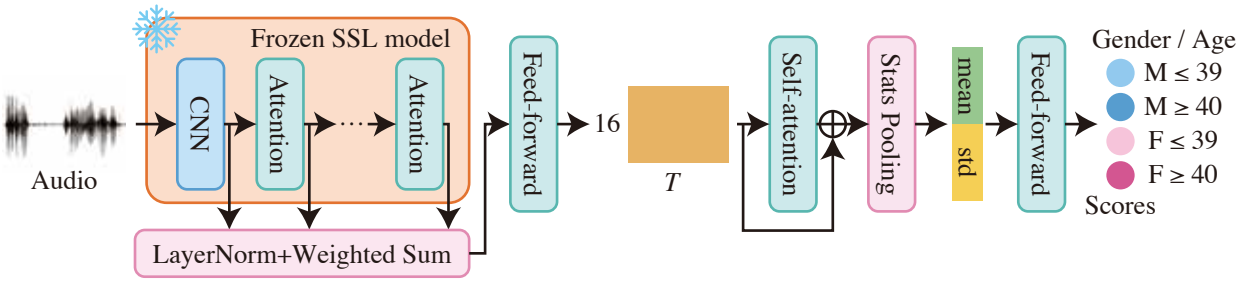

Fig. 1　音声 SSL モデルを用いた好感度予測モデルの構成．Transformer 層の出力の重み付き和を特徴量として，self-attention と stats pooling 層を介し，性別と年代にしたがった聴取者の属性ごとに評点を予測する．

Table 1　CocoNut-Humoresque の評価セットに対する，異なる特徴量間での音声好感度予測性能の比較．太字は各評価指標で最も優れたシステムを示す．

特徴量	MSE↓	LCC↑	SRCC↑	KTAU↑
くしなだ（HuBERT）	**0.077**	**0.611**	0.644	0.466
w/o 二乗誤差 $\mathcal{L}^{\text{reg}}$	0.236	0.326	0.362	0.256
w/o 対照損失 $\mathcal{L}^{\text{con}}$	0.079	0.601	0.636	0.460
いざなみ（Wav2vec2）	0.107	0.493	0.517	0.364
くしなだ+いざなみ	0.078	0.609	**0.645**	**0.466**
XLS-R 300M	0.091	0.541	0.557	0.398
XLS-R 2B	0.101	0.464	0.496	0.349
Log Mel-spectrogram	0.098	0.468	0.509	0.359

モデルは検証セットに対する最大 SRCC 規準で選択した．式 (1)–(3) 中のハイパーパラメータは，$\tau = 0.1$，$\alpha = 0.1$，$\beta = 1$，$\gamma = 0.5$ とした．学習時のデータ拡張として，学習音声に対して ±12 dB 以内の音量の変更と位相反転を無作為に行った．

4　実験結果

CocoNut-Humoresque の評価セットに対する音声好感度予測性能を Table 1 に示す．MSE と LCC に関してはくしなだを利用した場合，SRCC と KTAU に関してはくしなだといざなみを同時に利用した場合で，最も高い予測性能が示された．くしなだといざなみを同時に利用した場合，くしなだのみを用いた場合と比較して予測モデル全体のパラメータ数がおよそ 2 倍になるため，SRCC や KTAU の向上の程度を考えれば，実用上はくしなだのみの利用が有効である．

くしなだを利用した場合，対照損失の導入によりすべての評価指標において改善し，UTMOS と同様に好感度予測においても対照損失の有効性が示された．

日本語音声のみで学習されたくしなだと，多言語音声で学習された XLS-R を比較すれば，好感度予測においては前者を利用したモデルが有効であった．そのため，音声好感度の言語依存性が示唆された．

メルスペクトログラムを特徴量として利用した場合と比較して，くしなだを利用した場合はすべての評価指標において予測性能が高く，音声好感度予測における音声 SSL モデルの有効性が示唆された．

音声 SSL モデルにおいては，出力特徴量における支配的な音声情報が Transformer 層によって異なり，目的タスクによって有効な特徴量が異なることが知られている [11]．くしなだを用いた好感度予測モデルにおいて，各層の出力の重み付き和を計算する際の，重みを分析した結果を Fig. 2 に示す．話者認識や音素認識の場合と異なり，好感度予測においては全層にわたって重みの大きな差は見られず，上層と下層の特徴量がともに利用されている．そのため，好感度予測においては，話者情報や言語情報にわたる多様な音声情報を同時に扱う手法が有効である．これは，話者情報を主とするメルスペクトログラムを利用したモデルとの予測性能の差からも推測される．

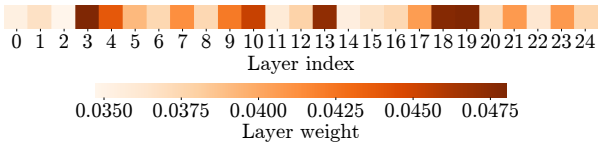

Fig. 2　くしなだを用いた音声好感度予測モデルにおいて，Transformer 層出力の重み付き和を得る際に適用される重み．横軸は層のインデックスを表す．

5　おわりに

本研究では，音声 SSL モデルから抽出された特徴量を利用した，音声好感度予測モデルを構築し，その有効性を評価した．好感度予測における対照損失の有効性が示されたほか，話者情報と言語情報を同時に獲得できる音声 SSL モデルの有効性が示された．さらに，音声好感度における言語依存性が示唆され，主観的なパラ言語・非言語情報における発話言語の事前情報の重要性が示唆された．

謝辞　本研究は，産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」の支援を受けて実施した．

参考文献

[1] Weiss *et al.*, Eds., Springer, 2021.
[2] Suda *et al.*, In Proc. Interspeech, 2024.
[3] Suda *et al.*, In Proc. Interspeech, 2025.
[4] Yang *et al.*, In Proc. Interspeech, 2021.
[5] Saeki *et al.*, In Proc. Interspeech, 2022.
[6] Leng *et al.*, In Proc. ICASSP, 2021.
[7] Watanabe *et al.*, In Proc. ASRU, 2023.
[8] 瀧澤，他，音講論（春），2025.
[9] 瀧澤，他，音講論（秋），2023.
[10] Babu *et al.*, In Proc. Interspeech, 2022.
[11] Chen *et al.*, JSTSP, 16 (6), 1505–1508, 2022.