自己教師ありモデル特徴量から音声波形を生成するニューラルボコーダの実験的評価*

◎中田 亘，佐伯 高明，齋藤 佑樹，高道 慎之介，猿渡 洋（東大院・情報理工）

1 はじめに

近年，自己教師あり学習 (Self-Supervised Learning: SSL) を用いて事前学習された SSL モデル [1–3] が，音声認識 [1], 音声変換 [4]，音声合成 [5]，音声強調 [6]，音声翻訳 [7] などの音声情報処理の幅広い分野において有用であると報告されている．SSL モデルを用いた音声情報処理では，SSL モデル全体を所望のタスクにファインチューニングするか，SSL モデルを用いて音声から得られる特徴量を用いて所望のタスクを解くための機械学習モデルを構築する．しかしながら，それぞれの SSL モデルは異なる事前学習規範により得られており，この SSL モデルの差異により下流タスクの性能が変わることが報告されている [3].

そのため，様々な音声言語タスクごとに SSL モデル特徴量に関する調査が行われてきた．先行研究 [8, 9] では，Canonical Correlation Analysis [10] を用いて SSL モデルの各層出力にどのような情報が含まれているか調べている．また，先行研究 [3] では，各 SSL モデル特徴量の各隠れ層の重み付き和を用いて SUPERB ベンチマーク [11] の下流タスクを解くことにより，タスクにより重要な特徴量を保持する隠れ層が異なることを明らかにしている．これら既存の多くの調査が対象とするタスクとは異なり，音声変換やテキスト音声合成などの合成タスクへの応用では，SSL モデルから高品質な音声波形を合成することが求められる．この SSL 特徴量からの音声波形合成に関する研究も行われているものの [12]，近年のデファクトスタンダードな SSL モデルを包括的に議論した先行研究は限定的である．そこで，本研究では，SSL モデル特徴量から音声波形を再合成するニューラルボコーダを学習し，合成音声の客観評価と主観評価を通じて，複数の SSL モデル及び特徴量抽出のための層の選択が合成音声品質に与える影響を包括的に調査する．調査結果より，一部の条件ではメルスペクトログラムとくらべ SSL モデル特徴量を中間特徴量として使用した場合に音声の自然性が有意に改善することが確認された．多くの条件ではメルスペクトログラムの方が高い性能を示したものの一部の SSL モデル特徴量がメルスペクトログラムを上回る性能を示すことが確認された．なお，調査に使用したコード及び学習済みモデルは Github 上で公開している[1].

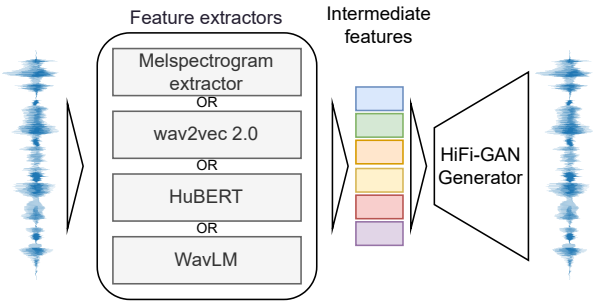

Fig. 1: SSL モデル特徴量及び，メルスペクトログラムから HiFi-GAN を用いて行う音声再合成

2 実験

本研究では図 1に示すようにメルスペクトログラムから音声を再構成する一般的なニューラルボコーダ (*mel*) に加えて，SSL モデルの各隠れ層出力による音声の変化を調査するために，SSL モデルとして広く使われている wav2vec2-base の第 {3, 6, 9,} 層目隠れ層出力 (wav2vec2-base-l{3, 6, 9}) 及び，最終層出力 (wav2vec2-base) を用いた音声の再構成を行った．wav2vec2-base は全部で 12 層隠れ層をもつため，{3, 6, 9} や最終層はそれぞれ全体の 1/4, 1/2, 3/4, 1 に対応している．その後，SSL モデルごとの差異を確認するために，{wav2vec2,hubert,wavlm}-{base,large} の最終層出力と base モデルでは第 3 層目出力 ({SSL モデル名 }-base-l3)，large モデルでは第 6 層目出力を用いた音声の再構成（{SSL モデル名 }-large-l6）を行った．全部で 15 モデルの学習を行った．

2.1 実験条件

データセットには，LibriTTS [13] の train-clean100, train-clean-360 サブセット及び VCTK-Corpus [14], LJSpeech [15] を使用した．それぞれ，1151 名（245 時間），109 名（44 時間），1 名（25 時間）の音声コーパスであり，合計 314 時間の音声で学習を行った．各音声は 22.05[kHz] にリサンプリングを行った．*mel* の入力特徴量や各モデルの損失関数で使用されているメルスペクトログラムは，フレーム長を 1024 サンプル，フレームシフトを 256 サンプル，次元数を 80 次元として生成した．

---

*An Empirical Study of Self-Supervised Learning Model Features for Speech Waveform Reconstruction
NAKATA, Wataru, SAEKI, Takaaki, SAITO, Yuki, TAKAMICHI, Shinnosuke, SARUWATARI, Hiroshi (The University of Tokyo).

使用したニューラルボコーダの基本構造は HiFi-GAN [16] を使用した．メルスペクトログラムと SSL モデル特徴量の異なる周期の特徴量を入力とするため，メルスペクトログラムを使用したモデルでは，$\frac{22050}{256}$ 倍のアップサンプルを実現するために転置畳み込み層のストライドを入力から近い順に $7, 7, 3, 3$ とした．SSL モデル特徴量を使用した場合には $8, 8, 2, 2$ とし，50 倍のアップサンプルを行い音声波形を生成した．

最適化手法には Generator 及び Discriminator 双方において AdamW($\mathrm{lr} = 2\times10^{-3}, \beta_1 = 0.8, \beta_2 = 0.99$) を使用した．また，学習率は毎反復ごとに 0.999998 倍にし，指数的に減少させた．

2.2 評価指標

評価指標には客観評価として 3 つ，主観評価として 2 つを使用した．

2.2.1 客観評価

客観評価は，後述する指標をもとに，各モデルから合成された音声と原音声を比較することで，再合成音声品質を評価した．評価には，CMU ARCTIC [17] の各話者 100 発話 (arctic) , JVS Corpus [18] の parallel100 サブセット (jvs)，JNV Corpus [19] (jnv), Laughterscape Corpus [20] (laughterscape), PNL100 Nonspeech Sounds [21] (pnl) を使用した．最初のコーパスは学習データと同じく英語読み上げであり，次いで，日本語読み上げ，日本語非言語音声，日本語笑い声，非音声のコーパスである．

Mel-Cepstral Distorion　原音声に近い音声が合成できていることを確認するために，Mel-Cepstral Distortion (MCD) を使用した．MCD の計算に際しては，FastDTW [22] を用いて原音声と合成音声の系列アライメントを取得したのちに，計算した．

Log-F0 RMSE　原音声と合成音声の F0 の違いを評価するために Log-F0RMSE (Root Mean Squared Error) を使用した．F0 の抽出には，WORLD ボコーダ [23] を使用し，原音声から抽出された F0 と合成音声から抽出された F0 の RMSE を計算した．

X-vector コサイン類似度　原音声と合成音声の話者性の変化を確認するために，話者認証で使用されている x-vector [24] のコサイン類似度を使用した．x-vector 抽出には Github で公開されている JTube-Speech [25] で学習されたモデル[2] を使用した．

2.2.2 主観評価

Mean Opinion Score (MOS) に基づく主観評価では JVS コーパス [18] を使用し，JVS コーパスの男性話者 5 名，女性話者 5 名の計 10 名を評価に使用した．話者の選択には，JVS コーパスに付随する F0 のレンジをもとに男女別に k 平均法を用いて話者のクラスタリングを行い，各クラスタの重心に近い話者を選択した．最終的に選択された話者は，jvs{001, 005, 014, 017, 021, 022, 024, 035, 038, 047} である．

Naturalness MOS　合成音声の自然性を評価するために，自然性 MOS を評価した．評価者数は 60 人とし，各評価者は 20 サンプルに対して 5 段階 (1: とても低い–5：とても高い）の自然性評価を行った．

Speaker similarity MOS　合成音声においてどれだけ原音声に近い話者性が実現されているか評価するために，話者類似度 MOS を評価した．自然性 MOS 同様，評価者数は 60 人とし，各評価者は 20 サンプルの合成音声と，対応する原音声を提示された上で，どれだけ合成音声の話者性が原音声に似ているかを 5 段階（1: とても異なる–5: 非常に似ている）で評価した．

3 結果と考察

3.1 使用する隠れ層出力による合成音声の変化

図 2に使用する隠れ層出力による音声の変化に関する，客観評価結果を示す．いずれの場合も mel が最良の評価結果となっており，中間特徴量としてメルスペクトログラムが非常に有効であることがわかる．また評価データセットに関しては，arctic, jvs, jnv, laughterscape, pnl の順に悪くなっていることがわかる．これは，学習データとして使用した多話者英語音声（LibriTTS, LJSpeech, VCTK Corpus) との乖離が大きくなるためと考えられる．隠れ層数に関しては，浅い層においていずれの評価指標においても良い結果となった．一方で，x-vector コサイン類似度の結果では，層数が増えるほど mel よりも話者類似度が低い結果になるものの，評価データセットのうち音声を使用したものでは，非常に高い類似度が確認された．このことから，SSL モデル特徴量には話者性に関する情報が多分に含まれることがわかる．

図 4,5にそれぞれ自然性，話者類似度に関する主観評価結果を示す．自然性に関しては，mel と wav2vec2-base-l3 の間で有意な差が見られなかった．このことから，wav2vec2-base の第 3 層目出力には，メルスペクトログラムと同程度に自然生の高い音声再合成が可能と考えられる．また，隠れ層出力が入力に近い程，自然性が改善する傾向が見られた．

話者類似度に関しては，mel が有意に他のモデルよりも良い結果となった．このことから wav2vec2-base-l3 には高い自然性を実現するための情報は含まれているが，話者性が欠落していると考えられる．使用した隠れ層出力に関しては，自然性同様入力に近い

[1] https://github.com/Wataru-Nakata/ssl-vocoder
[2] https://github.com/sarulab-speech/xvector_jtubespeech

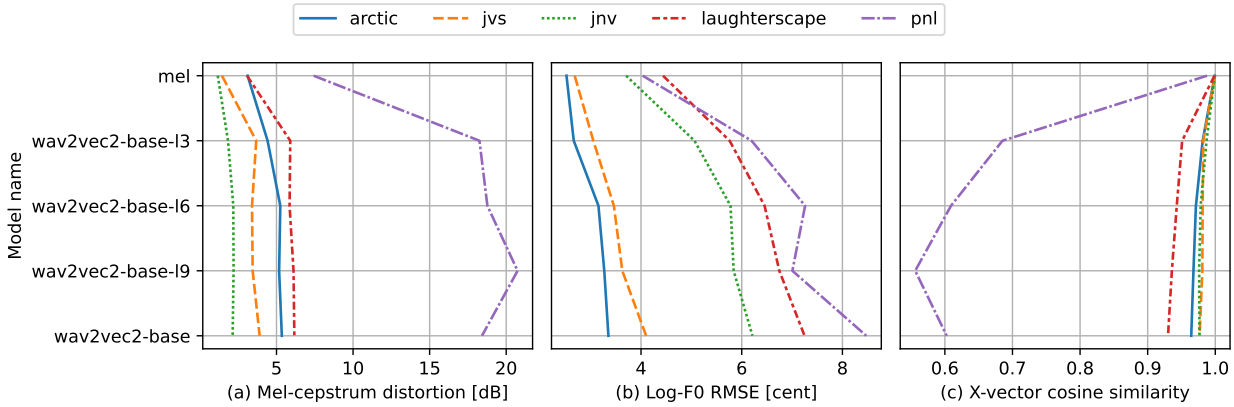

Fig. 2: 隠れ層ごとの比較の客観評価結果

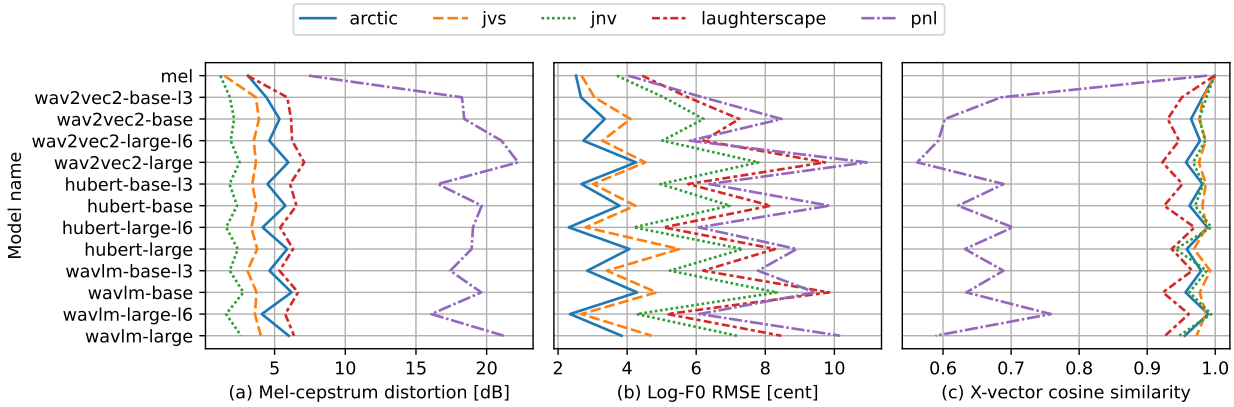

Fig. 3: SSL モデルごとの比較の客観評価結果

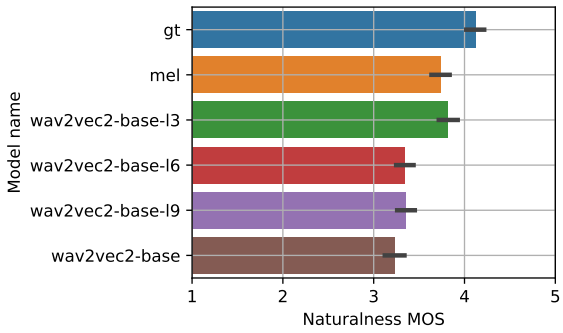

Fig. 4: 隠れ層ごとの比較の自然性 MOS 主観評価結果 エラーバーは 95%信頼区間を示す.

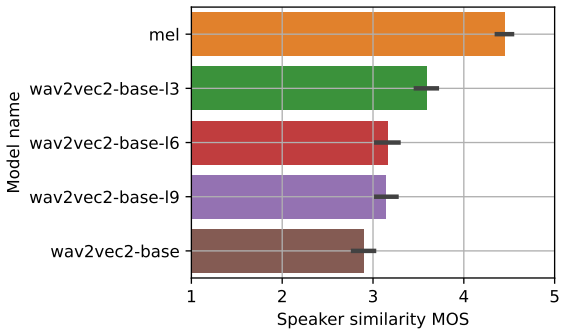

Fig. 5: 隠れ層ごとの比較の話者類似度 MOS 主観評価結果 エラーバーは 95%信頼区間を示す.

ほど，話者類似度が改善する結果が見られた．以上のことから，SSL モデルの隠れ層が入力に近いほど，より音声にの再合成に必要な音響情報を含んでいると考えられる．

3.2 モデル変化による合成音声の変化

図 3に，使用するモデルごとの音声の変化に関する客観評価結果を示す．こちらも図 2同様に，mel が一番原音声に近い結果となっている．また，いずれのモデルにおいてもより入力に近い層から得られる特徴量から合成された音声がより原音声に近い傾向が確認された．

図 6,7にそれぞれ自然性，話者類似度に関する主観評価結果を示す．自然性に関しては，wavlm-large-l6, hubert-large-l6 において原音声との有意差が確認されず，また mel と比べて有意に改善している結果となった．このことから，これらの SSL モデルの第 6 層目出力には，原音声の自然性に関する情報が多く含まれていると考えられる．一方で話者類似度に関しては，いずれの SSL モデル特徴量よりも mel が有意に良い結果となった．このことから，wavlm-large-l6 や hubert-large-l6 には自然性に関する情報はメルスペクトログラムより多く含まれているものの，話者性に関する情報は多く含まれていないことが確認された．

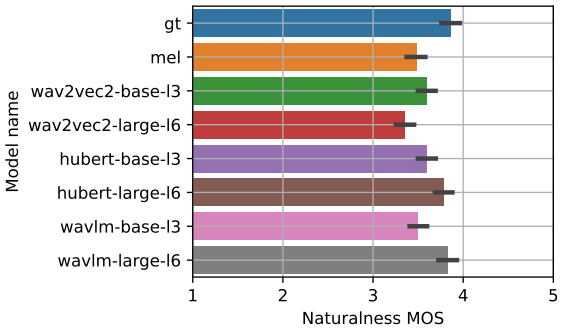

Fig. 6: SSL モデルごとの比較の自然性 MOS 主観評価結果 エラーバーは 95%信頼区間を示す.

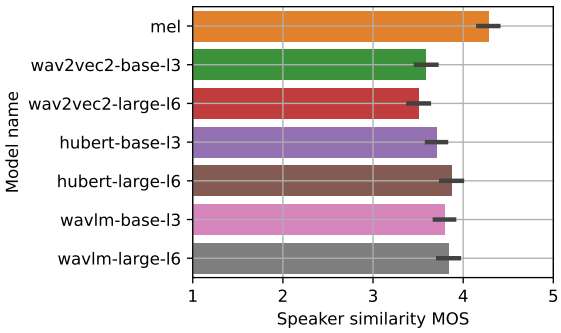

Fig. 7: SSL モデルごとの比較の話者類似度 MOS 主観評価結果 エラーバーは 95%信頼区間を示す.

4 まとめ

本研究は，自己教師ありモデル特徴量に含まれている情報の理解を目的とし，SSL モデル特徴量から音声を再構成することにより，SSL モデル特徴量にどのような情報が含まれているか調査した．結果から，自然性に関してはいくつかのモデルや隠れ層数において多くの情報が含まれているものの話者性に関する情報は欠落していることが確認された.

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.

参考文献

[1] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv preprint arXiv:2006.11477*, 2020.

[2] W.-N. Hsu et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *arXiv preprint arXiv:2106.07447*, 2021.

[3] S. Chen et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *arXiv preprint arXiv:2110.13900*, 2021.

[4] W.-C. Huang et al., “Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations,” *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5944–5948, 2020.

[5] H. Siuzdak et al., “Wavthruvec: Latent speech representation as intermediate features for neural speech synthesis,” in *Interspeech*, 2022.

[6] H. Song et al., “Exploring wavlm on speech enhancement,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 451–457.

[7] A. Lee et al., “Direct speech-to-speech translation with discrete units,” *arXiv preprint arXiv:2107.05604*, 2021.

[8] A. Pasad et al., “Layer-wise analysis of a self-supervised speech representation model,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[9] ——, “Comparative layer-wise analysis of self-supervised speech models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936. [Online]. Available: http://www.jstor.org/stable/2333955

[11] S. wen Yang et al., “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[12] A. Polyak et al., “Speech resynthesis from discrete disentangled self-supervised representations,” *arXiv preprint arXiv:2104.00355*, 2021.

[13] H. Zen et al., “LibriTTS: A corpus derived from librispeech for text-to-speech,” *ArXiv*, vol. abs/1904.02882, 2019.

[14] J. Yamagishi et al., “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” 2019.

[15] K. Ito, L. Johnson, “The LJ Speech Dataset,” https://keithito.com/LJ-Speech-Dataset/, 2017.

[16] J. Kong et al., “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *arXiv preprint arXiv:2010.05646*, 2020.

[17] J. Kominek, A. W. Black, “The cmu arctic speech databases,” in *Speech Synthesis Workshop*, 2004.

[18] S. Takamichi et al., “JSUT and JVS: Free japanese voice corpora for accelerating speech synthesis research,” *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[19] D. Xin et al., “JNV Corpus: A corpus of japanese nonverbal vocalizations with diverse phrases and emotions,” 2023.

[20] ——, “Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus,” *arXiv preprint arXiv:2305.12442*, 2023.

[21] G. Hu, D. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[22] S. Salvador, P. K. Chan, “Toward accurate dynamic time warping in linear time and space,” *Intell. Data Anal.*, vol. 11, pp. 561–580, 2007.

[23] M. Morise et al., “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[24] D. Snyder et al., “X-vectors: Robust dnn embeddings for speaker recognition,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[25] S. Takamichi et al., “Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification,” *CoRR*, vol. abs/2112.09323, 2021. [Online]. Available: https://arxiv.org/abs/2112.09323