2-Q-29

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)

学習済みモデル
学習用コード
推論デモ
公開中❤️

研究背景：自己教師あり学習（SSL）モデルを用いた音声情報処理の隆盛

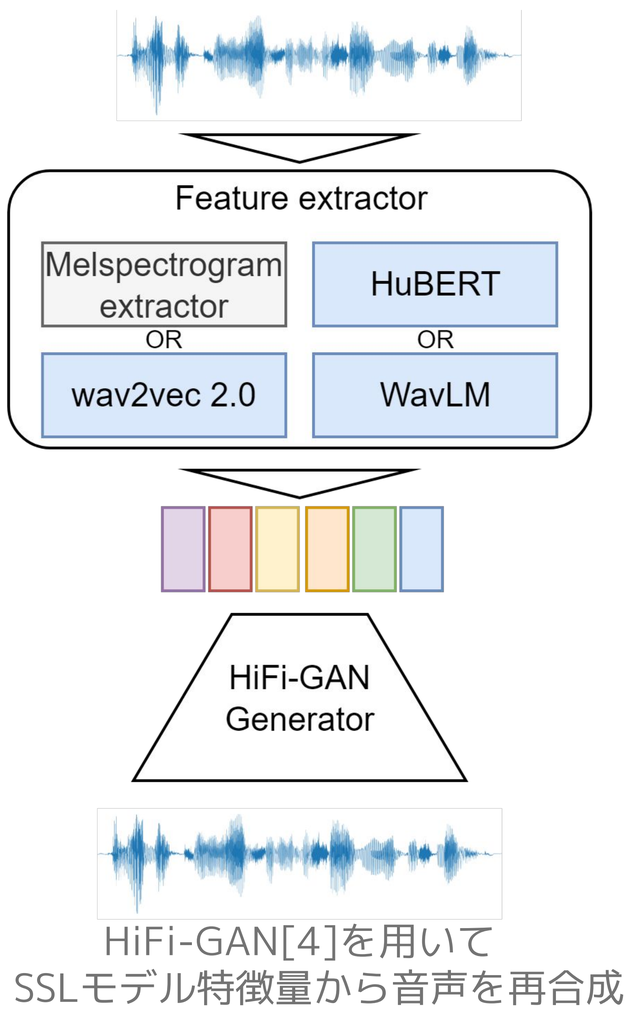
- **自己教師あり学習（SSL）モデル**
  - 画像、言語のみならず、音声においても有効
  - 大量の音声データを用いて事前学習
  - 幅広い音声処理タスクに対して有効
  - 使用する隠れ層の数やSSLモデルの種類により所望タスクにおける性能が変化

所望のタスクに対して適したSSLモデルを選択するためにSSLモデル特徴量の理解が必要

- **先行研究**
  - Canonical Corelation Analysisを用いたSSLモデルの解析[1]
  - 各隠れ層の重み付き和を用いてタスク事の適切な層を調査[2]
  - SSLモデル特徴量を用いた話者表現を調査[3]
- **本研究**
  - SSLモデル特徴量に音声合成に必要な音響的特徴がどの程度含まれているかを理解するために，再合成音声品質を調査
  - 様々な音声ドメインに対して、複数のSSLモデルの異なる隠れ層から再合成した音声を比較

実験条件

使用したボコーダ



HiFi-GAN[4]を用いて
SSLモデル特徴量から音声を再合成

使用したデータセット

学習用データセット

データセット名	話者数	時間
LibriTTS[5] train-clean-100 train-clean-360	1151名	245時間
VCTK-Corpus[6]	109名	44時間
LJSpeech[7]	1名	25時間

評価用データセット

データセット名	概要
CMU ARCTIC[8] (arctic)	英語音声
JVS corpus[9] (jvs)	日本語音声
JNV corpus[10] (jnv)	日本語 非言語音声
Laughterscape corpus[11] (laughterscape)	日本語　笑い声音声
PNL100 Nonspeech Sounds[12] (pnl)	非音声

評価指標

客観評価
- Mel-cepstrum distortion (MCD)
- Log F0 RMSE
- X-vector cosine similarity

主観評価
- Naturalness mean opinion score (MOS)
- Speaker similarity MOS

学習したモデル

近年盛んに使用されている
- wav2vec 2.0[13]
- HuBERT[14]
- WavLM[2]

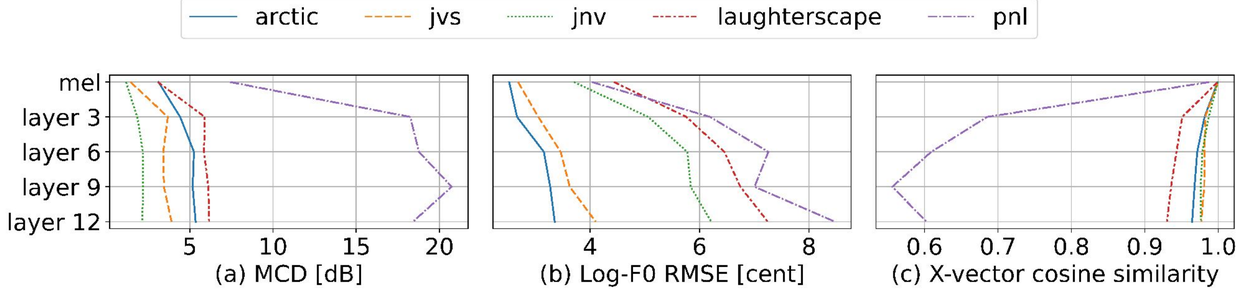
のbaseモデル（12層）largeモデル（24層）から得た特徴量から音声を再合成
全15モデルを学習

主要な評価結果（他の結果は予稿を参照）

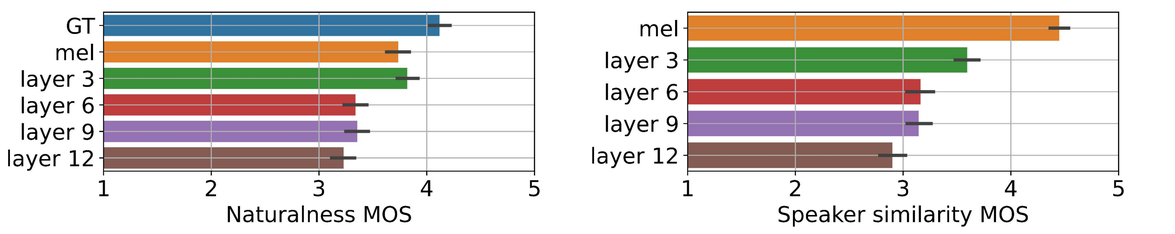
使用する隠れ層に関する評価

wav2vec 2.0-baseモデルに関して第3,6,9,12層を比較

客観評価結果



主観評価結果
評価者数60人　一人当たり20サンプルを評価



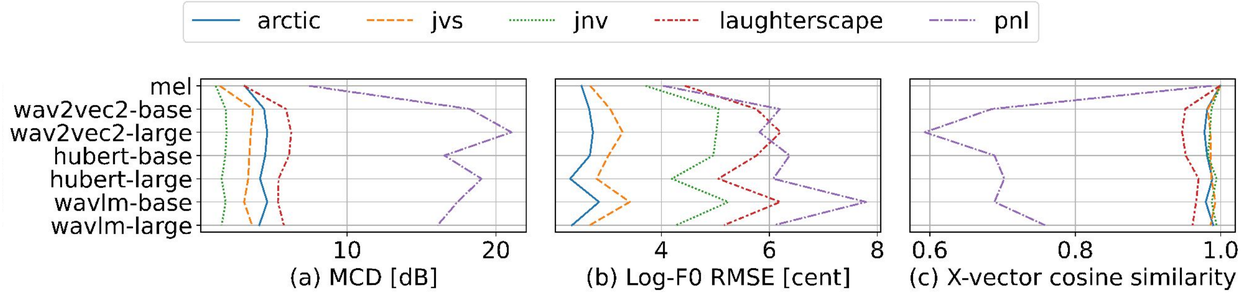
エラーバー: 95%信頼区間

- メルスペクトログラムを使用したモデルの高い汎化性能を確認
- より入力に近い隠れ層において高い再構成性能
- 自然性と比べ、話者性ではmelとの乖離が大きい

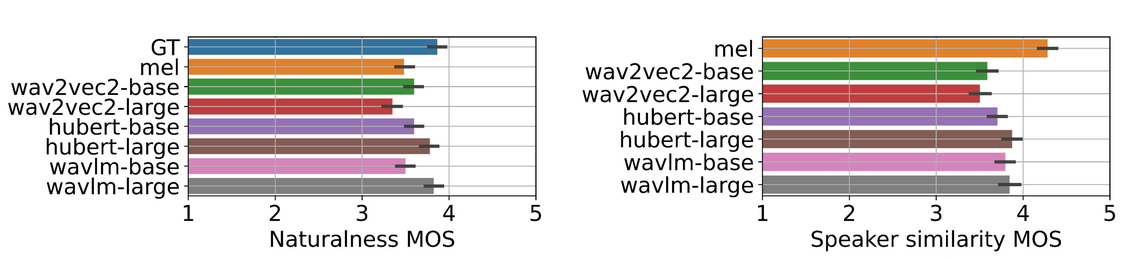
SSLモデルの種類に関する評価

wav2vec 2.0, HuBERT, WavLMの3種類のSSLモデルを比較
baseモデルでは第3層、largeモデルでは第6層特徴量を使用

客観評価結果



主観評価結果
評価者数60人　一人当たり20サンプルを評価



エラーバー: 95%信頼区間

- hubert-large, wavlm-largeがmelより高いMOS値
- 話者類似度に関しては、メルスペクトログラムより劣る

まとめ

- より浅い隠れ層に置いて音声再合成に適した表現が獲得可能
- 一部の条件では、メルスペクトログラムより高い自然性

今後の予定

- SSLモデルを用いたテキスト音声合成、音声変換への応用
- SSLモデル特徴量に特化したボコーダの検討

References

[1] A. Pasad et al., Proc. ASRU, 2021
[2] S. Chen et al.,JSTSP, 2021
[3] K. Fujita et al., Proc. ICASSP, 2023
[4] J. Kong et al., Proc. NeurIPS, 2020
[5] H. Zen et al., Proc. Interspeech, 2019
[6] J. Yamagishi et al., CSTR, 2019
[7] K. Ito et al., 2017
[8] J. Kominek et al., Proc. SSW, 2004
[9] S. Takamichi et al., JST, 2020
[10] D. Xin et al., 2023
[11] D. Xin et al., arXiv, 2023
[12] G. Hu et al., IEEE, 2023
[13] Baevski A. et al., Proc. NeurIPS, 2023
[14] Hsu, W. et al., IEEE TASLP, 2021

2-Q-29

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)

研究背景：SSLモデル特徴量の理解が求められている

- **自己教師あり学習（SSL）モデル**
  - 大量の音声データを用いて事前学習
  - 幅広い音声情報処理に対して有効
  - 使用する隠れ層や事前学習規範により所望タスクにおける性能が変化

所望のタスクに対して適したSSLモデル選択するためにSSLモデル特徴量の理解が必要

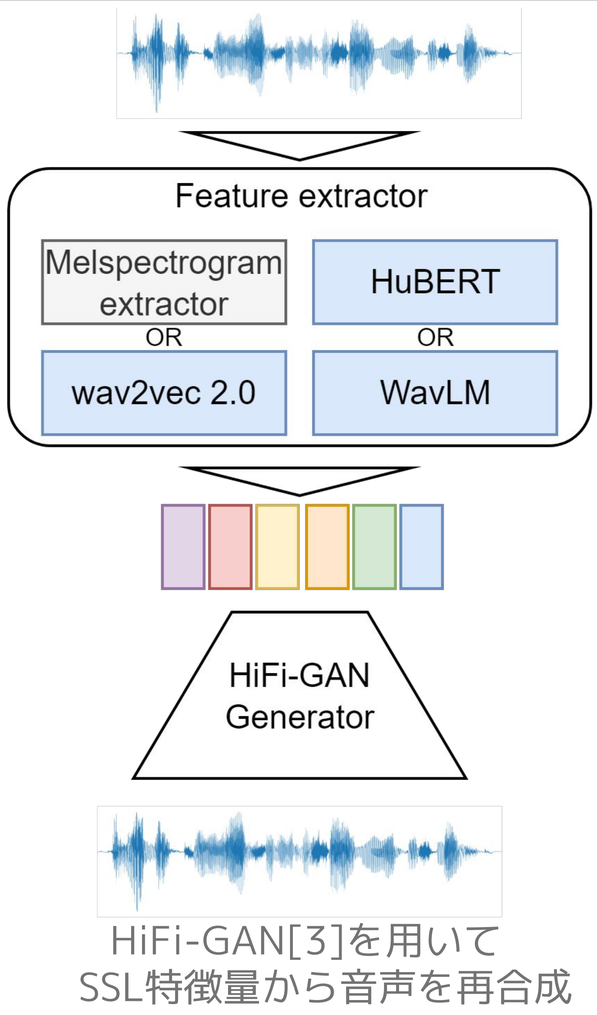
- **先行研究**
  - Canonical Corelation Analysisを用いたSSLモデルの解析[1]
  - 角隠れ層の重み付き和を用いてタスク事の適切な層を調査[2]

- **本研究**
  - テキスト音声合成、音声変換で重要となる再合成音声の品質を調査
  - 様々な音声ドメインに対する再合成音声を比較

オープンソースです❤️

実験条件

使用したボコーダ



HiFi-GAN[3]を用いてSSL特徴量から音声を再合成

使用したデータセット

学習用データセット

データセット名	話者数	時間
LibriTTS[4] train-clean-100 train-clean-360	1151名	245時間
VCTK-Corpus[5]	109名	44時間
LJSpeech[6]	1名	25時間

評価用データセット

データセット名	概要
CMU ARCTIC[7] (arctic)	英語音声
JVS corpus[8] (jvs)	日本語音声
JNV corpus[9] (jnv)	日本語 非言語音声
Laughterscape corpus[10] (laughterscape)	日本語　笑い声音声
PNL100 Nonspeech Sounds[11] (pnl)	非音声

評価指標

客観評価
- Mel-cepstrum distortion
- Log F0 RMSE
- X-vector cosine similarity

主観評価
- Naturalness MOS
- Speaker similarity MOS

学習したモデル

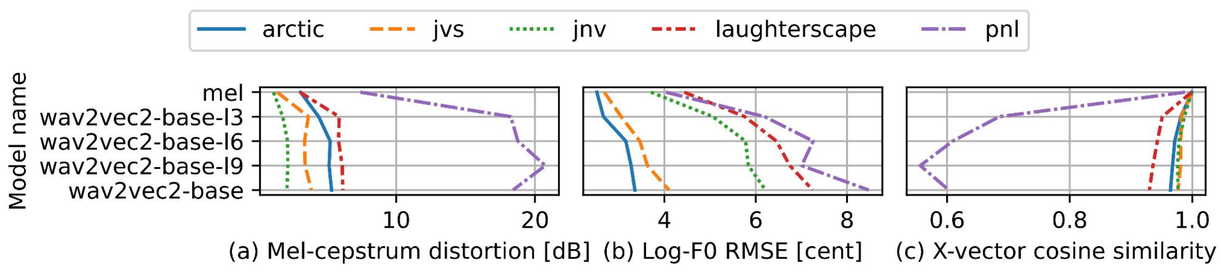
全15モデルを学習
命名規則
{SSLモデルの種類}-{SSLモデルのサイズ}-l{隠れ層数}

例: wav2vec2-base-l6
wav2vec2のbaseモデル第6層出力で学習

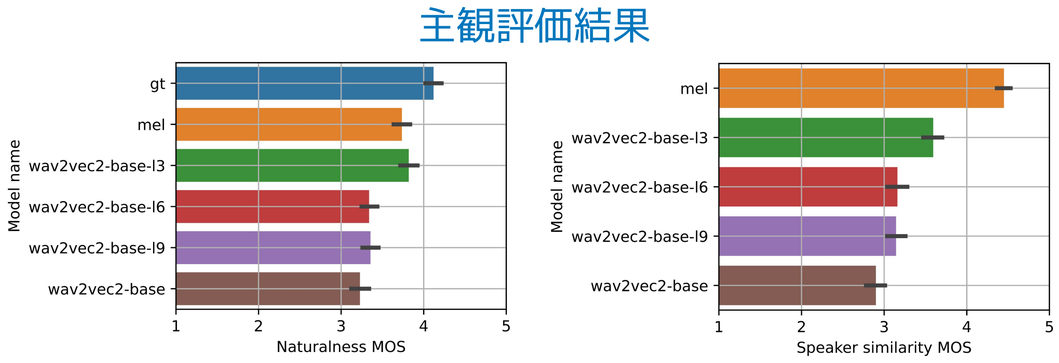
評価結果

使用する隠れ層出力による合成音声の変化

客観評価結果



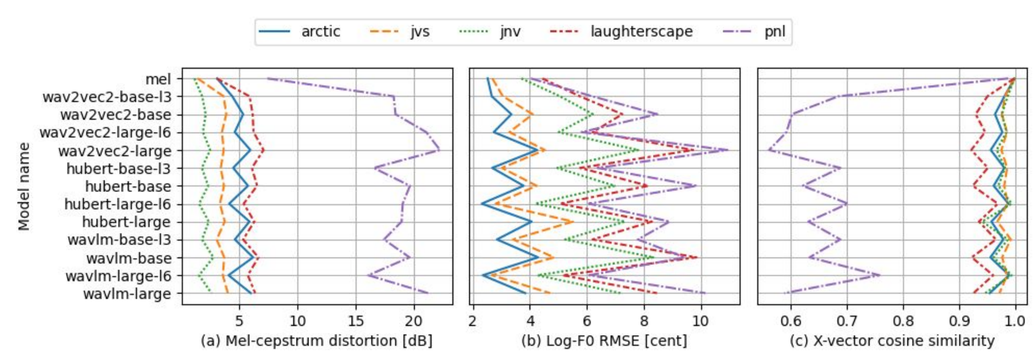
主観評価結果



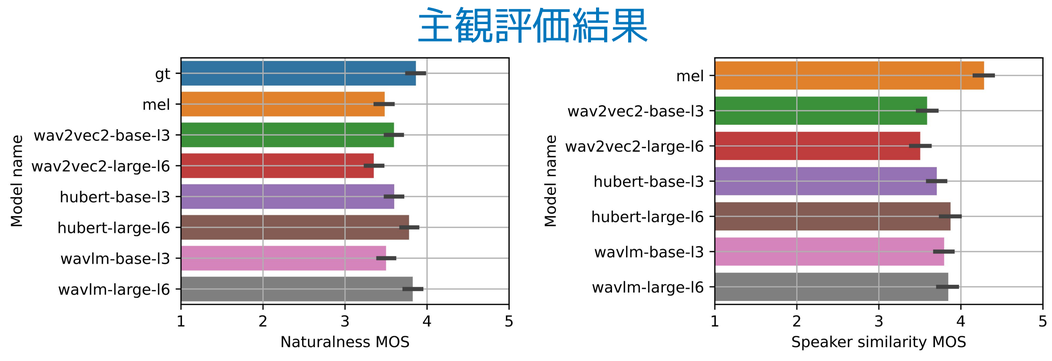
- メルスペクトログラムを使用したモデルの高い汎化性能を確認
- より入力に近い隠れ層において高い再構成性能

使用する隠れ層出力による合成音声の変化

客観評価結果



主観評価結果



- hubert-large-l6, wavlm-large-l6がmelより高い自然性
- 話者類似度に関しては、melより劣る

References

[1] A. Pasad et al., Proc. ASRU, 2021
[2] S. Chen et al., arXiv, 2021
[3] J. Kong et al., arXiv, 2020
[4] H. Zen et al., arXiv, 2019
[5] J. Yamagishi et al., CSTR, 2019
[6] K. Ito et al., 2017
[7] J. Kominek et al., Proc. SSW, 2004
[8] S. Takamichi et al., arXiv, 2020
[9]D. Xin et al., 2023
[10] D. Xin et al., arXiv, 2023
[11] G. Hu et al., IEEE, 2023