1-Q-37

NecoBERT：音声合成のために事前学習された自己教師あり学習モデル

◎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）

本研究の主眼：話者性や韻律のモデリングに適した音声 SSL モデルの構築

- **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
  - 現存のモデルは音声認識を念頭に設計
    - 話者性，韻律の情報が欠落しがち

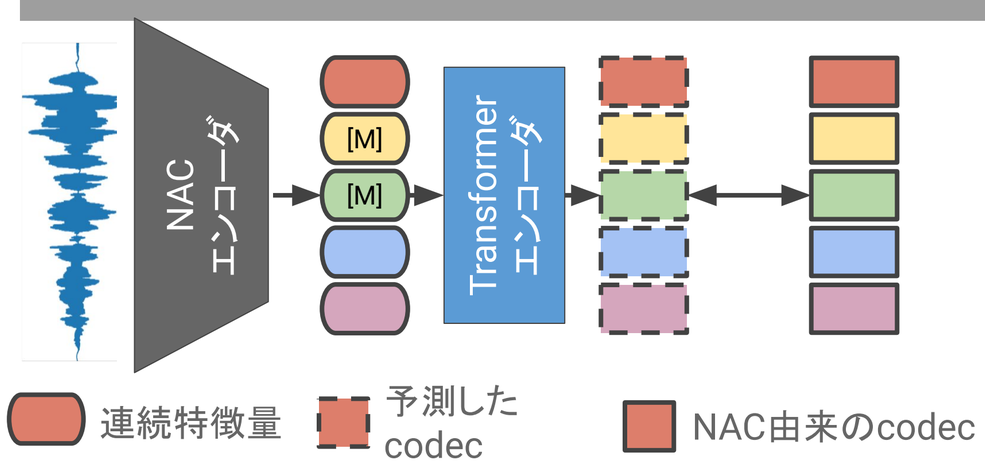
より幅広い応用先に対応できる
音声SSLモデルが求められている

- **先行研究**
  - Neural Audio Codec (NAC)[1]
    - 音声再合成に適した離散圧縮表現 (codec) を獲得
      - 韻律，話者性を正しく表現
      - 長い依存関係のモデリングは難しい
  - HuBERT[2]
    - Masked Language Modeling (MLM) で音声の文脈情報を獲得
    - ASRを念頭に設計するため，話者性，韻律が欠落[3]

- **本研究：NecoBERTの提案**
  - **Ne**ural Audio **Co**dec **BERT**
  - NAC の離散化前特徴量を用いた MLM で学習
  - 韻律，話者性を多分に含む
  - 様々な認識タスク & 音声再合成タスクで提案法の有効性を検証

提案法：NecoBERT のモデル構造と学習条件

モデル構造



- NACから得られるcodecを使用
- NACエンコーダ出力に対してMLM
- NACは音声のみで事前学習
  - c.f.) オリジナルの NAC は広範な音データで構築
  - 離散表現は音声空間に対応
- CrossEntropy Lossで学習

学習条件

NAC	Descript Audio Codec[1]
データセット	LibriSpeech[4] 960時間 多話者英語音声
モデルサイズ	層数12　隠れ層サイズ768 attention head数12
学習環境	8xA100 2日

認識タスクによる評価

タスク：SUPERBベンチマーク[5]

- 音声SSLモデルの評価に用いられるベンチマーク
  - SSLモデルから得られる特徴量が認識タスクに有用か評価
- 以下のタスクを使用
  - 話者認証 (ASV)
  - 感情認識 (ER)
  - 話者ダイアライゼーション(SD)

実験設定

- 比較したモデル
  - DAC encoder出力 (DAC)
  - HuBERT
  - NecoBERT
- 評価指標
  - ASV: Equal Error Rate (EER)
  - ER: ACCuracy (ACC)
  - SD: Diarization Error Rate (DER)

結果

Model	ASV (EER) ↓	ER (ACC) ↑	SD (DER) ↓
DAC	2.59	48.32	12.99
NecoBERT	**1.80**	53.57	9.96
HuBERT	5.11	**64.92**	**5.88**

ASV：HuBERTと比べ改善
他のタスク：HuBERTに劣る

音声再合成による評価

音声再合成 (neural vocoding)

- SSLモデル特徴量から音声を再合
  - NecoBERTから得られる特徴量に韻律，話者性の情報が含まれているか評価
  - ボコーダはHiFi-GAN[6]を使用

- 比較したモデル
  - メルスペクトログラム (Mel)
  - HuBERT-base (最終層)
  - HuBERT-large　(最終層)
  - NecoBERT (最終層)

- 評価指標
  - メルケプストラム歪み (MCD)
  - 対数F0の平方平均二乗誤差 (LogF0RMSE)
  - XVectorのコサイン類似度 (XVector-sim)

使用したデータセット

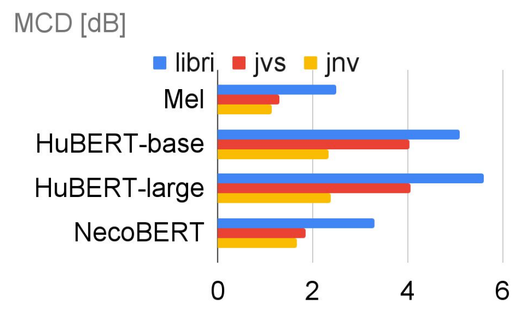
学習用データセット

データセット名	話者数	時間
LibriTTS[7] train-clean-100 train-clean-360	1151名	245時間
VCTK-Corpus[8]	109名	44時間
LJSpeech[9]	1名	25時間

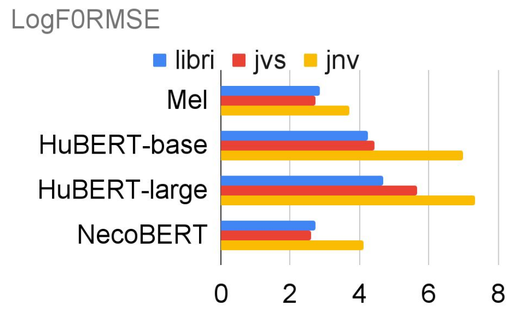
評価用データセット

データセット名	概要
LibriTTS[7] test-cleanサブセット (libri)	英語音声
JVS corpus[10] (jvs)	日本語音声
JNV corpus[11] (jnv)	日本語 非言語音声

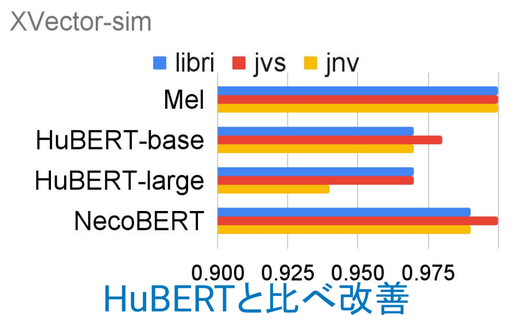
結果



HuBERTと比べ改善



Mel,HuBERTと比べ改善



HuBERTと比べ改善

まとめと今後の展望

- NecoBERTを提案
  - 認識タスク：ASVでHuBERTと比べ改善
  - 音声再合成：全指標でHuBERTと比べ改善

- 今後の予定
  - TTSへの適用
  - より広範な音声処理タスクでの評価

References

[1] R. Kumar et al., Proc. NeurIPS, 2023
[2] W. -N. Hsu et al., TASLP, 2021
[3] 中田 et al., 音響学会春季研究発表会 , 2023
[4] V. Panayotov et al., Proc ICASSP, 2015
[5] S.wen Yang et al., Proc Interspeech, 2021
[6] J. Kong et al., Proc NeurIPS, 2020
[7] H. Zen et al., Proc Interspeech, 2019
[8] J. Yamagishi et al., 2019
[9] K. Ito et al., 2017
[10] S.wen Yang et al., Proc Interspeech, 2021
[11] D Xin et al., Speech Communication, 2023