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）

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] を特徴量抽出器として利用することにより，様々な音声情報処理タスクで有効な特徴量が獲得できると報告されている [4]．これは SSL モデルが事前学習において音声に含まれる音響的な情報に加え，音素や単語の意味などを獲得しているからである [5]．

SSL モデル特徴量は音声合成でも有用であることが報告されており，これを用いた音声合成に関する研究がなされている [6, 7]．一方で音声合成の場合後段の層においては言語情報などは多分に含まれているが，一方で話者性，韻律などの音声固有の情報が欠落していることが報告されている [8]．これにより SSL モデルから得られる特徴量から音声波形を再構成する際に，音響特徴量を用いた場合と比較して高品質な音声波形の再構成が困難になることが報告されている [8]．先行研究 [9] では F0 や XVector [10] でニューラルボコーダを条件付けすることにより，音声波形を再構成している．これにより音声波形を再構成できることが報告されているが，後段の層において話者性，韻律などが欠落していることから SSL モデルではこれらの情報が十分にモデリングされていないと考えられる．より音声合成での SSL モデルの性能を向上するのためには，音声合成において重要な音響特徴量のモデリングを可能とする事前学習のタスク設計が期待される．

本研究では，これらの特徴量のモデリングを行う SSL モデルである NecoBERT (**Ne**ural audio **co**dec **BERT**) を提案する．NecoBERT は，音響特徴量を多分に含むことが期待される Neural Audio Codec (NAC) [11] から得られる特徴量を用いて事前学習を行う SSL モデルであり，マスク付き言語モデル学習を行うことにより NAC 特徴量に対して文脈情報が付加される．これにより，従来の SSL モデルでは欠落していた韻律や話者性などの情報が後段の層までモデリングされることが期待される．実験では，NecoBERT を音声再構成，音声言語モデリング [6]，認識タスク [4] において評価を行った．結果から認識タスクにおいては既存の SSL モデルよりも性能が劣化するものの，話者認証タスクや音声再合成タスクにおいては，既存の SSL モデルよりも高い性能を実現することが示された．

2 関連する研究

音声合成において音声表現として，兼ねてよりメルスペクトログラムなどの音響特徴量が用いられてきた [12,13]．近年では，SSL モデル特徴量が音声合成に対して広く用いられている [6, 7]．代表的な音声合成

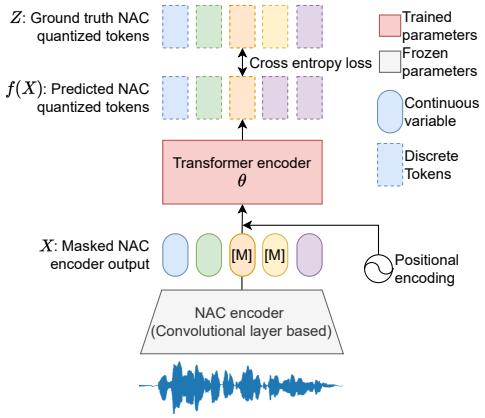

Fig. 1: 提案する NecoBERT のモデル構造及び学習の流れ．[M] で示されているのはマスクされた NAC エンコーダ出力である．

における SSL モデルから得られる特徴量の利用方法として SSL モデル特徴量に対して k 平均法を用いて離散化することにより離散化音声トークンを獲得し，Transformer encoder を Masked Language Modeling (MLM) [14] で学習することによりで音声 SSL モデルを構築している．離散化音声トークンの獲得には構築した音声 SSL モデルの中間特徴量を k 平均法を用いてクラスタリングすることにより得られる．しかし，このような音声トークンは音声に含まれている言語情報を主として表しており，公認室な音声を再構成するのに必要な話者性や韻律の情報が欠落している．

音声のみならず，音響信号の潜在表現の獲得手法として NAC が提案されている．これは主に Autoencoder 構造に情報離散化ボトルネックを導入することにより，より少ない情報量で高い Fidelity を持つ音声の表現を獲得することを目的としている [11, 15]．NAC を音に適用することにより，離散化音響トークンが獲得でき，新しい音声の表現として注目を集めている．しかしながら NAC では，音声の再構成により音響的な情報を忠実に表現する離散トークンが得られるものの，トークン間の時系列的な依存関係を捉えることは困難となり，多くのデータ量を学習に要する他 [16]，離散音声トークンを補助的に用いる必要がある [17, 18]．

3 手法

本研究では，音響情報・文脈情報の両方を捉えた離散化音声トークンの獲得に向けた NecoBERT を提案する．図 1に NecoBERT のモデル構造及び学習の流れを示す．NecoBERT は，NAC モデルである Descript Audio Codec (DAC) [11] から得られる離散化前の潜在表現を入力としてとり，出力として離散化後の音声トークンを予測する．通常の DAC のように音響信号

*NecoBERT: Self-supervised learning of speech representation for speech synthesis NAKATA, Wataru, SAEKI, Takaaki, SAITO, Yuki, TAKAMICHI, Shinnosuke, SARUWATARI, Hiroshi (The University of Tokyo).

データで学習するのではなく，音声コーパスを用いてDACを学習することにより，より音声に特化した離散化音声トークンを獲得する．これにより，話者性や韻律などの音声固有の表現を含めたモデリングを狙う．NecoBERTは，既存のSSLモデルと異なり，音声固有の情報が後段の層において欠落しにくいモデリングを行う．そのために事前学習タスクにおいて音響特徴量を多分に含むことが期待されるNAC由来の潜在表現を入力とし，離散化音響トークンを予測する．また入力にはランダムにマスキングを行うことでMLMを行う．これにより，音響情報を表現するNACを起点とし，MLMにより文脈情報をNAC特徴量に付加させる．

モデル構造としてTransformer encoder [19]を基本構造とする．これは，多くのSSLモデルで採用されている構造であり，その有効性が示されている [4]．マスキングでは先行研究 [2] 同様，ランダムな確率pで選ばれたトークンから，長さlの区間全てをマスキングする．またマスクされた特徴量に関しては，学習可能なパラメータを用いて置換を行う．また，学習に用いる損失関数は以下のように定義する．

$$L = \mathrm{CrossEntropyLoss}\left(f(X;\theta), Z\right) \qquad (1)$$

ここでLは損失関数，CrossEntropyLossは交差エントロピー損失関数，fはNecoBERTモデルのTransformer encoder，θはNecoBERTモデルの重みである．マスクされたSSL表現を$X = \{\boldsymbol{x}_n^M \in \mathbb{R}^d | n = 1, \cdots, N\}$とする．さらに，NACのresidual vector quantierから得られる離散tokenを$Z = \{\boldsymbol{z}_n \in \mathbb{R} | n = 1, \cdots, N\}$とする．これにより，Transformer encoderの最終層出力を音声の表現として使用することができる．得られる特徴量にはNAC由来の音響特徴量に加え，MLMにより文脈情報が含まれていることが期待される．

4 実験

本研究では提案するSSLモデルであるNecoBERTの学習を行い，その性能を音声再合成，音声言語モデリング，認識タスクの観点から評価を行った．

4.1 実験条件

データセットにはLibriSpeech [20]を使用した．これは960時間の多話者英語音声である．NACにはDescript Audio Codec (DAC) [11]を使用し，LibriSpeechにおいて150K stepsの学習を行ったものを使用した．DACおよびNecoBERT学習時には音声を24 kHzにリサンプリングを行った．

DACのEncoderのダウンサンプルレートは$2, 3, 4, 5, 5$とし，1/480のダウンサンプリングを行った．これにより24 kHzの入力において50 Hzの音響トークンが獲得できる．これは，HuBERT [2], WavLM [3], wav2vec 2.0 [1]と同じ周波数である．それ以外のパラメータに関しては，DACの公式実装[1]に従った．NecoBERTのTransformerの構造として，Transformer encoder（層数12 隠れ層サイズ768, アテンションヘッド数12）を使用した．これはwav2vec 2.0 [1], WavLM [3], HuBERT [2]-baseモデルと同一のモデル構造である．また，予測する音響トークンの抽出にはDACの最初のベクトル量子化結果のみを用いて学習を行った．

NecoBERTの最適化には，AdamW [21] ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-6}, \lambda = 1 \times 10^{-2}$)を使用した．学習率に関しては，最初の40kステップに関しては0から2×10^{-4}へ線形に増加させ，その後460kステップにわたって2×10^{-4}から0へ線形に減少させた．よって全体の学習ステップ数は500kステップである．

マスキングのパラメータに関しては，マスクする確率を$p = 0.08$マスクする区間を$l = 10$とした．

4.2 評価

本研究では評価手法として，音声言語モデリング [6]，音声再合成 [8]，認識タスクの3つの観点からNecoBERTの評価を行った．

4.2.1 音声言語モデリング

獲得された離散音声トークンに文脈情報が捉えられているかを確認するために，音声言語モデリングによる評価を行った．評価では，離散音声トークンを用いて言語モデルを学習する手法として知られているunit Language Model (uLM) [6]のperplexityを用いて評価を行った．評価にはHuBERTの第6層目特徴量，NecoBERT最終層特徴量，DAC特徴量を使用し比較を行った．まず，それぞれの特徴量に対してk平均法を用いて量子化を行った．量子化に使用したクラスタ数は1000である．その後，離散化された特徴量をGPT-neox [22]を使用してTransformerモデルを言語モデリングタスクにおいて学習を行った．評価には学習結果として得られたモデルのLibriTTS [23]のtestセットに対するperplexityを使用した．uLMの学習には，LibriTTS [23]のtrain-clean-100, train-clea-360, train-other-500サブセットを使用した．また，Transformerモデルのモデル構造として層数12，隠れ層サイズ1024，アテンションヘッド数16を使用した．ニューラルボコーダ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．また，各音声は22.05 kHzにリサンプリングを行った．

4.2.2 音声再合成

NecoBERTを用いて獲得された音声特徴量に，韻律や話者性などの情報が含まれていることを評価するために特徴量からの音声再合成を行った．音声再合成では，2つの条件を使用した．1つ目が離散化された特徴量からの再合成であり，2つ目が連続特徴量からの再合成である．両方の条件での音声再合成において，ニューラルボコーダであるHiFi-GAN [24]を用いた再合成を行なった．離散化した特徴量を入力する際には，One hot encodingを用いて離散特徴量を連続特徴量へと変換したのち合成を行なった．各条件において評価指標として対数F0の平方平均二乗誤差 (LogF0RMSE)，メルケプストラム歪 (MCD)，XVectorのコサイン類

[1] https://github.com/descriptinc/descript-audio-codec

Table 1: 音声言語モデリングにおける評価結果．太字は一番良い結果を示す．

離散音声トークン獲得に使用した特徴量	Perplexity
DAC	74.50
NecoBERT-L12	61.80
HuBERT-L6	**4.48**

似度 (XVector-sim) を使用した．

離散化された特徴量からの再合成では，データセットに LibriTTS [23] の train-clean-100, train-clean-360, train-other-500 サブセットを使用し HiFi-GAN を学習した．評価には test-clean サブセットを使用した．また，離散化の設定は 4.2.1と同一の設定を使用した．比較手法として，NecoBERT 最終層特徴量，DAC エンコーダ出力，HuBERT 第 6 層出力 (HuBERT-L6) に加え，それを HiFi-GAN を XVector で条件付けしたもの (HuBERT-L6+XVector)，さらにそれに加え F0 で条件づけしたもの (HuBERT-L6+XVector+F0) を使用した．XVector モデルには huggingface 上で公開されているモデルを使用した．[2]

連続特徴量からの再合成では，データセットに LibriTTS [23] の train-clean100, train-clean-360 サブセット及び VCTK-Corpus [25], LJSpeech [26] 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．比較手法として，NecoBERT 最終層特徴量に加え，メルスペクトログラム (Mel)，HuBERT-base,HuBERT-large のそれぞれの最終層特徴量からの音声再合成を行った．評価は LibriTTS の test-clean セット (libri) に加え，学習ドメイン外での評価のために，JVS コーパス [27] の parallel100 サブセット JVS001 010 の話者 (jvs)，JNV コーパス [28] (jnv)，PNL100 コーパス [29] (pnl) に対して行った．これらは順に英語音声，日本語音声，非言語発話，非音声である．

4.2.3 認識タスク

NecoBERT から得られる音声特徴量が認識タスクにおいて有用かどうかを評価するために認識タスクによる評価を行った．認識タスクには ASV (話者認証, Automatic Speaker Verification)，ER (感情認識, Emotion recognition)，SD (話者ダイアライゼーション, Speaker Diarization) の三つのタスクを使用した．評価条件は SUPERB [4] の条件を使用した．評価では NecoBERT から得られる特徴量に加え，DAC 特徴量を用意し比較を行った．また，HuBERT-base の結果は先行研究 [4] から引用している．ASV の評価指標としては Equal error rate (EER)，SD の評価指標としては Diarization error rate (DER)，ER の評価指標として精度 (Accuracy, ACC) をそれぞれ使用した．

5 結果と考察

5.1 音声言語モデリング

表 1に音声言語モデリングにおける評価結果を示す．結果から，NecoBERT-L12 を離散音声トークンの獲得に使用した場合ではでは DAC を使用した場合と比べて Perplexity が改善していることが確認できる．このことから，NecoBERT は MLM を用いた学習によって文脈情報を NAC 特徴量に付加できていることが確認できる．一方で HuBERT-L6 を離散音声トークンの獲得に使用した場合では，NecoBERT-L12 と比べ大きく perplexity が改善している．これは NecoBERT では韻律や話者などのより多様な音声の情報をモデリングしている一方で，HuBERT-L6 ではそれらの情報が含まれていないより言語情報に近い情報をモデリングしているため，音声言語モデルの学習が容易であるからと考えられる．

5.2 音声再合成

表 2に連続特徴量からの音声再合成評価結果を示す．結果から，NecoBERT を用いた場合では．HuBERT-base や HuBERT-large 特徴量を用いた場合と比べ各指標において改善していることが確認できる．加えて Mel と NecoBERT-base を比較すると libri や JVS などの音声に対して F0 の改善が見られた．これは，NecoBERT-base の最終層まで，音声固有の情報がモデリングされていることを示唆している．また，ドメイン外の音声である pnl や jnv といったデータに関して，NecoBERT は HuBERT-base,large と比べよりロバストに動作していることがわかる．これは，HuBERT の最終層では音響的な特徴量が失われている一方で，NecoBERT では音響的な特徴量が保持されているためと考えられる．

表 3に離散特徴量からの音声再合成評価結果を示す．結果から，NecoBERT 特徴量を用いた場合 F0 においては F0 で条件付けしているモデルである HuBERT-L6+F0+XVector を除き，最も良い結果となっている．これは，NecoBERT 特徴量にはより多くの音声固有の情報が含まれており，それが離散化によって失われていないことを示している．また，DAC と比較すると LogF0RMSE において NecoBERT の方がより良い結果を示している．

5.3 認識タスク

表 4に認識タスクにおける評価結果を示す．結果から，DAC 特徴量をそのまま使用するよりも NecoBERT を用いることで評価結果が改善することが確認された．これは，Transformer encoder を用いた MLM によりなんらかの情報をモデルが獲得していることを示す．一方で先行研究 [4] で報告されている結果と比べると，ER，SD などで劣化している一方で ASV では大きく改善する結果となっている．これは ER,SD などのタスクでは，音声に含まれる言語情報がタスクに対して有効である一方で，ASV では音声固有の情報が有効である．このため，言語情報

[2] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

Table 2: 連続特徴量からの音声再合成評価結果．太字は各指標，各コーパスにおいて一番良い結果を示す．

Model\Corpus	LogF0RMSE libri	jvs	jnv	pnl	MCD libri	jvs	jnv	pnl	XVector-sim libri	jvs	jnv	pnl
Mel	2.86	2.71	**3.70**	**4.03**	**2.49**	**1.31**	**1.13**	**7.37**	**1.00**	**1.00**	**1.00**	**0.99**
HuBERT-base	4.23	4.44	6.97	9.86	5.09	4.03	2.34	19.61	0.97	0.98	0.97	0.62
HuBERT-large	4.67	5.67	7.34	8.91	5.61	4.05	2.39	18.79	0.97	0.97	0.94	0.63
NecoBERT-base	**2.73**	**2.61**	4.10	5.22	3.31	1.85	1.66	9.96	0.99	1.00	0.99	0.89

Table 3: 離散化特徴量からの音声再合成評価結果太字は各指標，各コーパスにおいて一番良い結果を示す．

	LogF0RMSE	MCD	XVector-sim
DAC	4.88	**5.34**	**0.99**
HuBERT-L6	7.70	7.61	0.94
+XVector	5.64	6.28	0.97
+F0+XVector	**3.53**	5.79	0.98
NecoBERT	4.29	5.65	0.98

Table 4: 認識タスクにおける評価結果．↑で示されている列は高い値が良い結果を示しており，↓で示されている列は低い方が良い結果を示している．また，各タスクで一番良い手法を太字で示す．

Model	ASV (EER)↓	ER (ACC)↑	SD (DER)↓
DAC	2.59	48.32	12.99
NecoBERT	**1.80**	53.57	9.96
HuBERT [4]	5.11	**64.92**	**5.88**

を多分に含む HuBERT が ER，SD といったタスクで強く，音声固有の情報を多分に含む NecoBERT が ASV において頑健になっていると考えられる．

6 まとめ

本研究では，音声 SSL モデルである NecoBERT を提案した．提案する NecoBERT は NAC 由来の離散化音響トークンに対して MLM を行うことにより，文脈情報を付加することを狙った SSL モデルである．提案する NecoBERT の評価を認識タスク，音声言語モデリング，音声再合成の観点から評価を行った．結果から，認識タスクや音声言語モデリングでは既存の音声 SSL モデルと比べ性能が劣化する一方で，音声再合成においてはより原音声に近い再合成が可能であることが示された．今後の課題としては，より明示的な SSL モデル由来の離散化音声トークンの利用が挙げられる．

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," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 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] S. wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[5] T. Ashihara et al., "SpeechGLUE: How well can self-supervised speech models capture linguistic knowledge?" in *Proc. Interspeech 2023*, 2023.

[6] K. Lakhotia et al., "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[7] W.-C. Huang et al., "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," *ICASSP*, pp. 5944–5948, 2020.

[8] 中田亘 et al., "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," **日本音響学会** *2023* **年秋季研究発表会講演論文集**, Sep. 2023.

[9] A. Polyak et al., "Speech resynthesis from discrete disentangled self-supervised representations," in *Proc. Interspeech 2021*, 2021.

[10] 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.

[11] R. Kumar et al., "High-fidelity audio compression with improved RVQGAN," *ArXiv*, vol. abs/2306.06546, 2023.

[12] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.

[13] N. Li et al., "Neural speech synthesis with transformer network," in *Proc. AAAI*, Honolulu, U.S.A., July 2019, pp. 6706–6713.

[14] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[15] A. D'efossez et al., "High fidelity neural audio compression," *ArXiv*, vol. abs/2210.13438, 2022.

[16] C. Wang et al., "Neural codec language models are zero-shot text to speech synthesizers," *ArXiv*, vol. abs/2301.02111, 2023.

[17] Z. Borsos et al., "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2022.

[18] X. Zhang et al., "SpeechTokenizer: Unified speech tokenizer for speech large language models," *ArXiv*, vol. abs/2308.16692, 2023.

[19] A. Vaswani et al., "Attention is all you need," in *Neural Information Processing Systems*, 2017.

[20] V. Panayotov et al., "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, South Brisbane, Australia, Apr. 2015, pp. 5206–5210.

[21] I. Loshchilov, F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[22] A. Andonian et al., "GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch," 9 2023.

[23] H. Zen et al., "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[24] J. Kong et al., "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle et al., Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.

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

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

[27] S. Takamichi et al., "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, pp. 761–768, 2020.

[28] D. Xin et al., "JNV corpus: A corpus of Japanese nonverbal vocalizations with diverse phrases and emotions," *Speech Communication*, vol. 156, no. 103004, 2024.

[29] 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.