対照学習モデルによる音声-声質表現文の埋め込み表現獲得*

◎渡邊 亞椰，高道 慎之介，齋藤 佑樹，中田 亘，辛 徳泰，猿渡 洋 (東大院・情報理工)

1 はじめに

テキスト音声合成（text-to-speech: TTS）の声質制御手法として，声質を表現する自由記述（以降，声質表現文）による制御の需要は高く，昨今盛んに提唱されるようになった [1–3]．声質表現文を用いて制御する TTS の抽象化した構造を Fig. 1 に示す．

前報 [4] ではこのような TTS の研究に必要な音声・声質表現文ペアコーパスを Web データから構築した．最終的にはこのコーパスによる TTS 構築を目指すが，先んじて，声質表現文により声質を指定するモジュールの構築が必要である．このモジュールは，Fig. 1 のようなモデルで，声質に相当する埋め込み表現の獲得までを副問題として切り分ける場合に必要となる．

本研究では，対照学習モデルを用い，声質表現文と音声を同一空間に埋め込むことで，声質と結びついた声質表現文埋め込みを獲得することを目指す．期待通りに学習されたモデルは，声質表現文に対応する声質の音声を取得することが可能である．

構築方法として，CLAP [5] に基づく構造のモデルを提案する．対照学習をするのみではなく，主観評価に影響を及ぼす諸要素を明示的に学習させ，効率良く声質を学習させる．具体的には，性別や話速といった要素を損失関数として学習に取り入れる．

2 関連研究

2.1 声質表現文による TTS 制御

画像生成 [6] を発端としてテキストにより生成物を制御するマルチメディア生成技術が発展し，昨今は音声合成 [1–3]・声質変換 [7, 8] における声質制御への応用が盛んである．例として，PromptTTS [1] は，声質表現文を BERT [9] ベースのモデルでエンコードし，FastSpeech2 [10] ベースの多話者 TTS モデルに入力することで声質を制御している．この際，声質表現文のエンコーダは，声質表現文から性別や音高等の特徴を識別できるようにファインチューニングされる．

多くの手法において学習のために使われるコーパスは，LibriTTS [11] 等の TTS 向け既存資源に人手で声質表現文を付与したもの [1]，少数話者の発話を収録した内製コーパス [2] などである．これらの問題点として，声質の多様性に欠けること，非公開コーパスが多いことが挙げられる．

2.2 Coco-Nut

2.1 節で示したコーパスの欠点に対応すべく，より多様で大規模な公開コーパスとして，Web 上の音声を収集した日本語コーパスである Coco-Nut を前報 [4] にて提案した．Coco-Nut は，YouTube から収集された音声と，クラウドソーシングで収集した自由形式の声質表現文を含む．このコーパスは既存の声質表現文コーパスと比較して，幅広い声質に対し，幅広い記述形式での声質表現がなされている．従って，声質表現の学習という観点で最適なコーパスであると言える．

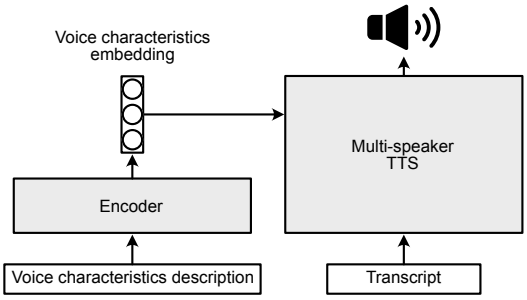

Fig. 1 声質表現文により制御する TTS の概図

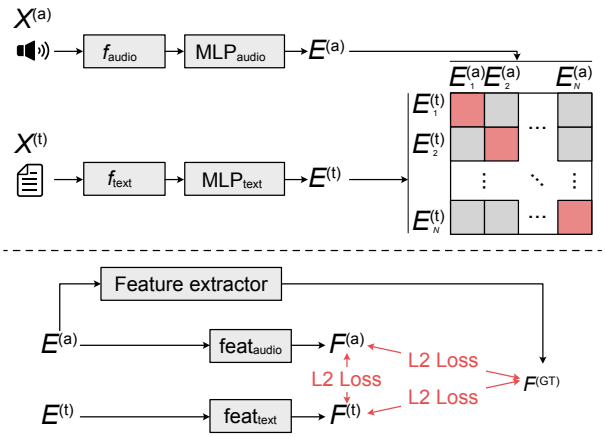

Fig. 2 上段: CLAP 下段: 音響特徴予測学習

2.3 対照学習

CLIP [12] に代表される対照学習は，異なる形式のメディアの対応を学習し，同一空間に埋め込むことを目標とする．例として，環境音およびそのキャプションについて対照学習を行う CLAP [5] の構造を紹介する．CLAP の構造の概要を Fig. 2 上段に示す．

CLAP は，音波形 $X^{(\mathrm{a})}$ および音を説明するキャプション $X^{(\mathrm{t})}$ を，それぞれ対応する事前学習済みエンコーダ $f_{\mathrm{audio}}(\cdot)$ および $f_{\mathrm{text}}(\cdot)$ によってエンコードし，多重線形層 $\mathrm{MLP}_{\mathrm{audio}}(\cdot)$，$\mathrm{MLP}_{\mathrm{text}}(\cdot)$ によって同一次元のベクトル $E^{(\mathrm{a})}$，$E^{(\mathrm{t})}$ に変換する．学習には音波形とキャプションのペアを用い，$(X_i^{(\mathrm{a})}, X_i^{(\mathrm{t})})$ を i 番目のペアをとすると，モデルは以下の損失関数

$$L_{\mathrm{CLAP}} = \frac{1}{2N}\sum_{i=1}^{N}\left(L_{(\mathrm{a}\to\mathrm{t})i} + L_{(\mathrm{t}\to\mathrm{a})i}\right) \tag{1}$$

の最小化によって学習される．ここで，

$$L_{(\mathrm{a}\to\mathrm{t})i} = \log\frac{\exp\left(E_i^{(\mathrm{a})}\cdot E_i^{(\mathrm{t})}/\tau\right)}{\sum_{j=1}^{N}\exp\left(E_i^{(\mathrm{a})}\cdot E_j^{(\mathrm{t})}/\tau\right)} \tag{2}$$

*Latent representation for pairs of speech and speech characteristics description trained by contrastive learning, by Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Wataru Nakata, Detai Xin, Hiroshi Saruwatari (The University of Tokyo).

$$L_{(\mathrm{t}\to\mathrm{a})i} = \log \frac{\exp\left(E_i^{(\mathrm{t})} \cdot E_i^{(\mathrm{a})}/\tau\right)}{\sum_{j=1}^{N} \exp\left(E_i^{(\mathrm{t})} \cdot E_j^{(\mathrm{a})}/\tau\right)} \quad (3)$$

である．

ここで，N は一度の損失計算に用いる学習用ペアの総数（ミニバッチ学習に置いてはバッチサイズ）を，τ は温度のパラメータを示す．学習により，正例ペアは近くに埋め込まれるようになる．

対照学習は自己教師あり学習であり，多量のペアデータから学習をする能力に長けるが，タスクの難易度やデータ量によっては効果に限界がある．しかし，別のタスクを併用することにより学習効率を上げ，欠点を補う先行研究は存在する [13]．

3 提案法

3.1 埋め込みモデル

声質表現文と音声を同一空間に埋め込むモデルを作成する．構築に当たり，対照学習の構造と損失関数を基としつつ，声質を効率に学習する目的で音響特徴予測学習を取り入れる．概要を Fig. 2 に示す．

3.1.1 モデル構造

モデル構造および学習方法は，**2.3** 節で示した CLAP [5] に準拠する．環境音に特化した HTS-AT [14] を f_{audio} に用いた CLAP に対し，提案法では HuBERT [15] を使用する．また，f_{text} には日本語のコーパスで学習された RoBERTa [16] を使用する．

3.1.2 音響特徴予測学習

提案モデルの最終目標は，人の持つ声質に対する知覚を学ぶことである．これを効率的に達成すべく，人間が着目しやすい声質要素と関連した音響特徴量を明示的に与え，予測する機構を導入する．特徴量予測モジュール $\mathrm{feat}_{\mathrm{audio}}(\cdot)$，$\mathrm{feat}_{\mathrm{text}}(\cdot)$ を導入し，モデルの出力 $E^{(\mathrm{a})}$ および $E^{(\mathrm{t})}$ それぞれを変換して予測特徴量 $F^{(\mathrm{a})}$，$F^{(\mathrm{t})}$ とする．学習コーパス内 i 番目のペアデータ $(X_i^{(\mathrm{a})}, X_i^{(\mathrm{t})})$ から得られる正解音響特徴量を $F_i^{(\mathrm{GT})}$，予測特徴量を $F_i^{(\mathrm{a})}$，$F_i^{(\mathrm{t})}$ とした際，音響特徴予測性能は損失関数 L_{feat} の最小化で学習される．

$$L_{\text{feat-audio}} = \sum_{i=1}^{N} ||F_i^{(\mathrm{GT})} - F_i^{(\mathrm{a})}||_2 \quad (4)$$

$$L_{\text{feat-text}} = \sum_{i=1}^{N} ||F_i^{(\mathrm{GT})} - F_i^{(\mathrm{t})}||_2 \quad (5)$$

$$L_{\text{feat-cross}} = \sum_{i=1}^{N} ||F_i^{(\mathrm{a})} - F_i^{(\mathrm{t})}||_2 \quad (6)$$

$$L_{\mathrm{feat}} = L_{\text{feat-audio}} + L_{\text{feat-text}} + L_{\text{feat-cross}} \quad (7)$$

このうち，式 (4) および式 (5) は $E^{(\mathrm{a})}$ および $E^{(\mathrm{t})}$ からの予測と正解特徴量との差を，式 (6) は $E^{(\mathrm{a})}$，$E^{(\mathrm{t})}$ 間の予測の差を示す．また，N は一度の損失計算に用いる学習用ペアの総数を示す．

特徴量予測学習は対照学習と同時に行われ，モデルの学習は式 (1) に示した対照学習の損失関数との和

$$L = L_{\mathrm{CLAP}} + \alpha L_{\mathrm{feat}} \quad (8)$$

の最小化として定義される．ここで，α は特徴量予測損失関数にかかる係数である．

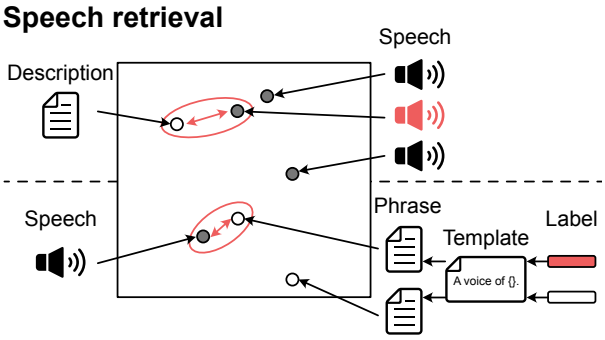

Fig. 3　音声取得および Zero-shot 音声分類の概要

3.2 モデル評価タスク

3.1 節で構築したモデルの検証のため，下流タスクを設定する．タスク概要を Fig. 3 に示す．

3.2.1 声質表現文による音声取得

[5] で示されているキャプションからの環境音取得と同様の方法で，声質表現文による音声の取得を行う．事前に **3.1** 節で構築したモデルを用いて取得対象の音声群を全て埋め込んだ上で，取得に用いる声質表現文の埋め込みを同モデルから獲得する．そして，音声埋め込みと声質表現文埋め込みとのコサイン類似度を算出し，音声と声質表現文との間の類似度とし，類似度の高い音声を取得する．

3.2.2 声質表現フレーズによる Zero-shot 音声分類

3.2.1 節の手法の応用として，声質表現フレーズを用いた Zero-shot 音声分類を行う．声質の特徴を示すラベル（e.g. 性別，感情等）群を，テンプレート文（e.g. ○○の声）を用いて声質表現フレーズに変換し，全て構築済みモデルで埋め込む．その後，識別対象の音声を同モデルで埋め込み，声質表現フレーズ群の埋め込みとの類似度を計算する．結果，音声は最も類似した声質表現文の元となったラベルへと分類される．

4 実験的評価

4.1 実験条件

モデルは LAION-AI の実装[1]をベースとして構築した．RoBERTa，HuBERT は日本語で事前学習されたモデル[2][3]を使用し，エンコーダの値はフリーズさせ，多重線形層を学習させた．学習には Coco-Nut [4] 学習セット[4]を用いた．出力次元数は 512 次元，温度パラメータ τ の初期値は $1/\log(1/0.07)$ とした．

音響特徴予測の対象には，F0 平均，エネルギー標準偏差，一秒あたり発話モーラ数を表す 3 次元ベクトルを用いた．これらは，前報 [17] での声質表現文頻

[1] https://github.com/LAION-AI/CLAP
[2] https://huggingface.co/rinna/japanese-roberta-base
[3] https://huggingface.co/rinna/japanese-hubert-base
[4] 音響特徴量抽出が不可能な一部データを除外した．以降，検証および評価セットについても同様に一部データを除外して用いた．

出表現調査から分析した，評価者に着目されやすい声質要素と結びつくものである．特徴量予測モジュールは ReLU 層を挟んだ 2 層の線形層とした．特徴量予測損失関数にかかる係数 α は，特徴量予測損失不使用としての 0.0，および 0.5，1.0 で学習し，比較した．

学習率は $5e^{-6}$，バッチサイズは 48 とし，学習エポック数は 90 とし，5 エポックごとに保存した．

事前に客観評価に基づく予備検討を行った上で，実験的評価として，音声取得結果に対する主観評価実験，音声分類結果に対する客観評価実験を行った．

予備検討では，Coco-Nut の検証セットに含まれる 593 件のペアデータを用いた．タスクは声質表現文による音声取得とし，取得結果上位 10 件における性別正解率を指標とした．音声取得に用いる声質表現文には検証セット内の全ペアから 1 文ずつランダム選択した計 593 文を，対象には同セット内の全音声 593 件を用いた．声質表現文の示す性別および音声の性別は，声質表現文の記述を元にラベル付けした[5]．保存した全てのモデルについて評価を行い，以降の実験的評価に用いるモデルを検討した．

音声取得結果への主観評価としては，声質表現文と取得結果音声との対応度に対する主観的評価を用いた．評価対象は Coco-Nut 評価セットとし，ランダムに選択した声質表現文 100 文で，評価セット内全音声 611 件を対象に音声取得を行った．使用した声質表現文と取得結果の音声とを組み合わせて提示し，その対応度合いを 1（全く対応していない）〜9（とてもよく対応している）の 9 段階で評価させた．この際，1 位の取得結果だけでなく，2-5 番目，6-10 番目の取得結果についても実験を行った[6]．また，評価セットからランダムに選んだ音声と声質表現分のペア，評価セット内の正解ペアについても実験を行った．聴取者はランサーズ[7]を通じて雇用し，聴取者 1 人あたり 16 件，1 ペアあたり 20 人の聴取者を割り当てた．

客観評価は，性別ラベルを用いて話者属性把握性能を評価した．評価対象は JVS [18] の parallel100 セットであり，49 名の男性話者と 51 名の女性話者の合計 9997 発話[8]を用いた．タスクは Zero-shot 性別二値分類とし，「男性が喋っている。」「女性が喋っている。」という声質表現フレーズを用いた．

4.2 予備検討結果

結果を Fig. 4 に示す．ここで，横軸は学習エポック数を，縦軸は取得結果上位 10 件の平均性別正解率を各性別の声質表現文ごとに平均したものである．

結果より，$\alpha = 1.0$ の場合と $\alpha = 0.5$ の場合は，性能の上限は概ね同様である．また，$\alpha = 0.0$ で音響特徴予測学習を用いない場合は，用いた場合に比べて性能が劣る．$\alpha = 1.0$ の場合は概ね 35，$\alpha = 0.0$ は 60 エポックで性能が上限に達したと見られる．

以降，音響特徴予測学習使用モデルとして $\alpha = 1.0$ の 35 エポック学習時の，不使用モデルとして $\alpha = 0.0$ の 60 エポック学習時のモデルを使用する．

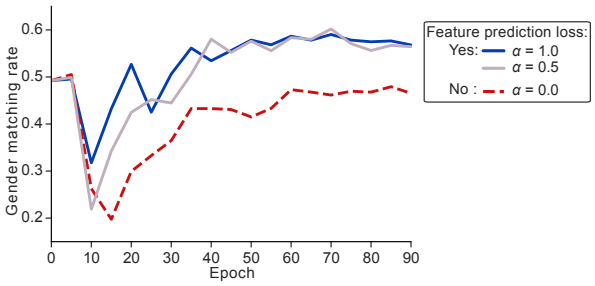

Fig. 4　予備検討結果（性別一致度）

4.3 実験結果と考察

4.3.1 声質表現文による音声取得

主観評価結果を Fig. 5 に示す．結果より，取得結果 1 位のみならず，取得結果 10 位まで全体の傾向として，音響特徴予測学習を導入した方が評価結果が高い．原因として，Fig. 6 に示したペアごとの評価平均の分布からわかるように，音響特徴予測学習不使用時には低評価なペアが多数存在するが，使用時には中程度評価のペアの割合が高いことが挙げられる．声質に関わる要素を学習させたため，違和感のある取得結果が生じにくくなっていると言える．

取得結果の例を Table 1 に示す．評価平均 (GT) は正例ペア音声との組に対する評価の平均を意味する．3 行目の例のように，声質表現文と性別不一致の音声への評価は他要素一致に関わらず性別一致の音声のものより低く，Fig. 4 に示す性別一致率向上は全体的な評価の安定に貢献したと推測できる．

4.3.2 声質表現フレーズによる Zero-shot 音声分類

正解の性別ごとの分類正答率を示したものを Table 2 に示す．性別要素についての二値分類という単純なタスクであれば音響特徴予測学習の有無によらず高精度で正解するが，全体として音響特徴予測学習を使用した場合のほうが正答率が高い．

5 まとめ

本研究では，声質表現文と音声を同一空間に埋め込み，声質表現文による音声取得をするモデルを構築した．対照学習をベースとしつつ，より声質学習を効率化すべく，音響特徴量の予測によって性能を向上させた．今後はこのモデルを組み込んだ TTS モデルの構築を目指す予定である．また，提案モデルの性能には未だ改善の余地があり，予測する音響特徴量の再選定，ChatGPT[9]等の言語モデルによるリフレージングを用いた学習コーパスの拡張による改善を検討する．

謝辞: 本研究は科研費 21H04900, 22H03639, 23H03418，JST 創発的研究支援事業 JP23KJ0828，ムーンショット JPMJPS2011 の助成を受けたものです．

参考文献

[1] Z. Guo et al., “Promptttts: Controllable text-to-speech with text descriptions,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[5] 声質表現文が「男」という字を含む場合は男性と，「女」という字を含む場合は女性とした．二値分類とするため，「男」「女」両方の字を含む，またはどちらも含まない声質表現文は除外した．

[6] 2-5 番目，6-10 番目については，1 声質表現文の取得結果についてランダムに 1 件選択した．

[7] https://www.lancers.jp/

[8] 一部破損したファイルを除いた．

[9] https://chat.openai.com

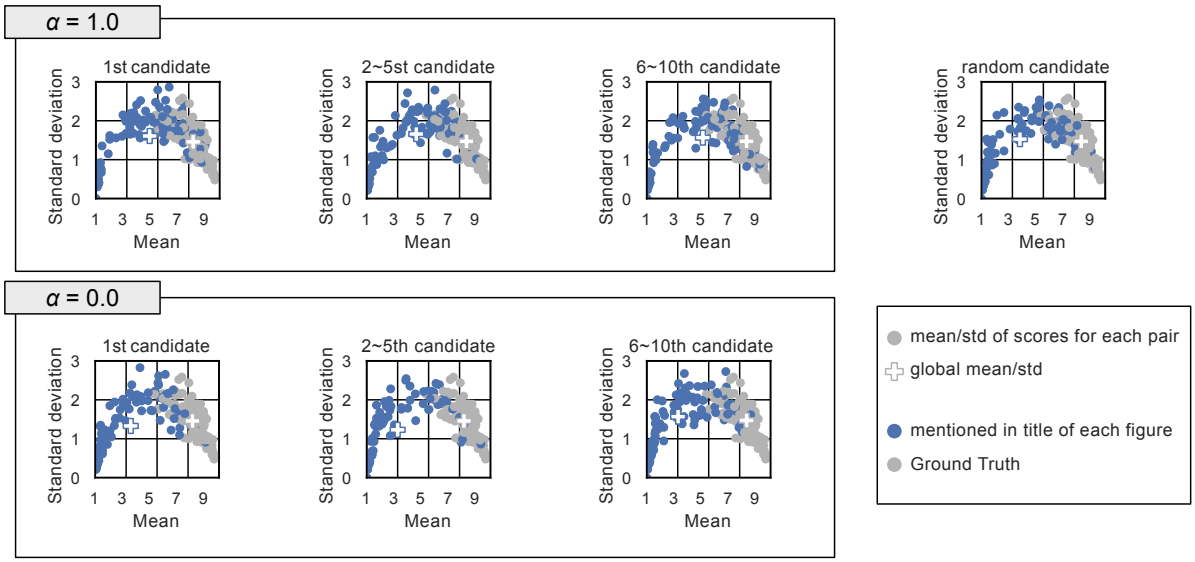

Fig. 5　主観評価結果（散布図）各プロットは 1 ペアについての 20 人分の評価の平均および標準偏差を示す．

Table 1　音声取得結果例．取得結果として示す文は取得音声の正例ペアとなる声質表現文である．

声質表現文	評価平均 (GT)	取得結果 1 位 ($\alpha = 0.0$)	評価平均	取得結果 1 位 ($\alpha = 1.0$)	評価平均
20 代くらいの若い女性が楽しそうな声で訴えるようなしゃべり方をしている。	6.75	20 代くらいの女性が、コソコソした声で、指示するように話している。	4.86	若い女性が、明るくはきはきした声で、少年のように喋っている。	5.74
若い男性が、早口で、何かを説明しながら喋っている。	8.55	３０代男性が抑えた口調で子供に語りかけている。	5.90	50 代の男性が、明瞭な声で少し迷いながらも落ち着いて説明している。	6.57
10 代くらいの若い男性が友達と話すような声のトーンで楽しそうに喋っている。	7.58	若い女性がのんびりした声で、楽しそうに自分の話に笑いながら喋っている。	1.95	20 代の男性が感情を押さえながら話している。	3.19

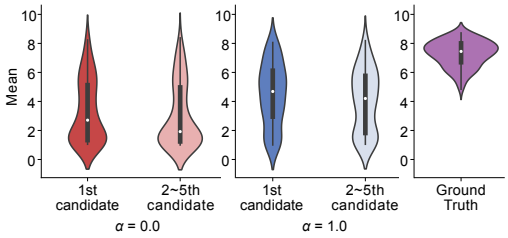

Fig. 6　主観評価結果（評価平均分布）

Table 2　Zero-shot 性別分類結果

α	正解性別	正答率
0.0	男性	0.998
	女性	0.941
1.0	男性	0.986
	女性	1.000

[2] D. Yang et al., “InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt,” *arXiv:2301.13662*, 2023.

[3] Y. Zhang et al., “Promptspeaker: Speaker generation based on text descriptions,” *arXiv preprint arXiv:2310.05001*, 2023.

[4] A. Watanabe et al., “Coco-nut: Corpus of Japanese utterance and voice characteristics description for prompt-based control,” *arXiv preprint arXiv:2309.13509*, 2023.

[5] B. Elizalde et al., “CLAP: Learning audio concepts from natural language supervision,” *arXiv:2206.04769*, 2022.

[6] A. Ramesh et al., “Zero-shot text-to-image generation,” *arXiv:2102.12092*, 2021.

[7] J. Yao et al., “Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts,” *arXiv preprint arXiv:2309.09262*, 2023.

[8] C.-Y. Kuan et al., “Towards general-purpose text-instruction-guided voice conversion,” *arXiv preprint arXiv:2309.14324*, 2023.

[9] J. Devlin et al., “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[10] Y. Ren et al., “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[11] H. Zen et al., “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in *Proc. Interspeech*, 2019, pp. 1526–1530.

[12] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. PMLR*, vol. 139, 18–24 Jul 2021, pp. 8748–8763.

[13] C.-F. Yeh et al., “Flap: Fast language-audio pre-training,” *arXiv preprint arXiv:2311.01615*, 2023.

[14] K. Chen et al., “Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection,” in *Proc. ICASSP*. IEEE, 2022, pp. 646–650.

[15] W.-N. Hsu et al., “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[16] Y. Liu et al., “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[17] 渡邊亞椰 et al., “Coco-nut: 自由記述文による声質制御に向けた多話者音声・声質自由記述ペアデータセット,” **日本音響学会第** *130* **回（***2023* **年秋期）研究発表会**, pp. 1133–1136, 2023.

[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.