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 *

☆渡邊 亞椰，高道 慎之介，齋藤 佑樹，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）の研究では，音質を向上する研究のみならず，声質や感情などを再現・制御する研究も盛んに行われてきた．既存の手法では話者 ID [1] や話者属性 [2,3]，その他様々な観点による制御が行われていたが，制御の自由度も直感度も限定されていた．そこで，より自由で直感的な制御として，自由記述による制御が考えられる．自由記述による制御は画像生成 [4] を中心に各分野において成功し，社会に浸透し始めている．つまり，音声においても実現可能，かつ扱いやすく，多様な声を再現できると考えられる．読み上げ文と区別するため，以降は声質を制御するための自由記述を声質制御文と呼ぶこととする．そして，声質制御文による声の制御が可能な TTS を Prompt TTS と定義する．Prompt TTS の概要を Fig. 1 に示す．

Prompt TTS を学習させるためには，読み上げ文と音声に加え，声質制御文が含まれたコーパスが必要となる．しかしながら，そのようなコーパスは現存せず，また，多様な声質を含むコーパス構築論も確立されていない．本研究では，Prompt TTS 用コーパスである Coco-Nut を構築する方法論を報告する．

2 関連研究

2.1 画像生成用データセット

自由記述による画像生成モデルの学習には，画像およびその内容を説明するキャプションの対が必要である．DALL-E [4] は，MS-COCO [5] と Web データ [6] を併用している．MS-COCO は画像キャプショニングのためのデータセットであり，必要なデータ対を収録している．DALL-E では MS-COCO と豊富な Web データを使用することにより多様な画像を生成している．HTML 内の画像には alt タグが付随するため，Web からテキスト－画像ペアを比較的簡単に収集することができる。しかし，Web データにはノイズが多く，フィルタが必要がある．データのフィルタには事前学習済み CLIP (contrastive language-image pretraining) モデル [7] での対応度定量評価がよく用いられる．このように，データの多様性を確保すること，対照学習によりテキストとマルチメディアとの対応を定量化することは，テキストによる生成タスクにおいて意義がある．この事実は，声質制御でも成り立つと予想される．

2.2 環境音合成・楽音合成用データセット

自由記述による環境音合成においても，画像生成と同種のデータ対が要請される．代表例として AudioCaps [8] や Clotho [9] が挙げられる．さらに，CLIP の環境音版である，テキストと環境音の対照学習モデルの CLAP [10] をフィルタに使用することもある [11]．MuLan [12] による楽音生成では，Web 上の楽音付き動画を検索し，動画の説明文が楽音の説明になっているかを機械学習モデルで判定する．この考え方は楽音以外でも使用できる可能性が高い．これらの研究により，音の生成においてもデータの多様性や対照学習が重要であることが示唆されている．

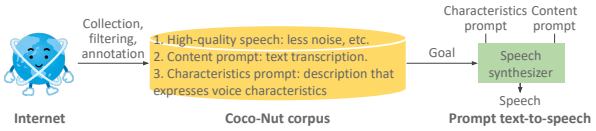

Fig. 1 Prompt TTS の概要

しかし，Prompt TTS に利用できるデータセットは現状限られている．内製の小規模コーパスに声質表現文を付与する例 [13,14] はあるものの，多くの TTS 用コーパス [15,16] は限定的な声質の音声しか含んでいない．より多様な声質を含むためには，**2.1** 節の例のように Web データを利用できるようにすることが重要である．また，声質表現文付きのオープンコーパスが存在していないことも問題である．以上のことから，より多様で大規模なコーパスを構築するための手法を確立する必要がある．Web 上の豊富なデータを利用して声質の多様なコーパスを作成し，オープンなコーパスとして整備することで，Prompt TTS の研究がより進展することが期待できる．

2.3 系列データの生成

音や動画等の系列データの生成では，系列全体を表す概念 (全体概念) と系列内で変化する概念 (変化概念) の両方を示す必要がある．テキストによりこれらを制御する方法には，両概念を区別せず 1 文で表現する方法と別々のテキストで表現する手法とで，大きく分けて 2 つある．両概念を別々のテキストで表現する手法の例として “bat hitting” (全体概念) と “ki-i-i-n” （変化概念）の入力による環境音生成 [17] 等がある．言語的な内容と声の特性を別々に制御する必要がある TTS にはこちらが適しており [13,14]，読み上げ文と声質制御文を別々に収集すべきである．

3 コーパス構築

本研究では日本語のコーパスを構築するが，多言語でも同様の手順で構築可能である．

3.1 構成要素

コーパスは以下の対データから構成される．

1. **音声：** TTS に利用できる，高品質な (例えばノイズの少ない) 音声．

* Coco-Nut: **Co**rpus of **co**nnecting **N**ihongo **u**tterance and **t**ext toward prompt-based control of voice characteristics, by Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Wataru Nakata, Detai Xin, Hiroshi Saruwatari (The University of Tokyo).

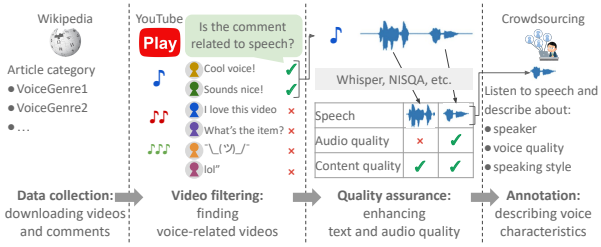

Fig. 2　コーパス構築手順の概要

2. **読み上げ文 (内容プロンプト)：** 音声の読み上げ文．**2.3** 節における変化概念に相当．
3. **声質制御文 (声質プロンプト)：** 声質の自由記述．感情等の発話スタイルも含む．**2.3** 節における全体概念に相当．

先行手法 [13,14] では，高品質の音声と読み上げ文から構成される既存の多話者 TTS コーパスに声質プロンプトを追加していた．しかし，**2** 節で説明したように，既存の多話者コーパス [15,16] においては話者数に限りがあり，声質の多様性が制限される．これを解消するため，Web 上のデータを利用する．

コーパス構築は Fig. 2 に示す 4 ステップで行う．

1. **データ収集：** Web 上の音声データを収集．
2. **動画フィルタ：** 「特徴的な声」を収集する．この「特徴的な声」は，Web 上でその声質について多く言及されている声を指す．
3. **品質保証：** 音と発話内容それぞれの品質を定量化し，低品質なデータを削除する．
4. **人手アノテーション：** 人手により声質プロンプトを付与する．

3.2　データ収集

前報 [18] と同様，対象言語 Wikipedia カテゴリの声に関連するフレーズを使用して，動画共有サイト（例えば YouTube）を検索する．ヒットした動画の ID，音，動画タイトル，視聴者コメントを取得する．

3.3　動画フィルタ

コメントに対するキーワードマッチングの後，機械学習に基づいて「声に関係する」コメントを抽出することで，特徴的な声を含む動画を選択する．

1. **ルールベースフィルタリング：**声に関係する単語セットを用意し，コメントに対するキーワードマッチングを行う．各動画でマッチしたコメント数が閾値以上であれば，当該動画を残す．
2. **機械学習による識別：**前報 [18] の機械学習を用いて，各コメントが声に言及しているかを判定する．また，単語サブセットによるキーワードマッチングを併用する．声に関係すると判定されたコメント数が閾値以上であれば，当該動画を残す．

3.4　品質保証

Web データには低品質なものが含まれるため，以下の処理によりデータ品質を保証する．

3.4.1　音声の品質保証

以下の処理で音声品質を保証する．

1. **音声区間検出 (voice activity detection: VAD)：**inaSpeechSegmenter [19] による VAD で，音声全体から音声セグメントを抽出する．
2. **音声抽出：**音源分離モデル Demucs[1] を使用し，音声セグメントから音声を抽出する．
3. **音質自動評価：**音声品質予測モデル NISQA [20] による定量的な音質評価を行う．各音声セグメントに対して NISQA スコアを計算し，閾値よりも低いスコアのセグメントを除外する．
4. **継続長および音量によるフィルタリング：**過度に長い・短い音声を取り除くため，2 秒から 10 秒の範囲の音声のみを残す．また、音量の閾値を設定し、音量の小さい音声を除外する．
5. **複数話者音声および歌唱音声の除去：**TTS に適していない，特に歌声や複数話者音声 (相槌など) を手動で判定し，除去する．
6. **声質多様性評価：**コーパスには多様な声質が含まれることが望ましい．これを実現するため，各音声セグメントの x-vector [21] を抽出し，その距離を利用してウォード法に基づく階層的クラスタリング [22] を行う．類似した声質のセグメントは同じクラスタに属すると考えられるため，各クラスタにつき 1 セグメントをランダム抽出する．

3.4.2　発話内容の品質保証

以下の処理で内容プロンプトを獲得し，発話内容の品質を保証する．

1. **発話内容の書き起こしと言語識別：**音声認識モデル Whisper [23] で発話内容を認識し，内容プロンプトを獲得する．同時に，Whisper を使用して発話言語を識別し，対象言語以外の音声セグメントを除外する[2]．
2. **不適切表現の削除：**内容プロンプトが公序良俗に反する場合，その音声セグメントを削除する．不適切表現の単語セットを用いたキーワードマッチングと，人手による除外を行う．
3. **非言語音声の削除：**TTS は非言語音声 (例えば叫び声) を扱わないため，当該セグメントを除外する．BERT [24] に基づく MLM スコア [25] を内容プロンプトごとに算出する．これはマスクされたトークンをそれ以外のトークンから予測する尤度であり，非言語音声の MLM スコアは高くなる[3]．MLM スコアが閾値よりも高い音声を除外する．

3.5　人手アノテーション

クラウドソーシングを使用し，音声セグメントに対して声質プロンプトを付与する．クラウドワーカは提示された音声を聞き，声質を自由記述する．記述には話者の属性、声質、発話スタイルを含めるように指示する．収集後，このコーパスで訓練されたモデ

[1] https://github.com/facebookresearch/demucs

[2] Whisper による言語識別だけでは対象言語以外の音声が含まれたため，人手の言語識別を行った．

[3] 例えば，「ああ [MASK] あああああ」と叫び声の一部がマスクされた場合，マスクされたトークン「[MASK]」は「あ」になることが容易に予測される．

ルが実在人物の名前から声を生成するのを防ぐため，固有名詞が含まれるものを手動で除外する．また，テキストの正規化も行う．

4　実験的評価

4.1　データ収集

データ収集は 2022 年 7 月から 2023 年 3 月までの期間に行った．詳細は前報 [18] を参照されたい．動画数は約 110 万個である．

4.2　動画フィルタ

キーワードマッチングには前報 [18] と同じ 8 単語を利用した．コメント数の閾値は 10 とした．機械学習による識別には，[18] で適合率上位 10 件として表示したモデルのうちタイトル併用の 7 つを使用し，いずれか 1 つのモデルにより「声に関係する」と判定されたコメントを声に関係するものとしてカウントした．コメント数の閾値は 10 件とした．

動画フィルタ適用後，1,523 件の動画が残った．

4.3　品質保証

VAD により，54,610 件の音声セグメントを取得した．NISQA スコアの閾値を 2.0 とした．音量の閾値を −55 dB とした．Whisper [23] には tiny モデルと large モデルの両方を使用した．これは，tiny モデルはより発話内容に忠実であり，large モデルはより文法的に正確であるためである．不適切内容の削除には MeCab[4] と不適切表現リスト[5] を使用した．非言語音声検出のための MLM スコア閾値は −0.01 とした．x-vector の抽出には xvector_jtubespeech[6] を使用した．11,000 のクラスタに分類した．選択後，公序良俗，使用言語，話者数，歌声か否かの観点で，さらに手動で識別した．最終的に 7,667 のセグメントが残り，継続長は合計で約 8 時間半となった．

4.4　人手アノテーション

クラウドソーシングには Lancers[7] を使用した．ワーカー 1 人あたり 10 セグメントを割り当て，報酬を 200 円とした．合計で 1,318 人のワーカーが評価を行った．

アノテーションの前に，学習，検証，評価セットへの分割を設計した．同一話者発話によるデータリークを避けるため，セット間で YouTube チャンネルに重複がないようにした．各セットにはそれぞれ 6,463、593、611 のセグメントが割り当てられた．

入力者ごとに記述が異なること考慮するため，既存の研究 [9,26] に倣い，検証および評価セットでは 1 セグメントに対し 5 件の声質プロンプトを付与した．

4.5　コーパス分析

構築したコーパスを，多様性を中心に評価する．

4.5.1　動画カテゴリ

YouTube で設定された動画カテゴリに基づき，音声セグメントをカテゴリごとに分類する．結果を Fig. 3 に示す．コーパスには 14 カテゴリが含まれており，幅広い分野を網羅している．上位 3 つのカテゴリ (エンターテイメント，教育，ゲーム) が約 70%を占めるが，科学と技術などのカテゴリも少数含まれている．

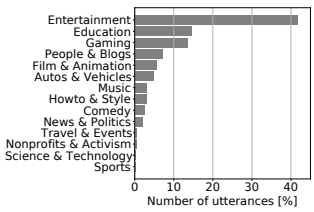

Fig. 3　動画カテゴリごとの音声セグメント数

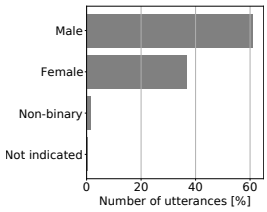

Fig. 4　性別ごとの音声セグメント数

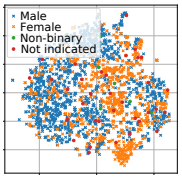

Fig. 5　性別ごとの x-vector 分布

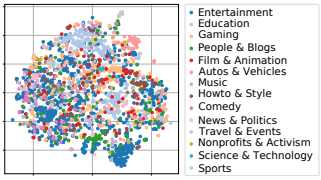

Fig. 6　動画カテゴリごとの x-vector 分布

4.5.2　性別の分布

話者の性別の多様性を分析するため，声質プロンプトに対して性別の手動アノテーションを行った。性別の分布を Fig. 4 に示す．大多数は男性または女性としてラベル付けされているが，non-binary および性別が記述されていない (not-indicated) プロンプトも存在する．Fig. 5 は各セグメントの x-vector を t-SNE で可視化し，性別で色づけしたものである．男性と女性はクラスタを形成しているが，non-binary と not-indicated はクラスタを形成せず，散在している．

4.5.3　カテゴリごとの声特徴分布

x-vector を t-SNE で可視化し動画カテゴリで色付けしたものを Fig. 6 に示す．エンターテイメントと教育のカテゴリにおいて，特に右下と中央上部の領域にクラスタが形成されていることがわかる．これは，各カテゴリに典型的な声の特徴があることを示唆している．一方，散布図の大部分では，顕著なクラスタは見られない，

4.5.4　収集声質プロンプトの頻出単語分析

声質プロンプトの単語頻度を品詞ごとに分析した．日本語で頻出する単語 (いる，する，ある等) と収集条件上頻出する単語 (性別，声，喋る等) は除外する．頻度を可視化したワードクラウドを Fig. 7 に示す．また，各品詞で頻度上位の「早口」「落ち着く」「若い」を含む例を Table 1 に示す．高頻度の単語は，コーパス内で頻出する声質特徴を示すと同時に声質表現で着目されやすい点を示す．例えば，名詞において「早口」の頻度が高いのは，YouTube ドメインにおいて早口の話者が多いこと，話速が着目されやすい観点であることを示している．また，複数の定性的特徴と関連する表現も頻出する．例えば「落ち着く」は「落ち着いた低い声」のように声の低さと，「落ち着いてゆっくり喋っている」と話速と結びつく．しかし「高い声で、落ち着いて」の例もあり，高くとも別の「落ち着いた」要素を含む声にも使用される．定性的特徴

[4] https://taku910.github.io/mecab/
[5] https://github.com/MosasoM/inappropriate-words-ja
[6] https://github.com/sarulab-speech/xvector_jtubespeech
[7] https://www.lancers.jp

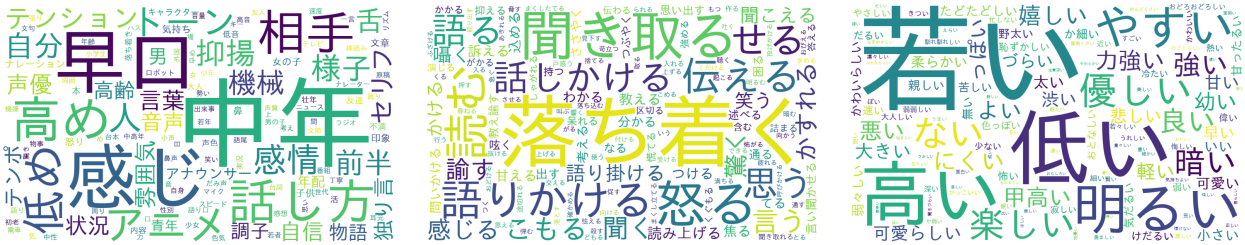

Fig. 7　声質プロンプトのワードクラウド (左から名詞・動詞・形容詞)

Table 1　声質プロンプトの例

単語	声質プロンプト
早口	中年の男性が、ハキハキした声で、早口で喋っている。 30 代の男性、会議で報告。早口口調です。 若い女性が少し慌てた様子で早口で話している。
落ち着く	30 代くらいの男性が落ち着いた低い声で慌てたように喋っている。 高齢の女性が低い落ち着いた声でゆっくり喋っている。 壮年の男性が、高い声で、落ち着いて説明するように喋っている。
若い	若い女性が若干低めの声で応答する感じでしゃべっている。 若そうな男性が、セリフのようなことをしゃべっている。 若い女性が少年のような口調で喋っている。

と関連しつつ完全に決定しない，評価者を跨ぎ頻出する表現から声質を予測する工夫が必要だろう．

5　まとめ

本研究では Prompt TTS 実現のため声質プロンプトを付与した Coco-Nut コーパスを構築した．今後 Coco-Nut を利用したモデルの構築を行う予定である．Coco-Nut はプロジェクトページ[8]にて配布する．

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．また，品質保証の人手評価において，研究室メンバーの佐藤 匡紀さん，山内 一輝さん，兵藤 弘明さん，武伯寒さん，中田 亘さんにご協力いただきました．

参考文献

[1] N. Hojo et al., “DNN-based speech synthesis using speaker codes,” *IEICE TRANSACTIONS on Information and Systems*, vol. E101-D, no. 2, pp. 462–472, 2017.

[2] D. Stanton et al., “Speaker generation,” in *Proc. ICASSP*, 2022, pp. 7897–7901.

[3] A. Watanabe et al., “Mid-attribute speaker generation using optimal-transport-based interpolation of gaussian mixture models,” in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

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

[5] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” *arXiv:1405.0312*, 2014.

[6] P. Sharma et al., “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *Proc. ACL*, Jul. 2018, pp. 2556–2565.

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

[8] C. D. Kim et al., “Audiocaps: Generating captions for audios in the wild,” in *NAACL-HLT*, 2019.

[9] K. Drossos et al., “Clotho: an audio captioning dataset,” in *Proc. ICASSP*, 2020, pp. 736–740.

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

[11] Y. Wu et al., “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” *arXiv:2211.06687*, 2022.

[12] Q. Huang et al., “MuLan: A joint embedding of music audio and natural language,” *arXiv:2208.12415*, 2022.

[13] Z. Guo et al., “PromptTTS: Controllable text-to-speech with text descriptions,” *arXiv:2211.12171*, 2022.

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

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

[16] 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.

[17] H. Ohnaka et al., “Visual onoma-to-wave: environmental sound synthesis from visual onomatopoeias and sound-source images,” in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[18] 渡邊亞椰 et al., “自由記述文による声質制御に向けた in-the-wild 文データ収集法,” **研究報告自然言語処理** *(NL)*, vol. 2023, no. 15, pp. 1–6, 2023.

[19] D. Doukhan et al., “An open-source speaker gender detection framework for monitoring gender equality,” in *Proc. ICASSP*. IEEE, 2018.

[20] G. Mittag et al., “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Proc. Interspeech*, 2021, pp. 2127–2131.

[21] D. Snyder et al., “X-vectors: Robust dnn embeddings for speaker recognition,” in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.

[22] J. H. W. Jr., “Hierarchical grouping to optimize an objective function,” *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[23] A. Radford et al., “Robust speech recognition via large-scale weak supervision,” *arXiv preprint arXiv:2212.04356*, 2022.

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

[25] J. Salazar et al., “Masked language model scoring,” in *Proc. ACL*, Jul. 2020, pp. 2699–2712.

[26] C. D. Kim et al., “Audiocaps: Generating captions for audios in the wild,” in *NAACL-HLT*, 2019.

[8] `https://sites.google.com/site/shinnosuketakamichi/research-topics/coconut_corpus`