社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

渡邊　亞椰† 　高道慎之介† 　齋藤　佑樹† 　猿渡　　洋†

† 東京大学大学院情報理工学系研究科　〒113–8656　東京都文京区本郷 7-3-1
E-mail: †aya-watanabe@g.ecc.u-tokyo.ac.jp, ††shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

あらまし　本稿では，音声を人工的に合成する音声合成タスクにおける，合成音声の声質を自由記述文で制御するための文データ自動収集法を提案する．自由記述による声質制御は，従来の制御法によりも汎用かつ複雑な声質を表現でき，また，昨今の言語モデルの影響を強く享受できると期待される．提案法ではまず，音声に関連すると思われる日本語動画とそのメタデータを自動収集する．次に，ルールと機械学習に基づいて，各動画の各コメントが声質や発話スタイルを表現するか否かを識別する．本稿ではその識別結果について報告するとともに，収集したコメント群について分析する．

キーワード　データセット構築、文章分類、機械学習、Web 応用、アノテーション、マルチモーダル、音声合成

Aya WATANABE†, Shinnosuke TAKAMICHI†, Yuki SAITO†, and Hiroshi SARUWATARI†

† Graduate School of Information Science and Technology, The University of Tokyo, Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8656 Japan.
E-mail: †aya-watanabe@g.ecc.u-tokyo.ac.jp, ††shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

1. はじめに

音声コミュニケーションにおいて話し手は，言語要素だけでなく音声要素（例えば，パーソナリティ，聴こえ，感情や態度）を聞き手に音声で伝達している．音声を人工的に合成するタスクである音声合成 (text-to-speech) では，音声要素をどのように制御するかが古くからの課題である．話者インデックス [1]，話者の属性 [2], [3]，母語 [4]，感情 [5]，パーソナリティ [6]，声質表現語 [7]，音声特徴量 [8] など，その制御軸は枚挙に暇がない．しかしながら，各軸は非常に狭くかつ単純な音声要素の範囲しか制御対象にとれず，その応用範囲は限定的である．

他方，text-to-{image, audio, music, video} 等の分野において合成対象メディア制御を自由記述文で制御する方式の発展が目覚ましい [9]〜[12]．自由記述文による制御は，複雑なメディア要素を扱えるポテンシャルがあり，また，今後も深化し続ける言語モデル [13][1]の恩恵を受けやすいメリットがある．前述した発展の背景には，foundation model からの転移学習や異メディア特徴量間の contrastive learning [10], [14] など機械学習の発展に加えて，ウェブ上に存在するダークデータ（利用可能性が未知の大量データ，in-the-wild データ）を収集し学習に利用可能にしたこと [15], [16] が挙げられる．自由記述文による制御は音声合成にも強い恩恵をもたらすと予想されるため，それに必要な機械学習論とデータ収集論の確立が必要である．

そこで本稿では，自由記述文を用いて音声要素を制御するための in-the-wild 文収集法を提案する．音声に関する動画候補及びその自由記述を YouTube から自動取得したのち，自由記述のうち音声に関する記述を抽出する．本稿ではその抽出結果と傾向分析について報告する．

(注1)：https://github.com/hwchase17/langchain/

2. 関連研究

2.1 テキスト-画像の対データの収集

Text-to-image の草分けとして知られる DALL-E [9] は MS-COCO [17] とウェブデータ [18] を学習に使用している．MS-COCO は，画像に対し人手で説明文を付与した画像キャプショニング用データセットである．ウェブデータのフィルタリングは，テキストと画像それぞれに対するフィルタリングと，それらの対データとしてのフィルタリングから成る．対データとしてのフィルタリングは，学習済み CLIP (contrastive language-image pretraining) モデル [14] などを用いた対応度定量化でも可能であり [16]，その対象として HTML の画像とそれに付属する alt 文を使用することが多い．ウェブデータの使用は多様な画像の生成に強く寄与することが知られている．

2.2 テキスト-音の対データの収集

Text-to-audio における学習データセットの典型例もキャプショニング用データセット [19], [20] である．音キャプショニング用データセットは画像のそれに比べて小規模であるため，

当該音を表す単語群 [21] から文を擬似的に生成する場合もある．また，CLIP のテキスト–音版と言える CLAP (contrastive language-audio pretraining) モデル [10] を用いてデータセットを作成する方法もある [15].

Text-to-music においては，MuLan [11] がウェブ上のミュージックビデオを利用する方法を提案しており，当該ビデオに付されているテキストが音楽を説明する文か否かを識別する機械学習モデルを作成している．この方法論は，楽音以外にも適用できる可能性が有る．

しかしながら，音声を対象としたデータベースは非常に限定的である[2]．小規模な内製データに記述文を追加した既存研究 [22], [23] は存在するが，Section 2. 1 に記述したウェブデータの貢献を鑑みれば，音声においても同様のデータ収集論の確立が必要である．

2.3 系列データのテキスト表現

映像や音などの系列データを生成するタスクでは，系列の全体を表す概念と系列の変化を表す概念をそれぞれ決定しなければならない．テキストを用いてこれらの概念を記述する方法は大きく 2 つ存在する．1 つ目は，両方の概念を単一のテキストで記述する方法である．動画生成の “Wooden figurine surfing on a surfboard in space.” [12] や楽音生成の “Hip-hop features rap with an electronic backing.” [11] がこれに類する例である[3]．この方法は，よりラフな記述のみから系列を生成する応用に適している．2 つ目は，それぞれの概念を別々のテキストで記述する方法であり，環境音生成の “bat hitting” (全体概念) と “ki-i-i-n” (変化概念) [25] や，動画生成の “A toy fireman is lifting weights” (変化概念) [26] がこの例である[4]．この方法は，系列をきめ細やかに制御する応用に適しており，発話スタイルと発話内容を別々に制御するケースの多い音声合成は，この方法を採用するのが好ましい [22], [23].

3. 提 案 法

3.1 コーパスの構成要素

自由記述文による声質・発話スタイル制御のために，以下の対データから成るコーパスを構築する．

（1） **声質・発話スタイルを表す文**：音声の要素のうち言語的な発話内容ではなく声質や発話スタイルを表した文（Section 2. 3 で述べた全体概念に対応）

（2） **発話内容を表す文**：音声の要素のうち言語的な発話内容を表す文（Section 2. 3 で述べた変化概念に対応）

（3） **音声**：上記の 2 つに対応する音声

これを構築する方法の一つは，(2) と (3) の対データを有する既存音声コーパス（例えば，音声認識用 [28], [29]，音声合成用 [30]）に対し，新たに (1) を付与する方法である．Section 2. 1 の整理を踏まえると，良質だが小規模なコーパスをこの方法で構築することに加え，ノイジーだが大規模コーパスの構築が必要である．そのようなコーパスの構築は，字幕・説明文・コメントを有する，YouTube などの動画サイトの利用が有効である．しかしながら，動画サイトから収集したテキストと音声は非常にノイジーであるため，(1)(2)(3) の各データ自体の質に加え，対データとしての質を評価・担保しなければならない．本節では，(1) のデータ自体の質に焦点を当て，データ収集法と洗練法を提案する．なお，(2)(3) それぞれのデータの質，および対データとしての質については，今後の予定である．

3.2 データ収集

3.2.1 動画検索フレーズの作成

動画サイトの検索エンジンに入力する検索フレーズを作成する．対象言語（本論文では日本語）の Wikipedia から音や声に関係するカテゴリを選択し，そのカテゴリに属する Wikipedia 記事のタイトルを検索フレーズとする．また，当該検索フレーズに関連すると思われる語との組も検索フレーズに追加する（例えば，「○○ (記事タイトル) 切り抜き」）．このフレーズを検索しヒットした動画 ID を取得する．

3.2.2 コメントの事前フィルタ

動画 ID からその動画に対するコメントを取得したのち，対象言語以外のコメントを削除する．また，コメントに対し字数制限を設け，極端に短いあるいは長いコメントを削除する．

3.3 声関係サブセット作成

声質や発話スタイルを表すコメントを豊富に含むコメント群（声関係サブセット）を自動作成するため，Section 3. 2 で収集したコメントに対しルールと機械学習を併用した処理を行う．これらに先んじて，機械学習の学習データとして用いるために，コメントに対する人手ラベル付けを行う．このデータ洗練はノイジーで大量のデータから正例の多い（すなわち，声質を表すコメントの多い）データセットの作成を目的としているため，最終的に得られるデータセットの正例の割合の高さを重視し，洗練の過程で多くの正例を除外しても構わないものとする．

3.3.1 単語を用いた動画単位フィルタリング

声に関係する単語を少数選定して動画単位フィルタとして用いる．このフィルタにより，選定した単語が 1 つでも含まれるコメントが一定件数以上ついている動画のみを残し，その動画についたコメントのみを使用する．

3.3.2 クラウドソーシングを用いた人手ラベル獲得

コメントの一部に対し，クラウドワーカにコメント及びそのコメントが付いた動画のタイトルを提示した上で，当該コメントが「声に関係する」「歌声に関係する」「そのどちらにも関係しない」のラベル付けを依頼し，人手ラベルを獲得する．動画タイトルを提示するのは，コメント内容が音声を指しているか否かを明確にするためである[5].

（注2）：音キャプショニングのデータセット [19], [20] の中に，環境音としての音声（すなわち，言語内容を強く指定しない音声）は含まれる．

（注3）：MusicLM [24] では，記述を一定時間（論文では 15 秒）で切り替えることで，やや微細な変化の制御を可能にしている．

（注4）：LAION-Audio-630K [27] では，非音声の環境音に全体記述を，音声の環境音に変化記述を用いている．

（注5）：例えば，「美しい」のコメントだけではそれが音声に関するか否かを判断できないが，「心安らぐ朗読」のタイトルも提示すれば，当該コメントが声に関すると判断できると期待される．

表 1　データ収集の結果

Retrieved item	Value
#categories	180 categories
#search-phrases	0.10M phrases
#videos found in the search	1.14M videos
Audio duration	0.30M hours
#comments	24.2M sentences

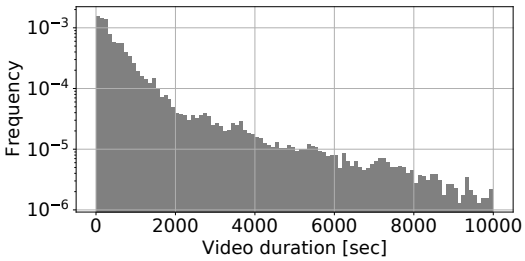

図 1　各動画の継続長のヒストグラム

3.3.3　単語を用いたコメント単位フィルタリング

Section 3.3.1 で使用したものと同じ単語群から一つ，あるいは複数の単語を選択し，選択した単語のいずれかを含むコメントのみを残すことでコメント単位のフィルタリングを行う．この際，すべての可能な単語の組み合わせについて比較・検討する．また，人名を含むコメントを削除する．これは，最終的に構築される音声合成システムが人名の入力からその人物の声色を合成しないようにするためである．

3.3.4　機械学習による「声に関係する」コメントの識別

声に関係するか否かを識別する機械学習モデルを学習する．人手ラベルのうち，「声に関係する」を正例，「歌声に関係する」「そのどちらにも関係しない」を負例とする，二値分類器を学習する．この分類器に用いる特徴量として，コメントのみを用いる場合と，コメントと動画タイトルを用いる場合について検討する．

4.　実験的評価

4.1　データ収集の結果

データ収集期間は 2022 年 7 月から 10 月であった．既存論文 [28] と異なり，字幕の有無に関わらず検索を実施した．収集時間コストを抑えるため，各動画について取得する最大コメント数は，「いいね」数の多い上位 100 コメントとした．使用するコメントは，3 文字以上 50 文字以下であり，かつ，平仮名あるいは片仮名を含むものとした．なお，本論文では使用しないが，付録 1. に示す情報も同時に収集した．

Table 1 に収集結果を示す．180 の Wikipedia カテゴリから約 10 万の検索フレーズを作成し，約 110 万の動画を取得した．動画の平均継続長は 0.27 時間，平均コメント数は 20.9 であった．継続長とコメント数のヒストグラムをそれぞれ Figure 1 と Figure 2 に示す．継続長は区分線形的に変化しているのに対し，コメント数は 0(すなわち，コメントなしの動画) を除き，動画間で極端な変化は見られない．

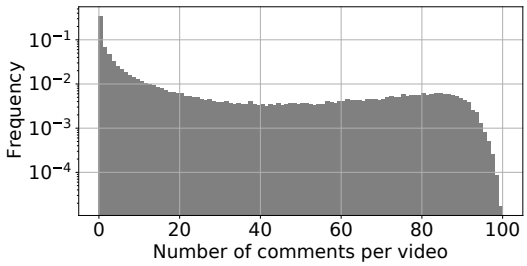

図 2　各動画のコメント数のヒストグラム

表 2　単語によるコメント単位フィルタの結果．“Total” は動画フィルタのみを通した後のコメント数．“Annotated” は，そのうち人手ラベルのついたコメントの数である．

Filter			Annotated		
人名除外	単語	Total	声関係	歌声関係	その他
	-	3163046	11647	16008	29838
✓	-	2349182	6765	8984	16704
✓	声	173665	928	1536	283
✓	ボイス	6786	75	46	18
✓	ヴォイス	89	1	0	0
✓	響き	2477	5	16	3
✓	音	69292	475	335	235
✓	聴	115025	393	812	441
✓	聞	134749	633	813	561
✓	歌	253347	438	2559	774

4.2　声関係サブセット作成の結果

動画単位及びコメント単位フィルタに用いた単語は「声」「ボイス」「ヴォイス」「響き」「音」「聴」「聞」「歌」の 8 種類とした．動画単位フィルタでは各動画についてこれらの単語のいずれかを含むコメントの数を集計し，それが 10 個を上回る動画 49486 件についた 3163046 件のコメントのみを残した．また，人手ラベル付けの被験者はクラウドソーシングプラットフォームであるランサーズ[6]を通じて雇用した．被験者の数は約 2000 人であり，各被験者は 30 個のコメントについて回答した．人名の除外には，形態素解析器 MeCab [31] と辞書 Neologd [32] を使用した．人手ラベル付け及び単語を用いたコメント単位フィルタの結果を Table 2 に示す．なお，このデータは人手ラベル付けの集計であるため，同一動画の同一コメントに対して複数回のラベル付けされた場合，同一コメントを重複して集計している．

人名除外を行いコメント単位フィルタを行っていない場合について，Annotated 合計の 32453 件に対して「声に関係する」と人手ラベルの付いたコメントは 6765 件であり，20.8% である．各コメント単位フィルタの結果を見ると，「響き」「ヴォイス」の 2 種については，該当するコメント数が比較的少なかった．特に「ヴォイス」は人手評価されたコメントが 1 件のみであり，単独で機械学習と評価をすることは不可能で，他の単語と併用した場合も効果は低いと考えられる．「響き」についても，後述する設計上学習に使えるコメントは 15 件程度であり，学習は難しい．そのため，この 2 単語については機械学習による識別実験の対象からは除外する．

(注6)：https://www.lancers.jp

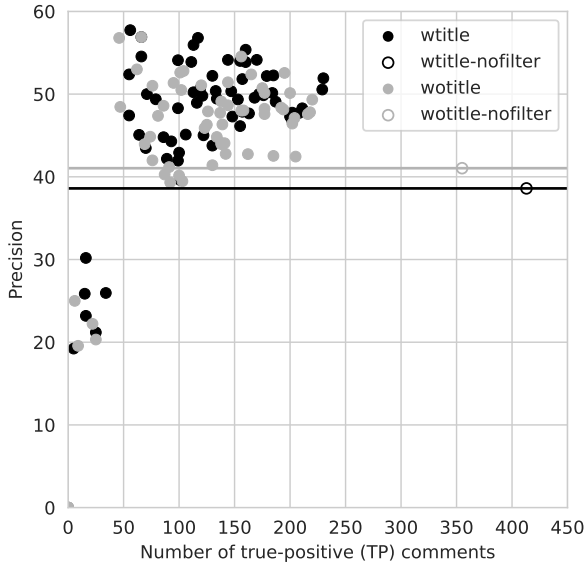

図 3　識別器の真陽性 (True positive, TP) ー適合率 (Precision)．各点は，コメント単位フィルタに用いる単語セットが異なる．wtitle, wotitle はそれぞれ，識別に動画タイトルを使用・不使用であることを表す．nofilter は，コメント単位フィルタ不使用を表す．

機械学習モデルは事前学習済みの BERT モデル[7]の出力層に全結合層 1 層を結合した 2 値分類器とした．コメント単位フィルタリングと機械学習の併用の影響を調査するため，フィルタリングに使用する単語セットを複数パターン用意し，各パターンに対して別々に機械学習モデルを学習する．単語セットのパターンは，「声」「ボイス」「音」「聴」「聞」「歌」から 1 6 個を採る全ての組み合わせとする．各パターンについて学習セット・検証セット・テストセットを 6 : 2 : 2 の割合で作成し，評価した．なお，エポック数は 3 であり，BERT のパラメータも更新した．BERT への入力に動画タイトルを使用する場合には，タイトルとコメントを特殊トークン [SEP] で接続した．コメントに対する事前の文正規化として，絵文字を含むコメントの削除，半角カナの全角化，全角空白とアットマークの除去を行った．識別器を学習した場合の真陽性と適合率の散布図を Figure 3 に示す．なお，参考として コメント単位フィルタを使用しない場合（Table 1 において，人名除外のみを施した場合に対応）の値を実線で示している．

また，適合率の上位 10 位となる単語及び動画タイトルの使用不使用について，真陽性と適合率の値，及び学習した識別器を用いてラベルのないコメント[8]について推論し「声に関係する」と識別されたコメントの数を Table 3 に，1 単語のみを用いたコメント単位フィルタにおける適合率を Table 4 に示す．比較としてコメント単位フィルタを使用しない場合の値も併記する．結果から，Table 2 から算出できるように人名除外時点で 20.8% だった声に関係するコメントの割合を「ボイス」「聞」によるコメント単位フィルタと動画タイトルを用いる機械学習モデルを使用することで，57.7 %まで改善できることが分かる．また，Figure 3 に示す通り，コメント単位フィルタを用いない場合の適合率が 40% 前後であるのに対し，コメント単位フィルタを利用すると 40% を概ね上回る．つまり，有効な単語セットであればコメント単位フィルタは有効であることが確認できる．タイトルの有無については，一部の例外（例えば，Table 1 の 8 位）が認められるものの，Figure 3 において動画タイトル使用の適合率が不使用のそれを概ね上回り，また，Table 3 の 7 つが動画タイトル使用であることから，動画タイトル使用による適合率改善傾向が見られる．Table 4 からは，単独の場合「音」「声」「聞」の適合率が高いこと，「聴」「歌」についてはタイトル不使用では適合率が 0.0% になってしまうもののタイトルを使用すれば 20% 前後の適合率は達成できること，「ボイス」についてはタイトルを使用しても適合率が 0.0% であることが確認できる．

(注7)：https://huggingface.co/cl-tohoku/bert-base-japanese-v2
(注8)：Table 2 の Total にあたる．なお，人名除外は適用しているものとする．

表 3　各単語の組み合わせによるコメント単位フィルタと併用して識別器を学習した際の真陽性及び適合率，及び Table 2 において人名除外のみを施したと示したコメント全てを識別した場合に「声に関連」と推論されたコメント数 (predicted positive)．適合率の順にソートしている．（適合率上位の 10 通り）

Word Filter	Title	TP	Precision [%]	Predicted Positive
ボイス，聞	✓	56	57.7	32159
音	✓	66	56.9	31091
音		66	56.9	29841
声	✓	117	56.8	57585
ボイス，聞		46	56.8	24994
音，聞	✓	113	55.9	52560
声，ボイス，音	✓	160	55.4	82622
声，ボイス，聞		156	54.5	79214
ボイス，音	✓	66	54.5	34044
声，音，歌	✓	170	54.1	85072
-	✓	413	38.6	285916
-		355	41.0	244303

表 4　1 単語のみを用いたコメント単位フィルタを併用した識別器のタイトル使用・不使用の場合における適合率 (Precision) [%]．

Word Filter	Without Title	With Title
声	52.8	56.8
ボイス	0.0	0.0
音	56.9	56.9
聴	0.0	25.9
聞	48.5	52.4
歌	0.0	19.2

4.3　収集コメントの分析

最後に，本稿で収集したコメントについて分析した結果を述べる．

4.3.1　収集コメント全体に対する分析

声についての感想を述べている場合，「好き」「良い」といった単純な好悪を述べるもの，「かっこいい」「かわいい」といった多分に主観的な表現など，表現方法に偏りがある．

4.3.2 人名除去についての分析

名字，フルネーム，ニックネーム，架空のキャラクタなどの人名を，その表記ゆれを含めておよそ除外できており，例えば，「○○（声優の名前）に声が似ている」などのコメントを除外できている．ただし，一般名詞と判別の難しいニックネーム（ジャガーさん，山ちゃん等）などは意図と反して残ってしまう．

4.3.3 「響き」「ヴォイス」を含むコメントについての分析

「ヴォイス」は「ハスキーヴォイス」や「ハイトーンヴォイス」等，声の特徴を示す形容詞と共起するケースが多く見られる．類似する単語である「ボイス」は，より一般的な表現であるからか，「ボイス販売」「ボイスサンプル」「ボイスドラマ」のように声の特徴に関係のない単語と共起する傾向にあるため，「ヴォイス」のほうが声を修飾する表現を集めやすいと考えられる．そのため，「ヴォイス」を含む時点でコメントを正例として扱うことも妥当であろう．ただし，「ヴォイス」を含むコメントは絶対数が少ないため，曲名やドラマのタイトルなどの該当単語を含む固有名詞が存在する場合，それらが無視できない量のノイズとなる可能性には留意したい．また，「響き」は「心に響く」に類する用例が多く，「魂に響く」等の類例を含めると，人手評価されたコメントの中では声関係コメントで 5 件中 4 件，歌声関係コメントで 16 件中 12 件，その他では 3 件中 2 件であった．つまり，「響き」を含むという基準でフィルタしたコメントを声に関係するコメントとして扱うことは難しいと考えられる．

4.3.4 識別器の適合率を上げるコメント単位フィルタについての分析

タイトルを用いる場合はタイトルを用いない場合よりも適合率が上がる傾向にあり，図 3 に示すタイトル使用 63 種のうち 43 種について上昇が確認される．また，「聞」と「聴」という意味の近い単語について，「聞」の使用が「聴」よりも適合率が高い傾向にある．特に，Table 4 より，コメント単位フィルタに「聞」のみを利用する場合と「聴」のみを利用する場合に顕著で，「聞」についてはタイトル有無に問わず 50% 前後であるが，「聴」はタイトルを使用して 20% ほど，タイトルを使用しないと適合率が 0% になる．これは，Table 1 に示すように「聴」は歌声に関係する場合が多く，これを負例として学習する際の学習難易度が向上するためだと推測される．

4.3.5 識別器による自動識別結果の分析

適合率上位の識別器のみの適用ではあるが，適合率約 50% を期待できる，最大で 8 万件以上のデータセットを作成することができると確認した．同一単語群によるコメント単位フィルタを用いた識別器のタイトル使用不使用についての比較では，一つの動画についているコメント群について，タイトル不使用では負例と識別されるものが多い場合であってもタイトル使用では全て正例と識別される場合，あるいはその逆（タイトル使用で負例だったものが不使用で正例）が散見される．タイトル使用不使用両方が適合率上位 10 位以内である，「ボイス，聞」および「音」をコメント単位フィルタに用いた識別器について，同一動画についた識別対象のコメント全てに同じ評価がついた場合の総数を Table 5 に示す．両方の場合について，10 件以上あるコメントが全て同じ識別結果となる例がタイトルを使用した場合に多く生じていることが分かる． この結果については，タイトルを直接入力に使い，明示的に補助入力であることを示さなかったことにより，タイトルに含まれるトークンが支配的に影響するようになったためである可能性がある．ただし，Section 3.3.2 で示している通り，正解ラベルとして利用している人手ラベルの収集では，タイトルとコメントのみを提示してのラベル付けを行っている．従って，正解ラベル自体がタイトルの強い影響を受けて作成されているのは自然であり，正解ラベルから学習したモデルがタイトルの影響を強く受けているのもまた自然であって，モデル構造上の欠陥とは断定できない．詳細な調査は今後の予定とする．

表 5 「ボイス，聞」および「音」について，それぞれの単語の組み合わせによるコメント単位フィルタを併用してタイトル使用・不使用について識別器を学習し Table 1 において人名除外のみを施したと示したコメント全体への識別を行った際，同一動画に対して全てが「声に関連」と推論される，または全てが「声に関連しない」と推論された場合の合計数．1 動画について識別したコメントの総数が 10 を超えているもののみ数える．Total は各コメント単位識別対象のうち，1 動画について識別したコメントの総数が 10 を超えているものの合計を示す

	With title	Without title	Total
ボイス，聞	729	492	1431
音	301	88	920

5. おわりに

本研究では，自然言語による声質及び発話スタイル制御システムを構築するための大規模学習データセットを動画サイトから自動収集する方法について検討した．提案法により，大規模なダークデータ群である YouTube コメントから，50% 程度声に関連する，つまり，声の対として使える文が含まれるサブセットを作成することができた．また，その文の内容について分析を行った．今後は獲得した文の対になる音声データを動画から取得する方法，文と音声のペアデータによる自動記述文制御可能な音声合成の学習プロセス等について検討する予定である．

謝辞：本研究は，JSPS 科研費 19H01116, 21H04900, 21H05054 の支援を受けた．

文　献

[1] N. Hojo, Y. Ijima, and H. Mizuno, “An investigation of DNN-based speech synthesis using speaker codes,” in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2278–2282.

[2] Daisy Stanton, Matt Shannon, Soroosh Mariooryad, RJ Skerry-Ryan, Eric Battenberg, Tom Bagby, and David Kao, “Speaker generation,” in *Proc. ICASSP*, 2022, pp. 7897–7901.

[3] Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Detai Xin, and Hiroshi Saruwatari, “Mid-attribute speaker generation using optimal-transport-based interpolation of gaussian mixture models,” *arXiv:2210.09916*, 2022.

[4] Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R.J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran, “Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning,” in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2080–2084.

[5] Rui Liu, Berrak Sisman, and Haizhou Li, “Reinforcement learning

for emotional text-to-speech synthesis with improved emotion discriminability," in *Proc. Interspeech*, Aug. 2021, pp. 4648–4652.
[6] Joakim Gustafson, Jonas Beskow, and Eva Szekely, "Personality in the mix - investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis," in *Proc. Speech Synthesis Workshop*, 2021, pp. 48–53.
[7] Kumi Ohta, Tomoki Toda, Yamato Ohtani, Hiroshi Saruwatari, and Kiyohiro Shikano, "Adaptive voice-quality control based on one-to-many eigenvoice conversion," in *Proc. Interspeech*, 2010, pp. 2158–2161.
[8] Tuomo Raitio, Ramya Rasipuram, and Dan Castellani, "Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features," in *Proc. Interspeech*, 2020, pp. 4432–4436.
[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," *arXiv:2102.12092*, 2021.
[10] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "CLAP: Learning audio concepts from natural language supervision," *arXiv:2206.04769*, 2022.
[11] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis, "MuLan: A joint embedding of music audio and natural language," *arXiv:2208.12415*, 2022.
[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans, "Imagen Video: High definition video generation with diffusion models," *arXiv:2210.02303*, 2022.
[13] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, "Training language models to follow instructions with human feedback," *arXiv:2203.02155*, 2022.
[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds. 18–24 Jul 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR.
[15] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," *arXiv:2211.06687*, 2022.
[16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," *arXiv:2210.08402*, 2022.
[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár, "Microsoft COCO: Common objects in context," *arXiv:1405.0312*, 2014.
[18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018, pp. 2556–2565, Association for Computational Linguistics.
[19] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "Audiocaps: Generating captions for audios in the wild," in *NAACL-HLT*, 2019.
[20] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, "Clotho: an audio captioning dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.
[21] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.
[22] Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan, "PromptTTS: Controllable text-to-speech with text descriptions," *arXiv:2211.12171*, 2022.
[23] Dongchao Yang, Songxiang Liu, Rongjie Huang, Guangzhi Lei, Chao Weng, Helen Meng, and Dong Yu, "InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt," *arXiv:2301.13662*, 2023.
[24] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "MusicLM: Generating music from text," *arXiv:2301.11325*, 2023.
[25] Hien Ohnaka, Shinnosuke Takamichi, Keisuke Imoto, Yuki Okamoto, Kazuki Fujii, and Hiroshi Saruwatari, "Visual onoma-to-wave: environmental sound synthesis from visual onomatopoeias and sound-source images," *arXiv:2210.09173*, 2022.
[26] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv:2302.01329*, 2023.
[27] Yusong Wu, Ke Chen, Tianyu Zhang, Marianna Nezhurina, and Yuchen Hui, "LAION-Audio-630K," 2022, `https://github.com/LAION-AI/audio-dataset`.
[28] Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe, "JTubeSpeech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification," *arXiv:2112.09323*, 2021.
[29] Seiji Fujimoto Yue Yin, Daijiro Mori, "ReazonSpeech: A free and massive corpus for Japanese ASR," in **言語処理学会 第 *29* 回年次大会**, 2023.
[30] Shinnosuke Takamichi, Ryosuke Sonobe, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.
[31] Taku Kudo, "Mecab: Yet another part-of-speech and morphological analyzer," *http://mecab. sourceforge. net/*, 2005.
[32] 奥村学 佐藤敏紀, 橋本泰一, "単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討," in **言語処理学会第 *23* 回年次大会** *(NLP2017)*. 2017, pp. NLP2017–B6–1, 言語処理学会.

付　　録

1. 収集したメタ情報

本論文では YouTube 動画のコメントのみを使用したが，以下の情報も合わせて収集した．これは，Section 2. 2 で述べたように，コメント以外の情報も利用できる可能性があるためである．

- チャンネル ID，チャンネル名，フォロワー数
- タイトル，説明文章，タグ，カテゴリ
- 再生回数，評価回数
- アップロード日，ライブ動画か否か