2023-09-28 日本音響学会第150回（2023年秋季）研究発表会　3-9-3

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 德泰, 猿渡 洋(東大院・情報理工)

概要

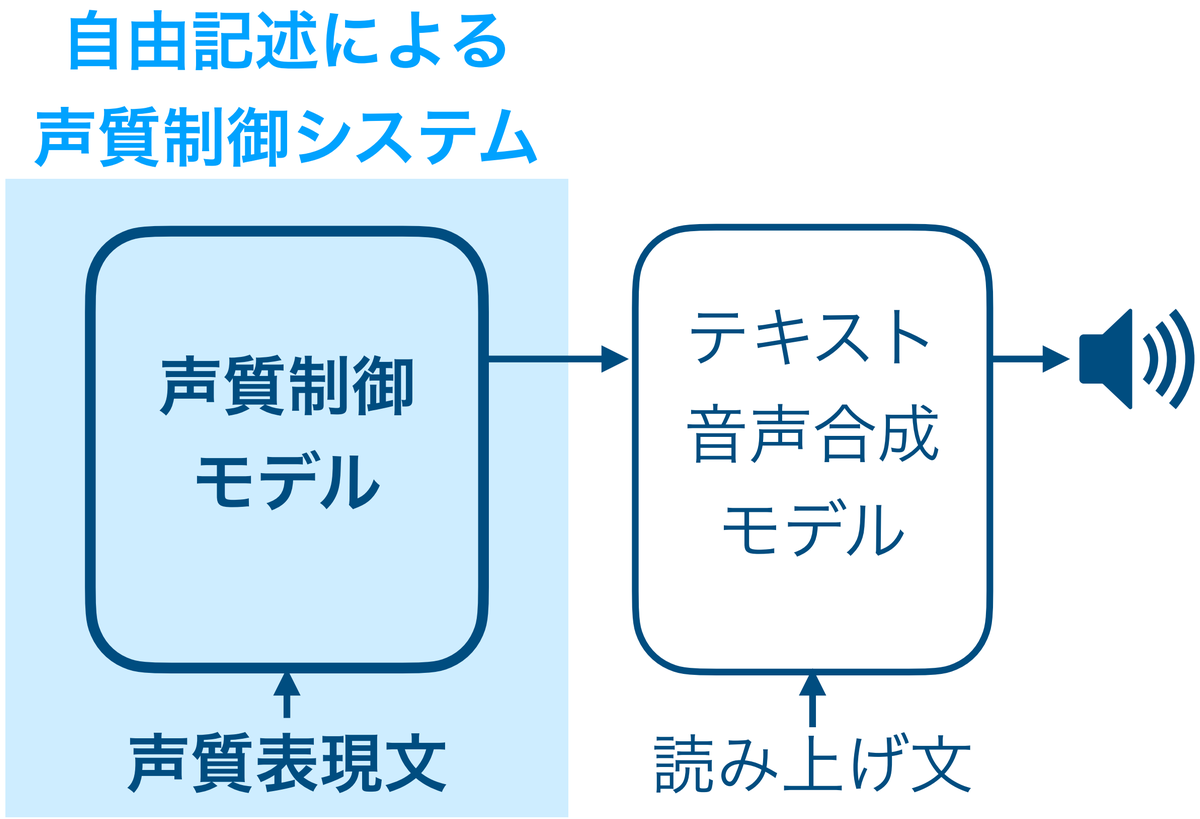
自由記述による声質制御の可能性

- 自由度が高い，直感的な操作を可能に
- 大規模でオープンなコーパスが必要

音声-声質表現文ペアコーパスの作成

- 多話者音声データの自動収集・書き起こし
- 音声への声質表現文付与フロー確立
- 約8000セグメント，8時間強の音声と声質表現文を含むコーパスが完成

近日公開予定



例: 大人びた低い声で落ち着いて話す，など

概要

ASRU版 (on arXiv)

- 対照学習モデルによる定量評価あり

関連研究

多話者・多スタイルテキスト音声合成における声質制御

- 話者選択[Hojo et al., 2017]，パラメータ（感情[Liu et al., 2021]等）調整: **低自由度**
- 自由記述による声質制御 [Guo et al., 2022, Yang et al., 2023]
  - 音声合成用に収録した少数話者コーパスがベース・非公開

    →多様な話者・誰でも研究に使えるコーパス（および構築手法）が必要

Text-to-Imageにおける多様・公開コーパスの貢献

- **MS-COCO** [Lin et al., 2014]
  - 人手アノテーションされたキャプション-画像ペアコーパス
  - 「80 object categories」で「>200K」の画像にキャプション付与
  - テキスト画像生成における利用 (e.g., DALL-E [Ramesh et al., 2021])

    →多様データに人手アノテーションされたコーパスの公開が研究を促進

コーパス構成要素

音声 + 書き起こし文（内容プロンプト） + 声質表現文（声質プロンプト）

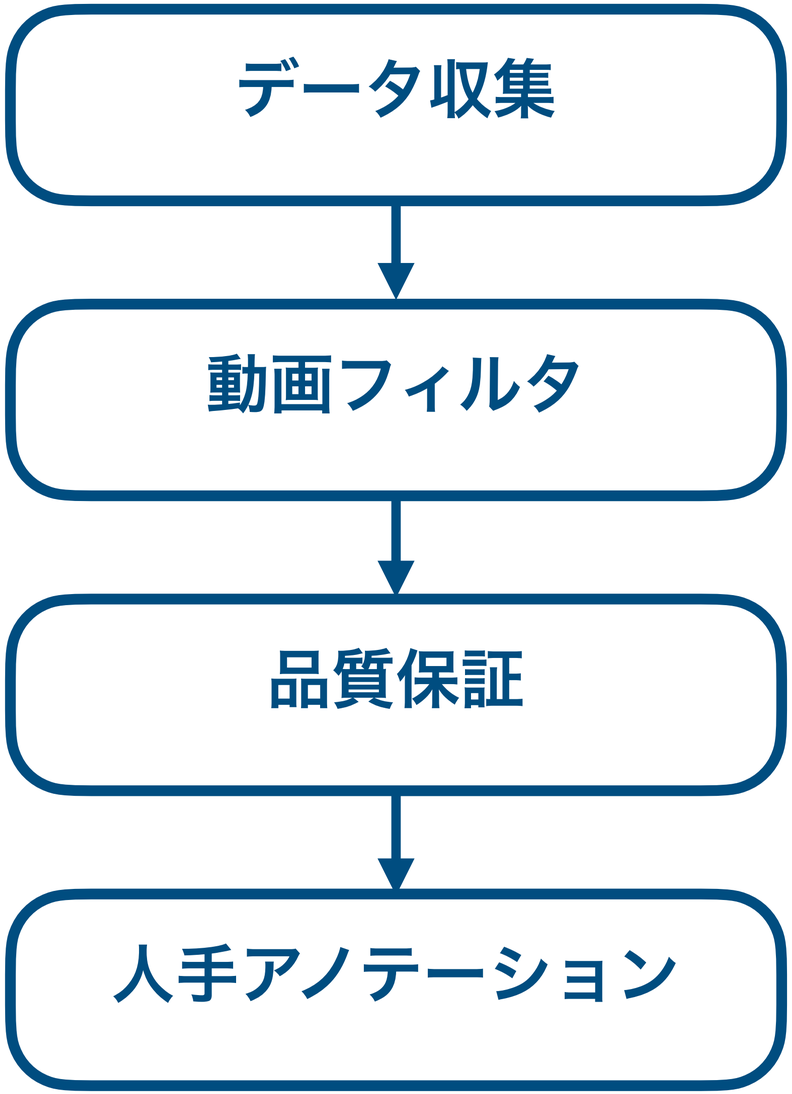
例:

募集させていただくこととなりました。今回はYouTubeなどのメディア露出可能な……

中年の男性が、わかりやすい口調で丁寧に話をしている。

コーパス構築手法

コーパス構築: 概要



- Web上の動画を**自動収集**（**多数**の話者の音声を収集）
- **声に関係する**キーワードで動画検索

- **音声を含まない・音声を含むが特徴の薄い声**の動画を除外
- 声に関係する視聴者コメントの数が基準

- 自動収集した音声を音響的・言語的に品質保証し，**コーパスの質を担保**

- 収集音声に**声質表現文**を付与
- クラウドソーシングを通じた**不特定多数からの声質表現文収集**フロー確立

コーパス構築: データ収集・動画フィルタ

声に関係するフレーズを用いたYouTube検索

- Wikipedia内声関係カテゴリ記事を使用したフレーズ収集 [渡邊 et al., 2023]

声関係コメント識別 [渡邊 et al., 2023] に基づく声関係動画抽出

- 声に関係するコメントが付く ＝ （言及したくなる特徴のある）音声を含む
- 一定数以上の「声に関係する」**視聴者コメント**が付いた動画のみ使用
- 「声に関係する」コメントの自動識別 [渡邊 et al., 2023]

コーパス構築: 品質保証

音声品質保証: 音としての品質保証

- 音声区間セグメント検出・分割，音声強調，低音質音声除去，短すぎる/長すぎる音声除去，複数話者音声除去，歌声除去，声質多様性最大化による発話選択
  - 声質多様性最大化による発話選択: x-vector [Snyder et al., 2018] による声質ベクトル化に基づき，ウォード法に基づく階層的クラスタリング [Ward Jr. et al., 1963] のあと各クラスタからランダムに発話を選択

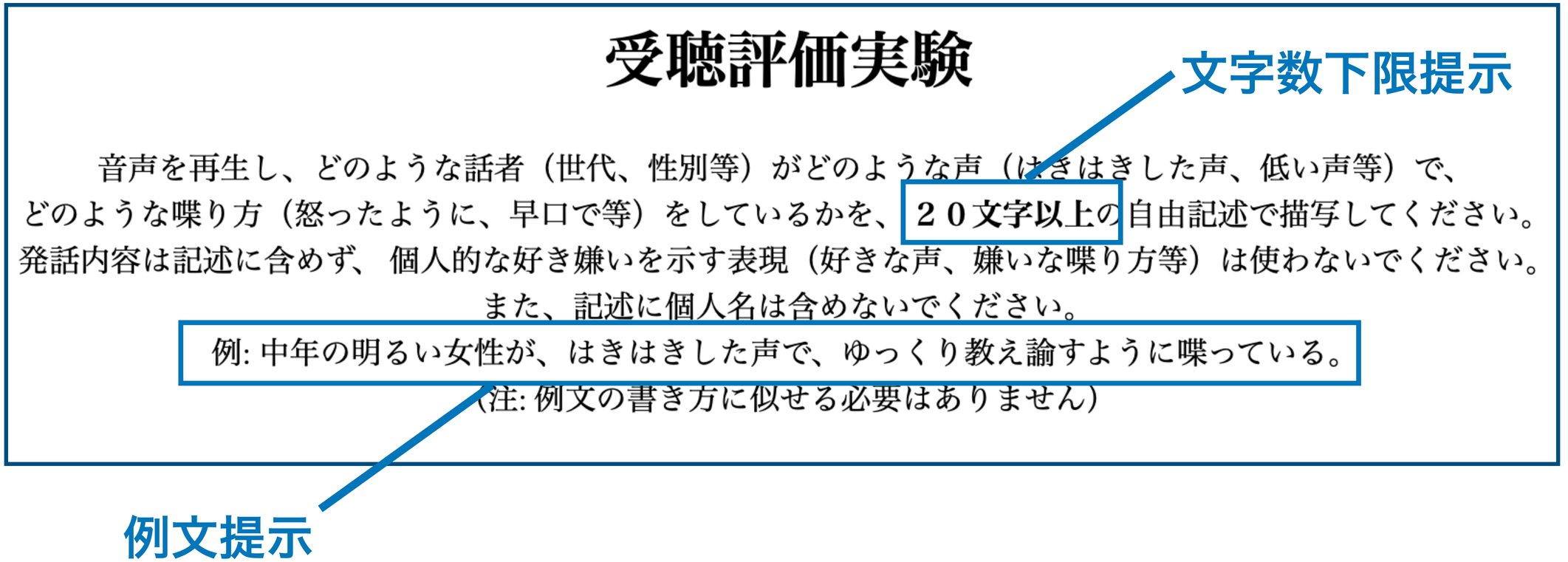
内容品質保証: 発話内容の品質保証

- 書き起こし: 自動音声認識 + 手動修正
- 非対象言語音声除去，不適切表現（NSFW）除去，ノンバーバル音声除去

音声・内容両側面からの品質保証により，自動収集した音声の質を保証

コーパス構築: 人手アノテーション

クラウドソーシングによる声質表現文収集



コーパス分析

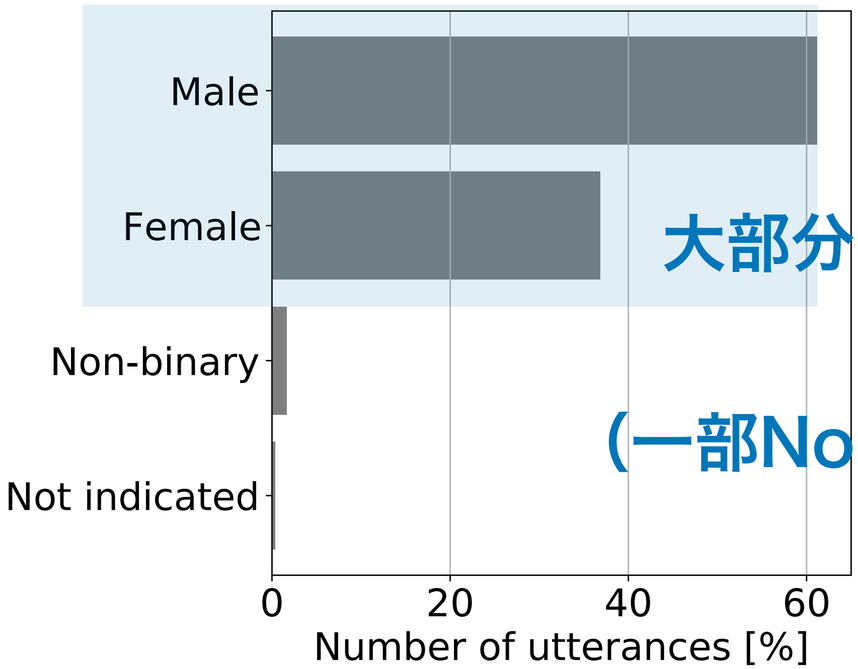
コーパス分析: データ収集概要

検索による獲得動画数	約110万 [渡邊 et al., 2023]
動画フィルタ後の動画数	1,523
セグメント数	7,667
合計継続長	約8時間半
声質表現文付与数/音声	学習セット: 1 / 検証, 評価セット: 5
継続長範囲	2-10秒

コーパス分析: 性別の分布

声質表現プロンプト内での「男」「女」の表記に基づき分類・集計

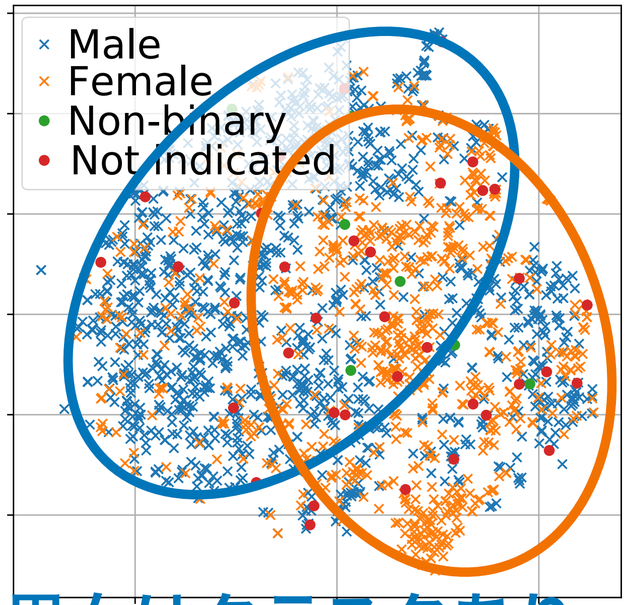
両方の表記があればNon-binary, 両方なければNot indicated



大部分が分類可能

（一部Non-binary等）

性別ごとの音声セグメント数



男女はクラスタあり

（Non-binary等はクラスタ無く分布）

性別ごとのx-vector分布（t-SNEで可視化）

コーパス分析: 声質表現文例

	30代くらいの男性の声。ゆっくりと穏やかな話し方でした。苦悩に満ちた、けだるそうな声でした。
	元気な男性が明るい声で、テンション高く発表をするように喋っている。
	若い女性が、抑揚のある声で、ゆっくりと喋っている。

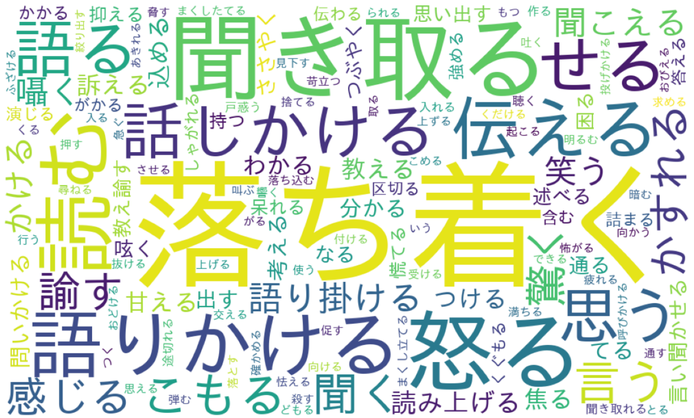
コーパス分析: 収集声質プロンプトの頻出単語分析

単語の出現頻度をワードクラウドで図示（大きく書かれた単語ほど頻出）

頻出単語は頻出属性・認識されやすい要素を示す

名詞

形容詞



動詞

- **YouTubeドメインには早口の話者が多い**
- **話速は声質の中でも認識しやすい要素である**

まとめ

自由記述で声質コントロール可能なテキスト音声合成のためのコーパス作成

- 音声-声質自由記述文の対
- 公開されており，多様な声質の話者を含み，人手による（=自動収集ではなく，質が高い）自由記述が付与されたコーパス

YouTubeでの音声データ収集 + 自動・人手品質保証

- 声に関係する動画の選別
- 音声・内容両側面からの品質保証
- 人手による自由記述付与

- 今後，本コーパスを用い，声質表現文で声質制御可能なテキスト音声合成モデルの構築を行う予定

Appendix: 動画フィルタの設定

声関係コメント数閾値	10件
キーワードマッチング 使用単語	声，ボイス，音，聴，聞，歌
タイトル併用	あり
BERTベース識別機使用 単語組み合わせ	「ボイス・聞」「音」「声」「音・聞」 「声・ボイス・音」「ボイス・音」「声・音・歌」

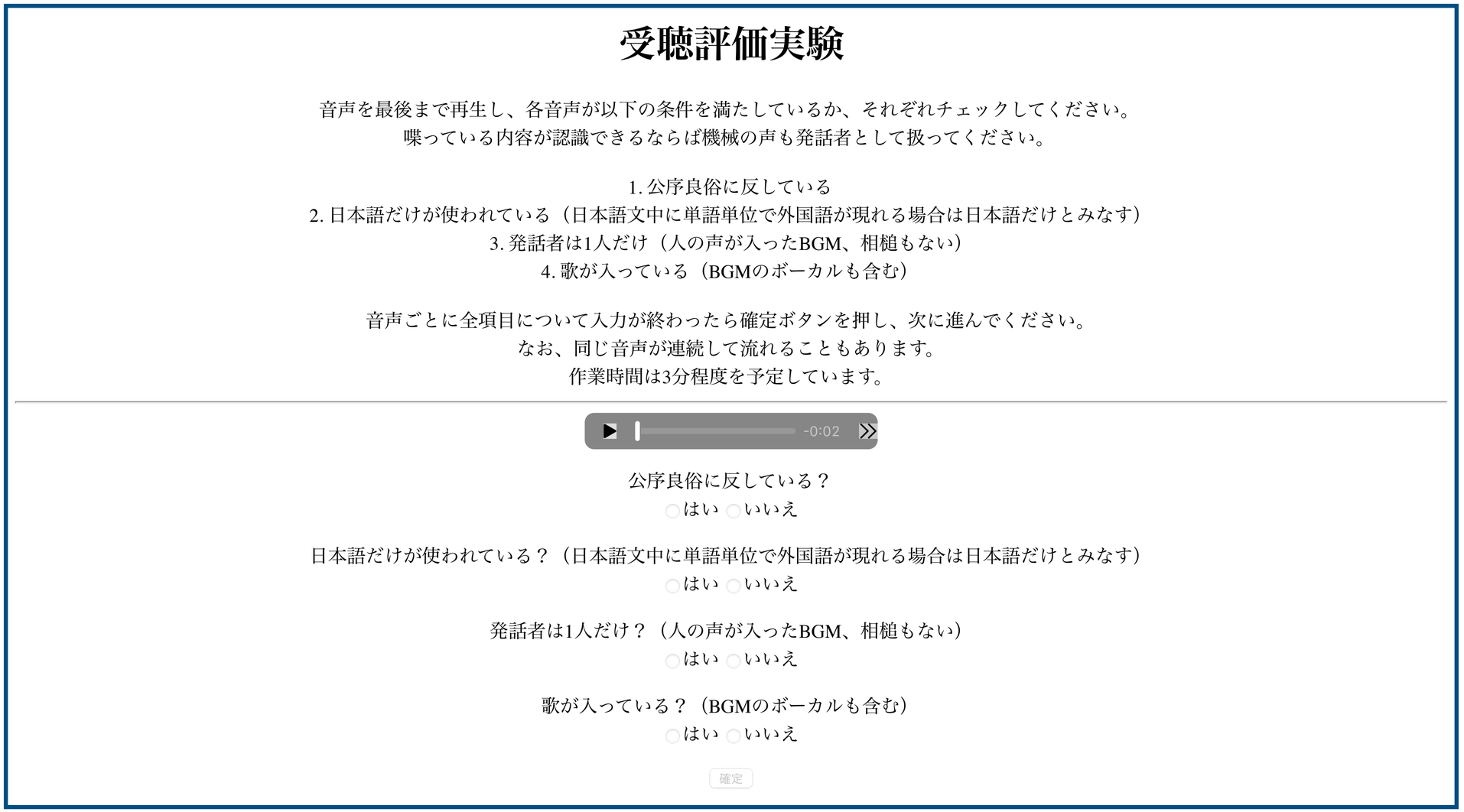
Appendix: 音声品質保証

- 音声区間検出・セグメント抽出 : inaSpeechSegmenterによるVAD
- 音声強調 : Demucs
- 低音質音声除去 : NISQAによる自動MOS評価
- 継続長フィルタ : 2~10秒
- 低音量音声除去 : 正規化後の音量-55db
- 複数話者音声除去 : 人手品質保証
- 歌声除去 : 人手品質保証
- 声質多様性最大化による発話選択 : scikit-learnを使用

Appendix: 内容品質保証

- 非対象言語音声除去
  : Whisperによる使用言語推定
  ＋ 人手品質保証
- 不適切表現（NSFW）除去
  : NSFW辞書によるキーワード検索
  ＋ 人手品質保証
- ノンバーバル音声除去
  : BERTによるMLMスコア評価
  （ノンバーバル音声はMLMスコアが極端に高くなる傾向がある(e.g., “あああああ”)）

Appendix: 人手品質保証のためのクラウドソーシングUI



Appendix: 人手品質保証のためのクラウドソーシング設定

プラットフォーム	Lancers (https://www.lancers.jp/)
評価する音声ファイル数 /ワーカー	10
予想所要時間 /ワーカー	3分
報酬 /ワーカー	¥60

Appendix: 人手アノテーションのためのクラウドソーシング設定

プラットフォーム	Lancers (https://www.lancers.jp/)
評価する音声ファイル数 /ワーカー	10
予想所要時間 /ワーカー	10分
報酬 /ワーカー	¥200

Appendix: 書き起こし文の人手修正例

- 自動書き起こしはWhisperで行う
- Typo，Whisperモデル構造由来のループ，切れている部分の自動予測されてしまった箇所などを人手修正

グリーン車は足元の**重量**乗車口4号車と
5号車でお待ちください

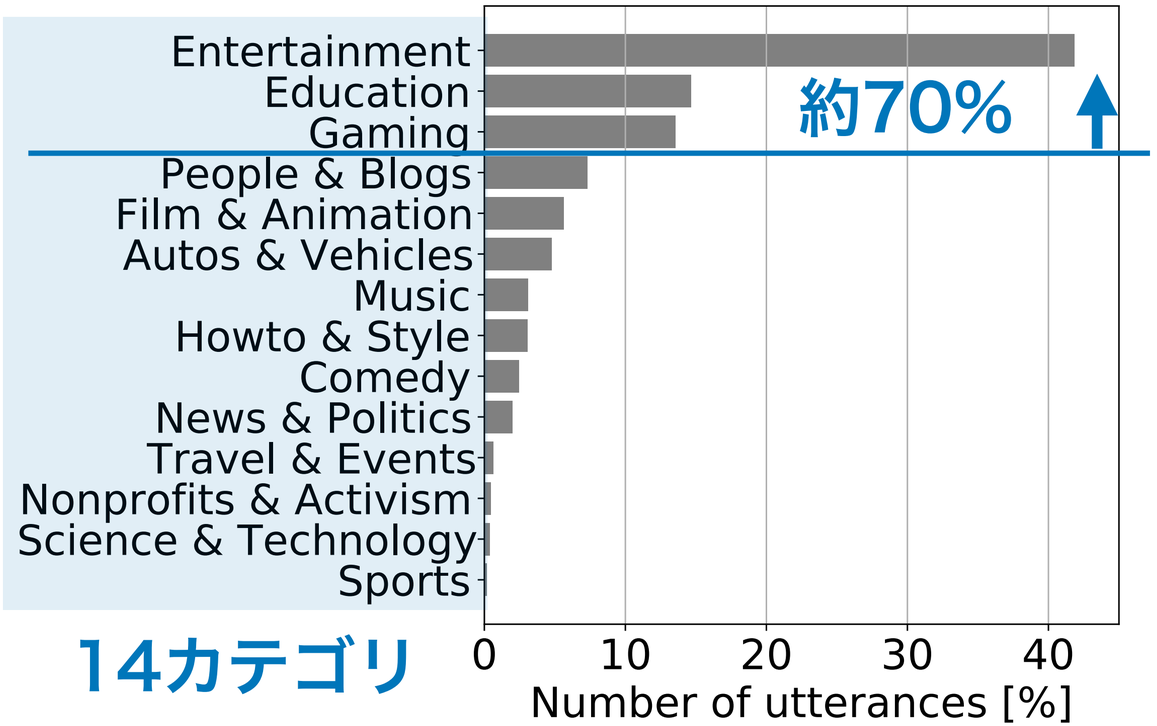
→

グリーン車は足元の**10両**乗車口4号車
と5号車でお待ちください

Appendix: 動画カテゴリの分布

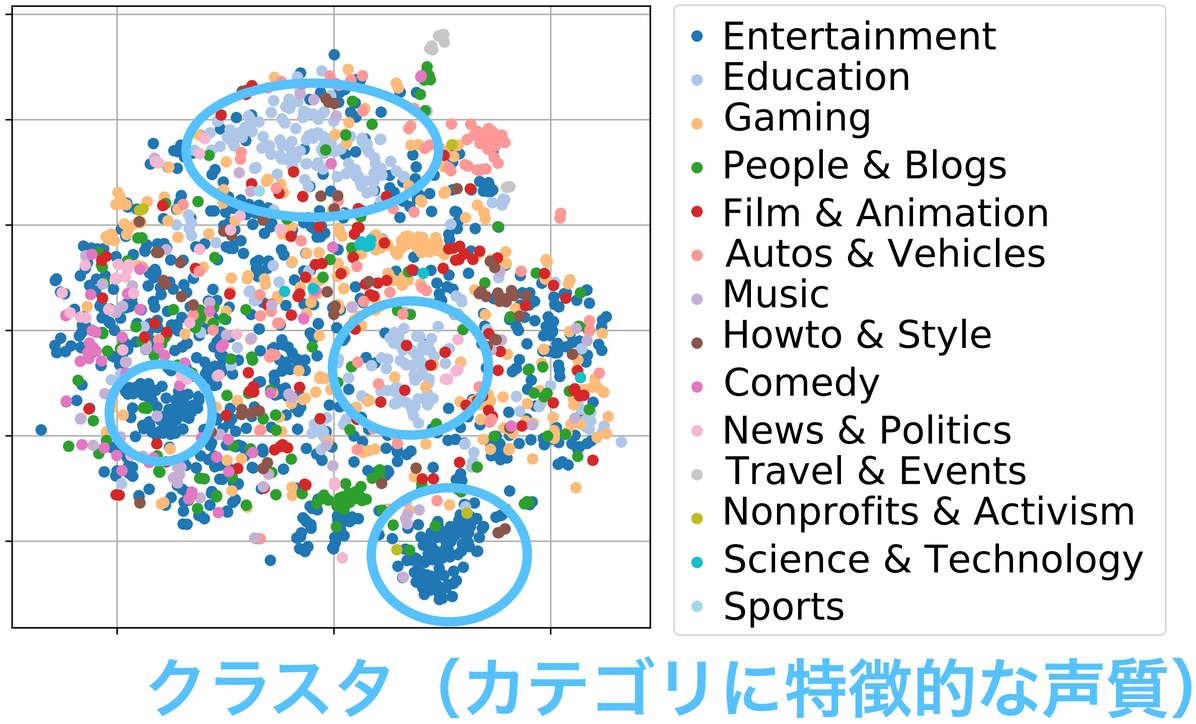
YouTubeで設定されている動画カテゴリに基づき分類し集計

カテゴリごとの声質の特徴をx-vectorによって定量化し，t-SNEで可視化



14カテゴリ

動画カテゴリごとの音声セグメント数



クラスタ（カテゴリに特徴的な声質）

動画カテゴリごとのx-vector分布