Coco-Nut: Corpus of Japanese Utterance and Voice Characteristics Description for Prompt-based Control

Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Wataru Nakata, Detai Xin, Hiroshi Saruwatari
(The University of Tokyo, Japan.)

Summary: Novel corpus for prompt-based control over voice characteristics on text-to-speech (TTS)

Various voice characteristics TTS

- Existing voice characteristics controllable TTS
  - Speaker ID [1], speaker attributes [2, 3], etc.
    - **Can only choose small range of voice!**
- **Potential of prompt-based control [4, 5]**
  - Control over wider range of characteristics

Issue with existing corpora [4, 5]

- Only a few speaker: lack of variation
  - Based on existing TTS corpora
    - The data size per speaker is more prioritized than the number of speakers.
- Internal: low reproducibility

→ Need of a public corpus with diverse speakers

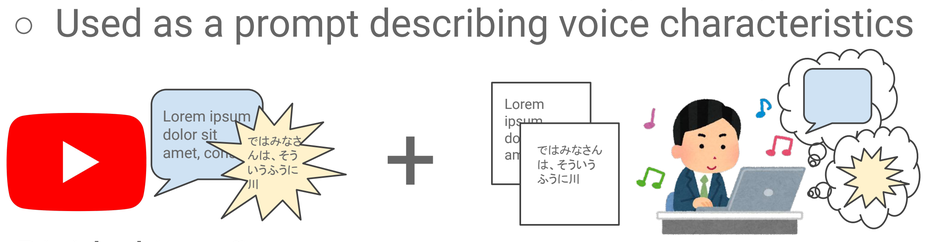
Novel corpus to promote research

- **Corpus construction method**
  - Based on speech collected through web
    - Diverse, but noisy
      - → Quality assurance is necessary
- **Coco-Nut: based on proposed method**
  - Diverse voice characteristics
  - Public (QR code is at the bottom.)

Corpus of connecting *nihongo* (= "Japanese" in Japanese) utterance and text and its construction method

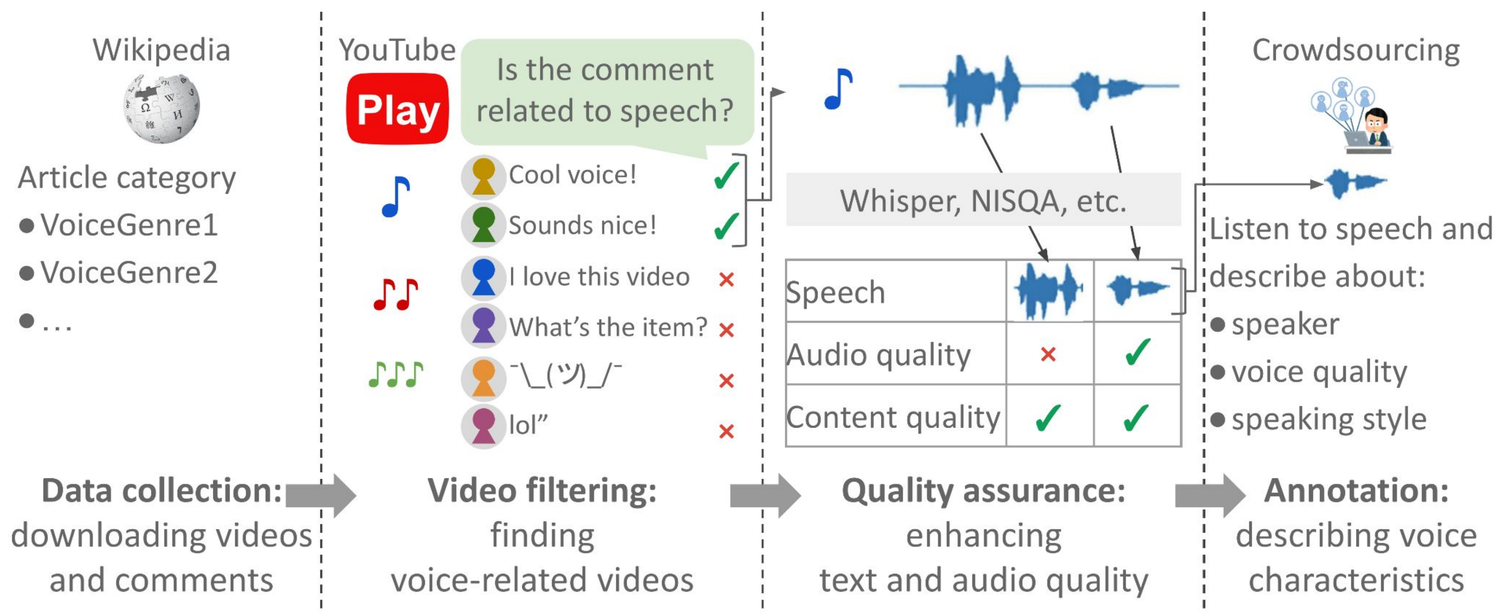
Corpus specification

- **Speech segments**
  - Approximately 8 hours, 8000 segments
- **Transcriptions**
  - With manual correction
- **Voice characteristics descriptions (prompt)**
  - Used as a prompt describing voice characteristics



- Divided into 3 sets
  - train:validation:test = 8:1:1
  - 1 description / speech for train, 5 for validation/test

Construction method overview



Data collection + video filtering

- **YouTube search with voice-related keywords**
  - Wikipedia article titles of voice-related categories
- **Video filtering with voice-related comment detection**
  - Viewers tend to comment on attractive voices.
  - BERT [6]-based classification to viewer comments
    - Trained on a few manual annotations

Quality assurance

- **Acoustic quality assurance**
  - Speech enhancement, perception-aware quality scoring, x-vector [7]-based data selection, etc.
- **Content quality assurance**
  - Based on transcriptions by Whisper [8]
  - Remove non-target lang., NSFW and non-verbal

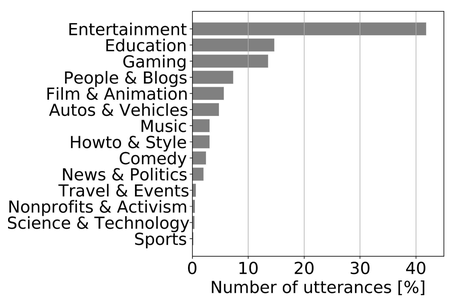
Annotation

- **Add descriptions on speech**
  - Using crowdsourcing
    - Workers listen to speech

ex) "A male voice in his thirties. He spoke slowly and calmly."

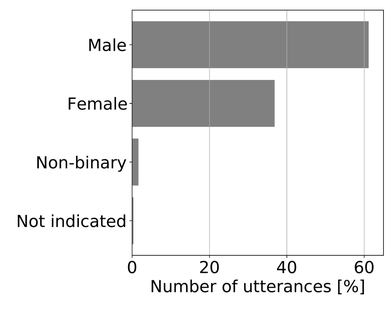
Corpus analysis

- **Video categories**



Based on video categories whom speech derived from (defined by YouTube)

- **Speech gender (referred description)**



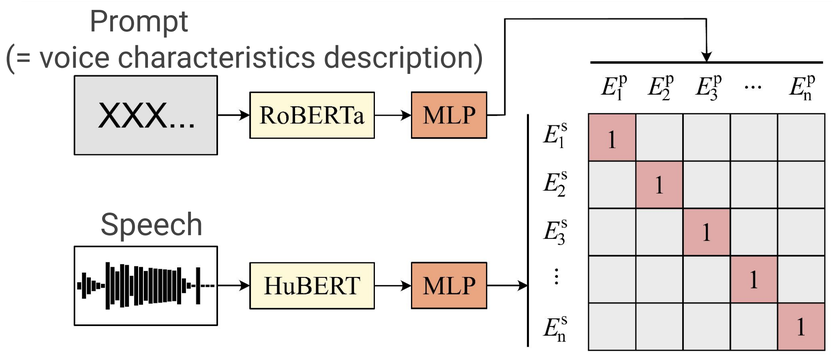
Non-binary means both "male" & "female" appear in the description.

- **Wide range of speech**
  - Multiple categories
  - Including non-binary voices

Machine learning baseline for prompt-based speech retrieval

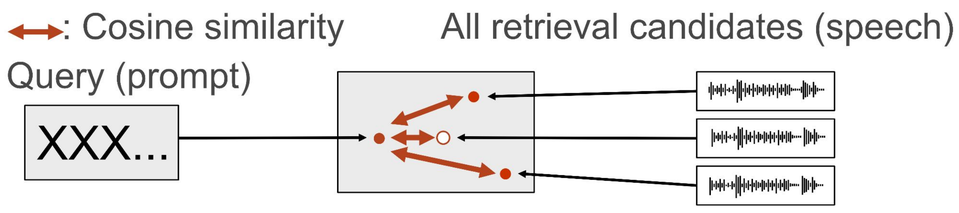
- Prompt-based speech retrieval can be considered as one step prior to prompt-based TTS.

CLAP [9]-based embedding model



Speech retrieval from descriptions

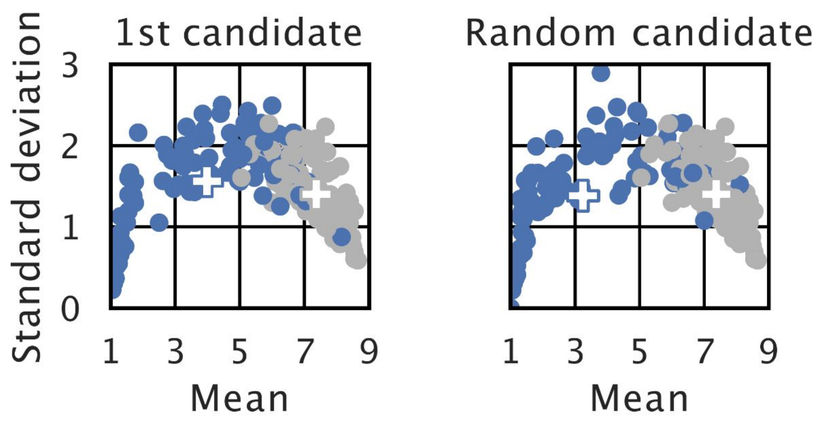
- Retrieve close speech to the query prompt in embedding space.
  - On cosine similarity



Subjective evaluation: degree of relevance of retrieved speech

- **How well do the retrieved speech match subjectively to the query prompt?**
  - Retrieved candidate, random, ground truth
  - Query: 100 prompts (ground truth speech exist in candidates)
  - Ask 20 crowdworkers per 1 prompt-speech pair, with a scale of 1 to 9.

Subjective evaluation result



● :mean/std of scores for each pair
✚ :global mean/std
Blue: mentioned in the title of each figure
Gray: ground truth (common across figures)

✚goes right → better scores in average
✚goes up → score varies among annotators.

- 1st > random → ability to learn voice characteristics space relative to human perception
- 1st << ground truth → room for improvement e.g., model architecture, training strategy

Project page

- Distribution pages are linked from here.
  - Downloading speech requires registration.

Reference
[1] Hojo et al., IEICE TRANSACTIONS 2017.
[2] Stanton et al., ICASSP 2022.
[3] Watanabe et al., ICASSP 2023.
[4] Guo et al., ICASSP 2023.
[5] Yang et al., arXiv 2023.
[6] Devlin et al., arXiv 2018.
[7] Snyder et al., ICASSP 2018.
[8] Radford et al., arXiv 2022.
[9] Wu et al., ICASSP 2023.

MOONSHOT RESEARCH & DEVELOPMENT PROGRAM
AVATAR SYMBIOTIC SOCIETY