Constructing an In-the-Wild Spoken Dialogue Dataset Based on YouTube Dialogue Videos

Yuki Sato*, Sanae Yamashita*, Shinnosuke Takamichi†‡, Ryuichiro Higashinaka*
* Nagoya University, Japan. higashinaka@i.nagoya-u.ac.jp
† Keio University, Japan.
‡ The University of Tokyo, Japan.

***Abstract*—The development of a spoken dialogue system requires a large-scale spoken dialogue corpus. However, existing Japanese spoken dialogue corpora are limited in size. In this study, we propose a method for constructing a large-scale Japanese spoken dialogue dataset by collecting dialogue videos from YouTube. Since YouTube hosts videos across various genres, recording environments, and speakers, it provides diverse spoken dialogue data that can be utilized for developing spoken dialogue systems. The constructed dataset primarily consists of two-speaker dialogues but also includes many multi-party dialogues involving five or more participants. Additionally, it contains only a small amount of non-dialogue speech, making it suitable for the development of multi-party dialogue systems. Furthermore, since the dataset includes many relatively long dialogues, it is expected to be useful for long-duration dialogue applications.**

I. Introduction

Spoken dialogue systems are expected to provide humans with a natural and convenient interface [1]. To build such systems, not only automatic speech recognition (ASR), spoken language understanding (SLU), natural language generation (NLG), and text-to-speech synthesis (TTS) are required but also more advanced modules such as turn-taking [2], voice activity projection [3], backchanneling [4], [5], and non-verbal vocalization [6] need to be taken into account.

Because most modules today consist of deep learning models, a corresponding spoken dialogue corpus is needed. Traditional corpora such as Switchboard [7], Fisher [8], Callhome [9], and the Corpus of Everyday Japanese Conversation (CEJC) [10] exist, but they remain too small in scale for unlocking the full performance of data-hungry deep learning models. Moreover, the recent emergence of end-to-end spoken dialogue models [11]–[14], which directly map audio input to audio output within a unified architecture, makes the limited scale of existing corpora an even more pressing problem.

On the other hand, looking at the development of various speech technologies (e.g., ASR and TTS), large-scale corpora created by collecting *in-the-wild* data have made significant contributions [15]–[17]. For instance, Whisper-v2[1] collects a massive amount of in-the-wild data and then quantitatively assesses the degree of correspondence between transcription and audio to acquire large-scale data for training ASR models. If a methodology can be established for collecting *in-the-wild dialogue-oriented data*, it could greatly accelerate the (pre-)training of various spoken dialogue models [18].

[1]https://github.com/openai/whisper

In this study, we propose a method for automatically constructing a spoken dialogue corpus from YouTube videos and build a large-scale Japanese spoken dialogue dataset. Among the many types of videos on YouTube, such as conversations, monologues, music, and nature, we identify those that contain dialogue and extract the specific time segments where dialogue occurs. The methodology consists of 1) classifying videos into dialogue or non-dialogue at the video level and 2) identifying dialogue or non-dialogue segments at the time-segment level. In the experiments, we verify the effectiveness of the proposed method and provide an analysis of the collected data.

II. Related work

A. Spoken dialogue corpus

Some languages already have several spoken dialogue corpora. For example, there are the Switchboard [7] and Fisher [8] corpora in English and HKUST for Chinese [19]. For Japanese, there are Callhome JPN [9], CEJC [10], and the Tabidachi travel dialogue dataset [20]. Although these corpora were recorded under the supervision of researchers and thus offer high-quality data, each one is limited to specific dialogue tasks or acoustic environments.

Some studies have attempted to build spoken dialogue corpora from in-the-wild data. For example, Siegert [21] use user-agent dialogue audio recorded via Amazon Alexa, defining those segments with minimal variation in speaker embeddings as high-quality dialogue. J-CHAT [17] primarily collects speech from podcasts. However, these corpora and their construction methodologies are fundamentally limited by their assumption that the in-the-wild data source already consists entirely of dialogue speech. Such sources impose significant constraints on the resulting data: Alexa recordings are restricted to brief, task-oriented exchanges in controlled home environments, while podcast speech typically features structured, professional discourse with predictable speaker roles. These constraints severely limit the acoustic diversity (recording conditions, background noise), conversational variety (topic range, interaction styles), and social dynamics (speaker relationships, register variation) that characterize truly natural dialogue. Consequently, while these datasets may be labeled as "in-the-wild," they fail to capture the full spectrum of spontaneous human conversation found in diverse real-world settings. In contrast, this paper leverages YouTube videos as the data source, which naturally encompasses a much

979-8-3315-7206-8/25/$31.00

broader range of dialogue scenarios, acoustic conditions, and conversational styles. Although YouTube videos may contain both dialogue and non-dialogue speech, our proposed method effectively handles this mixture.

B. Collecting in-the-wild data

Methods for extracting desired data from diverse sources can be categorized into those based on assessing the quality of paired data (i.e., speech with transcription) and those based on evaluating text or speech alone. For the first approach, some studies quantify how well the transcription aligns with the speech and include only data with high alignment scores [16], [22]. Other methods use pre-trained ASR or TTS models to generate pseudo-data for measuring alignment or measure the overall diversity of the paired data [23], [24].

As for the second approach, examples include selecting only high-quality speech [25]–[27] or quantifying a text's "domain-likeness" [28], [29]. Our proposed method is similar to this latter approach of domain quantification. We treat "dialogue" as a domain and, from a mixture of dialogue and non-dialogue YouTube videos, extract those with high "dialogue-likeness."

III. Proposed method

We propose a method for collecting a vast amount of spoken dialogue data by utilizing YouTube videos. Since the videos cannot be used as they are, we perform various steps to obtain the segments containing spoken dialogue interactions.

A. Pre-filtering at video level

As described later, this study uses a large language model (LLM) to determine whether a video is a dialogue video. However, running LLM inference on a large pool of candidate videos requires immense computational or financial resources. Therefore, before using the LLM for classification, we use rule-based filtering to remove possible non-dialogue videos. Concretely, we use only videos that pass all of the following criteria:

1) 10+ `min.`: The video's duration is at least 10 minutes.
2) `Target lang.`: The video's description is written in the target language.
3) 1+ `comments`: There is at least one user comment on the video.
4) 1+ `Likes`: The video has at least one "Like."
5) `No unrelated tag`: The video does not contain any tags unrelated to spoken dialogue, such as music videos, commentaries, and anime.

The above criteria can be strict; however, here, we value precision over recall for quality assurance.

B. LLM-based classification at video level

Based on the video's metadata, a video is determined to be either a dialogue video or not. Metadata here refers to the video's title, description, subtitles, and user comments. In addition to the instructions for labeling and the definition of what constitutes a dialogue video (refer to Section IV-A2 for the definition), the LLM is given the input "*[prompt] [title] [abstract] [caption] [comment]*," and it outputs a binary label indicating whether the video is a dialogue video or a non-dialogue video. For cost considerations, we set a maximum character limit for each of the title, description, subtitles, and user comments.

Here, *[caption]* refers to the transcription of utterances (without timestamps) concatenated with whitespace. We concatenate utterances in ascending order of their timestamps, stopping when the result would exceed the maximum character limit. Likewise, *[comment]* refers to user comments (without usernames) concatenated with whitespace; we randomly select comments for inclusion until adding another comment would exceed the maximum character limit.

C. Speech-based segmentation

We download videos classified as dialogue videos using the process detailed in the previous section and then extract segments deemed to be dialogue. Even if a given video centers on dialogue, it may not be entirely dialogue; for example, it could contain monologues at the beginning, between topic segments, or at the end. Therefore, we first perform speaker diarization on the speech to determine the start and end timings of each utterance and to assign speaker labels. On the basis of those results, we extract dialogue intervals in accordance with the following steps:

1) Treat all utterances in the entire video as one speech segment.
2) Compute a turn-transition count matrix for this segment. This matrix is a square matrix sized by the number of distinct speakers within the segment. The element in row i, column j indicates how many times two consecutive utterances by speaker i and speaker j occurred.
3) Gradually reduce the number of utterances in the segment until none of the off-diagonal elements in the matrix are zero or until the total number of utterances is two or fewer. The former condition implies that consecutive utterances are observed for every possible speaker pair within the segment.
4) If the segment contains 1) more than two utterances, 2) at least two different speakers, or 3) every element of the turn-transition count matrix is non-zero, consider that segment to be a dialogue interval. Otherwise, treat it as a non-dialogue interval.
5) From all utterances in the entire video, take those not yet judged as dialogue or non-dialogue intervals, treat them as a new speech segment, and return to step 2. Repeat this until all utterances in the video have been classified.

Figure 1 presents an example of interval classification results obtained using this method. The URL of the relevant video is shown in the footnote[2]. At intervals (a) and (c), a conversation occurs between speaker 0 and speaker 1, both classified as dialogue intervals. Interval (b) is an intermission between dialogues (in this example, a scene change), and it is largely classified as a non-dialogue interval, although part of

[2]https://www.youtube.com/watch?v=-Z2yhimqs-I

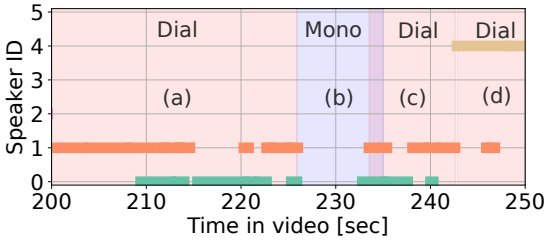

Fig. 1. Example of speech-based segmentation. "Dial" and "Mono" are dialogue and non-dialogue segments, respectively.

> Please evaluate whether the YouTube video, given the following title, summary, subtitles, and comments, might be a "dialogue video."
> A dialogue video is defined as footage in which multiple people engage in a conversation. Use the presence of dialogue elements and the extent to which the content proceeds in a conversational format as your criteria.
> When providing an answer, first write the reasoning for the judgment. Then, at the end, output it in the form "Judgment: XX." If it is a dialogue video, output "True," and if it is not a dialogue video, output "False."

Fig. 2. Text prompt to let LLM classify videos.

a conversation around 232 seconds is also classified as non-dialogue. Interval (d) is a scene featuring speaker 4; note that this interval is separately identified from interval (c).

IV. Experimental evaluation

We performed an experiment to verify our proposed method. After explaining the experimental settings, we describe the results of the proposed method at each step, followed by the description and analysis of the collected dataset.

A. Experimental settings

1) Implementation: We used OpenAI GPT-4o[3] as the LLM. When classifying dialogue videos with the LLM, we used the prompt shown in Figure 2. Our access period to GPT-4o was from September 2024 to December 2024. For the human evaluation described later, we used the Crowdworks[4] platform. We used jtubespeech scripts[5] and Selenium[6] for video searches. We used Langdetect[7] for language identification in the pre-filtering step. The target language of this paper was set to Japanese.

2) Datasets: In this paper, we prepared two sets of videos.

Small labeled set. This set was used for evaluating the video classification by the LLM. Using crowdsourcing, we first collected YouTube channels that post a large number of dialogue videos and then downloaded videos from those channels. We also searched for videos using keywords related to dialogue (e.g., "conversation," "chat," and "talk" in Japanese). The total number of retrieved videos was 363. We had crowdworkers watch each video and label it as either dialogue or non-dialogue. We defined a dialogue video as:

- For roughly 80% or more of its total duration, it contains voices of two or more people conversing.
- The contained dialogue is not based on a prepared script.
- The main language used is in the target language (Japanese).

Large unlabeled set. We used this set to construct the spoken dialogue dataset. We obtained the candidate videos in the same manner as in the work by Watanabe et al. [29], resulting in approximately one million videos (URLs) in total. The manual collection and keyword search approach used for the small labeled set was impractical for collecting the large number of videos needed here, as it would require prohibitive human effort to scale up.

TABLE I
Performance of LLM-based Dialogue/Non-Dialogue Labeling. "Char. Lim." Indicates the Upper Limit of Input Characters of Each of Description (Desc.), Caption, and Comment.

Input				Result			
Char. lim.	Desc.	Caption	Comment	Acc.	Prec.	Recall	F1
500	✓	✓	✓	0.882	0.921	0.920	0.919
		✓	✓	0.863	0.918	0.896	0.905
	✓		✓	0.841	0.935	0.849	0.887
	✓	✓		0.890	0.912	0.943	0.926
1000	✓	✓	✓	0.892	0.927	0.928	0.926
		✓	✓	0.878	0.918	0.919	0.917
	✓		✓	0.834	0.931	0.843	0.882
	✓	✓		0.896	0.912	0.951	0.930

B. LLM-based classification at video level

Using the small labeled set, we evaluated the classification performance of the LLM. The results are shown in Table I. Regardless of whether the maximum text length was set to 500 or 1000 characters, the classification that used both the video description and subtitles yielded the highest F1 score. In particular, focusing on the F1 score, it reached 0.92–0.93, indicating that the LLM can distinguish dialogue videos from non-dialogue videos with sufficiently high accuracy. Notably, removing subtitles from the input led to a significant drop in performance, suggesting that the LLM primarily relies on subtitles for classification of dialogues. On the other hand, including user comments in the input caused a slight decrease in performance, implying that viewer comments can interfere with accurate classification.

C. Pre-filtering

First, we checked how many videos from the large unlabeled set remained after applying each pre-filter. The results are shown in Table II. Among the individual filters, the one that reduced the largest number of videos was 10+ `min.`. We believe this is because the proposed search method excluded all "shorts videos," resulting in such an outcome. A total of 189k videos passed all filters. Of these, we used 134k of those for which Japanese subtitles were available for download as of December 2024 for LLM-based classification.

D. LLM-based classification

We classified 134k videos using the LLM, resulting in approximately 66k being labeled as dialogue videos and ap-

[3]https://openai.com/index/hello-gpt-4o/
[4]https://crowdworks.jp/
[5]https://github.com/sarulab-speech/jtubespeech
[6]https://www.selenium.dev/
[7]https://pypi.org/project/langdetect/

TABLE II

OUT OF 1 MILLION VIDEOS, NUMBER OF VIDEOS THAT PASSED EACH PRE-FILTER. "SEQ." REFERS TO NUMBER OF VIDEOS THAT PASSED ALL FILTERS LISTED ABOVE IN TABLE. FOR EXAMPLE, 266,975 IS NUMBER OF VIDEOS THAT PASSED ALL OF THESE FILTERS: 10+ MINUTES, TARGET LANGUAGE, AND 1+ COMMENTS. "SINGLE" INDICATES NUMBER OF VIDEOS THAT PASSED ONLY THAT FILTER.

	#videos	
Filter	Passed (seq.)	Passed (single)
10+ min.	424,086	424,086
Target lang.	348,151	890,701
1+ comments	266,975	757,536
1+ Likes	229,665	916,095
No unrelated tag	189,133	955,889

TABLE III

NUMBER OF VIDEOS FOR EACH VIDEO CATEGORY.

Categories	#video
Entertainment	25,172
People & Blogs	11,216
Sports	6,375
Gaming	5,612
News & Politics	3,504
Film & Animation	3,420
How-to & Style	2,645
Comedy	2,323
Education	1,886
Music	1,482
Autos & Vehicles	840
Travel & Events	668
Nonprofits & Activism	296
Science & Technology	247
Pets & Animals	135

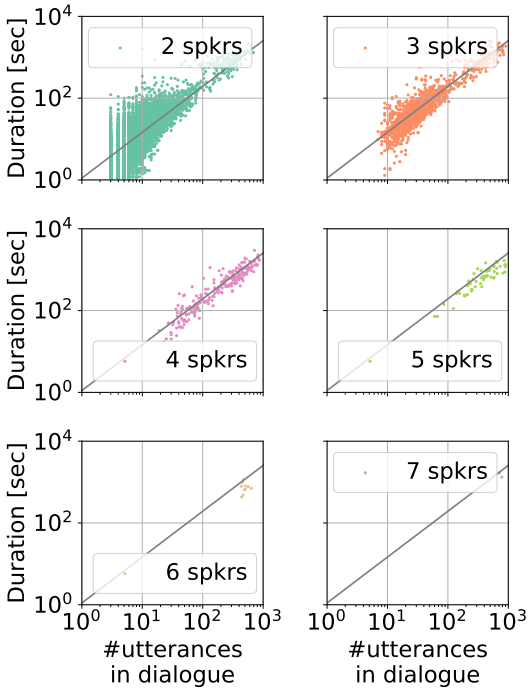

Fig. 3. Double-logarithmic plot of number of utterances versus duration of each dialogue segment. Each point corresponds to single dialogue segment. Parameters of regression line were computed using least-squares error criterion based on utterances from all dialogue segments.

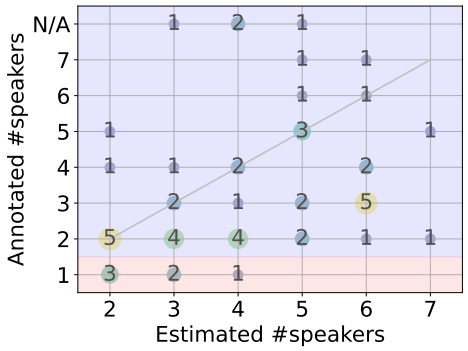

Fig. 4. Confusion between automatically estimated number of speakers and manually assigned number of speakers. "N/A" denotes dialogue segment that comprises multiple different dialogues—like compilation video—even though it is recognized as dialogue segment.

proximately 68k as non-dialogue videos. Table III shows the categories of the videos classified as dialogue videos. The "Entertainment" category was particularly prominent, accounting for roughly 37% of the total. Meanwhile, many videos also fell into a variety of categories. These results demonstrate that the proposed method can collect content with considerable variety.

E. Speech-based segmentation

Utilizing the LLM-filtered list, we downloaded approximately 4,000 videos (about 2,000 hours of content) as our dataset; retrieving the complete set remains a subject for future work. Using publicly available pre-trained speaker diarization models (PyAnnote) [30], we performed speech-based segmentation on 800 of these videos, and we analyzed the segmentation results from two perspectives: the number of speakers and whether the obtained segments actually constitute dialogue.

1) Number of speakers: Figure 3 shows a scatter plot of the number of utterances and the duration of each dialogue segment, grouped by the estimated number of speakers.

First, although two-speaker dialogues account for the majority of this dataset, there are also dialogues with five or more speakers. To verify the correctness of the speaker estimates, we randomly selected 10 segments for each estimated speaker count (except for a speaker count of 7, where we selected only 2 segments) and manually annotated the true number of speakers. The results are shown in Figure 4. Note that the estimated speaker counts often deviate from the actual counts, generally tending to overestimate the number of speakers. Nevertheless, it is clear that there are many multi-speaker dialogues (with five or more speakers) and almost no non-dialogue speech (corresponding to 1 on the vertical axis of the figure). Hence, while re-estimation of the number of speakers or a training algorithm robust to such label noise is necessary, this dataset is expected to be useful for multi-party dialogue systems. In addition, there are many dialogue segments lasting longer than 10^2 seconds, suggesting its potential for long-duration dialogue use cases.

2) Dialogue validity: We manually verified whether the dialogue segments obtained ultimately constitute genuine dialogue segments (i.e., whether they meet the requirements for dialogue videos). The evaluation target consisted of 150

segments selected from those estimated by speaker diarization to be dialogues between two speakers. We focused on two-speaker segments to facilitate verification and because we considered it important to ensure that two-person dialogue videos are reliably obtained for applications such as spoken dialogue modeling. This experiment was conducted under ethical review by our institution.

We recruited 20 crowdworkers to conduct the evaluation. Each segment was evaluated by two workers. Workers viewed each segment (video with audio) and responded to four evaluation criteria: (1) whether it contains dialogue with two or more speakers (yes/no), (2) whether it is natural dialogue, not based on scripts or manuscripts (yes/no), (3) whether the language used is primarily Japanese (yes/no), and (4) how many speakers are present (0, 1, ..., 5, or more). The first two criteria verify whether the dialogue video requirements are satisfied. The third criterion confirms whether the language filter functions properly. The final criterion verifies how many speakers are actually engaged in dialogue among those estimated to be two speakers. While we confirmed this with a small number of samples in the previous section, here we validate with a larger sample size. For speaker count, we counted only those who actually spoke as distinct individuals. Additionally, since YouTube videos may be deleted, workers were instructed to indicate when videos were deleted and could not be played.

Table IV presents the evaluation results regarding dialogue video validity. Eight segments were found to be deleted and unavailable. Regarding whether segments contain dialogues between two or more speakers, the results strongly support this classification. For natural dialogue not based on scripts or manuscripts, approximately 60% of evaluations responded "yes," though this represents lower accuracy than the LLM's performance. This suggests that the LLM may not perform appropriately for certain video populations. Particularly for lists composed of miscellaneous videos unrelated to manual curation or keyword searches, the accuracy of LLM filtering should be considered with appropriate caution. All segments were confirmed to be in Japanese, demonstrating that the language filter functioned effectively.

Figure 5 shows the number of speakers present in the dialogue segments. Note that these segments were specifically those identified by speaker diarization as containing two speakers. The results show that two-speaker responses account for half of the evaluations, indicating that the speaker diarization results are reasonable. However, segments estimated to contain two speakers were sometimes found to actually contain one or three speakers, which aligns with the findings from the previous section. Monologues were infrequent, and there appears to be a tendency to underestimate speaker counts.

These results demonstrate that the dialogue segments obtained through our proposed method appropriately capture genuine dialogues to a certain extent. The proper exclusion of scripted dialogue and high-precision speaker count estimation remain areas for future improvement.

TABLE IV
HUMAN EVALUATION RESULTS FOR DIALOGUE SEGMENTS. SINCE EACH SEGMENT WAS EVALUATED BY TWO PEOPLE, THE TOTAL NUMBER OF EVALUATIONS IS 300 (150 × 2).

Evaluation criteria	Yes	No	Deleted
(1) Dialogue with two or more speakers	253	31	16
(2) Natural dialogue	196	88	16
(3) Language is Japanese	284	0	16

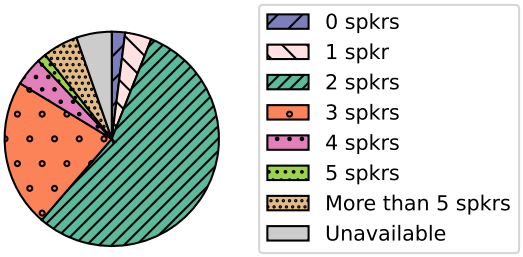

Fig. 5. Number of speakers present in dialogue segments.

V. CONCLUSION

In this study, we proposed a method for collecting an in-the-wild spoken dialogue dataset by retrieving dialogue videos from YouTube. We verified that the constructed dataset primarily consists of two-speaker dialogues but also includes many multi-party dialogues. Additionally, it contains only a small amount of non-dialogue speech, making it suitable for the development of multi-party dialogue systems.

For future work, we plan to extend the dataset. Our experiments used only a portion of the video lists we created. The expected size is 20k hours for our current list. We also want to extend the lists to include more data, with the aim of applying them to various tasks and modeling for spoken dialogue, including voice activity projection and spoken dialogue modeling. Eventually, we would like to release the dataset in the form of video IDs with timestamps for dialogue segments.

ACKNOWLEDGMENT

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011.

REFERENCES

[1] M. F. McTear, Z. Callejas, and D. Griol, *The conversational interface*. Springer, 2016, vol. 6.

[2] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 220–230.

[3] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," in *Proceedings of the 23rd Interspeech Conference*, 2022, pp. 5190–5194.

[4] P. M. Clancy, S. A. Thompson, R. Suzuki, and H. Tao, "The conversational use of reactive tokens in English, Japanese, and Mandarin," *Journal of pragmatics*, vol. 26, no. 3, pp. 355–387, 1996.

[5] Y. Den, N. Yoshida, K. Takanashi, and H. Koiso, "Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations," in *Proceedings of the 2011 International Conference on Speech Database and Assessments*, 2011, pp. 168–173.

[6] D. Xin, S. Takamichi, A. Morimatsu, and H. Saruwatari, "Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus," in *Proceedings of the 24th Interspeech Conference*, 2023, pp. 17–21.

[7] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 517–520.

[8] C. Cieri, D. Graff, K. Owen, M. Dave, and W. Kevin, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004.

[9] W. Barbara, M. Kaneko, and M. Kobayashi, *CALLHOME Japanese Speech, LDC96S37, Linguistic Data Consortium*, 1996.

[10] H. Koiso, H. Amatani, Y. Den, *et al.*, "Design and evaluation of the corpus of everyday Japanese conversation," in *Proceedings of the 13th Language Resources and Evaluation Conference*, 2022, pp. 5587–5594.

[11] A. Défossez, L. Mazaré, M. Orsini, *et al.*, "Moshi: A speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[12] B. Veluri, B. N. Peloquin, B. Yu, H. Gong, and S. Gollakota, "Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 21 390–21 402.

[13] Q. Zhang, L. Cheng, C. Deng, *et al.*, "OmniFlatten: An end-to-end GPT model for seamless voice conversation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 14 570–14 580.

[14] A. Ohashi, S. Iizuka, J. Jiang, and R. Higashinaka, "Towards a Japanese full-duplex spoken dialogue system," in *Proceedings of the 26th Interspeech Conference*, 2025, pp. 1783–1787.

[15] R. Ardila, M. Branson, K. Davis, *et al.*, "Common Voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[16] G. Chen, S. Chai, G. Wang, *et al.*, "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[17] W. Nakata, K. Seki, H. Yanaka, Y. Saito, S. Takamichi, and H. Saruwatari, "J-CHAT: Japanese large-scale spoken dialogue corpus for spoken dialogue language modeling," *arXiv preprint arXiv:2407.15828*, 2024.

[18] M. Cekic, R. Li, Z. Chen, Y. Yang, A. Stolcke, and U. Madhow, "Self-supervised speaker recognition training using human-machine dialogues," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6132–6136.

[19] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "HKUST/MTS: A very large scale Mandarin telephone speech corpus," in *Proc. ISCSLP*, 2006, pp. 724–735.

[20] M. Inaba, Y. Chiba, Z. Qi, *et al.*, "Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 9, 2024.

[21] I. Siegert, ""Alexa in the wild" – collecting unconstrained conversations with a modern voice assistant in a public environment," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 615–619.

[22] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "JTubeSpeech: Corpus of Japanese speech collected from YouTube for speech recognition and speaker verification," *arXiv preprint arXiv:2112.09323*, 2021.

[23] Y.-A. Lai, X. Zhu, Y. Zhang, and M. Diab, "Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 1739–1746.

[24] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[25] P. Baljekar and A. W. Black, "Utterance selection techniques for TTS systems using found speech," in *Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop*, 2016, pp. 184–189.

[26] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proceedings of the 22nd Interspeech Conference*, 2021, pp. 2127–2131.

[27] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6493–6497.

[28] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "MuLan: A joint embedding of music audio and natural language," in *Proceedings of the 23rd International Siciety for Music Information Retrieval Conference*, 2022, pp. 559–566.

[29] A. Watanabe, S. Takamichi, Y. Saito, W. Nakata, D. Xin, and H. Saruwatari, "Coco-Nut: Corpus of Japanese utterance and voice characteristics description for prompt-based control," in *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–8.

[30] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proceedings of the 24th Interspeech Conference*, 2023, pp. 3222–3226.