Scalable Spontaneous Speech Dataset (SSSD): Crowdsourcing Data Collection to Promote Dialogue Research

Zaid Sheikh[1], *Shuichiro Shimizu*[2], *Siddhant Arora*[1], *Jiatong Shi*[1], *Samuele Cornell*[1], *Xinjian Li*[1], *Shinji Watanabe*[1]

[1]Carnegie Mellon University, USA
[2]Kyoto University, Japan

zsheikh@cs.cmu.edu, shinjiw@ieee.org

Abstract

This paper introduces the Scalable Spontaneous Speech Dataset (SSSD) project, comprising 727 hours of spontaneous English conversations between two randomly-matched, anonymous participants on Amazon Mechanical Turk (MTurk) crowd-sourcing platform. The dataset features conversations averaging 25-30 minutes, covering a wide range of everyday topics. A key innovation of this work is our approach to maximizing the number of MTurk workers concurrently participating in our task, enabling more effective randomized matching and live two-person conversations. Data quality is ensured through a two-tiered task structure: a qualification round to select reliable workers, followed by the main recording sessions. We detail our methodology for collecting and recording spontaneous voice conversations, present analyses of the conversational content and speech quality of the dataset in comparison to other datasets, and discuss potential usage.

Index Terms: speech resources, crowdsourcing

1. Introduction

The development of robust and accurate models for speech recognition [1, 2], synthesis [3], and dialogue systems [4] heavily relies on high-quality and diverse data. While several conversational speech datasets, such as Switchboard [5] and Fisher [6], there remains a strong need for resources that feature spontaneous speech in a wider range of settings and with higher audio fidelity than these traditional telephone-based narrow-band corpora. Other existing large-scale speech datasets consist mainly of read speech [7, 8] or speech from formal settings such as news broadcasts [9], parliament proceedings [10, 11], limiting their applicability to casual and conversational speech. Web crawling is another approach to building large audio datasets [12, 13]. However, while these datasets may contain some conversational speech, the proportion is unclear due to a lack of multi-speaker annotations.

To address these limitations, this paper introduces the Scalable Spontaneous Speech Dataset (SSSD)[1], a 727 hours spontaneous English conversation dataset that covers over 70 topics. Unlike previous speech datasets, SSSD was collected through a scalable crowdsourcing approach, where anonymous participants were randomly paired for unscripted voice conversations. Gathering such data presents unique logistical challenges, particularly when live interaction between participants is required. Crowdsourcing platforms are optimized for tasks that can be completed individually and asynchronously. Conducting live, two-person tasks introduces significant hurdles related to scheduling, participant matching, and ensuring consistent engagement. While text-based chat datasets have been constructed through crowdsourcing [14], large-scale spoken dialogue datasets generated through this method are rare due to the inherent challenges. Our approach offers a scalable framework for collecting diverse, high-quality conversational speech, providing the community with a reproducible protocol for implementing this challenging approach. We chose Amazon Mechanical Turk (MTurk) for its vast and diverse worker pool, increasing the probability of finding multiple participants available concurrently. In this paper, we detail our methodology for collecting and recording spontaneous voice conversations via crowdsourcing, present analyses of the conversational content and speech quality of the dataset in comparison to other datasets, and discuss potential usage of the dataset.

Table 1: *Comparison with other conversational speech datasets*

Corpus	Speakers	Sample Rate	License	Hours
Switchboard [5]	543	8 kHz	LDC	260
Fisher [6]	11,917	8 kHz	LDC	1,960
CANDOR [15]	1,456	48 kHz	By request	850
SSSD	209	48 kHz	Permissive	727

2. Related Work

The primary source of spontaneous conversation speech corpora has been telephone conversations [5, 6, 16]. They have been invaluable for research on spontaneous speech and are still widely used, e.g., for the training of state-of-the-art dialogue models [4]. Beyond these established telephone-based corpora, more recent datasets have explored different conversational settings and styles, such as games [17], meetings [18, 19, 20], or radio programs [21]. While our dataset focuses on clean two-party conversations, other work [22, 23, 24] has examined multi-party interactions, which are often noisy and involve more than two participants. CANDOR [15] is another corpus focusing on naturalistic and unscripted conversations, with a key focus on multimodality, as opposed to our focus on simple conversational speech. Synthetic spoken conversation corpora [25, 26] have recently emerged, but they often lack natural human speech phenomena (disfluencies, backchannels etc.).

SSSD distinguishes itself from prior works by its scalable crowdsourcing approach and high fidelity (48 kHz) audio, contrasting with telephone conversation corpora [5, 6] with 8 kHz sampling rate. SSSD also features longer conversations averaging 25-30 minutes, which allows for the study of discourse phenomena over extended periods. Finally, while we provided predefined topics as conversation starters, participants were allowed to deviate freely, resulting in a broader, more natural range of conversations. Table 1 highlights key differences between SSSD and other conversational datasets.

[1] https://wavlab-speech.github.io/SSSD

Our data collection method shares similarities with a concurrent work, CASPER [27], in using React-based web application with Daily.co for audio capture. However, unlike their recruitment of participants via emails and flyers, our approach leverages crowdsourcing, potentially impacting the diversity of participant demographics and motivation.

3. Dataset Construction

SSSD was collected through a novel crowdsourcing approach on MTurk. This section outlines the data collection process[2].

3.1. User Interface

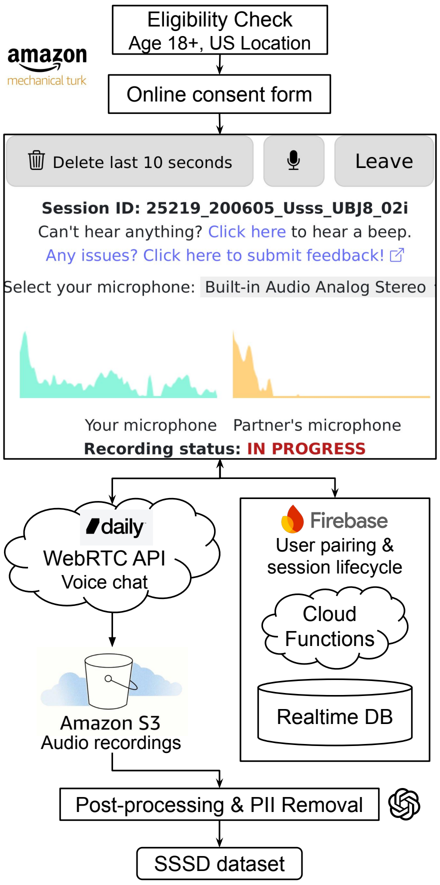
We use a simple easy-to-use React[3] web application to facilitate and record web-based audio calls. The web app is embedded within the MTurk environment using an iframe (an HTML element that embeds a webpage within another), enabling MTurk workers to quickly preview the task and begin working without having to go through a separate signup process on an external site. The web UI connects to a Firebase Realtime Database,[4] which manages user sessions and conversation topics, ensuring real-time synchronization of data. Additionally, Firebase Cloud Functions handle server-side logic, such as pairing users anonymously and managing session lifecycles. For voice communication, the system uses the WebRTC-based audio chat APIs provided by Daily.co.[5] The recorded audio is automatically stored in a private Amazon S3[6] bucket.

3.2. MTurk Qualification Tests

Before participating in the main recording tasks, all MTurk workers were required to complete a brief qualification test. MTurk's built-in `Locale` qualification type was used to restrict participation to the US (a few states were excluded from participation due to legal and privacy considerations). The qualification test involved a simple voice recording task where workers were instructed to enable their microphone and say a few random words. This step ensured that all participating workers had access to a functional microphone, could use the application without issues, and could successfully establish WebRTC calls (which can be affected by restrictive networks, firewalls, or browser extensions). The recorded audio samples were also manually reviewed by the research team to filter out workers with poor audio quality (e.g., excessive background noise, distorted audio). In total, 1,324 workers were approved after successfully completing the qualification test.

3.3. Consent and Privacy

Prior to participation, all workers were required to review and agree to a consent form outlining call recording procedures, data collection details (including collector identity and dataset purpose), and our privacy and anonymization measures. To aid in data categorization, participants could optionally provide demographic information, such as age range and accent type. No personally identifiable information (PII) was collected beyond MTurk Worker IDs, which were stored securely and used only for participant compensation. During recording sessions, the app clearly indicated that recording was active, and workers were instructed to avoid sharing PII. To further protect privacy, participants could mute themselves or delete the last 10 seconds of the recording to remove any inadvertently shared PII. Workers retained full control over their recordings and could choose to submit or delete them upon completion of the call.



Session ID: 25219_200605_Usss_UBJ8_02i
Can't hear anything? Click here to hear a beep.
Any issues? Click here to submit feedback!
Select your microphone: Built-in Audio Analog Stereo
Recording status: IN PROGRESS

Figure 1: *System overview for SSSD construction.*

3.4. Main Recording Sessions

An average of four 1-hour recording sessions were held per day, five days a week. These sessions were strategically scheduled throughout the day to maximize concurrent user availability and accommodate all time zones. Worker preferences were regularly polled via a When2meet[7] link (using pseudonyms). Workers were notified of upcoming recording sessions via MTurk notification emails and a shared calendar, and were advised that their chances of being matched with another worker quickly is highest at the start of the session. Approved workers could accept the task through MTurk during the scheduled session window, with a limit of one call per 1-hour session to minimize over-representation, but no limit on the number of sessions they could join. Upon accepting the task, workers were directed to a virtual waiting room where they were randomly paired with another worker. Once paired, a secure and private audio call is initiated between the two workers. The conversation was automatically recorded throughout the session. As the approved worker pool grew during the data collection period, average number of concurrent call participants also grew from 18 to 41.

[2] approved by CMU IRB (STUDY2022_00000291)
[3] `https://react.dev/`
[4] `https://firebase.google.com/docs/`
[5] `https://docs.daily.co/reference/rest-api`
[6] `https://aws.amazon.com/s3/`
[7] `https://www.when2meet.com/`

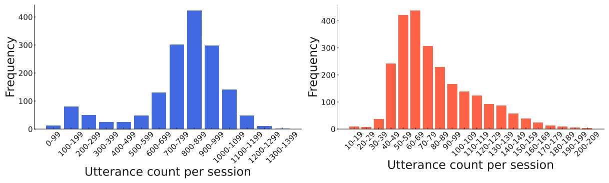


Figure 2: *utterance count per session (SSSD: left, SWBD: right)*

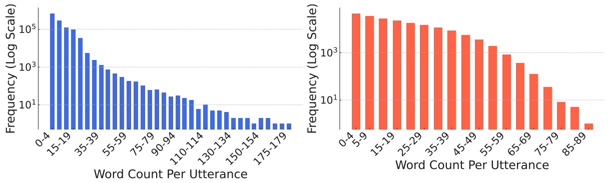


Figure 3: *words per utterance (SSSD: left, SWBD: right)*

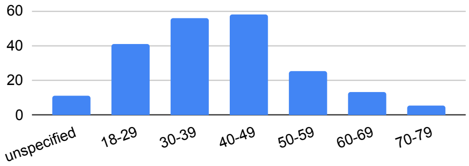


Figure 4: *Age distribution of the participants.*

3.5. Data Validation and Post-Processing

The audio from each speaker was recorded separately using Daily.co's API at a 48-kHz sample rate in WebM (Opus) format, and securely uploaded to a private Amazon S3 bucket. Recordings were transcribed using Whisper-large [1] and checked for quality using heuristics such as minimum conversation length (3 min) and minimum words-per-minute rate (40). Flagged recordings were manually reviewed, and workers were warned about quality issues. Overall 8.1% of recordings were removed in this manner. Post-processing steps included redacting or muting any user-requested portions of the audio. Finally, the two speakers' recordings were time-aligned. This involved adding silence to the beginning of one or both recordings, as necessary, using timing metadata included in the original per-speaker recordings from Daily.co. The time-aligned recordings were then combined into a single interleaved stereo file with each speaker assigned to a separate channel, and encoded as 16-bit FLAC (lossless format) at 48 kHz.

To detect any remaining PII, we fed the transcriptions into OpenAI's GPT-4o [28] (`gpt-4o-2024-08-06`) with the following prompt:

```
Analyze the following transcript and identify
any personally identifiable information (PII)
present.  List the PII detected or state
'No PII found' (only state this and nothing
else) if none is present.  PII includes
names, specific addresses (general location
is OK), phone numbers, and email addresses.
Transcript:  {transcript_text}.
```

To validate this approach, we randomly sampled 50 conversations from the dataset and manually reviewed the transcripts. For the 47 flagged as PII-free, no PII was detected. Among the 3 flagged as containing PII, only 1 truly contained a potential PII. This suggests that the method has high recall, which is desirable here; it may flag some non-PII content, but it reliably identifies potential PII. Recordings predicted to contain PII were excluded from the dataset (11.7% of the total). We plan to incorporate these recordings in a future release after eliminating the specific portions containing PII.

4. Analysis of the Dataset

4.1. Overall Statistics

The dataset comprises of 1,640 audio recordings of two-person conversations, most fall within the 25-30 minute range, with a small subset of 243 recordings (14.8% of the total corpus) having durations between 5-25 minutes.

Since workers were allowed to participate in multiple sessions, individual participation ranged from 1 to 195 instances (mean = 15.7, median = 1) across the 1,640 total sessions. Participant pairing followed a random, first-come-first-served process, and since we did not explicitly avoid repeated pairings, 276 of the 723 unique pairings were observed more than once.

Based on the transcripts of each speaker's audio obtained using Whisper-large [1], utterance and word distributions are calculated (Figures 2 and 3). Figure 2 shows the frequency distribution of the number of utterances per session. The mean and standard deviation is 779.0 ± 241.4. Compared to Switchboard (80.34 ± 31.75), SSSD contains a higher number of utterances per session with a greater variance. Figure 3 shows the frequency distribution of the number of words per utterance in log scale. Both SSSD and Switchboard roughly show exponential decay, indicating that shorter utterances are exponentially more frequent, a common characteristic of conversational speech.

4.2. Speaker Demographics

Demographic information was provided by the majority of participants (94.7%). Among these, 94.4% identified their accent as "US English". Of the 209 unique participants, 101 identified as female, 83 as male, 11 as non-binary or transgender, and 14 declined to disclose their gender. Of the 1,376 conversations where gender information for both participants was available, over half (778) were mixed-gender. Age distribution (Figure 4) peaks in middle age, with a significant number of young adults also participating, but fewer individuals in older categories.

The released dataset includes pseudonymized participant IDs per recording, the suggested initial topic, and, if provided, participant demographic information (age, gender, and accent).

4.3. Conversational Content

Speakers were prompted with 73 diverse everyday topics but were allowed to deviate freely. We used GPT-4o (`gpt-4o-2024-08-06`) to check if conversations matched their given topic: 43% did; the rest diverged into other casual topics, reflecting their open-ended nature.

Table 2 presents the semantic quality trends of utterances in the SSSD dataset. We compare these results with those from another spoken dialogue corpus, Switchboard, tokenizing the transcripts using Whisper's tokenization ("whisper_en") for a fair comparison. To assess semantic coherence, we compute perplexity using GPT-2 [29]. The results show that SSSD utterances exhibit lower perplexity, indicating greater semantic and grammatical coherence than those in Switchboard. Additionally, following the approach in Dialog GSLM [30, 31], we compute VERT to evaluate coherence and diversity in system outputs, along with Self BLEU-2 [32, 33] and Auto BLEU-2 to measure response diversity across and within utterances. Our findings indicate that diversity across utterances is higher

Table 2: *Linguistic Complexity & Diversity: SSSD vs SWBD*

Metric	SSSD	SWBD
Perplexity (GPT-2)	2069.6	5346.1
Perplexity (GPT-2) (>5 words)	128.8	197.8
Self BLEU-2	30.1	37.5
Auto BLEU-2	1.5	7.2
VERT	6.7	16.4

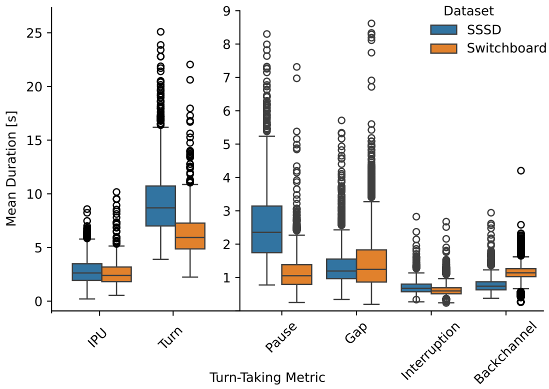


Figure 5: *Distribution of turn-taking events duration for SSSD and Switchboard (SWBD) datasets.*

in SSSD compared to Switchboard. Notably, the lower Auto BLEU-2 score suggests fewer repetitions within sentences, reflecting more varied and natural responses. These results highlight that the SSSD dataset contains more coherent and diverse conversational outputs than prior spoken conversation corpora.

Figure 5 presents the distribution of turn-taking events [31] duration across SSSD and Switchboard. We observe that SSSD exhibits overall slightly longer pauses and speaker turns, indicating that each speaker speaks for a longer duration without interruption. In contrast, the durations of IPUs (interpausal units, i.e., utterances), interruptions, and backchannel responses are comparable between the two datasets. As expected, the two datasets are generally comparable in terms of turn-taking events, as they feature a similar conversational setting.

4.4. Speech Quality

In this section, we compare the audio quality of the SSSD dataset with the Switchboard Eval2000 subset using multiple perceptual metrics. Having a substantial corpus of high-quality, wide-band conversational speech is particularly appealing for training dialogue systems [4]. Since Eval2000 has a sampling rate of 8 kHz, we upsample it to 16 kHz for metric calculations. For a fair comparison, we also analyze a downsampled version of SSSD (SSSD^{-} in Table 3), which is first downsampled to 8 kHz and then upsampled again to 16 kHz. All metrics are computed using the VERSA toolkit [34]. We use 16 kHz as current non-intrusive speech quality measures only support such sampling frequency. The evaluation results in Table 3 indicate that SSSD outperforms Switchboard in terms of quality, intelligibility, noise robustness, and signal-level clarity. These advantages are not only due to the narrow-band nature of Switchboard. In fact, even SSSD^{-} outperforms Switchboard overall. The latter suffers from potentially more degraded speech as a result of being based on analog telephone transmissions. We plot the DNSMOS histogram for more detailed analyses. As illustrated in Figure 6, the distribution of DNSMOS in SSSD is left-skewed, suggesting a rather stable quality of speech in the SSSD dataset.

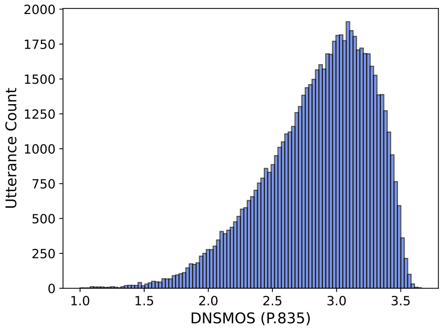


Figure 6: *Histogram of DNSMOS (P.835) for SSSD utterances.*

Table 3: *Audio quality comparison between SSSD and Switchboard Eval2000. SSSD^{-} stands for SSSD that is downsampled to 8kHz. Details of the metrics are discussed in Sec. 4.4.*

Metric	SSSD	SSSD^{-}	Eval2000
DNSMOS (P.835) [36]	2.86 ± 0.41	2.87 ± 0.66	2.57 ± 0.63
UTMOS [37]	2.55 ± 0.72	2.23 ± 0.40	2.20 ± 0.63
SHEET [38]	3.54 ± 0.60	3.51 ± 0.52	2.89 ± 0.78
Squim-STOI [39]	0.94 ± 0.07	0.93 ± 0.08	0.89 ± 0.15
Squim-PESQ [39]	2.61 ± 0.74	2.48 ± 0.71	2.21 ± 0.72
Squim-SI-SDR [39]	15.32 ± 7.08	14.26 ± 7.24	11.90 ± 9.83

Table 4: *ASR Performance*

ASR Model	**SWBD WER (↓)**	**CH WER(↓)**
OWSM (3.1)	11.3	16.7
S3D finetune (w/ SWBD)	9.5	15.6

We also estimated effective audio bandwidth using the technique outlined in [35]. Over 50% of SSSD recordings exceed 32 kHz, compared to 25% in CommonVoice.

5. Usage of the Dataset

We explore the potential of our dataset for building robust speech processing systems. Specifically, we leverage transcripts obtained from Whisper as pseudo-labels and combine them with the Switchboard dataset to train an ASR system. Using this combined dataset, we fine-tune the OWSM 3.1 [40] model and evaluate its performance on conversational speech, specifically the Switchboard (SWBD) / Callhome (CH) test sets. Our approach yields a notable relative improvement of nearly 15.9% / 5.4% in WER on SWBD / CH respectively (Table 4), indicating the effectiveness of our dataset for training ASR systems.

6. Conclusion

This paper introduces SSSD, a valuable new resource comprising 700+ hours of crowdsourced, spontaneous English conversations. SSSD offers several advantages over existing conversational speech corpora: it features higher-fidelity 48kHz audio from web-based calls, contrasting with the telephone-quality audio of datasets like Switchboard and Fisher. The average conversation length of 25-30 minutes allows for analysis of discourse phenomena over extended periods. Participants were given 73 pre-defined topics as conversation starters, but were free to deviate, resulting in a broad range of everyday subjects. We believe SSSD will be a valuable asset for researchers working to improve the robustness, naturalness, and overall capabilities of speech-based systems.

7. Acknowledgements

This work is supported by Meta, led by Prof. Emmanuel Dupoux and his team, whom we thank for their technical support and insightful discussions. We also thank Dr. Leibny Paola Garcia and her team at JHU for sharing valuable details related to the experimental protocols.

8. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[2] Y. Peng, J. Tian, B. Yan, D. Berrebbi, X. Chang, X. Li *et al.*, "Reproducing Whisper-Style Training Using An Open-Source Toolkit And Publicly Available Data," in *ASRU 2023*, 2023.

[3] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu *et al.*, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[4] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou *et al.*, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[5] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *ICASSP 1992*, 1992.

[6] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *LREC 2004*, 2004.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP 2015*, 2015.

[8] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *LREC 2020*, 2020.

[9] J. Kong and D. Graff, "TDT4 multilingual broadcast news speech corpus," 2005, LDC2005S11.

[10] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza *et al.*, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL 2021*, 2021.

[11] G. V. Garcés Díaz-Munío, J.-A. Silvestre-Cerdà, J. Jorge *et al.*, "Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data Filtering/Verbatimization," in *Interspeech 2021*, 2021.

[12] G. Chen, S. Chai, G. Wang, J. Du, W. Zhang, C. Weng *et al.*, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Interspeech 2021*, 2021.

[13] X. Li, S. Takamichi, T. Saeki, W. Chen, S. Shiota, and S. Watanabe, "Yodas: Youtube-Oriented Dataset for Audio and Speech," in *ASRU 2023*, 2023.

[14] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing Dialogue Agents: I have a dog, do you have pets too?" in *ACL 2018*, 2018.

[15] A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson *et al.*, "The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation," *Science Advances*, 2023.

[16] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech," 1997, LDC97S42.

[17] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The sheffield wargames corpus," in *Interspeech 2013*, 2013.

[18] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain *et al.*, "The AMI meeting corpus: a pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, 2005.

[19] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan *et al.*, "The ICSI Meeting Corpus," in *ICASSP 2003*, 2003.

[20] A. Vinnikov, A. Ivry, A. Hurvitz, I. Abramovski, S. Koubi *et al.*, "NOTSOFAR-1 Challenge: New Datasets, Baseline, and Tasks for Distant Meeting Transcription," in *Interspeech 2024*, 2024.

[21] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab *et al.*, "100,000 Podcasts: A Spoken English Document Corpus," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

[22] J. W. D. Bois, W. L. Chafe, C. Meyer, S. A. Thompson, R. Englebretson, and N. Martey, "Santa Barbara Corpus of Spoken American English, Parts 1–4," 2000–2005.

[23] D. Garcia-Romero, D. Snyder, S. Watanabe, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker Recognition Benchmark Using the CHiME-5 Corpus," in *Interspeech 2019*, 2019.

[24] M. V. Segbroeck, A. Zaid, K. Kutsenko, C. Huerta, T. Nguyen, X. Luo *et al.*, "DiPCo — Dinner Party Corpus," in *Interspeech 2020*, 2020.

[25] J. Ao, Y. Wang, X. Tian, D. Chen, J. Zhang, L. Lu *et al.*, "SD-Eval: A Benchmark Dataset for Spoken Dialogue Understanding Beyond Words," *arXiv preprint arXiv:2406.13340*, 2024.

[26] Z. Xie and C. Wu, "Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming," *arXiv preprint arXiv:2408.16725*, 2024.

[27] C. Xiao, R. Liang, X. Zhang, M. E. Tiryaki, V. Bae *et al.*, "CASPER: A Large Scale Spontaneous Speech Dataset," 2025. [Online]. Available: https://sites.google.com/view/casual-casper

[28] OpenAI, "GPT-4o System Card," 2024. [Online]. Available: https://cdn.openai.com/gpt-4o-system-card.pdf

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[30] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, 2021.

[31] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu *et al.*, "Generative Spoken Dialogue Language Modeling," *Transactions of the Association for Computational Linguistics*, 2023.

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL 2002*, 2002.

[33] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, "Texygen: A benchmarking platform for text generation models," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[34] J. Shi, H.-j. Shim, J. Tian, S. Arora, H. Wu, D. Petermann *et al.*, "VERSA: A versatile evaluation toolkit for speech, audio, and music," *arXiv preprint arXiv:2412.17667*, 2024.

[35] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-fi multi-speaker english tts dataset," in *Interspeech 2021*, 2021, pp. 2776–2780.

[36] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022*, 2022.

[37] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Interspeech 2022*, 2022.

[38] W.-C. Huang, E. Cooper, and T. Toda, "MOS-bench: Benchmarking generalization abilities of subjective speech quality assessment models," *arXiv preprint arXiv:2411.03715*, 2024.

[39] A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," in *ICASSP 2023*, 2023.

[40] Y. Peng, J. Tian, W. Chen, S. Arora, B. Yan, Y. Sudo *et al.*, "OWSM v3.1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer," in *Interspeech 2024*, 2024.