# J-CHAT: Japanese Large-scale Spoken Dialogue Corpus for Spoken Dialogue Language Modeling

**Wataru Nakata**[1*], **Kentaro Seki**[1*], **Hitomi Yanaka**[1]
**Yuki Saito**[1] **Shinnosuke Takamichi**[1,2], **Hiroshi Saruwatari**[1]
[1]The University of Tokyo, Japan
[2]Keio University, Japan
nakata-wataru855@g.ecc.u-tokyo.ac.jp
*Equal contribution.

**Abstract**

Spoken dialogue is essential for human-AI interactions, providing expressive capabilities beyond text. Developing effective spoken dialogue systems (SDSs) requires large-scale, high-quality, and diverse spoken dialogue corpora. However, existing datasets are often limited in size, spontaneity, or linguistic coherence. To address these limitations, we introduce J-CHAT, a 76,000-hour open-source Japanese spoken dialogue corpus. Constructed using an automated, language-independent methodology, J-CHAT ensures acoustic cleanliness, diversity, and natural spontaneity. The corpus is built from YouTube and podcast data, with extensive filtering and denoising to enhance quality. Experimental results with generative spoken dialogue language models trained on J-CHAT demonstrate its effectiveness for SDS development. By providing a robust foundation for training advanced dialogue models, we anticipate that J-CHAT will drive progress in human-AI dialogue research and applications.

**Keywords:** spoken dialogue corpus, speech datasets, Japanese, spoken language modeling, language resources

## 1. Introduction

To realize human-AI interaction through spoken language, the development of spoken dialogue systems (SDSs) is crucial. Traditional SDSs have been constructed by cascading multiple modules, such as speech recognition, response generation, and speech synthesis (Huang et al., 2024). However, this cascaded SDS approach struggles to account for subtle nuances and nonverbal vocal expressions, which are often lost in transcription yet are critical elements that distinguish spoken dialogue from text-based chat. Recently, it has been demonstrated that end-to-end SDSs can be realized using machine learning methods trained on large-scale data, and several such methods have been proposed (Nguyen et al., 2023; Mitsui et al., 2023). This approach is gaining attention because it is expected to enable smoother human-AI interaction by fully leveraging the information conveyed through spoken language.

The realization of end-to-end SDS requires large-scale datasets. For instance, previous studies on synthesizing dialogue speech using end-to-end SDS have utilized 20k hours of speech (Nguyen et al., 2023) and 100k hours of speech (Borsos et al., 2023), suggesting that tens of thousands of hours of dialogue speech corpora are necessary to achieve this approach. In addition to size, the quality and diversity of the corpus are also crucial. In dialogue systems, since the final output is speech, it is desirable for the training data to be clean and free of noise. Furthermore, machine learning models tend to degrade in performance when applied to out-of-domain data. Therefore, it is desirable to have a diverse dataset, particularly including spontaneous speech, which is often not covered by existing studio-recorded corpora.

However, large-scale open-source dialogue speech corpora remain scarce, and methods for their construction are underdeveloped. The largest available resource, Seamless Interaction (Agrawal et al., 2025), contains over 4,000 hours of dialogue speech. While this corpus is sufficiently large for training SDSs, its manual recording under controlled conditions makes it financially impractical to reproduce in other languages. Although several studies have investigated automatic methods for constructing large-scale speech corpora in recent years, these efforts have primarily targeted speech recognition (Chen et al., 2021; Takamichi et al., 2021) or speech synthesis (Seki et al., 2023), typically segmenting speech into short utterance-level units. Consequently, such corpora lack key characteristics of dialogue, including linguistic coherence across turns and natural turn-taking dynamics. To date, no general method for automatically constructing a large-scale dialogue speech corpus has been explored.

To address this issue, we propose a method for constructing large-scale dialogue speech corpora and have publicly released the constructed corpus, named “Japanese Corpus for Human-AI Talks” (J-CHAT)[1]. The corpus is distributed under CC BY-

[1]https://huggingface.co/datasets/sarulab-speech/J-CHAT



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 5649–5656
11–16 May 2026. 

NC 4.0 license for "information analysis" which is defined in Japanese copyright act article 30-4[2]. Our corpus construction method automatically filters and collects data on a dialogue-unit basis from the internet. Since the entire process is automated, it is highly scalable. The constructed J-CHAT corpus contains 76k hours of Japanese dialogue speech data, which is comparable to the existing large-scale corpora for the development of end-to-end SDSs. Additionally, background music removal procedures ensure the acoustic cleanliness of the dataset. Moreover, by collecting wild data from multiple domains, the corpus includes spontaneous speech and consists of diverse data.

Our contributions are as follows:

- We propose a method for constructing a large-scale corpus for end-to-end SDSs. Our proposed method constructs a corpus that is acoustically clean, diverse, and includes spontaneous speech.
- Our proposed method is an automated and language-independent approach, making it easy to construct corpora for other languages.
- We constructed and released the Japanese dialogue speech corpus, J-CHAT. Also, we experimentaly validated its effectiveness for end-to-end SDSs.

## 2. Related Work

A comparison between existing corpora and our J-CHAT corpus is shown in Table 1. The STUDIES corpus (Saito et al., 2022) is an open-source conversational speech corpus; however, since it consists of scripted speech recorded in a studio, it does not include spontaneous speech and covers only limited domains. The DailyTalk corpus (Lee et al., 2023) is also an open-source conversational speech corpus, but it contains only about 20 hours of data, which is relatively small in scale. The CallHome Japanese subset (hereinafter referred to as "CallHome-JP") is the largest spontaneous conversational speech dataset in Japanese. Although it is open-source, its scale remains insufficient for training end-to-end SDSs. The MultiDialog corpus (Park et al., 2024), and the SSSD (Sheikh et al., 2025) corpus are English spontaneous conversational speech datasets, but their size are still insufficient for end-to-end SDS. The Fisher corpus (Cieri et al., 2004) is a sufficiently large conversational speech dataset, but it is not open-source, making experimental reproducibility challenging. The LibriTTS corpus (Zen et al., 2019) and GigaSpeech corpus (Chen et al., 2021) are open-source corpora; however, as it is designed for speech synthesis and recognition, respectively, its data is segmented at the utterance level, lacking conversational context-specific information. Recently, Seamless Interaction (Agrawal et al., 2025) has been released which is over 4k hours in duration making it a viable training dataset for SDSs. However, their corpus construction methodology was based on manual recording in a controlled environment limiting the scalability required to reproduce SDSs in other languages.

In contrast, our J-CHAT corpus was collected from in-the-wild sources, allowing for the inclusion of diverse speech. Furthermore, using in-the-wild sources allows scalability of the dataset size making the construction SDSs in many languages possible. With a scale of 76k hours, which is over 1,000 times larger than CallHome-JP, it provides a dataset large enough for the development of end-to-end SDSs. Furthermore, as an open-source resource, it is expected to facilitate research and development in SDSs. Unlike utterance-level segmentation, J-CHAT is segmented at the dialogue level, which enables modeling of discourse phenomena such as contextual coherence and turn-taking behavior.

## 3. Corpus Construction Methodology

Figure 1 illustrates the overall workflow of our corpus construction process. To build a spoken dialogue corpus for a target language (Japanese in this study), we collected audio data from the internet. We then removed inappropriate content by performing language identification, dialogue extraction, and background-noise removal. Finally, we transcribed the spoken content using an automatic speech recognition (ASR) model.

### 3.1. Data Collection

Since previous research (Takamichi et al., 2021) has demonstrated that a diverse range of speech can be collected from YouTube, we used it as one of our primary data sources. We searched YouTube using randomly selected Wikipedia page titles as keywords, following previous research (Takamichi et al., 2021), which resulted in approximately 600k audio files totaling about 180k hours of speech.

YouTube contains not only dialogue-dominant videos but also non-speech videos like music and monologue videos like game commentaries. This resulted in a low proportion of dialogue data, making it challenging to secure a sufficient amount of data. To address this, we also collected data from podcasts to expand the scale of our dataset. Podcasts are speech platforms, making it efficient to gather speech data from them. Furthermore, Pod-

[2] https://laws.e-gov.go.jp/law/345AC0000000048#Mp-Ch_2-Se_3-Ss_5-At_30_4

Table 1: Comparison of speech corpora related to this study, sorted by size.

| Corpus name | Size(hours) | Open-source | Dialogue | Spontaneous | Clean speech |
|---|---|---|---|---|---|
| STUDIES (Saito et al., 2022) | 8.2 | ✓ | ✓ | | ✓ |
| DailyTalk (Lee et al., 2023) | 20 | ✓ | ✓ | ✓ | ✓ |
| CallHome-JP (Alexandra and Zipperlen, 1996) | 49 | ✓ | ✓ | ✓ | ✓ |
| MultiDialog (Park et al., 2024) | 340 | ✓ | ✓ | ✓ | |
| LibriTTS (Zen et al., 2019) | 585 | ✓ | | | ✓ |
| SSSD (Sheikh et al., 2025) | 727 | ✓ | ✓ | | ✓ |
| Fisher (Cieri et al., 2004) | 2$k$ | | ✓ | ✓ | ✓ |
| Seamless Interaction (Agrawal et al., 2025) | 4$k$ | ✓ | ✓ | ✓ | ✓ |
| GigaSpeech (Chen et al., 2021) | 33$k$ | ✓ | | ✓ | |
| **J-CHAT (This study)** | 76$k$ | ✓ | ✓ | ✓ | ✓ |

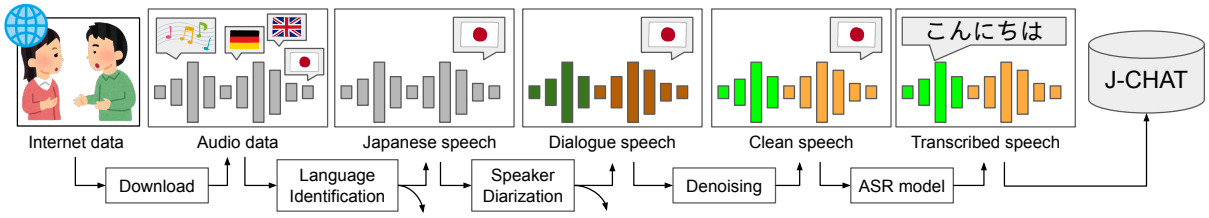


Figure 1: Corpus construction methodology proposed in this work.

castIndex[3] provides extensive metadata, including labels that indicate which language the content is. We retrieved RSS feed URLs of all podcast stations labeled as Japanese from PodcastIndex. Subsequently, we searched and downloaded for any audio URLs listed on the collected RSS feeds. As a result, we obtained approximately 880k audio files totaling around 140k hours of audio data.

## 3.2. Data Selection

### 3.2.1. Extracting Japanese Speech Data

To filter out non-Japanese audio, we used Whisper's language-identification model (Radford et al., 2023) and retained segments only when the probability of Japanese speech ($p$) exceeded 0.8. This process retained $55.7\%$ of the YouTube data and $84.7\%$ of the podcast data.

### 3.2.2. Extracting Dialogue Speech Data

From this Japanese speech dataset, we specifically extracted dialogue segments. This requires detecting conversation regions within the audio. Speaker diarization (SD) techniques are commonly used to identify who is speaking and when. SD analyzes the audio data to detect speech segments and links segments spoken by the same speaker, outputting pairs of speech segments and speaker IDs. In this study, we used a publicly available pre-trained SD model, PyAnnote (Plaquet and Bredin, 2023), to obtain speech segments with speaker IDs.

[3] https://podcastindex.org/

Based on the speech segments (hereafter referred to as turns), we split the audio data into separate dialogues at gaps of 5 seconds or more. Next, we filtered out dialogues where a single speaker's turns account for more than $80\%$ of the time, ensuring that only valid conversations are selected. This treats dialogues dominated by a single speaker as monologues. Here, we allowed dialogues with two or more distinct speaker IDs as valid types of conversations.

Through the above process, we obtained pairs of dialogue-speech data and their labels (each turn's duration and speaker ID). The proportion of Japanese speech data containing dialogues was $41.9\%$ for the YouTube data and approximately $45.0\%$ for the podcast data.

## 3.3. Data Cleansing

YouTube and podcasts often include background music (BGM), which acts as noise for speech generation models, so it needs to be removed. Techniques for extracting speech from data mixed with BGM are studied in the fields of speech enhancement and source separation, and recently, machine learning models have achieved high performance in this area (Koizumi et al., 2023; Rouard et al., 2023). We applied a pre-trained speech enhancement model Demucs (Rouard et al., 2023) for data cleansing to all the collected audio and obtained the J-CHAT corpus.

Table 2: Corpus statistics by its subsets, YouTube and Podcast. # means "number of".

| feature | YouTube | Podcast | Total |
|---|---|---|---|
| total duration[hr] | 11,017 | 65,019 | 76,036 |
| # dialogue | 1,015,109 | 4,409,405 | 5,424,514 |
| mean duration [s] | 39.07 | 53.11 | 50.23 |
| mean # turns | 7.58 | 10.68 | 10.10 |
| mean # speakers | 3.23 | 3.12 | 3.14 |

### 3.4. ASR

Some SDSs require transcription of the speech for training (Défossez et al., 2024). To meet this need, we created the transcription using ASR model. For ASR model, we used reazonspeech-nemo-v2[4]. For each subword in transcript, the alignment information is also provided.

## 4. Corpus Analysis

### 4.1. Dataset Size

Table 2 shows the statistical analysis of J-CHAT corpus. J-CHAT consists of a total of 76k hours of Japanese speech data, with the YouTube subset accounting for 11k hours and the podcast subset for 65k hours. The corpus also provides predefined train/valid/test/other splits for each YouTube and Podcast subsets for the reproducibility of research. For the YouTube subset, the durations of the train/valid/test/other splits are 10872.5/108.7/1.2 hours respectively. For the Podcast subset the duration of train/valid/test/other splits are 57291.5/575/1.3 hours respectively.

A notable difference between the subsets is that the average duration of dialogues in the podcast subset is approximately 1.4 times longer, which is attributed to a higher number of turns per dialogue. On the other hand, the average number of speakers per dialogue is nearly the same for both subsets.

### 4.2. Acoustic Cleanliness

To ensure that the J-CHAT corpus is not significantly affected by noise, we evaluated its acoustic cleanliness using NISQA (Mittag et al., 2021) and compared it with CallHome-JP (telephone recordings) and STUDIES (studio recordings).

The average quality scores were as follows: $4.01$ for STUDIES, $1.98$ for CallHome-JP, $2.37$ for the YouTube subset of J-CHAT, and $2.99$ for the Podcast subset of J-CHAT. Although there is a quality gap between J-CHAT and the studio-recorded STUDIES corpus, J-CHAT's quality is superior to that of CallHome-JP, the largest spontaneous dialogue corpus for Japanese. Furthermore, the Podcast subset of J-CHAT achieved a higher score than the YouTube subset, indicating the effectiveness of using podcasts as a data source for constructing dialogue corpora. This suggests that the noise level in the Podcast subset is sufficiently low for use in speech synthesis.

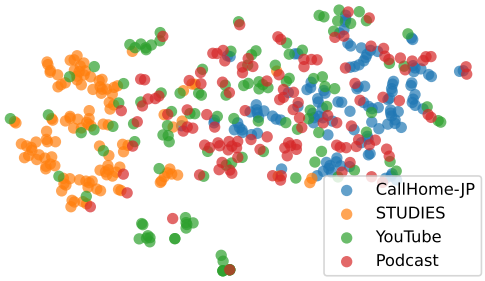


Figure 2: t-SNE visualization of sentence embeddings extracted from the ASRed transcripts from Podcast and YouTube subset of J-CHAT, STUDIES (Saito et al., 2022) and CallHome-JP (Alexandra and Zipperlen, 1996) corpus. For each corpus, random 120 dialouges are extracted for calculation.

### 4.3. Dataset Diversity

**Linguistic Diversity:** As a dialogue corpus, J-CHAT is expected to cover a wide range of topics. To analyze its topic diversity, we compared the distribution of sentence embeddings from J-CHAT transcriptions with those from STUDIES (scripted) and CallHome-JP (spontaneous). We randomly selected $120$ dialogues from the YouTube subset of J-CHAT, and the Podcast subset of J-CHAT as well as the all dialogues in STUDIES. For CallHome-JP, we extracted all $120$ dialogues. Subsequently sentence embeddings are extracted using a pre-trained sentence embedding model[5].

The t-SNE plot of the embeddings is shown in Figure 2. The results indicate that each subset of J-CHAT exhibits a broader distribution compared to CallHome-JP. Note that, although the figure based on this comparison at the same scale may make J-CHAT appear sparse, in reality, J-CHAT is significantly larger than CallHome-JP, ensuring that the region covered by CallHome-JP also contains a sufficient amount of data. Additionally, the distribution of STUDIES is distinct from those of CallHome-JP and J-CHAT because the dialogue topics of STUDIES are limited in conversations between a teacher and students in school.

Furthermore, to conduct a quantitative analysis,

[4] https://huggingface.co/reazon-research/reazonspeech-nemo-v2

[5] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

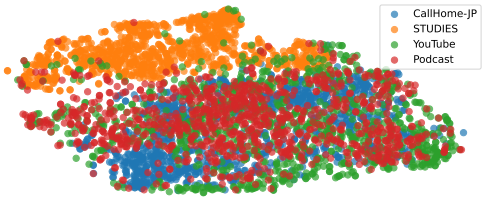


Figure 3: Distribution of HuBERT (Hsu et al., 2021) features extracted from J-CHAT (ours), STUDIES (simulated dialogue), and CallHome-JP (real dialogue).

we calculated the average pairwise cosine similarity for each dataset. The values for CallHome-JP, STUDIES, the YouTube subset, and the Podcast subset are $0.6164$, $0.5186$, $0.2390$, and $0.3457$, respectively. Since a lower cosine similarity indicates lower similarity, these results quantitatively demonstrate that the J-CHAT corpus covers a more diverse range of topics.

**Phonetic Diversity:** To demonstrate that J-CHAT includes spontaneous speech, we conducted an analysis of acoustic features to confirm that J-CHAT contains expression characteristics of spontaneous speech. In this analysis, we used HuBERT (Hsu et al., 2021) features, which are expected to capture phonetic information, and compared them with the STUDIES corpus, which consists of scripted dialogue speech, and CallHome-JP, which consists of spontaneous dialogue speech.

We randomly sampled 1,000 frame-wise HuBERT features from each subset of J-CHAT, as well as from STUDIES (Saito et al., 2022) and CallHome-JP (Alexandra and Zipperlen, 1996). For each sampled point, we computed HuBERT features by randomly selecting 5-second intervals.

The t-SNE plot is shown in Figure 3. The distributions for STUDIES were found to be limited to specific regions. In contrast, the subsets of J-CHAT and CallHome-JP encompassed both regions. These results indicate that J-CHAT covers phonetic diversity in spontaneous dialogue speech.

## 5. Experiments

To validate that J-CHAT corpus is a large-scale corpus suitable for training dialogue-oriented spoken language models, we trained and evaluated the performance of dialogue-generative spoken language model (dGSLM, (Nguyen et al., 2023)). dGSLM follows a framework proposed by (Lakhotia et al., 2021) which consists of three distinct modules: speech-to-unit, unit language model, and unit-to-speech (vocoder).

Our experiments include three models and resynthesis samples for comparison: a resynthesized J-CHAT test subset utilizing speech-to-unit and vocoder (resynth), dGSLM on the YouTube subset of J-CHAT (dGSLM-YouTube), dGSLM trained on the podcast subset (dGSLM-podcast), and dGSLM trained on all subsets of J-CHAT (dGSLM-J-CHAT). Trained models, generated samples and training data are available[1].

### 5.1. Experimental Conditions

We split the dialogues in the J-CHAT corpus into two channels by swapping the output channel of speech according to the turn-taking event. We used the train/valid/test sets explained in Section 4.1. Then, we performed discretization of speech using k-means clustering on HuBERT-extracted features using train set from both YouTube and Podcast subsets. The number of clusters for k-means was set to 1,000.

For the vocoder used for speech generation from the discretized speech, we used HiFi-GAN (Kong et al., 2020) conditioned with speaker information from XVector (Snyder et al., 2018), as used in previous work (Kharitonov et al., 2022).

### 5.2. Training Details

For the implementation of dGSLM, we used official implementation of the model[6]. For HuBERT model used in the speech-to-unit, we used the pretrained Japanese HuBERT model (Sawada et al., 2024)[7]. For k-means clustering, we used the original implementation of the dGSLM except for the number of clusters which was set to $1,000$. The dGSLM model was configured as in the previous research (Nguyen et al., 2023).

For training dGSLM model, we used 32 NVIDIA V100 GPUs with the learning rate of $2 \times 10^{-4}$. The training was performed for $100,000$ steps. The batch size was $36,864$ tokens and the the maximum number of tokens per sample are set to $3,000$. This was equivalent to the $1$ minute of dialogue. For other hyper-parameters regarding the training, we followed the original paper (Nguyen et al., 2023).

For the XVector conditioned vocoder, we used the pretrained XVector model[8]. The vocoder was trained with the JVS (Takamichi et al., 2020) and JVNV (Xin et al., 2024) corpus. These two corpora include Japanese reading-style, studio-quality speech without/with non-verbal expression, respectively. Training of vocoder took 2 days with 4 NVIDIA V100 GPUs.

[6] https://github.com/facebookresearch/fairseq
[7] https://huggingface.co/rinna/japanese-hubert-base
[8] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

Table 3: MOS test results with their 95% confidence intervals.

| Model | Naturalness | Meaningfulness |
|---|---|---|
| resynth | $2.55 \pm 0.18$ | $2.48 \pm 0.18$ |
| dGSLM-Youtube | $1.44 \pm 0.13$ | $1.56 \pm 0.14$ |
| dGSLM-podcast | $1.44 \pm 0.13$ | $1.52 \pm 0.13$ |
| dGSLM-J-CHAT | $2.28 \pm 0.19$ | $2.18 \pm 0.19$ |

### 5.3. Evaluation

For evaluation, we performed subjective listening tests on the mean opinion score (MOS) on the naturalness and meaningfulness of dialogues following previous work (Nguyen et al., 2023). For each subjective listening test, we recruited 60 Japanese native speakers, and each listener evaluated 8 samples. For sample generation used in the evaluation, we used the first 5 seconds of the test set derived from J-CHAT to prompt the model, then performed inference to predict the next 25 seconds of the dialogue based on those initial 5 seconds. For the sampling method, we used the beam search with a beam size of 5. When synthesizing speech from discritized speech, we conditioned the vocoder with JVS001 (male) and JVS002 (female) speaker from the JVS corpus (Takamichi et al., 2020). This resulted in the male speaker's voice being heard from the first audio channel (left) and the female speaker's voice from the second channel (right).

### 5.4. Result

Table 3 shows the results of the subjective evaluation. From these results, it can be seen that dGSLM-J-CHAT achieves the best performance among the dGSLM generated samples in both naturalness and meaningfulness. This suggests that J-CHAT is a useful corpus for constructing generative dialogue language models. We can also see that there is no statistical significance between dGSLM-YouTube and dGSLM-podcast, despite dGSLM-podcast being trained on a dataset approximately four times larger in number of dialogues. This indicates that simply scaling up the dataset is not enough to enhance dGSLM performance. However, there is a significant difference between resynth and dGSLM-J-CHAT in terms of both naturalness and meaningfulness. The trained model occasionally produces sensible words, but the generated dialogue often lacks coherence. This might be improved with better modeling or by increasing the dataset size.

## 6. Conclusion

In this study, we presented J-CHAT, a large-scale Japanese spoken dialogue corpus, and proposed an automated and language-independent methodology for its construction. The experimental results demonstrated that training on diverse dialogue data from multiple data sources, such as YouTube and podcasts, significantly improves spoken dialogue systems models than collecting from single data source. Future work includes further improving dialogue modeling techniques to enhance the quality of generated speech.

## 7. Ethics statement

The published data were collected and distributed in Japan and therefore comply with Japanese law. The legality of the collection and distribution process of J-CHAT was reviewed by a third-party Japanese lawyer. Under Article 30-4 of the Japanese Copyright Act, copyrighted materials may be used without explicit permission when the purpose is limited to "information analysis."[2] Accordingly, the dataset is released strictly for purposes related to information analysis, and not for entertainment or content redistribution. To mitigate privacy concerns, we performed anonymization of source metadata. In addition, the original recordings were segmented into dialogue units, preventing reconstruction of the original content for entertainment purposes. We also implement an opt-out policy, following prior large-scale web corpus practices (e.g. Common Crawl), allowing individuals or rights holders to request removal of their data from the dataset.

## 8. Acknowledgements

This work was supported by AIST KAKUSEI project (FY2023), JST Moonshot JPMJMS2011 and JST FOREST JPMJFR226V.

## 9. Bibliographical References

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. 2025. Seamless interaction: Dyadic audiovisual motion model-

ing and large-scale dataset. *arXiv preprint arXiv:2506.22554*.

Canavan Alexandra and George Zipperlen. 1996. CALLHOME Japanese speech LDC96S37. In *Linguistic Data Consortium*.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023. Soundstorm: Efficient parallel audio generation.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*, pages 3670–3674.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher Corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv*, arXiv:2410.00037.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. LibriTTS-R: A restored multi-speaker text-to-speech corpus. In *Proc. INTERSPEECH 2023*, pages 5496–5500.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. DailyTalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Kentaro Mitsui, Yukiya Hono, and Kei Sawada. 2023. Towards human-like spoken dialogue generation between ai agents from written dialogue.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *Proc. Interspeech 2021*, pages 2127–2131.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. 2024. Let's go real talk: Spoken dialogue model for face-to-face conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16348. Asscociation for Computational Linguistics.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker

diarization. In *Proc. INTERSPEECH 2023*, pages 3222–3226.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yuki Saito, Yuto Nishimura, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari. 2022. STUDIES: Corpus of Japanese Empathetic Dialogue Speech Towards Friendly Voice Agent. In *Proc. Interspeech 2022*, pages 5155–5159.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13898–13905, Torino, Italia. ELRA and ICCL.

Kentaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari. 2023. Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Zaid Sheikh, Shuichiro Shimizu, Siddhant Arora, Jiatong Shi, Samuele Cornell, Xinjian Li, and Shinji Watanabe. 2025. Scalable spontaneous speech dataset (SSSD): Crowdsourcing data collection to promote dialogue research. In *Proc. Interspeech 2025*, pages 3963–3967.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe. 2021. JTubeSpeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification.

Shinnosuke Takamichi, Ryosuke Sonobe, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. 2020. JSUT and JVS: Free japanese voice corpora for accelerating speech synthesis research. *Acoustical Science and Technology*, 41(5):761–768.

Detai Xin, Junfeng Jiang, Shinnosuke Takamichi, Yuki Saito, Akiko Aizawa, and Hiroshi Saruwatari. 2024. JVNV: A corpus of japanese emotional speech with verbal content and nonverbal expressions. *IEEE Access*, 12:19752–19764.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.