Open-Source Full-Duplex Conversational Datasets for Natural and Interactive Speech Synthesis

Zhitong Zhou[1,2], Qingqing Zhang[2], Lei Luo[2], Jiechen Liu[2], Ruohua Zhou[1*]

[1]Beijing University of Civil Engineering and Architecture No.15 Yongyuan Road, Huangcun Town, Daxing District, Beijing 102616, China
[2]Magic Data 5th floor, Building G, No. 44 BeiSanHuan Middle Road, Haidian District, Beijing 100088, China

Keywords: Full-duplex conversation; Speech dataset; Dual-track recording; Speech naturalness

Citation: Zhou Z.T., Zhang Q.Q., Luo L., et al.: Open-Source Full-Duplex Conversational Datasets for Natural and Interactive Speech Synthesis. Data Intelligence, 2026, Vol. 8, Art. No.: 2026d03. DOI: https://doi.org/10.3724/2096-7004.di.2026.00d2

ABSTRACT

Full-duplex, spontaneous conversational data are essential for enhancing the naturalness and interactivity of synthesized speech in conversational TTS systems. We present two open-source dual-track conversational speech datasets, one in Chinese and one in English, designed to enhance the naturalness of synthesized speech by providing more realistic conversational data. The two datasets contain a total of 15 hours of natural, spontaneous conversations recorded in isolated rooms, which produces separate high-quality audio tracks for each speaker. The conversations cover diverse daily topics and domains, capturing realistic interaction patterns including frequent overlaps, backchannel responses, laughter, and other non-verbal vocalizations. We introduce the data collection procedure, transcription and annotation methods. We demonstrate the utility of these corpora by fine-tuning a baseline TTS model with the proposed datasets. The fine-tuned TTS model achieves higher subjective and objective evaluation metrics compared to the baseline, indicating improved naturalness and conversational realism in synthetic speech. All data[① ②], annotations, and supporting code for fine-tuning and evaluation are made available to facilitate further research in conversational speech synthesis.

*Corresponding author: Ruohua Zhou (E-mail: zhouruohua@bucea.edu.cn; ORCID: 0000-0002-2956-0680).

① https://magichub.com/datasets/multi-stream-spontaneous-conversation-training-datasets_chinese/.
② https://magichub.com/datasets/multi-stream-spontaneous-conversation-training-datasets_english/.

© 2026 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

1. BACKGROUND

Full-duplex conversation features natural phenomena such as interruptions, overlaps, and spontaneous interjections [1], which are crucial for improving the naturalness of synthesized speech. Unlike traditional half-duplex (turn-by-turn) interactions, full-duplex voice interaction enables more natural conversations with fluid backchannels, interruptions, and quicker turntaking, which features common in human conversation [2]. For example, researchers have found that more than 40% of speaker turns in natural conversation involve some overlapping speech [1]. Such overlaps include user interruptions (barge-ins), listener backchannels (e.g. "mm-hmm", "right" interjections), and other spontaneous reactions [3]. Moreover, human speech exhibits dynamic pacing, with turn transitions often occurring within just a few hundred milliseconds, requiring conversational system to flexibly adjust response timing so that interaction feels fluid rather than mechanical [4]. This underscores the difficulties of training conversational TTS models on data that reflect real dialog dynamics, including simultaneous speech.

Most existing speech corpora for TTS and conversational modeling are turn-based. For example, classic conversational datasets like Switchboard captured many telephone dialogues but were constrained to sequential turn exchanges with limited overlap [5]. Building on this, FTADsw adds fine-grained turn-taking annotations; its dialogues retain telephone speech characteristics and focus on turn and silence transitions rather than simultaneous speech [6]. Similarly, TTS-specific conversational datasets like DailyTalk, containing over 2,500 conversation turns recorded in high fidelity, aim to support contextual speech synthesis, but still rarely include spontaneous overlapping speech or annotated interruption dynamics [7]. In addition, CASPER contains approximately 200 hours of spontaneous English dialogues with device and demographic metadata; its language is characterized by casual spoken style but remains largely turn-by-turn [8]. Afrispeech-Dialog provides 7 hours of accented English dialogues across medical and nonmedical domains, reflecting regional accent variation but still dominated by sequential exchanges [9]. CS-Dialogue offers 104 hours of Mandarin–English code-switching involving 200 speakers, highlighting frequent language alternation while interaction remains primarily half-duplex [10]. Although these corpora expand language diversity, accent coverage, and interactional annotation, their dialogue structures remain predominantly half-duplex and lack natural conversational elements such as overlaps, interruptions, and laughter, leaving gaps in conversational naturalness and thereby limiting their utility for training full-duplex interaction models.

To address this gap, we develop two separate conversational datasets in Chinese and English, each capturing the dynamics of real-time bidirectional speech. The dataset consists of topicguided spontaneous conversations between native speakers of Chinese and English, with each speaker recorded independently on a separate mono track. This dual-track setup preserves natural overlap and interruption events, resulting in two datasets containing approximately 5 hours of English and 10 hours of Chinese conversational speech, respectively. Moreover, the dataset captures spontaneous and dynamic conversational behaviors,

including backchannels, mid-turn interruptions, laughter, non-verbal reactions, brief overlaps, and rapid turn-taking, all of which are rarely found in half-duplex corpora but are essential to reflect the true dynamics of realworld conversation. We anticipate that these datasets will improve the naturalness of synthesized speech, making it closer to real human conversations.

2. METHODS

We constructed the full-duplex conversational dataset with high-quality conversational speech in both Chinese and English. Two participants, located in separate rooms, use each mobile phone for recording. The speech was captured as speaker-isolated dual-mono tracks using on-device local recording or a split-track microphone setup. Specifically, we carefully designed the protocol to ensure ecological validity: familiar speaker pairs, free-topic discussions, and rigorous quality control together guarantee that the recorded dialogues reflect authentic and natural interaction logic.

2.1 Speaker Recruitment and Prompt Selection

We recruited native speakers of Mandarin Chinese and English through Magic Data crowdsourcing platform. Participants were recruited in pairs, with a preference for dyads who were already familiar with each other (e.g., friends, classmates, or family members). Familiarity was prioritized to reduce social inhibition, facilitate spontaneous exchanges, and elicit naturalistic conversational behaviors such as backchannels and interruptions. Each pair selected their own discussion themes without restriction, allowing the content to emerge organically from speakers' shared interests and experiences. This open-ended design encouraged expressiveness and variation in discourse structure, prosody, and affect, while avoiding priming effects associated with prescriptive prompts. Conversations proceeded freely and were self-paced; no scripted conversation or role assignment was imposed. The collection protocol emphasized ecological validity and conversational authenticity, aiming to approximate ordinary human-to-human interaction.

Then, a human-assisted quality control pipeline filtered out recordings with severe noise or hardware artifacts.

2.2 Transcription and Annotation

We adopted a process of manual fine-grained transcription combined with multiple rounds of verification to ensure annotation quality. All annotators were professionally trained, and each transcript was double-checked by an independent annotator, followed by random sampling inspection conducted by quality-control staff to guarantee the accuracy and consistency of the text and labels. Qualified recordings were manually transcribed by trained annotators. Each speech was annotated with speaker identity, gender, and precise start and end timestamps.

Concurrent speech was represented as overlapping time intervals on the two speaker tracks. In addition to verbatim transcription, the annotation scheme included both paralinguistic and interactional elements.

Annotators focus not only on character accuracy but also on the correctness of punctuation, sentence segmentation, and contextual elements. Particularly, preserving semantic integrity is a crucial and essential aspect of the current model training process. We require annotators to perform VAD (Voice Activity Detection) segmentation in accordance with semantic completeness, ensuring that the segmentation reflects semantic integrity. Such annotations are highly scarce and valuable, as they can be utilized not only for training TTS (Text-to-Speech) systems but also for semantic understanding tasks.

3. DATA RECORDS

The full-duplex conversational dataset as formatted in Figure 1 consists of two language subsets: Chinese and English. All recordings are released as dual-track WAV files (16-bit PCM, 16 kHz), with track 1 corresponding to Speaker A and track 2 to Speaker B. Each conversation is accompanied by a plain text transcript, where utterances from both speakers are merged and ordered chronologically. Each line includes the start and end timestamps, speaker ID, gender, and transcribed utterance. Both the audio and transcription files are organized per speaker. Each transcription file contains timestamped segments for the corresponding speaker's utterances, with "None" labels marking silent or listening intervals.

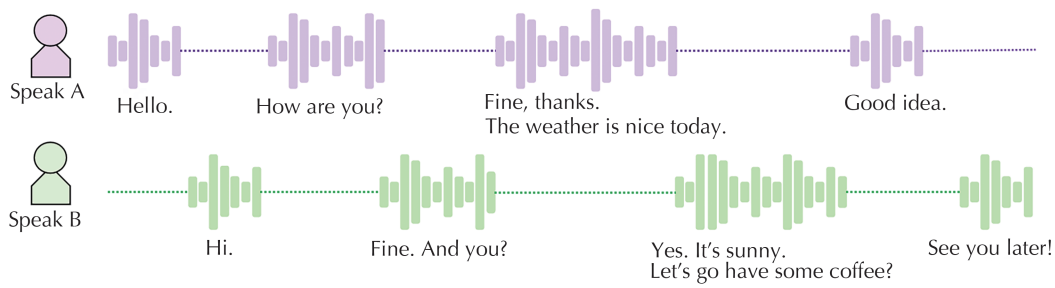


Figure 1. Illustration of Multi-Speaker, Multi-Stream Conversation Recording Format in the Dataset.

File names follow the convention A<SessionID>_S<TopicID>_0_G<SpeakerID>, where 'A' indicates group ID, 'S' is the topic number, '0' denotes the recording device, and 'G' refers to the speaker ID. The dataset includes a total of 35 recorded conversations (27 in Chinese and 8 in English), spoken by 14 unique speakers. A summary is shown in Table 1. More details can be found in [12] and [13].

Table 1. Dataset Statistics.

Language	Conversations	Unique Speakers	Utterance Segments	Total Duration
Chinese	27	6	2306	10 hours
English	8	8	4149	5 hours

4. TECHNICAL VALIDATION

To assess the quality and utility of our full-duplex Chinese–English conversational dataset, we fine-tuned a pretrained CosyVoice-300M model [11] separately on the Chinese and English datasets and conducted comprehensive objective and subjective evaluations. Unlike conventional single-speaker fine-tuning, we did not enforce cross-speaker alignment but instead preserved the original timing of each utterance. Thanks to the high quality of our dual-track recordings, which capture overlaps, interruptions, and backchannels with precise timestamps, the model can directly learn conversational prosody and natural interaction patterns from the dataset. CosyVoice-300M employs two distinct loss functions corresponding to its modular architecture. The text-to-token module is supervised by minimizing the following cross-entropy loss:

$$\mathcal{L}_{\text{LLM}} = -\frac{1}{L+1}\sum_{l=1}^{L+1}\log q(\mu_l) \tag{1}$$

where μ_l denotes the l-th target token (including EOS), and $q(\mu_l)$ is the predicted probability.

For the token-to-mel module, an optimal-transport conditional flow matching loss is utilized to learn the distribution of Mel spectrogram and generate samples from it with generated speech tokens as conditions, which is expressed as:

$$\mathcal{L}_{\text{OT-CFM}} = \mathbb{E}_{t,\, p_0(X_0),\, q(X_1)}\left[\left|\omega_t\left(\phi_t^{\text{OT}}(X_0,\, X_1)\,|\,X_1\right) - v_t\left(\phi_t^{\text{OT}}(X_0,\, X_1)\,|\,\theta\right)\right|\right] \tag{2}$$

where $\phi_t^{\text{OT}}(X_0,\, X_1)$ denotes the interpolation along the optimal transport path, ω_t is the target vector field, and v_t is the model-predicted vector field conditioned on speaker embedding, semantic tokens, and masked mel-spectrogram features. The two modules are trained independently and asynchronously. The learning rate and training epochs are set as 10^{-5} and 200. Chinese utterances were limited to approximately 30 characters, and English utterances to 15 words. If a segment fell below the minimum length threshold, it was merged with the subsequent utterance.

4.1 Objective Evaluation

We employed four core acoustic metrics to evaluate the similarity between synthesized and real speech: 1) Spectrum_l2 distance calculates the Euclidean distance between average melspectrograms of synthesized and real speech, providing a global measure of spectral (timbre) similarity. 2) F0 distance quantifies the distributional difference of fundamental frequency (F0), reflecting prosodic features such as

intonation and pitch variation; lower values indicate a closer match to natural speech rhythm; 3) Energy distance measures the discrepancy in loudness dynamics via root-mean-square (RMS) energy distributions, capturing how well the synthesized speech replicates natural amplitude variation; 4) ZCR distance evaluates the similarity in zerocrossing rate distributions, which relate to voicing patterns, speech rate, and the presence of unvoiced segments. F0, Energy and ZCR are measured in Wasserstein distance. Lower values across these metrics indicate more natural and human-like synthesis performance.

The evaluation revealed consistent improvements across both Chinese and English subsets. As summarized in Table 2, the Chinese model exhibited a significant reduction in F0 distance (7.08%) and mel spectrum distance (1.59%). Similar trends were observed for English, including a 3.67% reduction in F0 distance and a 7.44% reduction in ZCR distance, indicating improved prosody and temporal dynamics after fine-tuning. Beyond the numerical improvements, we observed that reductions in F0 and ZCR distances are particularly indicative of better prosodic modeling, suggesting that the fine-tuned models capture more natural rhythm and pitch variations that are typical of conversational speech.

Table 2. Objective Metric Improvements After Fine-Tuning (Chinese and English Subsets).

Language	Metric	Original	Fine-tuned	Difference	Improvement
Chinese	Spectrum_l2	28.5452	28.0907	0.4545	1.59%
	F0 Wasserstein	63.6259	59.1216	4.5043	7.08%
	ZCR_Wasserstein	0.0136	0.0134	0.0003	1.97%
	Energy_Wasserstein	0.0104	0.0103	0.0001	0.53%
English	Spectrum_l2	34.5732	34.8681	–0.2949	–0.85%
	F0_Wasserstein	55.8803	53.8320	2.0482	3.67%
	ZCR_Wasserstein	0.0074	0.0068	0.0005	7.44%
	Energy_Wasserstein	0.0172	0.0170	0.0002	1.33%

4.2 Subjective Evaluation

To further assess perceptual quality, we conducted the following subjective evaluation. Native Mandarin Chinese and English listeners (10 participants per language) were recruited to assess the naturalness and intelligibility of the synthesized speech samples. In each trial, participants heard two clips generated from the same text prompt, one from the original model (A) and one from the fine-tuned model (B) and judged which sounded more natural. Each listener evaluated the same 80 pairs of conversation. To minimize bias, we adopted a randomized singleblind protocol: the A/B presentation order was shuffled and the model identities were concealed. We used an A/B preference paradigm (A=original, B=fine-tuned). As shown in Figure 2, the aggregated results are summarized by stacked bars: for Chinese, 21.1% chose ‘similar’, 45.0% preferred B and 33.9% preferred A; for English, 12.9% chose ‘similar’, 46.4% preferred

B and 40.7% preferred A. Furthermore, as shown in Table 3, complementary MOS ratings further show small but consistent gains in both naturalness and intelligibility after fine tuning. In general, these findings suggest that fine-tuning produces perceptible improvements in perceived naturalness across both languages while maintaining intelligibility. In addition to quantitative scores, listeners consistently reported that fine-tuned samples conveyed more natural flow and stronger conversational engagement, especially in utterances involving overlaps or backchannel responses. These findings highlight the dataset's effectiveness in improving both perceived naturalness and interactivity, even though the improvements in MOS values appear modest.

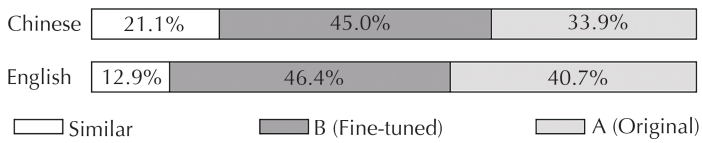


Figure 2. A/B preference distribution.

Table 3. Comparison of MOS Scores for Chinese and English.

	Chinese		English	
	Original (A)	Fine-tuned (B)	Original (A)	Fine-tuned (B)
Naturalness	4.37	4.40	4.47	4.51
Intelligibility	4.46	4.49	4.56	4.58

DATA AVAILABILITY STATEMENT

The dataset described in this study is publicly available to facilitate reproducibility and further research. The data can be accessed via ScienceDB at the following persistent link: https://doi.org/10.57760/sciencedb.j00104.00147 (Chinese Duplex Conversation Training Dataset), https://doi.org/10.57760/sciencedb.27628 (English Duplex Conversation Training Dataset). The dataset are released under CC BY 4.0, allowing reuse with proper attribution. If you use this dataset, please cite this paper using the following citation information:

Zhitong Zhou, Qingqing Zhang, Lei Luo, Jiechen Liu and Ruohua Zhou. Open-Source Full- Duplex Conversational Datasets for Natural and Interactive Speech Synthesis. Data Intelligence, Vol. 8, 2006.

REFERENCES

[1] G.-T. Lin, et al., "Full-duplex speech models—systems that can speak and listen concurrently—are reshaping the landscape of spoken human–computer interaction," arXiv preprint, arXiv:2507.23159, 2025.

[2] S. Arora, et al., "Talking Turns: Benchmarking Audio Foundation Models on Turn-Taking Dynamics," arXiv preprint, arXiv:2503.01174, 2025.

[3] Z. Ma, et al., "Language model can listen while speaking," arXiv preprint, arXiv:2408.02622, 2024.
[4] A. Reece, et al., "The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation," Sci. Adv., vol. 9, no. 5, Art. no. eadf3197, 2023.
[5] J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in Proc. ICASSP-92: 1992 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, San Francisco, CA, USA, 1992, pp. 517–520.
[6] Y. Chen, M. Ma, Y. Li, et al., "FTAD-sw: A Fine-grained Turn-taking Action Dataset based on Switchboard," in Proc. Interspeech 2021, Brno, Czech Republic, 2021, pp. 3964–3968.
[7] D. Lee, et al., "DailyTalk: Spoken Dialogue Dataset for Conversational Text-to-Speech," arXiv preprint, arXiv:2207.01063, 2023.
[8] M. Riviere, A. Baevski, S. Schneider, et al., "CASPER: A Large Corpus of Casual Speech for Conversational Modeling," arXiv preprint, arXiv:2506.00267, 2025.
[9] D. Adelani, P. Ogayo, J. Mukiibi, et al., "Afrispeech-Dialog: A Dialog Corpus for African English Accents in Clinical and General Domains," in Proc. NAACL 2025, Mexico City, Mexico, 2025, pp. 6789–6802.
[10] H. Li, X. Zhou, Y. Wang, et al., "CS-Dialogue: A Large-scale Mandarin–English Code-switching Dialogue Dataset," arXiv preprint, arXiv:2502.18913, 2025.
[11] X. Lyu, Y. Wang, T. Zhao, H. Wang, H. Liu, and Z. Du, "Build LLM-Based Zero-Shot Streaming TTS System with CosyVoice," in Proc. ICASSP 2025, 2025, pp. 1–2.
[12] Magic Data, "Multi-stream Spontaneous Conversation Training Dataset (Chinese)," MagicHub, Sep. 16, 2025. [Online]. Available: https://magichub.com/datasets/multi-stream-spontaneous-conversation-trainingdatasets_chinese.
[13] Magic Data, "Multi-stream Spontaneous Conversation Training Dataset (English)," MagicHub, Sep. 16, 2025. [Online]. Available: https://magichub.com/datasets/multi-stream-spontaneous-conversation-trainingdatasets_english.