

# LibriConvo: Simulating Conversations from Read Literature for ASR and Diarization

Máté Gedeon<sup>\*,†</sup>, Péter Mihajlik<sup>\*</sup>

<sup>\*</sup>Dept. of Telecommunications and Artificial Intelligence,  
Budapest University of Technology and Economics, Hungary

<sup>†</sup>Speechtex Ltd.

gedeonm@edu.bme.hu, mihajlik@tmit.bme.hu

## Abstract

We introduce LibriConvo, a simulated multi-speaker conversational dataset based on speaker-aware conversation simulation (SASC), designed to support training and evaluation of speaker diarization and automatic speech recognition (ASR) systems. Unlike prior resources that mostly rely on semantically disconnected utterances and implausible temporal gaps, LibriConvo ensures semantic coherence and realistic conversational timing. Our pipeline leverages CallHome with external VAD for reliable boundaries, applies compression to reduce unnaturally long silences, and organizes LibriTTS utterances by book to maintain contextual consistency. Acoustic realism is enhanced via a novel room impulse response selection procedure that ranks speaker-microphone configurations by spatial plausibility, balancing realism and diversity. The dataset comprises 240.1 hours across 1,496 dialogues with 830 unique speakers, split in a speaker-disjoint manner for robust evaluation. Baselines show that the sortformer model outperforms the pyannote pipeline in diarization, while a fine-tuned Fast Conformer-CTC XLarge with Serialized Output Training achieves 7.29% WER for ASR, surpassing zero-shot Whisper-large-v3. LibriConvo provides a valuable resource for advancing multi-speaker speech processing research with realistic conversational dynamics and controlled experimental conditions.

**Keywords:** conversation simulation, conversational speech, speech dataset, speech recognition

## 1. Introduction

Modern speech processing systems, such as end-to-end speaker diarization (EEND) and multi-speaker automatic speech recognition (ASR), require large amounts of annotated conversational data (Watanabe et al., 2020). However, collecting real multi-party conversations with precise speaker turn labels is costly and difficult. Consequently, research has turned to synthetic data generation as an alternative (Fujita et al., 2019). Synthetic mixtures of clean utterances allow precise ground-truth labels (*who speaks when and what*), and have become common in training diarization and ASR models (Yu et al., 2016; Kanda et al., 2020). For example, early EEND systems were pre-trained on simulated mixtures of two speakers due to the paucity of real conversational corpora. Such simulated datasets can significantly improve model robustness (Landini et al., 2022a), but naive mixing of utterances does not fully capture conversational dynamics. In particular, simple mixtures often lack realistic turn-taking patterns and may not preserve consistent speaker identities throughout a dialogue.

Recent work has therefore focused on more naturalistic conversation simulation. Yamashita et al. (2022) proposed a method that explicitly models turn-taking by defining different types of speaker transitions, producing synthetic dialogues whose silence and overlap statistics match real meetings. Landini et al. (2022b) likewise used statistics of pause and overlap distributions drawn from actual

conversations to generate speech segments that resemble real dialogues. Park et al. (2023) introduced a probabilistic *property-aware* simulator that dynamically controls silence and overlap amounts to closely match target distributions.

Building on this line of work, our previous study introduced the Speaker-Aware Simulated Conversation (SASC) framework (Gedeon and Mihajlik, 2025), which unified conversational dynamics into a single gap/overlap distribution, incorporated speaker-specific temporal variation, and modeled turn-taking using a Markov-chain process. The SASC framework achieved closer alignment with real conversational data across multiple intrinsic metrics, including gap statistics, pause-length correlation, and turn-taking entropy. However, that paper focused on theoretical evaluation, with no accompanying dataset released for broader community use.

In this paper, we address this gap by introducing LibriConvo, an open-source dataset of speaker-aware simulated conversations built upon the LibriTTS corpus. Following the SASC methodology, LibriConvo transforms independent read-speech utterances into coherent, temporally realistic dialogues that emulate natural human interaction. The corpus comprises 240.1 hours of audio across 1,496 simulated dialogues involving 830 distinct speakers, partitioned into speaker-disjoint training, validation, and test sets. Designed to support research in speaker diarization, ASR, and conversational modeling, LibriConvo enables reproducible

experimentation and standardized benchmarking under controlled yet realistic conversational conditions.

Our main contributions are as follows:

- We introduce LibriConvo, a synthetic conversational speech dataset designed for both ASR and speaker diarization. Each recording consists of multi-turn dialogues with ground-truth transcriptions and speaker turn annotations.
- We propose a methodology for constructing conversations that preserves semantic consistency across speaker turns.
- We present a strategy for selecting room impulse responses (RIRs) to better approximate realistic acoustic conditions.
- We report baseline results for both EEND and ASR on the dataset.

The dataset, comprising both audio recordings and complete metadata, is publicly available on Hugging Face<sup>1</sup> in two versions. The first version is segmented into clips of up to 30 seconds to facilitate ASR training and evaluation, while the second preserves the full-length conversations without segmentation.

The structure of the paper is as follows. Section 2 introduces the theoretical framework of the proposed methodology and its application to dataset generation. Section 3 presents the baseline diarization and ASR results. Finally, Section 4 concludes the paper by summarizing the findings and outlining directions for future work.

## 2. Methodology

### 2.1. Speaker-aware conversation simulation

The speaker-aware conversation simulation (SASC) method (Gedeon and Mihajlik, 2025) generates multi-speaker dialogues with temporal, structural, and acoustic properties that are modeled based on real conversations. Conversational timing is represented by a unified distribution of gaps  $\delta$ , where  $\delta < 0$  indicates overlap,  $\delta \geq 0$  indicates a pause, and the integral over the negative domain corresponds to the probability of overlap  $p_{\text{overlap}}$ . Instead of parametric or histogram-based approaches, kernel density estimation (KDE) is used to obtain smooth, continuous estimates of these gap distributions.

For timing consistency, two mean pause distributions are defined:  $\hat{D}_=$  for same-speaker mean

gaps (when no speaker change occurs between utterances) and  $\hat{D}_\neq$  for different-speaker mean gaps. For each speaker  $s$ , an initial base value is sampled from the appropriate distribution, while subsequent gaps are generated by adding a deviation:

$$\delta_n = \begin{cases} \mu_s^{\text{same}} + v & \text{if } X_n = X_{n-1}, v \sim V_=, \\ \mu_s^{\text{diff}} + v & \text{if } X_n \neq X_{n-1}, v \sim V_\neq. \end{cases}$$

Here  $V_=$  and  $V_\neq$  are zero-mean speaker deviation distributions that preserve local temporal consistency across turns.

Turn-taking is modeled by a first-order (generalizable to  $n$ -order) Markov chain with transition matrix  $P_{\text{turn}}$ , which defines the probability of selecting the next speaker given the previous one. All speakers are placed within a single acoustic environment by sampling a room from the available RIRs and assigning distinct positions within that room. After all utterances are concatenated with their respective gaps and overlaps, optionally background noise  $n \sim \mathcal{N}$  is added and scaled according to a sampled signal-to-noise ratio  $r \sim \mathcal{R}$  to produce the final mixture. A compressed version of the procedure is shown in Algorithm 2.1.

---

#### Algorithm 1 Simplified speaker-aware conversation simulation

---

- 1: Select  $N_{\text{spk}}$  speakers  $\mathcal{S}'$ ; assign RIRs (same room, distinct positions)
  - 2: Choose initial speaker  $X_1$
  - 3: **for**  $n = 1 \dots N_u$  **do**
  - 4:   **if**  $n > 1$  **then** sample  $X_n \sim P_{\text{turn}}(X_{n-1})$
  - 5:   Sample utterance  $u_n \in U_{X_n}$ , convolve with RIR  $\rightarrow y_n$
  - 6:   **if**  $n = 1$  **then** set  $\delta = 0$
  - 7:   **else if**  $X_n = X_{n-1}$  **then**
  - 8:     **if** first gap for  $X_n$  **then**  $\mu_s^{\text{same}} \sim \hat{D}_=$
  - 9:      $\delta = \mu_s^{\text{same}} + v, v \sim V_=$
  - 10:   **else**
  - 11:     **if** first gap for  $X_n$  **then**  $\mu_s^{\text{diff}} \sim \hat{D}_\neq$
  - 12:      $\delta = \mu_s^{\text{diff}} + v, v \sim V_\neq$
  - 13:   Mix  $y_n$  into conversation with gap  $\delta$
- 

### 2.2. Dataset generation

To implement the methodology, we first define the datasets and methods employed.

#### 2.2.1. Statistics

In their original work, Gedeon and Mihajlik (2025) noted that the annotations for Switchboard (Godfrey et al., 1992) exhibited inconsistencies, reporting an average gap corresponding to approximately half a second of overlap—an outcome that appears implausible in natural conversational settings. To

<sup>1</sup><https://huggingface.co/gedeonmate/datasets>

mitigate this issue, we opted to use the CallHome corpus (Canavan et al., 1997), complemented by an external voice activity detection (VAD) model – Silero VAD (Team, 2024) – to obtain more reliable temporal boundaries.

Even with CallHome, notable discrepancies emerged between the original annotations<sup>2</sup> and the VAD-derived speech segments, as illustrated in Figure 1. This contrast likely stems from the fact that the CallHome annotations were never intended for precise gap or overlap analysis, but rather to provide approximate timestamps suitable for diarization training.

Preliminary listening to reconstructed conversations revealed that the dialogues felt disjointed, with pauses longer than typical in spontaneous speech. To better emulate natural conversational timing, we made the subjective choice to apply a temporal compression to the detected pauses, which selectively shortens longer silences while preserving short gaps. The transformation increases compression strength with gap duration, pulling extreme values toward zero while maintaining relative proportions. This reduces unnaturally long pauses—often artifacts of segmentation or annotation—without removing genuine ones, resulting in smoother temporal flow and more natural-sounding simulated dialogues.

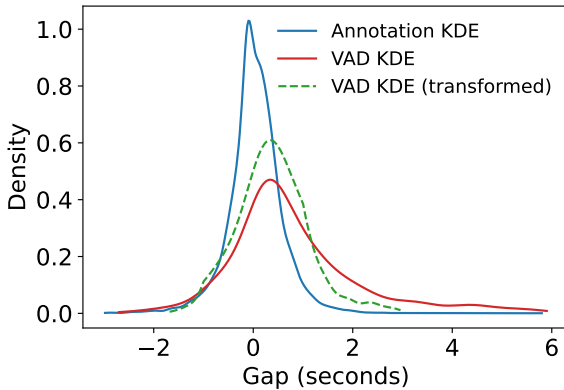


Figure 1: Comparison between original CallHome annotations and VAD-derived speech segments.

For turn-taking modeling, we used a Markov chain of order 1, based on the Callhome data.

### 2.2.2. Utterances

A common limitation in prior studies is that little attention was given to the semantic consistency of the texts used for simulated conversations. While this poses only a minor concern for EEND, it can significantly hinder ASR models, which can benefit from textual context to refine recognition. However,

<sup>2</sup><https://huggingface.co/datasets/talkbank/callhome>

sourcing independent spoken texts that are meaningfully aligned is challenging. To make sure utterances have some semantic common ground, we used utterances that are read parts of a book, with a fixed book for each simulated conversation. The utterances came from LibriTTS (Zen et al., 2019), which is already segmented to sentences. To make utterance length realistic in the conversations, we used utterances from 2 seconds to 10 seconds of length.

### 2.2.3. Room impulse responses

For the room impulse responses (RIRs), we used the BUT Speech@FIT Reverb Database (Szoke et al., 2019), which provides recordings from nine rooms with 31 microphones and about five speakers per room—an ideal configuration for our use case. A key consideration, however, is to select microphones that reflect a realistic acoustic setup—specifically, avoiding positions mounted on ceilings or walls, which would not typically occur in practical scenarios.

To achieve this, we designed a RIR selection procedure that ranks all speaker–microphone configurations within each room by their spatial plausibility. Each configuration is described by four geometric attributes—height, distance, elevation, and azimuth—relative to the microphone array. A realism score is then computed as a weighted sum of normalized deviations from idealized reference values: 1.5 m speaker (source) height, 1 m source–microphone distance, and 0° elevation. Lower scores indicate configurations closer to typical human speech positions.

To maintain spatial diversity, we selected multiple speakers with azimuths differing by at least 20°, avoiding collinearity and better approximating conversational scenes. For each selected speaker, an RIR was randomly drawn from its associated microphone positions. This realism- and diversity-driven strategy yields acoustically plausible and varied RIRs consistent with natural recording conditions.

Overall, the method acts as a lightweight spatial optimization that embeds perceptually motivated heuristics into the RIR sampling process. Unlike prior work optimizing placement for intelligibility or coverage (Morales et al., 2019), our approach targets achieving data realism after the recordings happened, automatically filtering implausible configurations without manual inspection or exhaustive simulation. We applied this procedure to 40% of the conversations.

### 2.2.4. Creating splits

To ensure robust evaluation and eliminate any speaker overlap between training and testing, we constructed speaker-disjoint splits of the simulated

conversations into training, validation, and test sets. Unique speakers were first extracted from the metadata, and all their associated conversations were identified. Since each conversation involved two participants, it was insufficient to simply distribute individual speakers into subsets; speaker pairings also had to be taken into account. Consequently, speakers were heuristically grouped and randomly assigned to one of three mutually exclusive subsets. Conversations were then allocated based on speaker membership to approximate an 80–10–10% train–validation–test ratio. This procedure preserves conversational integrity, maintains balanced data proportions, and ensures that no speaker identity appears in more than one subset, enabling fair generalization to unseen speakers. The resulting distribution is presented in Table 1.

Subset	Speakers	Conversations	Duration (h)
Train	580	1199	193.7
Validation	127	137	23.1
Test	123	160	23.4
<b>Total</b>	<b>830</b>	<b>1,496</b>	<b>240.1</b>

Table 1: Dataset split statistics.

### 3. Experiments

We conducted evaluations on the generated dataset, providing useful baselines for further research.

#### 3.1. Data preparation

To facilitate efficient processing and model training, simulated conversations were segmented into temporally coherent units of up to 30 seconds, matching the input limit of the `Whisper-large-v3`<sup>3</sup> ASR model. Utterances were added sequentially until the next would exceed this limit, at which point a new segment was initiated, omitting intervening silences. Segment times were redefined relative to onset while preserving absolute timestamps. This procedure maintains natural temporal continuity and provides consistent analysis units for training, evaluation, and error tracking.

#### 3.2. Diarization

We evaluated speaker diarization performance on our validation and test sets using two state-of-the-art (SOTA) diarization frameworks, both applied without additional fine-tuning to assess their generalization capabilities on our dataset. The goal of this evaluation was to establish baseline results

and analyze how different model architectures perform, when confronted with simulated multi-speaker speech.

The first baseline is the *pyannote* diarization pipeline (Bredin, 2023), a modular framework composed of neural components for speech segmentation, speaker embedding extraction, and clustering. We employed the `speaker-diarization-3.1`<sup>4</sup> model. The system detects short speech segments using a sliding-window segmentation module, extracts speaker-discriminative embeddings from each, and groups them via agglomerative hierarchical clustering. Finally, clustered segments are merged in a post-processing step to produce the diarization output.

The second baseline is *Sortformer* (Park et al., 2024), an encoder-based diarization model originally designed to supervise speaker tagging in speech-to-text systems. Unlike traditional methods relying on permutation-invariant loss, Sortformer introduces a *Sort Loss* that enforces a consistent ordering of speaker labels, jointly modeling speaker assignment and temporal continuity. This design improves both diarization accuracy and multi-speaker transcription by embedding speaker identity information directly into the ASR process. We used the `diar_sortformer_4spk-v1`<sup>5</sup> model.

Table 2 presents the average diarization error rate (DER) obtained with both models on the evaluation and test sets. As shown, Sortformer significantly outperforms the *pyannote* pipeline, achieving a considerably lower DER on both sets. This suggests that the end-to-end transformer architecture is more effective at disentangling overlapping speech and maintaining speaker consistency over longer segments.

Model	Validation (%)	Test (%)
Pyannote	25.6	24.4
Sortformer	<b>12.9</b>	<b>11.1</b>

Table 2: Comparison of diarization models.

To further analyze performance variability, Figure 2 visualizes the distribution of DER values across individual recordings in the test set. In addition to achieving a substantially lower mean error, Sortformer demonstrates much higher consistency, as indicated by the narrower spread of its error density. In contrast, the *pyannote* pipeline shows larger variance, often struggling with recordings containing high speaker overlap or rapid turn-taking.

<sup>3</sup><https://huggingface.co/openai/whisper-large-v3>

<sup>4</sup><https://huggingface.co/pyannote/speaker-diarization-3.1>

<sup>5</sup>[https://huggingface.co/nvidia/diar\\_sortformer\\_4spk-v1](https://huggingface.co/nvidia/diar_sortformer_4spk-v1)



Model	Validation			Test		
	WER ↓	cpWER ↓	SegAcc ↑	WER ↓	cpWER ↓	SegAcc ↑
Whisper-large-v3	9.41	9.25	-	7.46	7.30	-
Canary-1b-v2	<b>8.69</b>	8.69	-	7.59	7.52	-
fastconformer_l	22.57	22.41	-	23.14	23.07	-
fastconformer_xl	16.98	16.87	-	16.82	16.76	-
fastconformer_l (ft)	10.08	9.89	84.30	10.34	10.01	<b>82.70</b>
fastconformer_xl (ft)	8.88	<b>8.67</b>	<b>85.11</b>	<b>7.29</b>	<b>6.97</b>	82.38

Table 3: ASR baseline results (%) on validation and test splits, without and with fine-tuning.

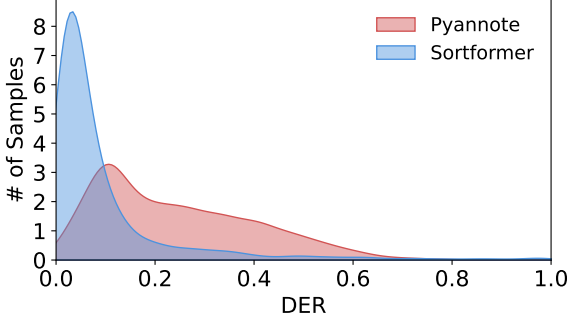


Figure 2: Distribution of DER values across test recordings.

### 3.3. ASR

To establish ASR baselines, we evaluated both state-of-the-art (SOTA) pretrained models—Whisper-large-v3 (Radford et al., 2022) and Canary-1B-v2 (Sekoyan et al., 2025)—without fine-tuning, as well as smaller architectures based on *FastConformer* (Rekesh et al., 2023) that were fine-tuned for our specific task. Training was performed using *Serialized Output Training* (SOT) (Kanda et al., 2020), where speaker changes are explicitly marked with a `<sc>` (speaker change) token.

Unlike the token-level variant (*t-SOT*) (Kanda et al., 2022), which signals every speaker change immediately on the token level with a `<cc>` (channel change) token, our approach treats overlapping speech differently. Specifically, we preserve each speaker’s utterance as a coherent unit, without splitting it when another speaker begins to overlap.

To illustrate the conceptual difference between *t-SOT* and our SOT approach, consider the following example, where the word “I’m” starts before “you” and “good” starts before “thanks”:

**Original:** – How are you? – I’m fine, thanks – good

**t-SOT:** How are `<cc>` I’m `<cc>` you? `<cc>` fine `<cc>` good `<sc>` thanks

**Ours:** How are you? `<sc>` I’m fine, thanks `<sc>` good

For evaluation, we computed both the conventional *Word Error Rate* (WER) and the *concatenated minimum-permutation WER* (cpWER), which accounts for all possible permutations of segments separated by `<sc>` tokens and reports the minimal achievable WER. In addition, for the fine-tuned models, we measured *segment accuracy* (SegAcc), defined as the percentage of cases where the model correctly predicted the number of segments separated by `<sc>` tokens. Table 3.2 summarizes the evaluation metrics and outcomes. The two models used for fine-tuning were `fastconformer-ctc-large`<sup>6</sup> and `fastconformer-ctc-xlarge`<sup>7</sup>.

As shown in Table 3.2, the SOTA pretrained models achieve strong performance without fine-tuning. However, after task-specific fine-tuning, the `fastconformer-ctc-xlarge` model achieves the best overall results. Notably, all models show improvement when evaluated with the cpWER metric, suggesting that while their transcriptions are accurate in content, sometimes the ordering of speaker segments may differ from the reference annotations.

## 4. Conclusion

This work presented a comprehensive methodology for generating realistic simulated multi-speaker conversational datasets with controlled temporal dynamics and acoustic characteristics. By addressing key limitations in existing approaches—particularly the lack of semantic coherence in utterance selection and the presence of implausible temporal gaps in annotated data—we developed a pipeline which aims to produce conversational simulations suitable for enhancing the training and evaluation of both diarization and ASR systems.

<sup>6</sup>[https://huggingface.co/nvidia/stt\\_en\\_fastconformer\\_ctc\\_large](https://huggingface.co/nvidia/stt_en_fastconformer_ctc_large)

<sup>7</sup>[https://huggingface.co/nvidia/stt\\_en\\_fastconformer\\_ctc\\_xlarge](https://huggingface.co/nvidia/stt_en_fastconformer_ctc_xlarge)

Our approach leverages the CallHome corpus with external VAD-based temporal boundary detection, applies temporal compression to reduce unnaturally long silences, and incorporates semantically coherent utterances from LibriTTS organized by source text. The integration of physically plausible room impulse responses through a realism-based selection strategy further enhances the acoustic authenticity of the generated data. The resulting dataset – named LibriConvo – comprises 240.1 hours of simulated conversations across 1,496 dialogues with 830 unique speakers, organized into speaker-disjoint train, validation, and test splits.

Baseline evaluations demonstrate the dataset’s utility for benchmarking state-of-the-art systems. For diarization, Sortformer substantially outperformed the pyannote pipeline, achieving both lower mean DER and greater consistency across recordings, highlighting the advantages of end-to-end transformer architectures for handling overlapping speech and speaker continuity. For ASR, our Serialized Output Training approach—which preserves utterance-level coherence rather than fragmenting at every overlap—proved effective when combined with fine-tuning. The `fastconformer_ctc_large` model achieved the best overall performance with 6.97% cpWER on the test set after fine-tuning, outperforming SOTA pre-trained models in our evaluation framework.

The strong baseline results and the dataset’s realistic conversational characteristics position it as a valuable resource for future research in multi-speaker speech processing. The methodology presented here is reproducible and can be extended to generate larger-scale datasets or adapted to incorporate additional acoustic conditions and conversational patterns. Future work will explore the utility of datasets generated this way, when mixing with authentic conversational training data.

## Acknowledgment

Project No. 2025-2.1.2-EKÖP-KDP-2025-00005 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the EKÖP\_KDP-25-1-BME-21 funding scheme.

## 5. Bibliographical References

- Hervé Bredin. 2023. [pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe](#). In *Interspeech 2023*, pages 1983–1987.
- Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. 2019. [End-to-end neural speaker diarization with permutation-free objectives](#). In *Interspeech*.
- Máté Gedeon and Péter Mihajlik. 2025. [From independence to interaction: Speaker-aware simulation of multi-speaker conversational timing](#).
- Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. 2020. [Serialized output training for end-to-end overlapped speech recognition](#). In *Interspeech*.
- Naoyuki Kanda, Jian Wu, Yu Wu, Xiong Xiao, Zhong Meng, Xiaofei Wang, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Takuya Yoshioka. 2022. [Streaming Multi-Talker ASR with Token-Level Serialized Output Training](#). In *Interspeech 2022*, pages 3774–3778.
- Federico Landini, Mireia Díez, Alicia Lozano-Diez, and Lukáš Burget. 2022a. [Multi-speaker and wide-band simulated conversations as training data for end-to-end neural diarization](#). *ICASSP 2023*, pages 1–5.
- Federico Landini, Alicia Lozano-Diez, Mireia Díez, and Lukáš Burget. 2022b. [From simulated mixtures to simulated conversations as training data for end-to-end neural diarization](#). In *Interspeech*.
- Nicolas Morales, Zhenyu Tang, and Dinesh Manocha. 2019. [Receiver placement for speech enhancement using sound propagation optimization](#). *Applied Acoustics*, 155:53–62.
- Tae Jin Park, Ivan Medennikov, Kunal Dhawan, Weiqing Wang, He Huang, Nithin Rao Koluguri, Krishna C. Puvvada, Jagadeesh Balam, and Boris Ginsburg. 2024. [Sortformer: A novel approach for permutation-resolved speaker supervision in speech-to-text systems](#).
- T.J. Park, H. Huang, C. Hooper, N.R. Koluguri, K. Dhawan, A. Jukić, J. Balam, and B. Ginsburg. 2023. [Property-aware multi-speaker data simulation: A probabilistic modelling technique for synthetic data generation](#). In *Proc. 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023)*, pages 82–86.
- Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Ras-torgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. [Less is More: Accurate Speech Recognition & Translation without Web-Scale Data](#). In *Interspeech 2024*, pages 3964–3968.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever.

2022. [Robust speech recognition via large-scale weak supervision](#).
- Dima Rekish, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. [Fast conformer with linearly scalable attention for efficient speech recognition](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.
- Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. [Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast](#).
- Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. <https://github.com/snakers4/silero-vad>.
- Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. 2020. [Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings](#). In *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, pages 1–7.
- Natsuo Yamashita, Shota Horiguchi, and Takeshi Homma. 2022. [Improving the naturalness of simulated conversations for end-to-end neural diarization](#). In *The Speaker and Language Recognition Workshop*.
- Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Højvang Jensen. 2016. [Permutation invariant training of deep models for speaker-independent multi-talker speech separation](#). *ICASSP 2017*, pages 241–245.
- John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1, ICASSP'92*, page 517–520, USA. IEEE Computer Society.
- Igor Szoke, Miroslav Skacel, Ladislav Mosner, Jakub Paliesek, and Jan Cernocky. 2019. [Building and evaluation of a real room impulse response dataset](#). *IEEE Journal of Selected Topics in Signal Processing*, PP:1–1.
- 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](#). In *Interspeech 2019*, pages 1526–1530.

## 6. Language Resource References

- Alexandra Canavan, David Graff, and George Zipperlen. 1997. [Callhome american english speech](#). Web Download. LDC Catalog No.: LDC97S42, ISBN: 1-58563-111-6, ISLRN: 952-976-147-406-5.