SOMMELIER: Scalable Open Multi-turn Audio Pre-processing for Full-duplex Speech Language Models

Kyudan Jung[1,♣] Jihwan Kim[1,♣] Soyoon Kim[2] Jeonghoon Kim [1,2] Jaegul Choo[1†] Cheonbok Park[1,2†]
{kyudan, jihvvan.kim}@kaist.ac.kr,
{soyoon.kim, jeonghoon.samuel}@navercorp.com,
jchoo@kaist.ac.kr, cbok.park@navercorp.com

[1]KAIST AI, [2]NAVER Cloud

Abstract

As the paradigm of AI shifts from text-based LLMs to Speech Language Models (SLMs), there is a growing demand for full-duplex systems capable of real-time, natural human-computer interaction. However, the development of such models is constrained by the scarcity of high-quality, multi-speaker conversational data, as existing large-scale resources are predominantly single-speaker or limited in volume. Addressing the complex dynamics of natural dialogue, such as overlapping and back-channeling remains a challenge, with standard processing pipelines suffering from diarization errors and ASR hallucinations. To bridge this gap, we present a robust and scalable open-source data processing pipeline designed for full-duplex model. Our code and project page are publicly available at sommelier.github.io.

1 Introduction

Recent advances in speech large language models (SLMs) have evolved from single short queries to multi-turn, open-ended conversations (Xu et al., 2025b; Goel et al., 2025). Yet most systems still operate in disjoint user and assistant turns through a cascaded ASR and TTS pipeline (Ye et al., 2025), which inherits latency, discards paralinguistic cues, and struggles with interruptions, overlapping speech, and backchanneling. Full-duplex system (Wang et al., 2024b; Roy et al., 2026) addresses these limitations by enabling the system to listen and speak simultaneously, supporting more fluid and human-like interaction.

Progress toward full-duplex SLMs has been facing bottlenecked by the lack of high-quality conversational data suitable for duplex training. While Moshi (Défossez et al., 2024) leverages millions of hours of unsupervised audio for pre-training, these sources are largely single-stream and provide limited supervision for overlapping speech. Consequently, overlap robustness relies on relatively small high-fidelity conversational corpora such as Fisher (Cieri et al., 2004), which is unlikely to meet the scale and diversity required for supervised fine-tuning (SFT) (Xu et al., 2025b).

Curating full-duplex training data from in-the-wild recordings is challenging because real conversations contain frequent overlaps, backchannels, and acoustic clutter, which amplify diarization and transcription errors (Wang et al., 2024a). In addition, long-form audio typically includes non-conversational or irrelevant regions (e.g., music, noise, long silences), requiring careful filtering and normalization while preserving speaker structure and multi-turn context. Finally, processing web-scale audio demands high throughput to make large-scale curation feasible under practical compute budgets in the real industry.

To address these challenges, we propose an open, robust, and scalable speech pre-processing pipeline designed for full-duplex SLMs. Our contributions are as follows:

- **The first scalable pipeline for full-duplex SLMs:** We release a scalable pipeline for curating multi-turn conversational speech suitable for full-duplex training, helping alleviate the community-wide data scarcity.
- **High-fidelity overlap processing:** We provide a detailed processing strategy that handles overlaps via rigorous diarization analysis and reduces ASR hallucinations using paralleled model ensembling and n-gram filtering.
- **Proven efficacy on a full-duplex model:** We validate our pipeline by fine-tuning the

♣This work was done during the residency program at NAVER Cloud.

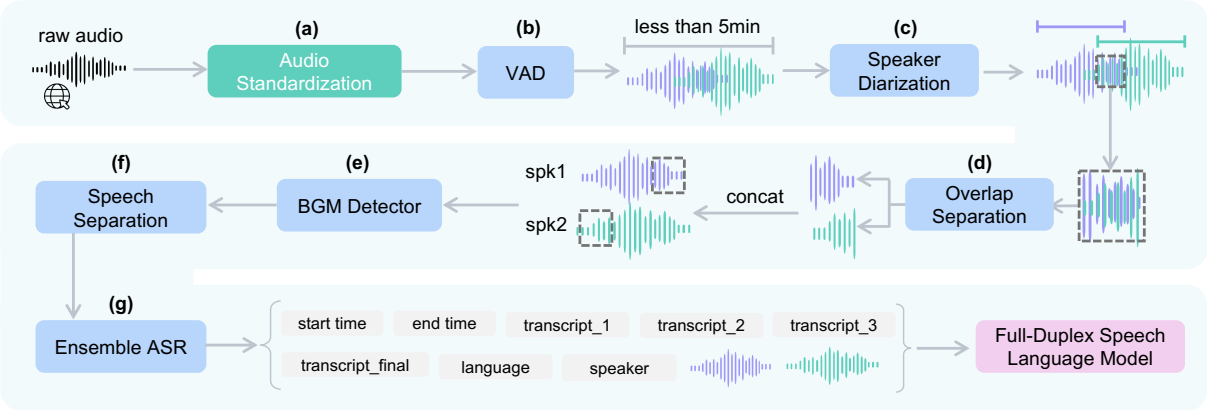


Figure 1: The overall pipeline of the SOMMELIER conversational audio pre-processing. Blue boxes denote neural model-based components, and a green box represent a algorithmic component.

full-duplex model Moshi on SOMMELIER-processed speech and analyze practical data requirements for stable full-duplex training.

2 Related Work

Speech Language Models and Full-Duplex Interaction. Spoken-language understanding has shifted from cascaded ASR→LLM→TTS stacks, which suffer from error propagation and loss of paralinguistics (Lee et al., 2025; Jung et al., 2024a), to End-to-End Speech Language Models (SLMs) that operate directly on acoustic tokens (Borsos et al., 2023; Rubenstein et al., 2023). Recent multimodal systems such as Qwen-Audio/Omni (Chu et al., 2023, 2024; Xu et al., 2025a,b), HyperCLOVA-X-Omni (Team, 2026) and Audio Flamingo (Goel et al., 2025) further bridge audio and language. While earlier SLMs were **half-duplex** (turn-based), recent systems such as Moshi (Défossez et al., 2024; Radhakrishnan et al., 2023a; Ko et al., 2024a; Hu et al., 2024b; Yang et al., 2025b) and GPT-4o (Hurst et al., 2024) target **full-duplex** communication in which listening and speaking occur simultaneously, which requires the model to handle overlapping speech, back-channels, and fluid turn-taking (Lin et al., 2025b,a). Training such models demands high-quality, multi-stream conversational data; the scarcity of such data is the bottleneck SOMMELIER targets.

Data and pipeline context. Existing large-scale corpora are mismatched for this regime: LibriSpeech (Panayotov et al., 2015) and GigaSpeech (Chen et al., 2021) are dominated by read or monologue speech, while classic multi-speaker resources like Fisher (Cieri et al., 2004) and Switchboard (Godfrey and Holliman, 1993) are limited to 8 kHz telephony. Web-scale pipelines such as WenetSpeech (Zhang et al., 2022) and Emilia (He et al., 2024) aggregate large single-stream corpora but treat overlapping speech as noise to be removed. In parallel, recent work couples ASR with LLMs for *generative error correction* (GER) (Radhakrishnan et al., 2023b; Ko et al., 2024b; Hu et al., 2024a; Yang et al., 2025c); we ship such a stage as optional (§4.4). A full treatment of datasets and prior pipelines is deferred to Appendix C.

3 Method

In this section, we present SOMMELIER, a robust data processing pipeline designed to transform raw, in-the-wild conversational audio into high-quality training corpora for full-duplex Speech Language Models (SLMs). Unlike traditional ASR pipelines that prioritize clean, non-overlapping speech, our design philosophy centers on preserving the chaotic yet rich dynamics of human dialogue, such as overlaps and backchannelings, while ensuring scalability for web-scale processing. The overall architecture, illustrated in Figure 1, is built as a modular framework where each component can be toggled or reconfigured, allowing researchers to adapt the trade-off between data purity and conversational authenticity.

SOMMELIER is designed to transform raw audio into clean, well-structured data while preserving the semantic context. The process begins with standardization, bringing diverse audio formats into a unified representation. We then segment the audio based on silence detection, followed by a Voice Activity Detection (VAD)

model (Team, 2024) that further partitions the content into chunks of less than five minutes, a practical constraint that prevents downstream models from running out of memory on lengthy recordings (§ 3.1). Speaker diarization (§ 3.2) follows, identifying who speaks when. Guided by these speaker boundaries, we separate and restore overlapping speech regions (§ 3.3), with optional removal of background noise and music (§ 3.4)depending on the use case. Finally, an ensemble of three Automatic Speech Recognition (ASR) models (§ 3.5) generates text transcripts and captions, leveraging model diversity to improve robustness. Each module in the pipeline can be toggled on or off to suit specific requirements.

Rather than stripping away speech overlaps and backchannelings, interruptions, and simultaneous speech that characterize real dialogue, we preserve them. This allows the duplex speech language model to learn not just what people say, but how conversations actually unfold.

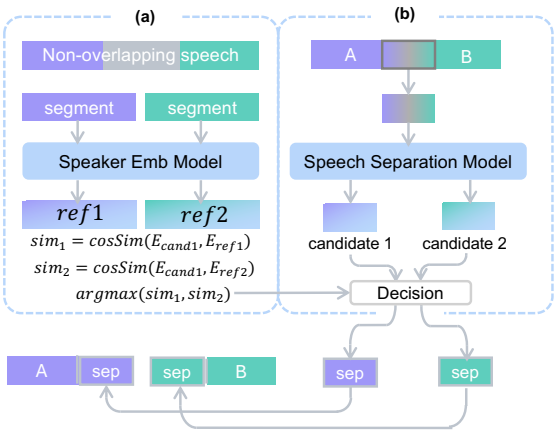


Figure 2: Illustration of the speech overlap separation process. **(a)** The process of calculating similarity to distinguish speaker identities using arbitrary independent speaker segments. **(b)** Separating overlapped regions and making identity decisions for candidates based on the similarity calculated in (a). Finally, the separated segments are concatenated with the original segments.

3.1 Audio Standardization

Since collected radio and podcast data vary in format and volume, we adopt the method of (He et al., 2024). Using the `pydub` [1] and `librosa` (McFee, 2025) libraries, we convert all audio to the standard format (16kHz, 16-bit, Mono) and perform loudness normalization to -20dBFS (He et al., 2024) as illustrated in Figure 1(a).

3.2 VAD & Speaker Diarization

To prevent out-of-memory issues with the diarization model, we split long audio files into units of less than 5 minutes as shown in Figure 1(b). To maintain conversational context, we use a VAD model to cut the audio at silence intervals.

For speaker diarization, as shown in Figure 1(c), instead of the commonly used pyannote `speaker-diarization-3.1` model (Bredin, 2023), we adopted *Sortformer* (Park et al., 2025) from NVIDIA. Section 4.2 presents a performance comparison between the two models, demonstrating Sortformer's superiority in robustly capturing very short utterances such as backchannelings.

3.3 Handling Overlapping Speech

Conversational audio features frequent turn changes and short utterances (Wang et al., 2025b). To systematically handle massive-scale industrial speech data, we categorized overlapping scenarios into four distinct cases, as shown in Figure 8. *Case 1* segments based on the overlap, yielding non-overlapping segments but losing full utterance information. *Cases 2* and *Case 3* assign the overlapping speech to one side, which risks ASR errors where utterances mix or transcripts fail. *Case 4* allows both segments to contain the overlap based on speaker identity, preserving full information despite sharing the ASR issues of *Cases 2* and *Case 3*.

We selected *Case 4* as our baseline, incorporating a module that performs two-speaker separation (Shin et al., 2024) on the overlapped intervals. We find that inputting only the duplicated part into separation model works better than using the entire segment.

Before separating overlapped speech, we extract non-overlapping parts longer than 2 seconds to generate embedding tuples $(e_{ref_1}, e_{ref_2}) = (\mathcal{M}_{emb}(a_0), \mathcal{M}_{emb}(a_1))$, where a denotes the audio segment and $\mathcal{M}$ represents the model. This process uses the speaker embedding model $\mathcal{M}_{emb}$ as shown in Figure 2(a). In parallel, the overlapped audio $a_{overlap}$ is fed into the speech separation model $\mathcal{M}_{sep}$ to produce candidates a_{cand1} and a_{cand2} as shown in Figure 2(b). To identify the speakers, we calculate the cosine similarity scores $S_1 = \text{sim}(\mathcal{M}_{emb}(a_{cand1}), e_{ref_1})$ and $S_2 = \text{sim}(\mathcal{M}_{emb}(a_{cand1}), e_{ref_2})$. The candidate

[1] https://github.com/jiaaro/pydub

with the higher similarity is assigned to the corresponding speaker (a_i), while the other candidate corresponds to the remaining speaker. Finally, we concatenate the non-overlapping parts with the separated segments to create single-speaker segments as shown in Figure 1(d).

3.4 Background Music Removal

In addition to multi-speaker overlaps, the diverse nature of industrial audio sources introduces another challenge. Audio from radio broadcasts or dramas contains background music (BGM), which may be undesirable for training speech language models. We employ PANNs (Kong et al., 2020) (Pre-trained Audio Neural Networks) to estimate the probability of background music presence in each segment. If the probability exceeds a threshold of 0.3, we apply the Demucs (Rouard et al., 2023; Défossez, 2021) model to extract the vocal track. Since music removal can degrade speech quality, we selectively apply it only to segments identified by PANNs, minimizing unnecessary processing as shown in Figure 1(e) and (f).

We find that feeding the entire audio context into Demucs yields substantially better separation performance than processing short segments in isolation. Therefore, we input full two-minute audio chunks into the model and subsequently extract only the required portions from the separated output. We also considered SAM-Audio (Shi et al., 2025) for music removal but excluded it due to its high inference latency (RTF 0.73 on A100), which limits its scalability for large datasets.

3.5 Ensemble-based ASR

High-quality ASR is essential for constructing large-scale datasets, as it generates the text labels necessary for model training. However, relying on a single model, even SOTA architectures like Whisper (Radford et al., 2022a), poses significant risks. These models are prone to hallucinations, particularly in silent or noisy segments, where they often generate repetitive or nonsensical text (Koenecke et al., 2024a; Barański et al., 2025; Mansoor et al., 2025). Such artifacts introduce noise into the training signal, causing the downstream model to mimic these pathological behaviors.

To mitigate this, we employ a Recognizer Output Voting Error Reduction (ROVER) (Fiscus, 1997) ensemble strategy combining outputs from three distinct SOTA models as shown in Figure 1(g). We align transcripts at the word level and apply a prioritized majority voting scheme: a word is accepted if predicted by at least two models; otherwise, we default to the prediction of our primary backbone, Whisper, to maintain consistency. Residual hallucinations are further pruned using a `RepetitionFilter` that discards samples with excessive n-gram ($n = 15$) repetitions (count ≥ 5) (Udandarao et al., 2025). Concurrently, we extract word-level timestamps via Whisper. Precision in temporal alignment is critical for modern streaming speech language models (as detailed in Section 4.1), which typically require strict synchronization between audio and text tokens.

4 Experiments

In this section, we validate the individual components of our proposed pipeline. First, we examine the practical utility of our approach by fine-tuning Moshi using the dataset preprocesed by our proposed pipeline (§ 4.1). We then quantitatively evaluate the diarization accuracy (§ 4.2), the audio quality following overlap separation (§ 4.3), and the accuracy of the ensemble-based ASR (§ 4.4). Furthermore, to provide a comprehensive analysis, we discuss the pipeline's latency (§ 4.5).

4.1 Effectiveness of SOMMELIER-Processed Data for Full-Duplex Models

To validate the effectiveness of our proposed SOMMELIER pipeline, we examine whether training a full-duplex model on the data processed by this pipeline yields performance improvements. To this end, we performed LoRA fine-tuning on `moshiko-pytorch-bf16` (Défossez et al., 2024) and evaluated its duplex performance using the Full-Duplex-Bench (Lin et al., 2025b,a).

Dataset. We find that long turn-taking in the training data for Moshi, specifically when a single speaker holds the floor for too long (more than a minute), leads to unstable loss reduction and degrades performance, causing the model to become unresponsive. Consequently, we selected segments from the SOMMELIER-processed data where each turn lasted no more than 10 seconds. We defined a valid region as a sequence of at least three consecutive turns, truncating the region if an utterance exceeding 10 seconds appeared. Additionally, we assigned only a single speaker to the left channel of the stereo training data. We confirmed that these selection criteria significantly

Model	Pause Handling		Backchannel			Smooth Turn Taking		User Interruption		
	Synthetic TOR ↓	Candor TOR ↓	TOR ↓	Freq ↑	JSD ↓	Candor TOR ↑	Latency ↓	TOR ↑	GPT-4o ↑	Latency ↓
`Moshi`	0.985	0.980	1.000	0.001	0.957	0.941	0.265	1.000	0.765	0.257
`Moshi`+SOMMELIER	1.000	1.000	**0.291**	**0.052**	**0.630**	**1.000**	0.344	0.858	**3.684**	1.065

Table 1: Full-Duplex-Bench 1.0 results for base Moshi and Moshi fine-tuned on 83 hours of SOMMELIER-processed data. Arrows indicate whether higher (↑) or lower (↓) values are better.

Model	DER (%)	JER (%)	DER (≤1.0s, %)	DER (turn, %)
pyannote3.1	8.40	17.68	20.21	0.051
sortformer_v1	**7.16**	**14.69**	**16.87**	**0.006**

Table 2: Diarization model ablation on VoxConverse (Chung et al., 2020) (common subset, ≤4 speakers). Lower is better for DER/JER and RTF.

SIR	OVL	WER (%) ↓			STOI ↑			UTMOS ↑		
		Ori	Sep	Orc	Ori	Sep	Orc	Ori	Sep	Orc
0 dB	0.2	10.5	**6.1**	4.8	.961	**.982**	1.00	3.04	**3.53**	3.88
	0.5	13.9	**7.9**	5.8	.888	**.969**	1.00	2.27	**3.32**	3.87
	1.0	48.9	**15.6**	5.3	.778	**.913**	1.00	1.70	**3.02**	3.84
5 dB	0.2	11.3	**7.6**	5.3	.961	**.978**	1.00	3.06	**3.47**	3.87
	0.5	18.8	**7.1**	4.3	.887	**.971**	1.00	2.34	**3.39**	3.91
	1.0	52.5	**9.1**	4.0	.761	**.936**	1.00	1.79	**3.12**	3.91
10 dB	0.2	12.6	**7.0**	5.6	.961	**.980**	1.00	3.26	**3.60**	3.98
	0.5	29.7	**10.1**	5.2	.877	**.956**	1.00	2.58	**3.21**	3.86
	1.0	51.0	**13.8**	4.8	.754	**.919**	1.00	2.17	**3.01**	3.92

Table 3: Speech quality for separated overlapped speech across metrics for **Ori**ginal audio, source **Sep**arated, and **Orac**le (pre-synthesis speech quality).

impact the training dynamics. The model configuration is detailed in Appendix F.

Results. As shown in Table 1, evaluation on Full-Duplex-Bench 1.0 demonstrated performance improvements across Backchanneling, Smooth Turn-Taking, and User Interruption handling. Regarding Pause Handling, however, we observe that the model performs comparably to base Moshi, exhibiting similar limitations. We hypothesize that this stems from the Moshi architecture or the absence of prompt audio, as proposed in Personaplex (Roy et al., 2026). Regarding latency, the base model exhibited notably short latencies simply because it failed to engage in backchanneling or interruption handling, reflecting suboptimal behavior where the model continued speaking regardless of user input. In contrast, after fine-tuning with Sommelier-processed data, the increased latency can be interpreted positively, as it indicates that the model is now actively processing user input and preparing appropriate responses for backchannels and interruptions. Detailed descriptions of the benchmark metrics are provided in Appendix F.

4.2 Diarization Model Choice

Using `Pyannote 3.1` has been widely regarded as the default for diarization models, a trend followed by recent works such as He et al. (2024) . In this study, however, we compare the performance of Sortformer (Park et al., 2025), which is adopted in our pipeline, against the Pyannote 3.1 (Bredin, 2023; Plaquet and Bredin, 2023) baseline.

Metrics. We evaluate speaker diarization quality using DER (Diarization Error Rate) and JER (Jaccard Error Rate). DER measures the fraction of speaker time that is incorrectly attributed, typically aggregating *missed speech*, *false alarm speech*, and *speaker confusion* within a tolerance collar. JER measures the average Jaccard distance between the reference and hypothesis speaker segments, and is known to be more sensitive to boundary quality and segmentation consistency. To stress-test challenging regimes, we additionally compute *DER on short-duration speech* by restricting evaluation to reference segments shorter than a threshold (≤0.5 s and ≤1.0 s), and *DER on turn-taking regions* by restricting evaluation to temporal windows around speaker change points (speaker alternations within a small gap). All metrics are reported on the VoxConverse (Chung et al., 2020) common subset containing recordings with at most four speakers.

Results and analysis. Table 2 demonstrates that Sortformer consistently outperforms the Pyannote 3.1 baseline across global metrics on the VoxConverse benchmark. More importantly, the gains are most pronounced in regimes critical for conversational modeling. Sortformer exhibits superior robustness in handling short utterances and rapid turn-taking, effectively reducing errors in brief interjections and speaker boundaries. These results confirm that Sortformer is better suited

Dataset	Model	WER (%)	Time (s)
LibriSpeech Test Clean	Whisper	3.63 ± 9.37	0.39
	MoE (Ours)	2.04 ± 6.50	1.40
LibriSpeech Test Other	Whisper	6.26 ± 11.63	0.35
	MoE (Ours)	3.92 ± 8.92	1.27
TEDLIUM3 Test	Whisper	12.19 ± 12.31	0.36
	MoE (Ours)	10.66 ± 11.73	1.33

Table 4: Evaluation results on LibriSpeech (Clean/Other) and TEDLIUM3. Whisper refers `Whisper-large-v3` model.

for processing highly interactive, overlapping dialogue than standard baselines.

4.3 Speech Quality of Overlap Separation

Processing overlapped speech is a critical step in constructing training datasets for full-duplex conversational models (Défossez et al., 2024). This is because full-duplex training requires speech segments to overlap freely, as in natural human conversations, while maintaining source-separated audio streams for each speaker.

Given two speech segments a_i and a_j that are sequentially overlapped, where a_i starts at t_{start}, a_j ends at t_{end}, and the overlap occurs from t_1 to t_2, we evaluate speech quality for each diarized speaker's utterance interval: $[t_{start}, t_2]$ for `speaker1` and $[t_1, t_{end}]$ for `speaker2`.

Dataset. To simulate diverse real-world overlap conditions, we synthesized 900 samples of two speaker mixtures from the LibriSpeech (Panayotov et al., 2015) test utterance by varying Signal-to-Interference Ratio (SIR $\in \{0, 5, 10\}$ dB) and overlap ratio ($\rho \in \{0.2, 0.5, 1.0\}$), forming nine different conditions. We also mix silence-trimmed sources to achieve the target overlap precisely.

Metrics. Evaluation is conducted across three conditions: (1) *Original*, which directly extracts time segments from the mixed signal, (2) **Sep**, which applies SepReformer (Shin et al., 2024)-based separation with speaker identity matching (see Section 3.3), and (3) *Oracle*, which uses clean source signals as an upper bound. Ground-truth diarization timestamps from data synthesis are used across all conditions to ensure fair evaluation of overlapped regions. We assess intelligibility using Word Error Rate (WER), acoustic quality using SI-SDR and STOI, and perceptual naturalness using UTMOS (Saeki et al., 2022).

Results and Analysis. The quantitative analysis presented in Table 3 reveals that while variations in the Signal-to-Interference Ratio (SIR) have a limited impact on performance, the overlap ratio serves as the critical determinant of task difficulty. As the overlap ratio increases, the baseline model suffers significant degradation; however, our proposed method (Sep) consistently outperforms the baseline across all experimental conditions, demonstrating robust separation capabilities even in highly overlapped scenarios. Most notably, in terms of perceptual quality (UTMOS (Saeki et al., 2022)), the proposed method achieves scores closely approximating the Oracle upper bound. This result strongly suggests that our model not only improves intelligibility but also preserves speech naturalness effectively, thereby guaranteeing the generation of high-quality samples suitable for use as training data. More detailed results and analysis are provided in Appendix D.

4.4 ASR Ensemble Performance

We compare the performance of the Whisper model against our proposed three-model ensemble method utilizing ROVER (Fiscus, 1997).

Metrics. To evaluate ASR performance, we measure the Word Error Rate (WER) using the LibriSpeech (Panayotov et al., 2015) test dataset (clean and other splits) and the TEDLIUM3 (Hernandez et al., 2018) test set. Since LibriSpeech `test-other` and TEDLIUM3 contain noisy conditions, these datasets allow us to assess the model's robustness against real-world scenarios.

Results and Analysis. The comparison results between the standalone Whisper Large v3 and the ensemble (Whisper + Canary + Parakeet (Sekoyan et al., 2025)) are presented in Table 4. In terms of WER, the proposed approach demonstrated a significant improvement of approximately 37%, reducing the error rate from 6.26% to 3.92% compared to the single `Whisper-large-v3` baseline. This gap was particularly evident in noisy data, demonstrating improved recognition accuracy in segments containing low volume or BGM.

Regarding inference time, the ensemble approach required approximately three times longer than the baseline. This latency is attributed to the inference speed of Canary, the slowest model among the three, rather than sequential execution.

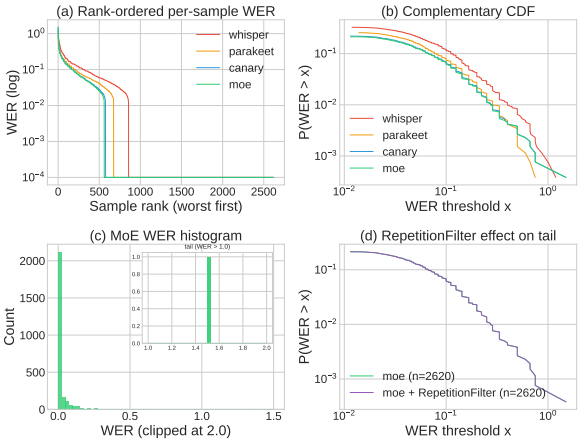


Figure 3: Per-sample WER distribution on LibriSpeech `test-clean` (2,620 samples). (a) rank-ordered per-sample WER on log-y: the tail collapses after rank $\sim$600; (b) complementary CDF $P(\text{WER} > x)$: below 10^{-3} at $x = 1$; (c) histogram with a zoomed inset of the WER > 1 region; (d) `RepetitionFilter` does not prune any sample on clean data.

Additionally, the concurrent loading and inference of three models may introduce slight overhead.

Heavy-tail analysis. Although the mean WER is low (2.04%), the reported standard deviation (6.50%) is driven by a small heavy-tailed set of failures rather than systematic accuracy loss. On `librispeech-test-clean`, the per-sample quantiles are P50=0.000, P95=0.125, P99=0.272, P99.9=0.698, max=1.50; only 6 samples (0.23%) have WER>0.5 and a single sample (0.04%) has WER>1.0 (full statistics in Appendix A). Figure 3(a,b) shows this: the rank-ordered curve collapses after rank $\sim$600 and the cCDF drops below 10^{-3} at x=1. The `RepetitionFilter` (n=15, threshold 5) prunes 0 samples on this split, consistent with its role as a safety net for realistic noisy web audio rather than a crutch for clean benchmarks.

Generative Error Correction as an optional stage. Following recent work that couples ASR hypotheses with an LLM to recover residual errors (Radhakrishnan et al., 2023b; Ko et al., 2024b; Hu et al., 2024a; Yang et al., 2025c), we implemented an optional *Generative Error Correction* (GER) stage that asks an instruction-tuned LLM (Qwen2.5-3B) to pick the most likely transcript from the three ensemble hypotheses. Zero-shot GER *degraded* WER on all three splits (librispeech-test-clean 2.04→33.4%, test-other 3.92→31.3%, TEDLIUM3 10.66→41.3%; Appendix B), because out-of-domain LLM paraphrases and punctuation drift outweigh the corrections on already-clean speech. We therefore ship GER as *optional*, to be enabled only when domain-matched LLM fine-tuning is performed as in prior work.

Qualitatively, we observed hallucinations in the Whisper outputs, such as repetitive generation (e.g., *"Yeah., Yeah., Yeah..."*). We confirmed that our method successfully corrected these errors by selecting the accurate transcript provided by Canary (e.g., *"Yeah, big decision for Dan"*).

Stage	Processing Time (s)	RTF
Audio Duration	120.00	–
VAD + Sortformer	1.91	0.0159
SepReformer Separation	0.15	0.0013
ASR ensemble	13.91	0.1159
FlowSE Denoising	4.99	0.0416
Total	**20.95**	**0.1746**

Table 5: Processing time breakdown for the proposed pipeline on a 120-second audio sample.

4.5 Latency

Data preprocessing is a computationally intensive task (He et al., 2024; Dua et al., 2025). Thus, minimizing latency in this phase is critical. As shown in Table 5, running a single process on an A100 (80GB) yields a total Real-Time Factor (RTF) of 0.1746. Excluding the optional FlowSE Denoising step further reduces the RTF to 0.133, with the primary bottleneck occurring in the ASR stage. Given the peak memory usage of 23GB, it is possible to allocate three concurrent processes on a single GPU, which effectively lowers the RTF to 0.0443 per GPU. Consequently, processing 10,000 hours of audio using eight A100 GPUs would take approximately 55 hours, demonstrating the practical feasibility of our approach.

5 Conclusion

We presented SOMMELIER, the first scalable, open-source pipeline for full-duplex SLMs. Our pipeline combines rigorous diarization, overlap handling, and ensemble-based ASR to improve overall transcript quality. We validated the overall utility of the SOMMELIER pipeline by fine-tuning Moshi on SOMMELIER-processed speech. We release our pipeline to support reproducible industrial research and to accelerate progress toward natural, real-time human–AI interaction.

Limitations

A limitation of our pipeline is its exclusive focus on processing speech data. While optimized for conversational dialogue, it does not explicitly account for non-speech acoustic events or general sound scenes, limiting its scope compared to omni-modal audio approaches. Although our overlap separation module effectively disentangles simultaneous speakers from single-stream recordings, the resulting audio fidelity is inevitably slightly inferior to datasets that are originally recorded with distinct, isolated channels (Oracle), as the artificial separation process may introduce minor acoustic artifacts.

Ethical Considerations

We developed the SOMMELIER pipeline with a strict adherence to open-source compliance and intellectual property rights. All software components, libraries, and pre-trained models integrated into our framework are governed by commercially permissive licenses, primarily MIT and Creative Commons (CC), allowing for broad academic and industrial application without legal ambiguity.

Furthermore, the podcast audio samples featured on our project demonstration page were exclusively selected from sources explicitly released under CC licenses. We have rigorously verified the usage terms of these recordings to ensure that no copyrighted material is infringed upon and that the original creators' rights are respected.

Beyond licensing compliance, we acknowledge the broader implications of releasing tools for high-fidelity speech processing. While our goal is to advance full-duplex interaction, we recognize that high-quality conversational datasets can potentially be misused for non-consensual voice cloning or deepfake generation. We urge the research community to utilize this pipeline responsibly, ensuring that any private data processed is done so with appropriate consent and privacy safeguards in place.

Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-00555621)

We would like to express our deepest gratitude to Taehong Moon.

References

Inclusion AI, Bowen Ma, Cheng Zou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, GuangMing Yao, and 1 others. 2025. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation. *arXiv preprint arXiv:2510.24821*.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of whisper asr hallucinations induced by non-speech audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: a language modeling approach to audio generation. *Preprint*, arXiv:2209.03143.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

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 *Interspeech 2021*. ISCA.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. 2020. Spot the conversation: speaker diarisation in the wild. *CoRR*, abs/2007.01216.

Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2004. Fisher english training speech part 1 transcripts. Linguistic Data Consortium. LDC2004T19.

Alexandre Défossez. 2021. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*.

Karan Dua, Puneet Mittal, Ranjeet Gupta, and Hitesh Laxmichand Patel. 2025. Speechweave: Diverse multilingual synthetic text & audio data generation pipeline for training text to speech models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, page 718–737. Association for Computational Linguistics.

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. *Preprint*, arXiv:2410.00037.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354. IEEE.

John J. Godfrey and Edward Holliman. 1993. Switchboard-1 release 2. Linguistic Data Consortium (LDC). LDC Catalog No.: LDC97S62. DOI: `https://doi.org/10.35111/sw3h-rw02`.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. *TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation*, page 198–208. Springer International Publishing.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and Eng Siong Chng. 2024a. Large language models are efficient learners of noise-robust speech recognition. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EnSiong Chng. 2024b. Large language models are efficient learners of noise-robust speech recognition. *Preprint*, arXiv:2401.10446.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kyudan Jung, Seungmin Bae, Nam Joon Kim, Hyun Gon Ryu, and Hyuk-Jae Lee. 2024a. Improving asr performance with ocr through using word frequency difference. In *2024 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4.

Kyudan Jung, Nam-Joon Kim, Hyongon Ryu, Sieun Hyeon, Seung jun Lee, and Hyeok jae Lee. 2024b. Texbleu: Automatic metric for evaluate latex format. *Preprint*, arXiv:2409.06639.

Yuka Ko, Sheng Li, Chao-Han Huck Yang, and Tatsuya Kawahara. 2024a. Benchmarking japanese speech recognition on asr-llm setups with multi-pass augmented generative error correction. *Preprint*, arXiv:2408.16180.

Yuka Ko, Sheng Li, Chao-Han Huck Yang, and Tatsuya Kawahara. 2024b. Benchmarking japanese speech recognition on asr-llm setups with multi-pass augmented generative error correction. *Preprint*, arXiv:2408.16180.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024a. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024b. Careless whisper: Speech-to-text hallucination harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1672–1681. ACM.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117.

Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho. 2025. Dittotts: Diffusion transformers for scalable text-to-speech without domain-specific factors. *Preprint*, arXiv:2406.11427.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, and Hung-yi Lee. 2025a. Full-duplex-bench v1. 5: Evaluating overlap handling for full-duplex speech models. *arXiv preprint arXiv:2507.23159*.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. 2025b. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*.

Harras Mansoor, Umer Abdullah, Shahryar Adil, Akhtar Jamil, Alaa Ali Hameed, and Faezeh Soleimani. 2025. Mitigating hallucinations in speech recognition systems for noisy data. In *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*, pages 1–5. IEEE.

Brian McFee. 2025. librosa/librosa: 0.11.0. Version 0.11.0.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Taejin Park, Ivan Medennikov, Kunal Dhawan, Weiqing Wang, He Huang, Nithin Rao Koluguri, Krishna C. Puvvada, Jagadeesh Balam, and Boris Ginsburg. 2025. Sortformer: A novel approach for permutation-resolved speaker supervision in speech-to-text systems. *Preprint*, arXiv:2409.06656.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022a. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022b. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér. 2023a. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 10007–10016. Association for Computational Linguistics.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez-Cabrero, and Jesper Tegnér. 2023b. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 23*.

Rajarshi Roy, Jonathan Raiman, Sang gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. 2026. Personaplex: Voice and role control for full duplex conversational speech models. *Preprint*, arXiv:2602.06053.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

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. *Preprint*, arXiv:2509.14128.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, and Ann Lee. 2025. Sam audio: Segment anything in audio. *Preprint*, arXiv:2512.18099.

Ui-Hyeop Shin, Sangyoun Lee, Taehan Kim, and Hyung-Min Park. 2024. Separate and reconstruct: Asymmetric encoder-decoder for speech separation. *Advances in Neural Information Processing Systems*, 37:52215–52240.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 15725–15788.

NAVER Cloud HyperCLOVA X Team. 2026. Hyperclova x 8b omni. *Preprint*, arXiv:2601.01792.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Anvarjon Tursunov, Soonil Kwon, and Hee-Suk Pang. 2019. Discriminating emotions in the valence dimension from speech using timbre features. *Applied Sciences*, 9(12):2470.

Vishaal Udandarao, Zhiyun Lu, Xuankai Chang, Yongqiang Wang, Violet Z. Yao, Albin Madapally Jose, Fartash Faghri, Josh Gardner, and Chung-Cheng Chiu. 2025. Data-centric lessons to improve speech-language pretraining. *Preprint*, arXiv:2510.20860.

Helin Wang, Jiarui Hai, Dading Chong, Karan Thakkar, Tiantian Feng, Dongchao Yang, Junhyeok Lee, Thomas Thebaud, Laureano Moro Velazquez, Jesus Villalba, and 1 others. 2025a. Capspeech: Enabling downstream applications in style-captioned text-to-speech. *arXiv preprint arXiv:2506.02863*.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024a. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12121–12125. IEEE.

Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Wei Xia, and Yuanjun Xiong. 2024b. A full-duplex speech dialogue scheme based on large language models. *Preprint*, arXiv:2405.19487.

Yiyang Wang, Chen Chen, Tica Lin, Vishnu Raj, Josh Kimball, Alex Cabral, and Josiah Hester. 2025b. Companioncast: A multi-agent conversational ai framework with spatial audio for social co-viewing experiences. *Preprint*, arXiv:2512.10918.

Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, and 1 others. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, Kaimeng Ren, Ming Yang, Mingxue Yang, Qiang Xu, Qin Zhao, Ruijie Xiong, Shaoxiong Lin, Xuezhi Wang, Yi Yuan, and 6 others. 2025. Ming-uniaudio: Speech llm for joint understanding, generation and editing with unified representation. *Preprint*, arXiv:2511.05516.

Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, Ke Li, Shuai Fan, Kai Yu, Wei-Qiang Zhang, Guoguo Chen, and Xie Chen. 2025a. Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement. *Preprint*, arXiv:2406.11546.

Zhengdong Yang, Zhen Wan, Sheng Li, Chao-Han Huck Yang, and Chenhui Chu. 2025b. CoVoGER: A multilingual multitask benchmark for speech-to-text generative error correction with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6314, Suzhou, China. Association for Computational Linguistics.

Zhengdong Yang, Zhen Wan, Sheng Li, Chao-Han Huck Yang, and Chenhui Chu. 2025c. CoVoGER: A multilingual multitask benchmark for speech-to-text generative error correction with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6314, Suzhou, China. Association for Computational Linguistics.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. *Preprint*, arXiv:2110.03370.

Appendix

We provide detailed supplementary materials organized as follows:

- **Appendix** A provides per-sample WER tail statistics across the three ASR test splits.
- **Appendix** B reports the full generative error correction (GER) ablation.
- **Appendix** C extends the main-body related-work discussion to large-scale speech datasets and automated processing pipelines.
- **Appendix** D presents additional experimental results regarding overlap separation.
- **Appendix** E illustrates specific cases for handling backchanneling and overlapping speech.
- **Appendix** F details supplementary results from the fine-tuning experiments.
- **Appendix** G describes the techniques employed for audio captioning.
- **Appendix** H provides examples of data processed by SOMMELIER.

A Per-sample WER distribution analysis

The variance statistics reported in Table 4 (Std) can be read as a heavy-tailed phenomenon: a small number of samples with WER $\gtrsim 0.5$ dominate the standard deviation while the median sample has WER $= 0$. Table 6–8 report per-model tail statistics across the three ASR evaluation splits; %>0.5 and %>1.0 are the fraction of samples exceeding those thresholds. On `librispeech` the ensemble pushes P99 below 0.28 and %>0.5 below 0.25%; the only sample with WER > 1.0 (max $= 1.50$) is a long hallucination from the Canary component that survives voting. On `librispeech-test-other` and `TEDLIUM3` the tail broadens as expected from acoustic conditions, but again concentrates in a small minority of utterances (Figure 4). Per-duration and per-reference-length binned statistics are included in the released CSVs (`exp_asr/results/analysis/`).

Table 6: Per-sample WER tail statistics — LibriSpeech `test-clean` ($N = 2{,}620$).

Model	N	Mean	Std	P50	P90	P95	P99	P99.9	%>0.5	%>1.0	Max
whisper	2620	0.036	0.086	0.000	0.111	0.167	0.400	0.808	0.61	0.04	1.195
parakeet	2620	0.024	0.061	0.000	0.077	0.143	0.286	0.538	0.11	0.00	0.750
canary	2620	0.021	0.066	0.000	0.067	0.125	0.273	0.698	0.23	0.04	1.500
moe	2620	0.020	0.065	0.000	0.067	0.125	0.272	0.698	0.23	0.04	1.500

Table 7: Per-sample WER tail statistics — LibriSpeech `test-other` ($N = 2{,}939$).

Model	N	Mean	Std	P50	P90	P95	P99	P99.9	%>0.5	%>1.0	Max
whisper	2939	0.063	0.119	0.000	0.182	0.286	0.524	1.000	1.02	0.07	1.739
parakeet	2939	0.050	0.105	0.000	0.150	0.238	0.500	0.812	0.65	0.07	2.000
canary	2939	0.040	0.090	0.000	0.133	0.200	0.371	1.000	0.41	0.03	1.500
moe	2939	0.039	0.089	0.000	0.131	0.200	0.375	1.000	0.41	0.03	1.500

B Generative Error Correction (GER) ablation

We study an optional GER stage that consumes the three ensemble hypotheses (Whisper, Parakeet, Canary) and asks Qwen2.5-3B-Instruct to emit a single corrected transcript. The prompt lists the three hypotheses as "Transcription from system A/B/C" and requests the most likely correct transcription. We enable `repetition_penalty=` 1.1 and `no_repeat_ngram_size=` 6 to suppress degenerate loops, and strip chat-template / tool-call tokens from the generated text. No LM fine-tuning is performed; GER is evaluated zero-shot.

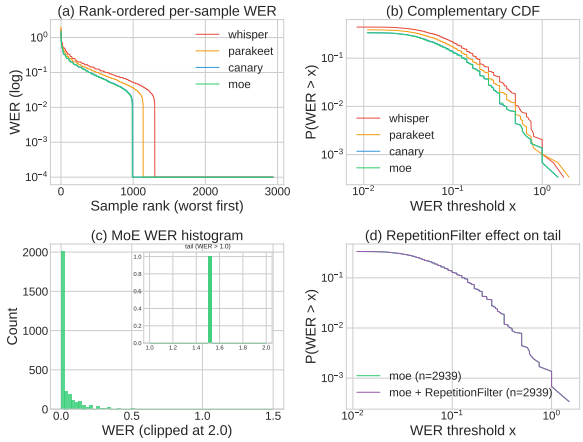
Table 9 reports mean, standard deviation, P95, P99, and hallucination rate (fraction of samples with WER > 0.5 or flagged by the 15-gram `RepetitionFilter`) for each configuration. Across all three splits, zero-shot GER *increases* WER and *broadens* the tail: on librispeech-`test-clean` the P99 jumps from 0.27 to 0.97, and the hallucination rate from 0.23% to 17.1%. The primary failure mode is LLM paraphrase and over-insertion (e.g., replacing `etcetera` with `et cetera`, adding punctuation, switching casing) that survives text normalisation only partially, combined with occasional Qwen-specific degeneracies (e.g., Chinese character leakage). Figure 5 shows the complementary CDF comparison across the three splits; the GER curve lies strictly to the right of the ensemble + RepetitionFilter curve everywhere.

C Related Work (continued)

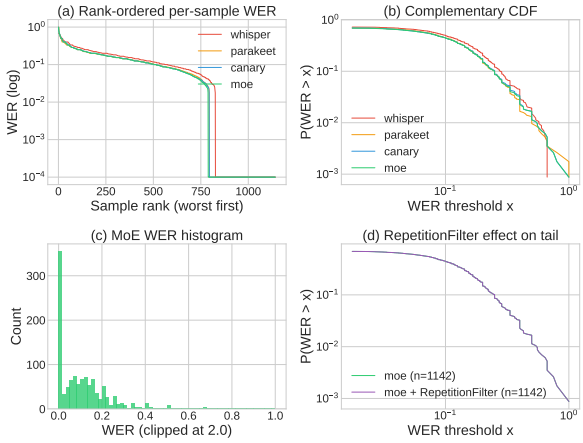
The short treatment in §2 is expanded here; we focus on topics that did not fit in the main body's page budget.

Table 8: Per-sample WER tail statistics — TEDLIUM3 `test` ($N = 1{,}142$).

Model	N	Mean	Std	P50	P90	P95	P99	P99.9	%>0.5	%>1.0	Max
whisper	1142	0.122	0.123	0.100	0.278	0.363	0.558	0.667	1.40	0.00	0.667
parakeet	1142	0.107	0.116	0.086	0.250	0.315	0.500	0.972	0.88	0.00	1.000
canary	1142	0.108	0.118	0.083	0.250	0.333	0.523	0.793	1.05	0.00	1.000
moe	1142	0.107	0.117	0.083	0.250	0.316	0.523	0.793	1.05	0.00	1.000



(a) librispeech-`test-other`



(b) TEDLIUM3 `test`

Figure 4: Per-sample WER distributions on the remaining two ASR splits. Panels follow Figure 3.

C.1 Large-Scale Speech Datasets

Existing speech datasets, despite their increasing volume, remain suboptimal for training full-duplex models that require rich interactional dynamics. Traditional ASR benchmarks like LibriSpeech (Panayotov et al., 2015) and GigaSpeech (Chen et al., 2021) are dominated by scripted read speech or solitary monologues, failing to capture the dynamic and interactive spontaneity of human dialogue. While conversational datasets such as Fisher (Cieri et al., 2004) and Switchboard (Godfrey and Holliman, 1993) offer multi-speaker interactions, they are severely limited by their archaic telephony quality (8kHz), narrow bandwidth, and relatively small scale (typically a few thousand hours), which is insufficient for modern large-scale pre-training (Radford et al., 2022b; Xu et al., 2025b). While recent web-scale initiatives like WenetSpeech (Zhang et al., 2022) and Emilia (He et al., 2024) have successfully aggregated massive datasets, their pipelines are heavily optimized for single-stream speech, thereby neglecting the concurrent dynamics required for full-duplex interaction. Crucially, their pre-processing pipelines treat overlapping speech as noise to be excised or ignored rather than a feature to be modeled. This structural limitation results in data that lacks the distinct multi-stream separation and essential acoustic collisions required for learning true full-duplex interaction.

C.2 Automated Speech Data Processing Pipelines

While open-source data processing pipelines have become the bedrock of Large Language Model (LLM) research, exemplified by transparent frameworks like Dolma (Soldaini et al., 2024), RedPajama (Weber et al., 2024), and FineWeb (Penedo et al., 2024), the domain of speech processing remains significantly opaque. Although model weights for Speech Language Models (SLMs) are frequently released, the intricate "data recipes" required to curate high-quality pre-training corpora remain proprietary "black boxes," impeding the community's ability to reproduce results or improve upon existing strategies.

This lack of standardized, open pipelines is particularly critical when addressing the technical demands of full-duplex communication. Current methodologies rely heavily on tools designed for single-stream processing (Dua et al., 2025; Yang et al., 2025a), which are ill-suited for capturing the concurrent dynamics of human dialogue. For instance, while speaker diarization is a prerequisite for multi-turn modeling, standard tools like Pyannote (Bredin, 2023; Plaquet and Bredin, 2023) often struggle in the complex acoustic environments of in-the-wild web videos. Crucially, these tools frequently misinterpret the *overlaps* and rapid *turn-taking*, essential features of full-duplex interaction, as segmentation errors or noise, thereby degrading the structural integrity of the conversational data.

Furthermore, the reliance on ASR models for transcription introduces the risk of hallucina-

Table 9: Ablation: ensemble stage vs. optional GER, across librispeech-`test-clean`, librispeech-`test-other`, TEDLIUM3. Mean, Std, P95, P99 are WER quantiles; Halluc. % is the fraction of samples with WER>0.5 or flagged by the 15-gram RepetitionFilter. GER consistently degrades WER on already-clean data, motivating our decision to ship it as optional.

Configuration	LS-test-clean					LS-test-other					TedLium3-test				
	Mean	Std	P95	P99	Halluc.%	Mean	Std	P95	P99	Halluc.%	Mean	Std	P95	P99	Halluc.%
best single (canary)	0.021	0.066	0.125	0.273	0.23	0.040	0.090	0.200	0.371	0.41	–	–	–	–	–
ensemble	0.020	0.065	0.125	0.272	0.23	0.039	0.089	0.200	0.375	0.41	0.107	0.117	0.316	0.523	1.05
ensemble + RF	0.020	0.065	0.125	0.272	0.23	0.039	0.089	0.200	0.375	0.41	0.107	0.117	0.316	0.523	1.05
ensemble + RF + GER (Qwen2.5-3B-Instruct)	0.334	0.405	0.811	0.970	17.14	0.313	0.244	0.800	0.941	15.58	0.413	0.262	0.871	0.955	30.82
best single (parakeet)	–	–	–	–	–	–	–	–	–	–	0.107	0.116	0.315	0.500	0.88

tions. Models like Whisper (Radford et al., 2022b), though powerful, are prone to generating repetitive loops or nonsensical text during silence or non-speech intervals, a critical instability highlighted in recent studies such as *Careless Whisper* (Koenecke et al., 2024b). Existing pipelines lack the robustness to filter these hallucinations or handle the multi-stream nature of duplex speech, underscoring the urgent need for a transparent, hallucination-aware processing framework tailored for conversational AI.

D Detail of Overlapping disentangle experiments

This section presents detailed experimental results on the efficacy of our overlap separation module. We evaluated performance across varying Signal-to-Interference Ratios (SIR) and Overlap Ratios using four key metrics: Word Error Rate (WER), Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Short-Time Objective Intelligibility (STOI), and UTMOS. The results, summarized in Tables 10 through 12, demonstrate that applying the separation module ('Sep') consistently improves signal quality compared to the baseline ('Base'). Notably, the performance gain is significantly larger for Speaker 2 (the interfering or secondary speaker) than for Speaker 1 (the primary speaker), particularly in challenging conditions with high overlap ratios as shown in Figure 6.

D.1 Analysis of Results

Asymmetric Gains (Spk1 vs. Spk2) Across all metrics, the gap between the 'Base' and 'Sep' conditions is most dramatic for Speaker 2. For instance, in the 0,dB SIR and 1.0 overlap condition, the WER for Speaker 2 improves drastically from 0.444 to 0.138 (Table 10), whereas Speaker 1 sees a relatively smaller, though still significant, improvement. This suggests that our module is particularly effective at recovering the subordinate or quieter speaker in a mixture, which is crucial for full-duplex conversational AI where both parties must be heard clearly.

SIR	Overlap	Speaker 1			Speaker 2		
		Base	Sep	Oracle	Base	Sep	Oracle
0 dB	0.2	0.109	**0.048**	0.034	0.100	**0.074**	0.061
	0.5	0.162	**0.094**	0.078	0.115	**0.065**	0.039
	1.0	0.535	**0.175**	0.058	0.444	**0.138**	0.048
5 dB	0.2	**0.080**	0.088	0.066	0.146	**0.065**	0.040
	0.5	0.099	**0.059**	0.036	0.277	**0.084**	0.049
	1.0	0.136	**0.069**	0.039	0.913	**0.113**	0.042
10 dB	0.2	0.059	**0.056**	0.058	0.194	**0.084**	0.053
	0.5	0.067	**0.064**	0.051	0.527	**0.138**	0.052
	1.0	0.096	**0.083**	0.051	0.923	**0.193**	0.044

Table 10: Word Error Rate (WER) comparison across different conditions. Lower is better.

SIR	Overlap	Speaker 1			Speaker 2		
		Base	Sep	Oracle	Base	Sep	Oracle
0 dB	0.2	0.959	**0.980**	1.000	0.964	**0.985**	1.000
	0.5	0.889	**0.969**	1.000	0.887	**0.968**	1.000
	1.0	0.785	**0.918**	1.000	0.771	**0.908**	1.000
5 dB	0.2	0.971	**0.980**	1.000	0.951	**0.976**	1.000
	0.5	0.929	**0.980**	1.000	0.844	**0.962**	1.000
	1.0	0.847	**0.955**	1.000	0.676	**0.917**	1.000
10 dB	0.2	0.985	**0.988**	1.000	0.938	**0.971**	1.000
	0.5	0.956	**0.981**	1.000	0.798	**0.931**	1.000
	1.0	0.901	**0.954**	1.000	0.608	**0.883**	1.000

Table 11: Short-Time Objective Intelligibility (STOI) scores. Higher is better.

Resilience to High Overlap The benefits of separation become more pronounced as the overlap ratio increases. In the worst-case scenario (1.0 overlap), the baseline UTMOS scores (Table 12) drop severely (e.g., $\approx$1.7), but the separation module restores quality to near-natural levels ($\approx$3.0). Similarly, STOI scores (Table 11) remain high ($>$0.9) even under full overlap when separation is applied, confirming that intelligibility is preserved.

the Baseline method exhibits competitive SI-SDR performance at $\rho \in \{0.2, 0.5\}$. However,

SIR	Overlap	Speaker 1			Speaker 2		
		Base	Sep	Oracle	Base	Sep	Oracle
0 dB	0.2	3.08	**3.56**	3.94	2.99	**3.50**	3.82
	0.5	2.28	**3.29**	3.85	2.25	**3.35**	3.90
	1.0	1.73	**3.05**	3.81	1.67	**2.99**	3.86
5 dB	0.2	3.12	**3.47**	3.83	3.00	**3.48**	3.90
	0.5	2.46	**3.51**	3.90	2.22	**3.28**	3.92
	1.0	1.86	**3.29**	3.94	1.72	**2.94**	3.87
10 dB	0.2	3.39	**3.69**	3.95	3.12	**3.51**	4.02
	0.5	2.82	**3.50**	3.91	2.35	**2.92**	3.82
	1.0	2.22	**3.29**	3.95	2.12	**2.73**	3.90

Table 12: UTMOS (MOS prediction) scores. Higher is better.

this comparison is inherently unfair: the Baseline extracts each speaker's time segment directly from the mixed signal, meaning that at 20% and 50% overlap ratios, the majority of each segment contains clean, non-overlapping speech. For instance, at $\rho = 0.2$, the actual overlap constitutes only 15.2% for S1 and 14.7% for S2, leaving approximately 7 seconds of clean audio per speaker. Consequently, ASR systems can effectively recognize the clean portions, resulting in artificially low WER for the Baseline.

E Overlap cases

In this section, we present two representative overlap cases. The first is *backchanneling*, where a speaker produces a short utterance while the other is speaking; in this case, one segment is fully contained within another segment. The second is *overlap*, where two segments partially overlap but neither segment fully contains the other.

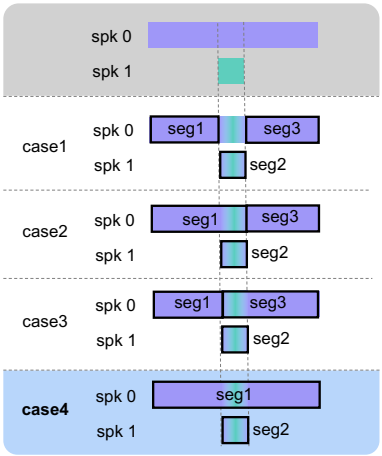


Figure 7: Four ways to handle backchanneling in overlapping speech.

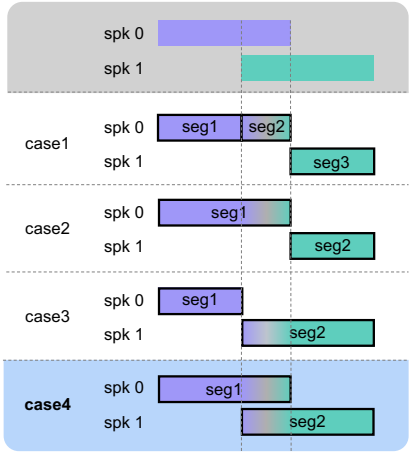


Figure 8: Four distinct types of separable cases in overlapping speech.

F Detail of Finetuning Experiment

The fine-tuning hyperparameters for Moshiko are listed in Table 13. We significantly benefited from the implementation provided at `https://github.com/kyutai-labs/moshi-finetune`.

Hyperparameter	Value
Total Data Duration	$\approx$ 83 hours
Training Steps	2,000
Hardware	8 $\times$ A100
Rank	128
Batch Size	16
Learning Rate	2e$-$6
Weight Decay	0.1

Table 13: Training hyperparameters and settings.

F.1 Dataset Statistics

This section provides statistics for the data fine-tuned in Section 4.1. Figure 9 illustrates that our training data originates from a wide range of conversational domains.

F.2 Full-Duplex-Bench 1.0: Metric Definitions

Full-Duplex-Bench 1.0 evaluates spoken dialogue models under full-duplex conditions, focusing on pause handling, backchanneling, smooth turn-taking, and user interruption handling. Across all tasks, we define latency for an instance i as $\Delta_i = t_{\text{start},i} - t_{\text{end},i}$, where t_{start} denotes the model's response onset and t_{end} denotes the end of the relevant user event.

Pause Handling. To evaluate whether the model incorrectly treats mid-utterance silence as a turn boundary, we use **Synthetic/Candor TOR** (Turn-Over Rate, ↓). This metric calculates the fraction of pause instances in which the model starts speaking. A failure is recorded if the model's output during a pause exceeds a minimal threshold (duration ≥ 1 second or > 3 words).

Backchanneling. We assess the model's ability to provide brief acknowledgements without seizing the floor using three metrics. **Backchannel TOR** (↓) measures the fraction of backchannel-eligible windows where the model produces a full turn (duration $\geq$ 3s or $\geq$ 1s with > 3 words). **Frequency** (↑) reports the number of backchannels normalized by total audio duration. Finally, **JSD** (↓) computes the Jensen–Shannon divergence between the model's backchannel timing distribution and human ground truth to evaluate timing naturalness.

Smooth Turn Taking. This task measures the model's promptness in responding after the user completes an utterance. We report **Candor TOR** (↑), defined as the fraction of user-turn endings where the model successfully begins speaking, and **Latency** (↓), measured only on instances where the model successfully takes the turn.

User Interruption. When a user interrupts the model, we evaluate the system's responsiveness and contextual adaptation. **Interruption TOR** (↑) measures the fraction of interruption events where the model responds. **Latency** (↓) tracks the time from the end of the interruption to the model's response. Additionally, we use a **GPT-4o relevance score** (↑) (0–5 scale) to assess whether the model's response is semantically relevant to the content of the interruption.

F.3 Results on Full-Duplex-Bench 1.5

Lin et al. (2025a) released Full-Duplex-Bench 1.5 as a successor to v1.0. We further evaluated the model fine-tuned on SOMMELIER-processed data using this updated benchmark.

The experimental results presented in Table 14 and Table 15 demonstrate that the SOMMELIER-fine-tuned Moshi model significantly outperforms the base model across all overlap scenarios. In terms of audio quality, the fine-tuned model exhibits superior signal fidelity and robustness, evidenced by substantial gains in PESQ and SI-SDR scores; notably, the SI-SDR for the 'Background Speech' scenario improved drastically from 5.43 dB to 20.76 dB. Furthermore, the latency analysis reveals a critical enhancement in conversational responsiveness, with both stop and response latencies reduced to sub-second averages in the majority of cases, thereby enabling more natural and immediate turn-taking interactions.

G Context Captioning

Speech data contains rich non-verbal information, such as timbre and emotion, beyond text semantics (Koolagudi and Rao, 2012; Tursunov et al., 2019; Lee et al., 2025; Jung et al., 2024b). Detailed captioning of this information serves as effective metadata for speech understanding and generation (Wang et al., 2025a; AI et al., 2025; Yan et al., 2025). Unlike other studies, we propose captioning audio segments using the Qwen3-Omni-Captioner (Xu et al., 2025b) model to generate rich metadata, including emotion, gender, age group, and situation descriptions.

However, captioning short segments individually can fail to capture context (e.g., sarcasm). To address this, we implemented context-aware captioning by providing the preceding two segments as audio prompts (In-Context Learning). Specifically, for consecutive audio segments a_1, a_2, and a_3, we calculate the conditional probability $P(C_3|I, a_1, a_2)$ to generate the caption C_3 for a_3.

H Example

This section presents real-world podcast examples. Figure 10 visualizes the data processed by SOMMELIER, followed by an example of the corresponding JSON file.

Table 14: Full-Duplex-Bench v1.5 Evaluation Results (Moshi vs. Fine-tuned Moshi). Comparison of post-distractor audio quality metrics and behavior classification across four overlap scenarios.

Full-Duplex-Bench v1.5: Overlap Handling Evaluation													
Scenario	**Audio Quality (Post)**				**Rate**	**Pitch**		**Intensity**		**Behavior Ratio**			
	STOI↑	PESQ↑	SI-SDR↑	UTMOS↑	WPM	μ	σ	μ	σ	RESP	RESU	UNCERT	UNK
Moshi (Base)													
Background Speech	0.79	2.19	5.43	1.86	75.9	85.9	11.5	-64.6	16.3	0.15	0.07	0.03	0.75
Talking to Other	0.90	2.55	12.64	**2.34**	124.5	96.6	16.0	-49.1	16.2	0.15	0.18	0.04	0.63
User Backchannel	0.63	1.60	-6.57	1.25	25.8	66.5	6.2	-87.0	16.0	0.01	0.06	0.01	0.92
User Interruption	0.94	2.87	16.07	**2.65**	149.0	111.1	21.8	-41.3	16.0	0.59	0.17	0.03	0.21
Moshi (Fine-tuned)													
Background Speech	**0.98**	**3.33**	**20.76**	**1.87**	157.5	88.3	13.7	-64.8	16.4	0.28	0.11	0.00	0.60
Talking to Other	**0.96**	**3.30**	**20.26**	2.30	146.4	96.4	17.0	-51.0	16.3	0.18	0.16	0.00	0.63
User Backchannel	**0.91**	**3.01**	**16.48**	**1.32**	132.5	72.0	7.9	-85.2	16.0	0.08	0.11	0.00	0.72
User Interruption	**0.97**	**3.27**	**20.26**	2.58	156.0	110.7	22.8	-43.8	15.7	0.51	0.12	0.00	0.36

Table 15: Full-Duplex-Bench v1.5 Latency Analysis (Moshi vs. Fine-tuned Moshi). Stop latency measures time from user speech onset to model speech cessation. Response latency measures time from user speech offset to model speech resumption.

Latency Analysis (seconds)						
Scenario	**Stop Latency↓**		**Response Latency↓**		**Sample Count**	
	μ	σ	μ	σ	Stop	Resp
Moshi (Base)						
Background Speech	1.02	0.55	2.90	2.05	150	89
Talking to Other	1.13	**0.57**	3.22	1.87	184	117
User Backchannel	1.30	0.42	2.38	1.84	113	29
User Interruption	1.30	0.72	1.99	2.24	391	237
Moshi (Fine-tuned)						
Background Speech	**0.68**	**0.48**	**0.73**	**0.49**	44	192
Talking to Other	**0.82**	0.65	**0.84**	**0.69**	47	188
User Backchannel	**0.70**	**0.40**	**1.12**	**0.85**	57	156
User Interruption	**0.89**	**0.64**	**0.66**	**0.55**	110	383


```
{
  "metadata": {
    "audio_duration_seconds": 120.0,
    "audio_duration_minutes": 2.0,
    "vad_sortformer": {
      "processing_time_seconds": 0.9742708206176758,
      "rt_factor": 0.008118923505147297
    },
    "whisper_large_v3": {
      "processing_time_seconds": 14.903292655944824,
      "rt_factor": 0.12419410546620686
    },
    "total_segments": 26,
    "whisperx_alignment": {
      "processing_time_seconds": 32.398998737335205,
      "rt_factor": 0.26999165614446,
      "enabled": true
    },
    "sepreformer_separation": {
      "processing_time_seconds": 0.12217402458190918,
      "rt_factor": 0.00101811687151591,
      "overlap_threshold_seconds": 0.2,
      "enabled": true
    },
    "flowse_denoising": {
      "processing_time_seconds": 3.8639004230499268,
      "rt_factor": 0.032199170192082724,
      "enabled": true
    }
  },
  "segments": [
    {
      "start": 0.0,
      "end": 0.64,
      "text": "Mr. Franklin?",
      "text_whisper": "Mr. Franklin?",
      "text_parakeet": "The Franklin?",
      "text_canary": "The Franklin",
      "speaker": "SPEAKER_00",
      "language": "en",
      "demucs": false,
      "is_separated": true,
      "sepreformer": false,
      "words": [
        {
          "word": "Mr.",
          "start": 0.0,
          "end": 0.171,
          "score": 0.414
        },
        {
          "word": "Franklin?",
          "start": 0.192,
          "end": 0.661,
          "score": 0.936
        }
      ]
    },
    {
      "start": 0.48,
      "end": 1.2,
      "text": "I'm ready.",
      "text_whisper": "I'm ready.",
      "text_parakeet": "I'm ready.",
      "text_canary": "I'm ready",
      "speaker": "SPEAKER_01",
      "language": "en",
      "demucs": false,
      "is_separated": true,
      "sepreformer": false,
      "words": [
        {
          "word": "I'm",
          "start": 0.48,
          "end": 0.861,
          "score": 0.611
        },
        {
          "word": "ready.",
          "start": 0.9039999999999999,
          "end": 1.221,
          "score": 0.748
        }
      ]
    },
    {
      "start": 1.12,
      "end": 2.5599999999999996,
```

```
      "text": "It’s Ira Glass here",
      "text_whisper": "Tyra Glass here.",
      "text_parakeet": "It’s Iraq Glass here.",
      "text_canary": "It’s Ira Glass here",
      "speaker": "SPEAKER_00",
      "language": "en",
      "demucs": false,
      "is_separated": true,
      "sepreformer": false,
      "words": [
        {
          "word": "Tyra",
          "start": 1.12,
          "end": 1.9220000000000002,
          "score": 0.686
        },
        {
          "word": "Glass",
          "start": 1.943,
          "end": 2.1900000000000004,
          "score": 0.801
        },
        {
          "word": "here.",
          "start": 2.21,
          "end": 2.5810000000000004,
          "score": 0.75
        }
      ]
    },
    {
      "start": 2.8,
      "end": 7.279999999999999,
      "text": "Oh you’re the MC on the show I read about Oh great I read I read I read",
      "text_whisper": "You’re the emcee on the show, Ira. Oh, great. Ira, are you Ira? Ira?",
      "text_parakeet": "Oh, you’re the MC on the show, Ira. Oh, great. Ira Iron.",
      "text_canary": "Oh you’re the MC on the show I read about Oh great I read I read I read",
      "speaker": "SPEAKER_01",
      "language": "en",
      "demucs": false,
      "is_separated": true,
      "sepreformer": true,
      "words": [
        {
          "word": "You’re",
          "start": 2.8,
          "end": 3.2439999999999998,
          "score": 0.498
        },
        {
          "word": "the",
          "start": 3.264,
          "end": 3.3649999999999998,
          "score": 0.685
        },
        {
          "word": "emcee",
          "start": 3.4459999999999997,
          "end": 3.7479999999999998,
          "score": 0.678
        },
        {
          "word": "on",
          "start": 3.7889999999999997,
          "end": 3.8489999999999998,
          "score": 0.932
        },
        {
          "word": "the",
          "start": 3.87,
          "end": 3.9299999999999997,
          "score": 0.977
        },
        {
          "word": "show,",
          "start": 3.9499999999999997,
          "end": 4.213,
          "score": 0.72
        },
        {
          "word": "Ira.",
          "start": 4.233,
          "end": 5.282,
          "score": 0.799
        },
        {
          "word": "Oh,",
          "start": 5.302,
          "end": 5.484,
          "score": 0.832
        },
        {
          "word": "great.",
          "start": 5.645,
          "end": 5.968,
          "score": 0.85
        },
        {
          "word": "Ira,",
          "start": 6.311,
          "end": 6.614,
          "score": 0.534
        },
        {
          "word": "are",
          "start": 6.634,
          "end": 6.695,
          "score": 0.211
        },
        {
          "word": "you",
          "start": 6.775,
          "end": 6.936999999999999,
          "score": 0.317
        },
        {
          "word": "Ira?",
          "start": 6.957,
          "end": 7.119,
          "score": 0.531
        },
        {
          "word": "Ira?",
          "start": 7.139,
          "end": 7.3,
          "score": 0.582
        }
      ],
      "flowse_denoised": true
    },
    {
      "start": 4.24,
      "end": 5.76,
      "text": "I am the MC on this show, yes.",
      "text_whisper": "I am the MC on this show, yes.",
      "text_parakeet": "I am the MC on the show, yes.",
      "text_canary": "I am the MC on this show yes",
      "speaker": "SPEAKER_00",
      "language": "en",
      "demucs": false,
      "is_separated": true,
      "sepreformer": true,
      "words": [
        {
          "word": "I",
          "start": 4.24,
          "end": 4.61,
          "score": 0.838
        },
        {
          "word": "am",
          "start": 4.63,
          "end": 4.7330000000000005,
          "score": 0.946
        },
        {
          "word": "the",
          "start": 4.774,
          "end": 4.877000000000001,
          "score": 0.215
        },
        {
          "word": "MC",
          "start": 4.897,
          "end": 5.123,
```

```
        "score": 0.683
      },
      {
        "word": "on",
        "start": 5.144,
        "end": 5.205,
        "score": 0.882
      },
      {
        "word": "this",
        "start": 5.226,
        "end": 5.329000000000001,
        "score": 0.287
      },
      {
        "word": "show,",
        "start": 5.349,
        "end": 5.514,
        "score": 0.455
      },
      {
        "word": "yes.",
        "start": 5.534000000000001,
        "end": 5.781000000000001,
        "score": 0.818
      }
    ],
    "flowse_denoised": true
  },
  {
    "start": 7.36,
    "end": 8.48,
    "text": "IRA. IRA.",
    "text_whisper": "IRA. IRA.",
    "text_parakeet": "Ira, IRA.",
    "text_canary": "IRA I-R-A",
    "speaker": "SPEAKER_00",
    "language": "en",
    "demucs": false,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "IRA.",
        "start": 7.36,
        "end": 8.003,
        "score": 0.779
      },
      {
        "word": "IRA.",
        "start": 8.024000000000001,
        "end": 8.501000000000001,
        "score": 0.648
      }
    ]
  },
  {
    "start": 8.48,
    "end": 11.2,
    "text": "Oh, great. Now hold on one second Larry. Don't don't go away.",
    "text_whisper": "Oh, great. Now, hold on one second there. Don't go away.",
    "text_parakeet": "Oh, great. Now hold on one second Larry. Don't don't go away.",
    "text_canary": "Oh great now hold on one second there don't go away",
    "speaker": "SPEAKER_01",
    "language": "en",
    "demucs": false,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "Oh,",
        "start": 8.48,
        "end": 8.825000000000001,
        "score": 0.766
      },
      {
        "word": "great.",
        "start": 8.906,
        "end": 9.292,
        "score": 0.951
      },
      {
        "word": "Now,",
        "start": 9.678,
        "end": 9.759,
        "score": 0.146
      },
      {
        "word": "hold",
        "start": 9.779,
        "end": 9.881,
        "score": 0.504
      },
      {
        "word": "on",
        "start": 9.941,
        "end": 10.002,
        "score": 0.515
      },
      {
        "word": "one",
        "start": 10.063,
        "end": 10.144,
        "score": 0.348
      },
      {
        "word": "second",
        "start": 10.165000000000001,
        "end": 10.327,
        "score": 0.295
      },
      {
        "word": "there.",
        "start": 10.347000000000001,
        "end": 10.693000000000001,
        "score": 0.322
      },
      {
        "word": "Don't",
        "start": 10.733,
        "end": 10.875,
        "score": 0.69
      },
      {
        "word": "go",
        "start": 10.896,
        "end": 10.997,
        "score": 0.406
      },
      {
        "word": "away.",
        "start": 11.017,
        "end": 11.22,
        "score": 0.513
      }
    ]
  },
  {
    "start": 12.4,
    "end": 13.44,
    "text": "Hello?",
    "text_whisper": "Hello?",
    "text_parakeet": "Hello?",
    "text_canary": "Hello",
    "speaker": "SPEAKER_01",
    "language": "en",
    "demucs": false,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "Hello?",
        "start": 12.4,
        "end": 13.461,
        "score": 0.415
      }
    ]
  },
  {
    "start": 14.48,
    "end": 18.080000000000002,
    "text": "Sheldon, call me after 3 o'clock. I've got great news for you. Ira...",
    "text_whisper": "Sheldon, call me after 3 o'clock. I've got great news for you. Ira...",
    "text_parakeet": "Sheldon, call me after three o'clock. I've got great news for you. Ira.",
    "text_canary": "Shelton McCoomy at three o'
```

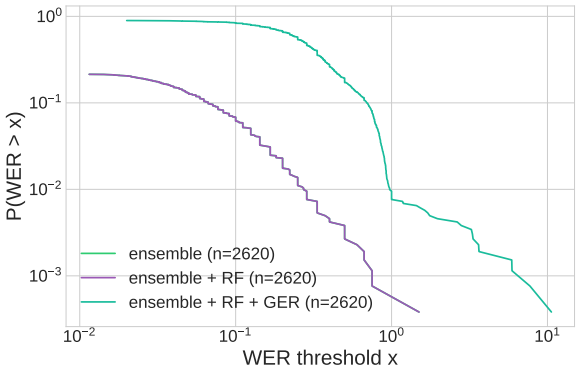
```
        clock got great news for you Irum",
    "speaker": "SPEAKER_01",
    "language": "en",
    "demucs": false,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "Sheldon,",
        "start": 14.48,
        "end": 15.046000000000001,
        "score": 0.336
      },
      {
        "word": "call",
        "start": 15.067,
        "end": 15.208,
        "score": 0.237
      },
      {
        "word": "me",
        "start": 15.228,
        "end": 15.329,
        "score": 0.714
      },
      {
        "word": "after",
        "start": 15.39,
        "end": 15.552,
        "score": 0.484
      },
      {
        "word": "3",
        "start": 15.572000000000001,
        "end": 15.754000000000001,
        "score": 0.444
      },
      {
        "word": "o'clock.",
        "start": 15.774000000000001,
        "end": 16.017,
        "score": 0.782
      },
      {
        "word": "I've",
        "start": 16.037,
        "end": 16.118000000000002,
        "score": 0.003
      },
      {
        "word": "got",
        "start": 16.138,
        "end": 16.199,
        "score": 0.242
      },
      {
        "word": "great",
        "start": 16.219,
        "end": 16.401,
        "score": 0.784
      },
      {
        "word": "news",
        "start": 16.442,
        "end": 16.624000000000002,
        "score": 0.566
      },
      {
        "word": "for",
        "start": 16.664,
        "end": 16.806,
        "score": 0.624
      },
      {
        "word": "you.",
        "start": 16.846,
        "end": 17.008,
        "score": 0.733
      },
      {
        "word": "Ira...",
        "start": 17.615000000000002,
        "end": 18.1,
        "score": 0.818
      }
    ]
  },
  {
    "start": 18.88,
    "end": 19.12,
    "text": "Yes.",
    "text_whisper": "Yes.",
    "text_parakeet": "Yeah.",
    "text_canary": "Yeah",
    "speaker": "SPEAKER_00",
    "language": "en",
    "demucs": false,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "Yes.",
        "start": 18.88,
        "end": 19.144,
        "score": 0.471
      }
    ]
  },
  {
    "start": 18.96,
    "end": 33.28,
    "text": "So uh listen, Tony if the phone rings, take it in the back and then tell me then come out and tell me who it is. is. Just Just say Joe's being with a camera crew. Just for about 10 minutes. We'll do about five minutes, ten minutes, right, Irv?",
    "text_whisper": "So, listen, Tony, if the phone rings, take it in the back, and then come out and tell me who it is. Just say Joe's being with a camera crew. Just for about ten minutes. We'll do about five minutes, ten minutes, right, Iris?",
    "text_parakeet": "So listen, Tony. If the phone rings, take it in the back and tell me, then come out and tell me who it is. Just say Joe's being with a camera crew. Just for about 10 minutes. We'll do about five minutes, ten minutes, right, Ivory?",
    "text_canary": "So uh listen Tony if the phone rings take it in the back and then tell me then come out and tell me who it is just say just say Joe's being with the camera crew just for about 10 minutes we'll do a five minute ten minutes right Irv?",
    "speaker": "SPEAKER_01",
    "language": "en",
    "demucs": true,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "So,",
        "start": 18.96,
        "end": 19.983,
        "score": 0.798
      },
      {
        "word": "listen,",
        "start": 20.685000000000002,
        "end": 20.905,
        "score": 0.823
      },
      {
        "word": "Tony,",
        "start": 20.946,
        "end": 21.186,
        "score": 0.799
      },
      {
        "word": "if",
        "start": 21.928,
        "end": 22.029,
        "score": 0.914
      },
      {
        "word": "the",
        "start": 22.069000000000003,
        "end": 22.169,
        "score": 0.872
```

```
      },
      {
        "word": "phone",
        "start": 22.229,
        "end": 22.51,
        "score": 0.66
      },
      {
        "word": "rings,",
        "start": 22.57,
        "end": 22.851,
        "score": 0.851
      },
      {
        "word": "take",
        "start": 23.573,
        "end": 23.753,
        "score": 0.923
      },
      {
        "word": "it",
        "start": 23.814,
        "end": 23.854,
        "score": 0.979
      },
      {
        "word": "in",
        "start": 23.894000000000002,
        "end": 23.934,
        "score": 0.989
      },
      {
        "word": "the",
        "start": 23.974,
        "end": 24.034,
        "score": 0.967
      },
      {
        "word": "back,",
        "start": 24.094,
        "end": 24.335,
        "score": 0.953
      },
      {
        "word": "and",
        "start": 25.157,
        "end": 25.438000000000002,
        "score": 0.708
      },
      {
        "word": "then",
        "start": 25.839,
        "end": 26.02,
        "score": 0.77
      },
      {
        "word": "come",
        "start": 26.28,
        "end": 26.381,
        "score": 0.906
      },
      {
        "word": "out",
        "start": 26.401,
        "end": 26.481,
        "score": 0.957
      },
      {
        "word": "and",
        "start": 26.521,
        "end": 26.581000000000003,
        "score": 0.993
      },
      {
        "word": "tell",
        "start": 26.601,
        "end": 26.722,
        "score": 0.854
      },
      {
        "word": "me",
        "start": 26.762,
        "end": 26.822000000000003,
        "score": 0.966
      },
      {
        "word": "who",
        "start": 26.862000000000002,
        "end": 26.962000000000003,
        "score": 0.963
      },
      {
        "word": "it",
        "start": 27.002000000000002,
        "end": 27.043,
        "score": 0.879
      },
      {
        "word": "is.",
        "start": 27.103,
        "end": 27.163,
        "score": 0.86
      },
      {
        "word": "Just",
        "start": 27.183,
        "end": 27.303,
        "score": 0.927
      },
      {
        "word": "say",
        "start": 27.343,
        "end": 27.444000000000003,
        "score": 0.782
      },
      {
        "word": "Joe's",
        "start": 28.306,
        "end": 28.527,
        "score": 0.608
      },
      {
        "word": "being",
        "start": 28.567,
        "end": 28.767000000000003,
        "score": 0.883
      },
      {
        "word": "with",
        "start": 29.229,
        "end": 29.349,
        "score": 0.822
      },
      {
        "word": "a",
        "start": 29.389000000000003,
        "end": 29.409,
        "score": 0.432
      },
      {
        "word": "camera",
        "start": 29.449,
        "end": 29.79,
        "score": 0.854
      },
      {
        "word": "crew.",
        "start": 29.810000000000002,
        "end": 30.011000000000003,
        "score": 0.585
      },
      {
        "word": "Just",
        "start": 30.633000000000003,
        "end": 30.773000000000003,
        "score": 0.973
      },
      {
        "word": "for",
        "start": 30.793,
        "end": 30.873,
        "score": 0.847
      },
      {
        "word": "about",
        "start": 30.893,
        "end": 31.034,
        "score": 0.98
      },
      {
        "word": "ten",
        "start": 31.074,
        "end": 31.234,
        "score": 0.669
```

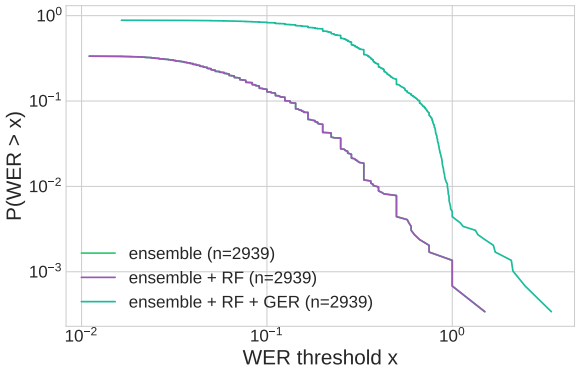
```
      },
      {
        "word": "minutes.",
        "start": 31.254,
        "end": 31.535,
        "score": 0.69
      },
      {
        "word": "We'll",
        "start": 31.555,
        "end": 31.816000000000003,
        "score": 0.449
      },
      {
        "word": "do",
        "start": 31.836,
        "end": 31.956000000000003,
        "score": 0.81
      },
      {
        "word": "about",
        "start": 31.976,
        "end": 32.077,
        "score": 0.188
      },
      {
        "word": "five",
        "start": 32.117000000000004,
        "end": 32.317,
        "score": 0.854
      },
      {
        "word": "minutes,",
        "start": 32.357,
        "end": 32.538,
        "score": 0.359
      },
      {
        "word": "ten",
        "start": 32.578,
        "end": 32.698,
        "score": 0.858
      },
      {
        "word": "minutes,",
        "start": 32.718,
        "end": 32.919,
        "score": 0.925
      },
      {
        "word": "right,",
        "start": 32.939,
        "end": 33.079,
        "score": 0.728
      },
      {
        "word": "Iris?",
        "start": 33.099000000000004,
        "end": 33.3,
        "score": 0.277
      }
    ]
  },
  {
    "start": 33.2,
    "end": 33.6,
    "text": "That's right.",
    "text_whisper": "That's right.",
    "text_parakeet": "Yep.",
    "text_canary": "That's fair",
    "speaker": "SPEAKER_00",
    "language": "en",
    "demucs": true,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "That's",
        "start": 33.2,
        "end": 33.444,
        "score": 0.426
      },
      {
        "word": "right.",
        "start": 33.489000000000004,
        "end": 33.622,
        "score": 0.164
      }
    ]
  },
  {
    "start": 40.08,
    "end": 41.04,
    "text": "Well you know what?",
    "text_whisper": "Well, you know what?",
    "text_parakeet": "Well you know what?",
    "text_canary": "Well you know what",
    "speaker": "SPEAKER_02",
    "language": "en",
    "demucs": true,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "Well,",
        "start": 40.08,
        "end": 40.455999999999996,
        "score": 0.865
      },
      {
        "word": "you",
        "start": 40.477,
        "end": 40.539,
        "score": 0.823
      },
      {
        "word": "know",
        "start": 40.580999999999996,
        "end": 40.684999999999995,
        "score": 0.822
      },
      {
        "word": "what?",
        "start": 40.79,
        "end": 41.061,
        "score": 0.358
      }
    ]
  },
  {
    "start": 41.6,
    "end": 49.68,
    "text": "Great Great thing about starting a new show is utter anonymity. Nobody really knows what to expect from you.",
    "text_whisper": "Great thing about starting a new show is utter anonymity. Nobody really knows what to expect from you.",
    "text_parakeet": "Great thing about starting a new show is utter anonymity. Nobody who really knows what to expect from you.",
    "text_canary": "The great thing about starting a new show is utter anonymity. Nobody really knows what to expect from you.",
    "speaker": "SPEAKER_02",
    "language": "en",
    "demucs": true,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "Great",
        "start": 41.6,
        "end": 42.042,
        "score": 0.86
      },
      {
        "word": "thing",
        "start": 42.062000000000005,
        "end": 42.203,
        "score": 0.797
      },
      {
        "word": "about",
        "start": 42.243,
        "end": 42.384,
        "score": 0.898
      },
      {
        "word": "starting",
        "start": 42.424,
        "end": 42.685,
```

```
        "score": 0.591
      },
      {
        "word": "a",
        "start": 42.705,
        "end": 42.726,
        "score": 0.669
      },
      {
        "word": "new",
        "start": 42.766,
        "end": 42.866,
        "score": 0.486
      },
      {
        "word": "show",
        "start": 42.906,
        "end": 43.107,
        "score": 0.976
      },
      {
        "word": "is",
        "start": 43.128,
        "end": 43.369,
        "score": 0.715
      },
      {
        "word": "utter",
        "start": 43.409,
        "end": 44.233000000000004,
        "score": 0.718
      },
      {
        "word": "anonymity.",
        "start": 44.273,
        "end": 45.298,
        "score": 0.852
      },
      {
        "word": "Nobody",
        "start": 46.866,
        "end": 47.107,
        "score": 0.66
      },
      {
        "word": "really",
        "start": 47.147,
        "end": 47.268,
        "score": 0.245
      },
      {
        "word": "knows",
        "start": 47.288000000000004,
        "end": 47.469,
        "score": 0.63
      },
      {
        "word": "what",
        "start": 47.57,
        "end": 47.75,
        "score": 0.788
      },
      {
        "word": "to",
        "start": 47.771,
        "end": 47.871,
        "score": 0.827
      },
      {
        "word": "expect",
        "start": 47.931000000000004,
        "end": 48.313,
        "score": 0.832
      },
      {
        "word": "from",
        "start": 48.353,
        "end": 49.399,
        "score": 0.958
      },
      {
        "word": "you.",
        "start": 49.459,
        "end": 49.7,
        "score": 0.861
      }
    ]
  },
  {
    "start": 50.64,
    "end": 53.04,
    "text": "This interviewee did not know us from Adam.",
    "text_whisper": "This interviewee did not know us from Adam.",
    "text_parakeet": "This interviewee did not know us from Adam.",
    "text_canary": "This interviewee did not know us from Adam",
    "speaker": "SPEAKER_02",
    "language": "en",
    "demucs": true,
    "is_separated": true,
    "sepreformer": false,
    "words": [
      {
        "word": "This",
        "start": 50.64,
        "end": 50.864,
        "score": 0.856
      },
      {
        "word": "interviewee",
        "start": 50.904,
        "end": 51.352000000000004,
        "score": 0.837
      },
      {
        "word": "did",
        "start": 51.393,
        "end": 51.494,
        "score": 0.847
      },
      {
        "word": "not",
        "start": 51.535000000000004,
        "end": 51.637,
        "score": 0.989
      },
      {
        "word": "know",
        "start": 51.698,
        "end": 51.86,
        "score": 0.847
      },
      {
        "word": "us",
        "start": 51.982,
        "end": 52.064,
        "score": 0.98
      },
      {
        "word": "from",
        "start": 52.206,
        "end": 52.409,
        "score": 0.956
      },
      {
        "word": "Adam.",
        "start": 52.43,
        "end": 53.06,
        "score": 0.447
      }
    ]
  },
  {
    "start": 54.96,
    "end": 63.04,
    "text": "Okay, well what? About a minute. We're one minute five into the new show right now. it is stretching in front of us. The perfect future.",
    "text_whisper": "Okay, well, what? About a minute. We're one minute five into the new show right now. It is stretching in front of us. The perfect future.",
    "text_parakeet": "Okay, well what? About a minute. We're one minute five into the new show. Right now, it is stretching in front of us. A perfect future.",
    "text_canary": "Okay well what about a minute well one minute five into the new show right now it is stretching in front of us a perfect future",
```

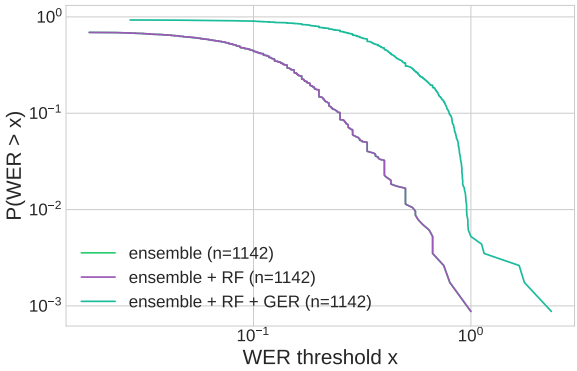
```
      "speaker": "SPEAKER_02",
      "language": "en",
      "demucs": true,
      "is_separated": true,
      "sepreformer": false,
      "words": [
        {
          "word": "Okay,",
          "start": 54.96,
          "end": 55.462,
          "score": 0.22
        },
        {
          "word": "well,",
          "start": 55.483000000000004,
          "end": 55.583,
          "score": 0.241
        },
        {
          "word": "what?",
          "start": 55.603,
          "end": 55.804,
          "score": 0.764
        },
        {
          "word": "About",
          "start": 56.086,
          "end": 56.246,
          "score": 0.999
        },
        {
          "word": "a",
          "start": 56.307,
          "end": 56.327,
          "score": 0.999
        },
        {
          "word": "minute.",
          "start": 56.407000000000004,
          "end": 56.648,
          "score": 0.943
        },
        {
          "word": "We're",
          "start": 56.668,
          "end": 56.829,
          "score": 0.377
        },
        {
          "word": "one",
          "start": 56.89,
          "end": 56.97,
          "score": 0.891
        },
        {
          "word": "minute",
          "start": 57.01,
          "end": 57.211,
          "score": 0.904
        },
        {
          "word": "five",
          "start": 57.251,
          "end": 57.774,
          "score": 0.805
        },
        {
          "word": "into",
          "start": 57.814,
          "end": 58.055,
          "score": 0.774
        },
        {
          "word": "the",
          "start": 58.075,
          "end": 58.156,
          "score": 0.828
        },
        {
          "word": "new",
          "start": 58.196,
          "end": 58.317,
          "score": 0.597
        },
        {
          "word": "show",
          "start": 58.337,
          "end": 58.538000000000004,
          "score": 0.791
        },
        {
          "word": "right",
          "start": 58.578,
          "end": 59.502,
          "score": 0.883
        },
        {
          "word": "now.",
          "start": 59.543,
          "end": 59.724000000000004,
          "score": 0.982
        }
      ]
    }
  ]
}
```



(a) librispeech-`test-clean`



(b) librispeech-`test-other`



(c) TEDLIUM3 `test`

Figure 5: GER effect on the WER tail (complementary CDF, log-log). The green `ensemble` and purple `ensemble + RepetitionFilter` curves overlap closely, whereas the teal `ensemble + RF + GER` curve is shifted right at every threshold — zero-shot GER makes the tail heavier, not lighter.

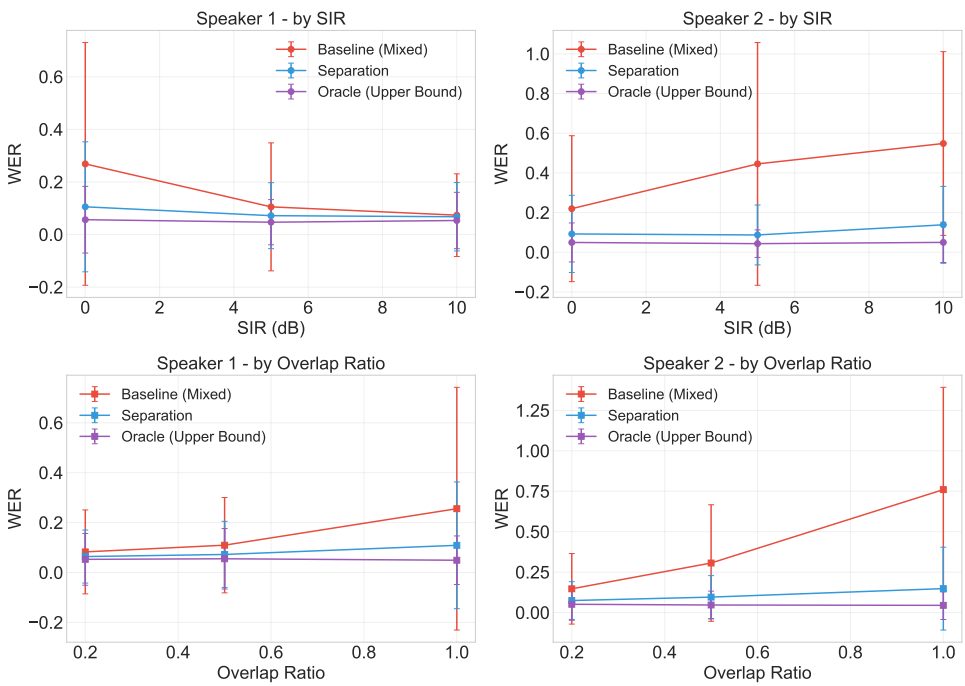


Figure 6: WER comparison by method, SIR, and overlap ratio for both speakers. Top: WER as a function of SIR (dB). Bottom: WER as a function of overlap ratio. Methods include Baseline (mixed), Separation, and Oracle. Error bars represent standard deviation.

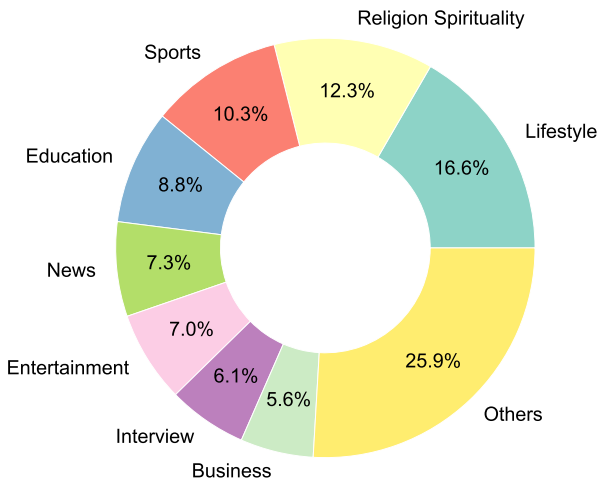


Figure 9: Category-wise statistics of the dataset used for Moshi fine-tuning experiments.

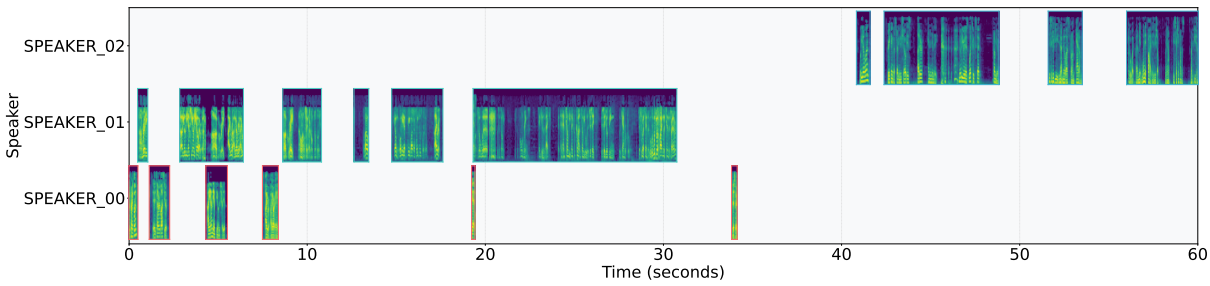


Figure 10: Visualization of preprocessing results for a 1-minute audio clip using a mel-spectrogram.