WESR: A Benchmark and Strong Baseline for Word-level Event-Speech Recognition

Chenchen Yang[1,2] **Kexin Huang**[1] **Liwei Fan**[1] **Qian Tu**[1]
Botian Jiang[1,2] **Dong Zhang**[1] **Linqi Yin**[1,2] **Shimin Li**[1]
Zhaoye Fei[1] **Qinyuan Cheng**[1] **Xipeng Qiu**[1,2*]

[1]Fudan University
[2]Shanghai Innovation Institute

ccyang6971@gmail.com, chengqy21@m.fudan.edu.cn

Abstract

Speech conveys not only linguistic information but also rich non-verbal vocal events such as laughing and crying. While semantic transcription is well-studied, the precise localization of non-verbal events remains a critical yet under-explored challenge. Current methods suffer from insufficient task definitions with limited category coverage and ambiguous temporal granularity. They also lack standardized evaluation frameworks, hindering the development of downstream applications. To bridge this gap, we first develop a refined taxonomy of 21 vocal events, with a new categorization into discrete (standalone) versus continuous (mixed with speech) types. Based on the refined taxonomy, we introduce **WESR-Bench**, an expert-annotated evaluation set (900+ utterances) with a novel position-aware protocol that disentangles ASR errors from event detection, enabling precise localization measurement for both discrete and continuous events. We also build a strong baseline by constructing a 1,700+ hour corpus, and train specialized models, surpassing both open-source audio-language models and commercial APIs while preserving ASR quality. We anticipate that WESR will serve as a foundational resource for future research in modeling rich, real-world auditory scenes. [1]

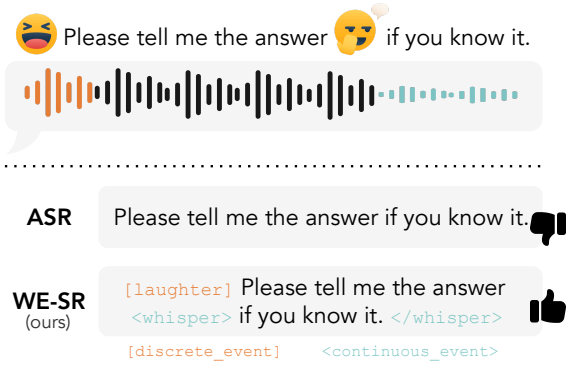

Figure 1: Task overview of our WESR that generates transcripts with explicit word-aligned continuous/discrete event tags.

1 Introduction

Speech, a crucial component of human communication, not only conveys textual information but also embeds rich non-verbal cues, such as speaker emotions, intonation, and diverse speech events like laughter, coughing, and whispering. These events reflect the speaker's intent and expressive style, significantly enhancing the contextual meaning and expressiveness of speech.

However, simply knowing that an event occurred is not enough; knowing exactly *where* it happened is just as important for understanding the meaning. Sentence-level detection often misses the specific relationship between the words and the non-verbal sounds. For instance, consider the difference between *"<laughing> I can't believe you did that! </laughing>"*, which sounds like a friendly joke, versus *"I can't believe you did <laughing> that </laughing>"*, where the laughter on a specific word might imply mockery or disbelief about the action itself. Even though the lexical content is the same, the position of laughing changes the meaning completely. Therefore, we need fine-grained, word-level modeling to capture these subtle differences and truly understand what the speaker means.

Despite this, achieving word-level vocal event transcriptions remains a significant challenge. Conventional ASR and VED systems cannot robustly perform such a task. On one hand, most ASR models focus solely on converting speech to plain text, ignoring non-verbal events as noise and thus losing crucial context. On the other hand, conventional audio event detection datasets and methods operate at the utterance or frame level classification, lacking fine-grained alignment with spoken words.

Recent work has introduced several word-level

* Corresponding author.
[1]Code and data: https://github.com/Cr-Fish/WESR

Findings of the Association for Computational Linguistics: ACL 2026, pages 3121–3134
July 2-7, 2026

Dataset	Lang.	# Cat.	Dur. (h)	Cont.	Multi.	Source	Annotation
NonverbalTTS (Borisov et al., 2025)	EN	10	17	✗	✓	VoxCeleb, Expresso	Pipeline
NVSpeech-170k (Liao et al., 2025)	EN, ZH	18	732	✗	✓	miHoYo, Emilia	Model
NonVerbalSpeech-38K (Ye et al., 2025)	EN, ZH	10	131	✓	✗	Web	Pipeline
SMIIP-NV (Wu et al., 2025)	ZH	3	33	✗	✓	In-house Recorded	–
Synparaspeech (Bai et al., 2025)	ZH	6	119	✗	✗	Synthesized	–
MNV-17 (Mai et al., 2025)	ZH	17	8	✗	✗	In-house Recorded	–
CapSpeech-SEDB (Wang et al., 2025)	EN	10	<1	✓	✗	Synthesized	–
WESR-Bench	EN, ZH	**21**	**3**	✓	✓	Web	Human
WESR-Train	EN, ZH	**21**	**1,767**	✓	✓	Web	Gemini

Table 1: Comparison of existing non-verbal speech datasets. "Lang." denotes language coverage. "# Cat." indicates the number of included categories. "Dur." denotes the total duration. "Cont." indicates whether the dataset contains continuous events. "Multi." indicates whether an utterance can be annotated with multiple event categories. "–" indicates that annotations are inherently available from the recording or synthesis process.

vocal event datasets for event-aware ASR, but they either cover only a small set of event categories, contain limited hours of audio, or exhibit highly imbalanced language proportions. In addition, while some corpora include continuous event tags, they typically treat them as overlapping sound events rather than as vocalizations that modulate the speech itself.

On the evaluation side, current event-aware ASR systems are usually assessed with word error rate (WER) together with sentence-level event classification accuracy, without explicitly accounting for the word-level positions of the events. A few studies further incorporate penalties for positional misalignment of tags within the sentence, but their metrics still cannot naturally handle multiple events occurring in the same utterance. There are also no effective evaluation methods for continuous events that span multiple words.

To address these gaps, we propose **Word-level Event-Speech Recognition (WESR)**. On the evaluation side, we introduce WESR-Bench, an expert-annotated evaluation set of 900+ utterances spanning 21 event categories (15 discrete, 6 continuous), along with a robust, position-aware evaluation protocol that decouples lexical errors from event localization and jointly scores event type and word-level alignment, with native support for multiple events in one sentence. On the modeling side, we train WESR models by constructing WESR-Train, a large-scale corpus totaling 1,700+ hours of speech with word-level event transcriptions. Through extensive experiments, we demonstrate the efficacy of our data-centric approach: our resulting models outperform strong open-source audio understanding baselines as well as commercial APIs, while preserving ASR quality. Together, our model and evaluation framework establish a reliable foundation for training and benchmarking event-aware ASR in bilingual, naturalistic speech.

In summary, our main contributions are:

- **Comprehensive Definition**: We formalize the Word-level Event–Speech Recognition (WESR) task by establishing a rigorous taxonomy of 21 vocal event categories and distinguishing between *discrete* and *continuous* attributes, providing a comprehensive framework for modeling paralinguistic information alongside lexical content.
- **Benchmarking**: We introduce WESR-BENCH, a 900+ samples, expert-annotated test set featuring bilingual, naturalistic speech based on our taxonomy.
- **A Strong Baseline for WESR**: We construct a WESR-specialized baseline that surpasses strong open-source audio language models and proprietary APIs on the WESR, serving as a convenient tool for the community.

2 Related Works

2.1 Automatic Speech Recognition (ASR)

ASR technology has achieved remarkable progress in recent years, driven by large-scale pre-training and multi-task learning. Notable models such as OpenAI's Whisper (Radford et al., 2023) support multilingual recognition and a variety of tasks including speech translation and alignment, demonstrating strong cross-lingual generalization. Similarly, models like SeedASR (Bai et al., 2024) and

FireRedASR (Xu et al., 2025b) are widely adopted in both industry and academia. Despite their impressive performance in linguistic content recognition, these models generally overlook non-verbal events and environmental information in speech.

2.2 Vocal Event Detection

Vocal event detection was initially a classification task, aiming to determine whether an audio segment contains specific types of vocal event. Early datasets such as ESC-50 (Piczak, 2015) and VocalSound (Gong et al., 2022) provide clear category labels, enabling single- or multi-label model training. These tasks typically focus on utterance-level event discrimination and struggle to accurately identify the temporal location of events. AudioSet (Gemmeke et al., 2017) expanded the task by introducing large-scale, multi-label weakly tagged audio and, in 2021, provided strong labels with precise timestamps, supporting temporal-aware event detection. Most existing methods, such as HTS-AT (Chen et al., 2022a), define event detection as utterance-level classification, while others like PANNs (Kong et al., 2020) support frame-level outputs with temporal information. However, these approaches lack integration with semantic information and joint modeling of the relationship between speech content and events.

2.3 Non-Verbal Speech Corpora

A series of recent datasets focus on the inline modeling of non-verbal events. Some methods produce word-level labels through a pipeline (frame-level event detection followed by word align to insert events into the transcription), such as NonverbalTTS (Borisov et al., 2025) and NonVerbalSpeech-38K (Ye et al., 2025). This approach ensures that event-containing speech segments can be mined from large-scale corpora, but the labeling accuracy and label coverage is limited by the event detection model, and the multi-stage process (event classification, ASR, and alignment) leads to error accumulation, while inherently limited by the performance of the annotation model (Mai et al., 2025).

Other methods, such as SynParaSpeech (Bai et al., 2025) and CapSpeech (Wang et al., 2025), obtain event-annotated data through synthesis. SynParaSpeech uses voice conversion to transform vocal event clips, while the speech portion is generated by TTS; CapSpeech relies on human annotators to insert event audio into speech recordings. These methods can achieve precise word-level labels (the insertion positions are explicitly specified), but they tend to suffer from unnaturalness in the resulting audio.

There are also approaches that rely on manual annotation. NVSpeech-170k (Liao et al., 2025) manually labeled 48k audio clips covering 18 event categories, then trained an ASR model on this annotated subset and used the model to automatically expand the dataset to 170k samples. SMIIP-NV (Wu et al., 2025) and MNV-17 (Mai et al., 2025) recruited participants to record speech with non-verbal events. SMIIP-NV collected 33 hours of manually recorded data with 3 event categories, and MNV-17 collected 7.55 hours of manually recorded Chinese data with 17 categories.

However, these methods share several common limitations: 1) the event category design is either too small or contains many only marginally distinctive tags (a summary of tags used in prior work is provided in Supplementary Table 9); 2) the amount of high-quality annotated data for robust evaluation remains limited; 3) continuous events are poorly defined—even when datasets include tags such as `<B>...</B>` to indicate continuity, they typically denote overlapping sounds rather than nonverbal speech coupled with the transcript; and 4) there is no widely accepted evaluation protocol: most event-aware ASR systems are evaluated only with classification F1 and WER/CER, which ignore word-level localization, while metrics such as TPD/NTD in NonVerbalSpeech-38K cannot handle multiple or continuous events within the same sentence, and utterance-level accuracy in MNV-17 obscures per-category behavior and still fails to capture word-level performance. Comparison of all datasets is shown in Table 1.

3 Task Formulation

To facilitate standardized evaluation of paralinguistic modeling, we define the task of WESR. Given an input audio stream, the objective of the WESR task is to generate a transcription that not only contains the spoken content (as in conventional ASR), but also annotates non-verbal vocal event tags at the correct word-level position of the transcription. Specifically, the output transcription should include: 1) **Discrete event tags:** These tags mark brief, discrete events such as `[laughs]` and `[clear_throat]`, which occur at specific time points within the audio. 2) **Continuous event**

Example in WESR-Bench	Tags	Type
好的，[clear_throat]那我们先来说一个比较轻松的话题，嗯，可能。	[clear_throat]	Discrete
Alexander! [laughs] Oh, my little warrior. Come here. Come on. [laughs]	[laughs]	Discrete
<laughing>诶，不是重点，然后我就想说</laughing>，那我就可以稍微减少一点点儿，就是碳水，然后呢可能多吃一点儿这个蛋白，然后多吃纤维。	<laughing>	Continuous
<singing> I wish you a Merry Christmas, I wish you a Merry Christmas and a Happy New Year. </singing>	<singing>	Continuous
<shouting>住手,快点住手</shouting>[giggle]那我走了。	<shouting>, [giggle]	Mixed
<crying> Oh my face, my face. </crying> <shouting> I brought sin into this world once. [inhale] I couldn't risk it again. </shouting> [sobbing]	<crying>, <shouting>, [inhale], [sobbing]	Mixed

Table 2: Examples from WESR-Bench demonstrating three event types: discrete events, continuous events, and mixed scenarios containing multiple event types within a single utterance.

tags: These capture vocal events that modulate the speech itself, such as speaking while laughing, or singing. Represented by pairs like <laughing>... </laughing> and <singing>... </singing>.

For category selection, we start from a broad list of vocal events reported in prior work and commonly observed in conversational media (e.g., podcasts, livestreams, audiobooks). We further refine this set through pilot annotation on a held-out subset of data, discarding labels that annotators find ambiguous or inconsistent, and arrive at a compact yet expressive taxonomy that covers both discrete events and continuous events while remaining feasible for large-scale labeling and reliable evaluation. The entire tag taxonomy is detailed in Supplementary Table 10.

Formally, given an input audio segment x, the model outputs a sequence $Y = (y_1, y_2, \ldots, y_n)$, where each y_i is either a word from the spoken content or an event tag representing a vocal event, with precise placement and, for continuous events, appropriate span marking. For example:

Input audio: [audio clip](A person whispers "hello", then laughs.) *Output transcription:* <whispering> hello </whispering> [laughs]

4 WESR-Bench

In this section, we present WESR-Bench, detailing the construction of our expert-annotated dataset and the standardized word-level evaluation metrics.

4.1 Data Construction

Data Curation We collect web-scale audio data from diverse sources, including movies, TV dramas, podcasts, and audiobooks. To ensure audio quality, we employ MossFormer2 (Zhao et al., 2024) for denoising and filter out samples with DNSMOS below 2.0, retaining only high-quality recordings suitable for rigorous evaluation.

Hybrid Retrieval Strategy Constructing an evaluation set rich in vocal events via random sampling is impractical due to the sparsity of such events in general speech. To address this, we develop a hybrid retrieval mechanism to efficiently mine target samples. We leverage two complementary models: 1) BEATs (Chen et al., 2022b), a pre-trained acoustic encoder that captures non-verbal audio patterns, used for audio-based vector search; 2) AF-CLAP (Ghosh et al., 2025), an enhanced CLAP-style encoder with strong audio-text alignment for audio events, used for text-based retrieval over ASR transcripts. For each label class, we design three representative queries in the text modality and three in the audio modality. By retrieving utterances with high similarity to these queries, we obtain a candidate set of 1,297 utterances for expert annotation in WESR-Bench.

4.2 Human Annotation

To ensure the benchmark serves as a reliable gold standard, we implement a rigorous human annotation process. We recruit three annotators who completed a mandatory training program and passed a qualification test before participating. The training includes a tutorial on event ontology and a guided exercise with feedback (annotation guidelines and interface are provided in Appendix A). Annotators were compensated at a rate of $30 per hour of audio annotated. During the annotation phase, they worked independently to insert event tags without referencing retrieval results in Section 4.1. Fi-

Figure 2: Tag distribution in WESR-Bench. The inner ring shows major event categories, each represented by a distinct color. The outer ring displays the specific tag instances within each category.

nally, a senior expert reviewed all samples to verify boundary precision and classification correctness to ensure quality.

4.3 Dataset Analysis

Statistics After filtering out utterances where annotators found no valid vocal events or ambiguous temporal boundaries, the final benchmark comprises 927 verified utterances. The distribution of tag categories is shown in Figure 2, and per-utterance statistics are detailed in Table 3. Notably, 29.13% of utterances contain multiple tags, and 24.60% feature multiple distinct event types, indicating a meaningful presence of diverse vocal events within single utterances. The corpus contains 1,918 tag occurrences in total, with a split of 58.8% continuous events and 41.2% discrete events. The dataset comprises 58% Chinese and 42% English by duration.

Case Study Table 2 presents cases in our proposed WESR-Bench, illustrating the diversity of events of our word-level event-speech recognition task. We showcase both discrete events, such as `[clear_throat]` and `[laughs]`, as well as continuous events, such as `<singing>` and `<whispering>`. We also demonstrate that mixed event tags can occur in one utterance. These examples demonstrate the ability of our method to capture fine-grained word-level boundaries and provide rich annotations for a wide range of vocal event types.

4.4 Evaluation Protocol

To rigorously assess the performance of WESR, we focus on two key aspects: whether all occurring vocal events are successfully detected and whether their predicted positions are accurate. Since direct comparison of event labels can lead to misalignment when predicted text differs from the ground truth, we first align the hypothesis and reference transcripts at the word level. Subsequently, we map and compare the event tags based on this alignment to calculate the final metrics as shown in Figure 3. The detailed process is described below:

Step 1. Event-Preserving Alignment To align the hypothesis and reference sequences without altering event tags, an event-preserving alignment procedure is applied. First, event tags are temporarily removed from the reference to obtain plain text for alignment, while being preserved in the hypothesis. Then, using SequenceMatcher, we align the hypothesis to the reference, generating edit operations (insert, delete, replace). Finally, the operations are executed that: insertions add reference words to the hypothesis; deletions remove only non-event words; for replacements, event tags within the replaced segment are extracted and preserved, the segment is replaced with the reference text, and the extracted event tags are re-inserted at the most similar positions in the new segment. The pseudocode for this process is shown in Appendix B.

Step 2. Mapping Events to Words To precisely locate each label, we introduce the concepts of "word positions" and "inter-word positions." Continuous events are assigned to all words within their

Metric	# Tags	# Utt.	%
Total Tags	1	657	70.87
	2	184	19.85
	≥ 3	86	9.28
Unique Tags	1	699	75.40
	2	180	19.42
	≥ 3	48	5.18

Table 3: Distribution of utterances by tag statistics. **Total Tags**: all event tags in an utterance (continuous pairs `< >...</>` counted as one). **Unique Tags**: distinct event categories (repeated tags counted once). "# Utt." indicates the number of utterances.

Figure 3: Illustration of Steps 2 and 3 of the WESR evaluation method after Event-Preserving Alignment. Event tags are extracted and mapped to their corresponding word or inter-word positions for metrics calculation.

span, indicating that the event persists across those words; discrete events are assigned to the positions between two words, marking a momentary event occurrence. Thus, for a sequence of N words, there are $2N + 1$ possible positions: N word positions and $N + 1$ inter-word positions (including start and end). This design captures both continuous intervals and exact insertion points, enabling unified evaluation of different label types.

Step 3. Metrics Calculation With the aligned sequence and unified word/inter-word positions, we compute true positives (TP), false positives (FP), and false negatives (FN) for each label type. TP are positions correctly labeled with the event, FP are positions incorrectly labeled with an event that should not be there, and FN are positions that should have an event label but are missing or incorrect. Note that true negatives are not counted, as most positions in conversational speech do not contain vocal events.

5 Building a Strong Baseline for WESR

In this section, we present a strong baseline for WESR by constructing the WESR-TRAIN dataset and training WESR-specialized models. We then evaluate these models against recent open-source ALMs and commercial APIs on our WESR-Bench.

5.1 WESR-Train

To facilitate training for word-level event-speech recognition, we construct a large-scale weakly labeled dataset called WESR-TRAIN.

Data Collection and Annotation For web-sourced data, we employ a similar data curation and hybrid retrieval strategy as described in Section 4.1. We use different text/audio queries from those used in constructing WESR-Bench to retrieve diverse samples. To ensure fair evaluation, we perform deduplication against WESR-Bench to prevent any overlap. The collected data is then automatically annotated using Gemini [2], the latest version available at the time of data collection, to generate word-level event annotations (see Appendix D for the prompt details).

Adaptation from other datasets For open-source datasets, we incorporate word-level vocal event data from NonverbalTTS, NVSpeech-170k, NonVerbalSpeech-38K, and SMIIP-NV, adapting their annotations to our format. We normalize their vocal event tags to match our WESR taxonomy through careful mapping, removing audio with tags outside our taxonomy. To ensure mapping quality, we conduct a manual review of the mapped annotations, with particular attention to NonVerbalSpeech-38K, which uses different continuous event definitions than ours. We also observe that audios annotated with continuous <B>...</B> tags in NonVerbalSpeech-38K contain significant annotation errors and thus exclude them from our training set. The final data distribution of WESR-Train is shown in Table 5.

5.2 Adaptation across Backbones

Leveraging the rich annotations provided by WESR-Train, we apply supervised fine-tuning to three distinct backbones: Whisper-Large-v3 (1.5B) (Radford et al., 2023), Kimi-Audio-7B-Instruct (7B) (KimiTeam et al., 2025), and Qwen3-Omni-Instruct (30B) (Xu et al., 2025a), and evaluate their performance on WESR-Bench. Our experiment setup is detailed in Appendix C.

As shown in Table 4, our approach achieves consistent performance across different architectures, with Macro F1 scores ranging from 37.7% to 38.0% despite significant variations in model size (1.5B to 30B parameters). This demonstrates that our training approach generalizes well across diverse model architectures and scales.

For the subsequent evaluation and comparison, we select our fine-tuned Qwen3-Omni as the representative model, given its consistently strong performance across different event categories.

5.3 Comparison with Other Baselines

To comprehensively evaluate our proposed methods, we compared their performance with

[2] gemini-2.5-pro

Tag	Kimi-Audio P / R / F1 (%)	Qwen3-Omni P / R / F1 (%)	Gemini-2.5-Pro P / R / F1 (%)	Gemini-3-Pro P / R / F1 (%)	WESR-Whisper P / R / F1 (%)	WESR-Kimi P / R / F1 (%)	WESR-Qwen P / R / F1 (%)
<crying>	28.9 / 04.6 / 08.0	55.3 / 11.7 / 19.4	59.4 / 80.3 / 68.3	66.2 / 73.7 / 69.7	65.1 / 89.4 / **75.3**	60.7 / 90.2 / 72.5	60.8 / 92.1 / 73.3
<laughing>	00.4 / 00.3 / 00.4	03.8 / 12.5 / 05.8	22.5 / 17.3 / 19.6	30.8 / 25.6 / **28.0**	29.8 / 12.5 / 17.6	26.3 / 15.7 / 19.7	35.4 / 22.4 / 27.5
<panting>	0 / 0 / 0	0 / 0 / 0	39.8 / 30.5 / 34.5	37.8 / 20.5 / 26.6	32.4 / 32.9 / 32.7	34.5 / 24.5 / 28.6	38.0 / 36.9 / **37.5**
<shouting>	31.5 / 17.2 / 22.2	73.0 / 41.6 / 53.0	77.9 / 53.7 / 63.6	62.8 / 68.4 / 65.5	72.1 / 61.7 / **66.5**	65.4 / 55.7 / 60.2	72.5 / 59.2 / 65.2
<singing>	57.1 / 00.5 / 00.9	96.3 / 75.9 / 84.9	98.5 / 87.2 / 92.5	97.3 / 81.6 / 88.7	99.6 / 92.4 / 95.9	99.4 / 93.7 / **96.5**	98.8 / 93.0 / 95.8
<whispering>	0 / 0 / 0	59.5 / 10.8 / 18.3	96.4 / 14.1 / 24.7	94.0 / 37.1 / 53.2	85.6 / 64.4 / 73.5	84.2 / 67.7 / 75.1	85.9 / 69.9 / **77.1**
[breathing]	10.3 / 06.2 / 07.8	03.3 / 02.1 / 02.6	05.0 / 02.1 / 02.9	16.4 / 22.9 / **19.1**	0 / 0 / 0	03.6 / 02.1 / 02.6	02.3 / 02.1 / 02.2
[chuckle]	0 / 0 / 0	40.0 / 03.1 / 05.7	12.8 / 32.3 / 18.3	20.0 / 12.3 / 15.2	16.8 / 52.3 / 25.5	20.9 / 52.3 / **29.8**	16.6 / 53.8 / 25.4
[clear_throat]	38.9 / 48.3 / 43.1	30.4 / 48.3 / 37.3	44.2 / 65.5 / 52.8	51.4 / 62.1 / 56.2	67.9 / 65.5 / 66.7	74.1 / 69.0 / **71.4**	65.6 / 72.4 / 68.9
[cough]	39.7 / 31.3 / 35.0	64.6 / 42.4 / 51.2	51.4 / 55.6 / 53.4	65.4 / 52.0 / 58.0	74.4 / 67.7 / **70.9**	71.8 / 56.6 / 63.3	69.6 / 55.6 / 61.8
[crowd_laughter]	0 / 0 / 0	0 / 0 / 0	66.7 / 15.4 / 25.0	29.0 / 23.1 / 25.7	42.9 / 15.4 / 22.6	44.4 / 10.3 / 16.7	56.2 / 23.1 / **32.7**
[cry]	11.4 / 08.3 / 09.6	00.5 / 02.1 / 00.9	16.2 / 22.9 / **19.0**	05.9 / 02.1 / 03.1	100 / 08.3 / 15.4	83.3 / 10.2 / 18.2	28.6 / 04.2 / 07.3
[exhale]	03.2 / 03.1 / 03.2	10.0 / 28.1 / 14.8	06.5 / 06.2 / 06.3	14.3 / 09.4 / 11.3	20.0 / 09.4 / 12.8	31.6 / 18.8 / **23.5**	15.0 / 09.4 / 11.5
[giggle]	16.7 / 10.7 / 13.0	02.4 / 07.1 / 03.5	10.5 / 28.6 / 15.4	20.0 / 22.2 / 21.1	25.6 / 39.3 / 31.0	28.6 / 35.7 / **31.7**	19.6 / 32.1 / 24.3
[inhale]	10.2 / 07.3 / 08.5	11.1 / 01.5 / 02.6	08.7 / 09.5 / 09.1	06.0 / 33.6 / 10.1	09.6 / 06.6 / 07.8	13.5 / 09.4 / 11.1	14.8 / 09.5 / **11.6**
[laughs]	20.2 / 58.8 / 30.0	33.3 / 38.2 / 35.6	33.8 / 35.9 / 34.8	30.8 / 61.8 / 41.1	45.5 / 42.0 / 43.7	49.6 / 45.6 / **47.5**	44.5 / 40.5 / 42.4
[roar]	0 / 0 / 0	0 / 0 / 0	16.7 / 25.0 / 20.0	0 / 0 / 0	50.0 / 25.0 / **33.3**	50.0 / 25.0 / **33.3**	20.0 / 25.0 / 22.2
[scream]	04.2 / 10.0 / 05.9	0 / 0 / 0	10.0 / 20.0 / 13.3	08.7 / 20.0 / 12.1	21.1 / 40.0 / 27.6	08.3 / 10.0 / 09.1	23.5 / 40.0 / **29.6**
[shout]	10.0 / 27.3 / 14.6	02.4 / 09.1 / 03.7	20.0 / 13.6 / 16.2	25.0 / 22.7 / **23.8**	40.0 / 09.1 / 14.8	40.0 / 09.1 / 14.8	50.0 / 13.6 / 21.4
[sigh]	07.4 / 05.9 / 06.6	09.9 / 23.5 / 13.9	12.1 / 47.1 / 19.3	26.8 / 44.1 / 33.3	27.0 / 58.8 / 37.8	35.9 / 67.6 / **46.9**	27.3 / 61.8 / 37.8
[sobbing]	05.4 / 19.3 / 08.5	00.6 / 03.5 / 01.0	12.8 / 40.4 / 19.5	09.6 / 42.1 / 15.7	12.6 / 50.9 / 20.2	14.1 / 48.2 / 21.9	13.7 / 57.9 / **22.1**
Micro avg.	19.4 / 04.5 / 07.3	47.8 / 27.0 / 34.5	64.3 / 46.4 / 53.9	63.8 / 54.7 / 58.9	72.5 / 68.9 / 70.6	71.3 / 69.7 / 70.5	71.2 / 71.7 / **71.4**
Macro avg.	14.1 / 12.3 / 10.3	23.6 / 17.2 / 16.9	34.4 / 33.5 / 29.9	34.2 / 35.1 / 32.3	44.7 / 40.2 / 37.7	44.8 / 38.9 / 37.8	40.9 / 41.6 / **38.0**

Table 4: Performance of various models on all vocal event categories in WESR-Bench. The first four columns show results using 2-shot prompting (Kimi-Audio, Qwen3-Omni, Gemini-2.5-Pro, Gemini-3-Pro), while the last three columns show results for our WESR-trained models (WESR-Whisper, WESR-Kimi, WESR-Qwen). "Micro avg." is computed by directly averaging across all samples, while "Macro avg." is computed by averaging per-category performance. The best and second-best F1 scores in each row are bolded and underlined, respectively.

Data Source	Dur. (h)	Language
NonverbalTTS	14	EN
NVSpeech-170k	332	EN, ZH
NonVerbalSpeech-38K	87	EN, ZH
SMIIP-NV	35	ZH
Gemini-annotated	1,299	EN, ZH
WESR-TRAIN	**1,767**	EN, ZH

Table 5: Composition of WESR-Train dataset. "Dur." denotes the total duration.

other baselines of several representative audio language models (ALMs) on WESR-Bench. ALMs considered in our evaluation are as follows: 1) Newest open-source ALMs such as Kimi-Audio (KimiTeam et al., 2025), MiMo-Audio (Zhang et al., 2025) and Qwen3-Omni (Xu et al., 2025a). 2) Commercial APIs, such as Gemini [3] and GPT [4]. We prompt these models using the same instruction (see Appendix D), which includes two in-context examples illustrating the use of discrete and continuous tags. The detailed results are presented in Table 4.

[3] `gemini-2.5-pro, gemini-3-pro`

[4] `gpt-4o-audio-preview`

As shown in Table 4, different models exhibit substantial differences in non-verbal event detection. Overall, our Qwen3-Omni fine-tuned WESR model achieves the highest F1 scores across all categories, reaching a macro F1 of 38.0%, with absolute improvements of 21.1% over Qwen3-Omni and 8.1% over Gemini-2.5-Pro. This improvement is particularly pronounced in challenging categories such as <panting>, <whispering>, and various laughter-related tags ([chuckle], [giggle], [laughs]), where few-shot models often achieve near-zero recall but fine-tuned models attain substantially higher detection rates. For instance, WESR-whisper achieves 64.4% recall on <whispering> compared to Gemini-3-Pro's 37.1%, and 52.3% on [chuckle] versus 12.3%. These results suggest that while large multimodal models possess some inherent capability for vocal event recognition through prompting, systematic exposure to labeled training data remains critical for achieving robust and reliable performance across diverse paralinguistic phenomena.

During evaluation, we also notice that GPT-4o and MiMo-Audio frequently refused to respond to the WESR prompt. In the rare cases where it did produce an output, it failed to insert any event tags

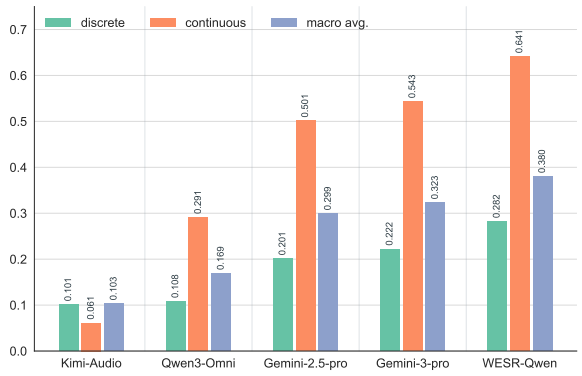

Figure 4: Performance comparison between continuous and discrete event tags on WESR-Bench.

into the transcription, rendering them unsuitable for our WESR task.

We also report aggregated results in Supplementary Table 7, where fine-grained tags are grouped into broader vocal event categories (e.g., grouping `[clear_throat]` and `[cough]` into *Cough*, detailed in Supplementary Table 10). The aggregated metrics further confirm the ability of our model on distinguishing different vocal event categories. We also observe highly synchronized performance trends: all models achieve their best performance on distinct, sustained events such as *Singing* (F1 > 94%) and *Whispering* (F1 >75%), while consistently facing challenges with subtle, low-energy acoustic features like *Breathing*.

5.4 Comparison on Discrete vs. Continuous Events

Continuous events (e.g., `<crying>...</crying>`) present a greater challenge than discrete events (e.g., `[cough]`). The model must not only infer the correct event type but also determine accurate span boundaries. To study all baselines' ability on discrete and continuous events, we calculate F1 scores separately for discrete and continuous events. Note that continuous F1 scores are generally higher than discrete ones: continuous tags allow partial credit through token-level overlap when boundaries are slightly misaligned (see Section 4.4), whereas discrete tags require exact position matching.

As shown in Figure 4, open-source models struggle significantly: Kimi-Audio achieves only 0.101 / 0.061 F1 (discrete / continuous), while Qwen3-Omni reaches 0.108 / 0.291. Proprietary models perform better, with Gemini-2.5-Pro at 0.201 / 0.501 and Gemini-3.0-Pro at 0.222 / 0.543. WESR-Qwen substantially outperforms all baselines on both event types, achieving 0.282 F1 on discrete events (+27% over Gemini-3.0-Pro) and 0.641 F1 on continuous events (+18% improvement). This demonstrates that our specialized training effectively enhances both precise event localization and accurate span boundary prediction.

5.5 Impact on ASR Performance

To explore the impact of WESR to the underlying ALM's ASR ability, we conduct experiments on Common Voice 15 (Ardila et al., 2020) test set, a widely used large-scale ASR benchmark. We compare the word error rate (WER) of the original model and our WESR-fine-tuned models on both English (en) and Chinese (zh-CN) test sets. Since Common Voice does not include vocal event annotations, we exclude vocal event tags from the WER computation to ensure a fair comparison. The results are shown in Table 6.

Model	en	zh-CN
Qwen3-Omni	7.2	6.0
WESR-Qwen	8.6	7.2

Table 6: Word error rate (WER, %) comparison of the original Qwen3-Omni and our WESR fine-tuned model on CommonVoice 15 test set (en and zh-CN). Lower WER indicates better performance.

Results show that our WESR fine-tuned model maintains competitive ASR performance with only modest increases in WER, demonstrating that WESR can be integrated without significantly compromising transcription accuracy.

6 Conclusion

In this work, we introduce the Word-level Event-Speech Recognition (WESR) task. We formalize a rigorous taxonomy categorizing 21 vocal events into *discrete* and *continuous* types, establishing a comprehensive framework for this task. We develop WESR-BENCH, an expert-annotated benchmark with a novel position-aware protocol that disentangles ASR errors from event detection, enabling precise localization measurement. We establish a strong baseline for WESR by constructing WESR-Train and training specialized models that, across different parameter scales, outperform both open-source audio-language models and commercial APIs while maintaining ASR quality. We believe WESR will serve as a foundational resource for future research in modeling rich, real-world auditory scenes.

Limitations

We acknowledge that our work may have the following limitations: 1) Language Coverage: Our system is currently verified only for English and Chinese. While the underlying methodology is language-agnostic, its generalization capabilities to languages with different morphological structures or lower resource availability remain unverified. 2) Resource Intensity: The construction of our high-quality dataset relies on the commercial Gemini API and expert annotation. While this ensures data quality, it presents a trade-off in terms of cost-efficiency and scalability. This may pose challenges for reproducibility in resource-constrained environments.

Ethical Considerations

We prioritize ethical standards throughout our data construction and modeling processes. Our datasets utilize publicly available audio under fair use principles. We ensured the welfare of our annotators through fair compensation ($30/h). Due to our large-scale annotation process, comprehensive manual verification of all data instances is not feasible. As such, the dataset may inadvertently include inappropriate content. We emphasize that any content appearing in the source audio or annotations does NOT reflect the perspectives, beliefs, or endorsements of the authors. We release our code, data, and models solely for academic use.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U24B20181 and 62525602).

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Bingsong Bai, Qihang Lu, Wenbing Yang, Zihan Sun, Yueran Hou, Peilei Jia, Songbai Pu, Ruibo Fu, Yingming Gao, Ya Li, and 1 others. 2025. Synparaspeech: Automated synthesis of paralinguistic datasets for speech generation and understanding. *arXiv preprint arXiv:2509.14946*.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, and 1 others. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Maksim Borisov, Egor Spirin, and Daria Diatlova. 2025. Nonverbaltts: A public english corpus of text-aligned nonverbal vocalizations with emotion annotations for text-to-speech. *arXiv preprint arXiv:2507.13155*.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022a. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022b. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

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.

Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu. 2025. Nvspeech: An integrated and scalable pipeline for human-like speech modeling with paralinguistic vocalizations. *arXiv preprint arXiv:2508.04195*.

Jialong Mai, Jinxin Ji, Xiaofen Xing, Chen Yang, Weidong Chen, Jingyuan Xing, and Xiangmin Xu. 2025. Mnv-17: A high-quality performative mandarin dataset for nonverbal vocalization recognition in speech. *arXiv preprint arXiv:2509.18196*.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

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

Zhuojun Wu, Dong Liu, Juan Liu, Yechen Wang, Linxi Li, Liwei Jin, Hui Bu, Pengyuan Zhang, and Ming Li. 2025. Smiip-nv: A multi-annotation non-verbal expressive speech corpus in mandarin for llm-based speech synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12564–12570.

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. 2025a. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025b. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*.

Runchuan Ye, Yixuan Zhou, Renjie Yu, Zijian Lin, Kehan Li, Xiang Li, Xin Liu, Guoyang Zeng, and Zhiyong Wu. 2025. A scalable pipeline for enabling non-verbal speech generation and understanding. *arXiv preprint arXiv:2508.05385*.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, and 1 others. 2025. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*.

Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. 2024. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10356–10360. IEEE.

A Annotation Details

Table 8 shows instrctions on vocal event taxonomy and examples shown to annotators. Figure 6 shows the annotation page.

B Evaluation Details

Algorithm 1 Event-Preserving Alignment

Require: Hypothesis sequence H, Reference sequence R
Ensure: Aligned hypothesis sequence H'

$R_{text} \leftarrow$ Remove all event tags from R
$H_{split} \leftarrow$ Split H into list (preserving event tags as separate elements)
$R_{split} \leftarrow$ Split R_{text} into list
$ops \leftarrow$ SequenceMatcher(H_{split}, R_{split}) {Get edit operations}
$H' \leftarrow H_{split}$
for each operation op in ops **do**
  if op is INSERT **then**
    Insert corresponding words from R_{split} into H'
  else if op is DELETE **then**
    Delete only non-event words from H'
  else if op is REPLACE **then**
    $events \leftarrow$ Extract event tags from replaced segment in H'
    Replace segment in H' with corresponding text from R_{split}
    Re-insert $events$ at most similar positions in new segment
  end if
end for
return H' {Aligned sequence matching reference in non-event content}

C Training Details

For Whisper, we fine-tuned whisper-large-v3 using a learning rate of 1×10^{-5}, global batch size of 8, warmup steps ratio of 0.1, and trained for 3 epochs on H100*8 for 4 hours.

For Kimi, fine-tuned Kimi-Audio-7B-Instruct using a learning rate of 1×10^{-6}, packing sequence length of 4,096 tokens, and trained for 3 epochs on H100*8 for 5 hours.

For Qwen3-Omni, we fine-tuned Qwen3-Omni-30B-A3B-Instruct with a learning rate of 1×10^{-6}, packing sequence length of 4,096 tokens, and trained for 3 epochs on H200*8 for 5 hours.

All hyperparameters were determined by conducting preliminary experiments with three different learning rates on a validation set, which was randomly sampled from the training set. Vocal event tags are added as special tokens.

D Prompt

```
"""
You are an expert in recognizing special
    patterns in audio. Please perform the
    following task:

Transcribe text from the given audio, and insert
     the following fine-grained vocal event tags
     into the precise position to the text
    according to what is detected in the audio.
The supported tags are: [laughs], [chuckle], [
    giggle], <laughing></laughing>, [
    crowd_laughter], [crying], [sobbing], <
    crying></crying>, [cough], [clear_throat], [
    shout], [scream], [roar], <shouting></
    shouting>, <whispering></whispering>, [
    inhale], [exhale], <panting></panting>, [
    breathing], [sigh], <singing></singing>.

Tag usage guidelines:

Square brackets [ ]: Insert the tag at the exact
     point where the event occurs, typically
    between words or within a sentence. For
    example:
[inhale] I don't think they saw us.
Angle brackets < >...</>: Use these tags to wrap
     around specific words or phrases that are
    spoken while the vocal event is happening.
    For example:
<laughing> Just like that! </laughing>

Annotation guidelines:

Only add event tags at the exact positions where
     vocal events occur.
The annotation should be precise to the position
     where the event occurs.
If there are no events to annotate, the output
    should be the ASR text only.
There can be audio without any vocal event. Do
    not insert event tags when there are no
    vocal events in the audio.
Your final output must only contain the text
    with event annotations. Do not include any
    other explanations or formatting.

Example:
"I don't think they saw us. [inhale] Let's keep
    moving."
or
"<laughing> Just like that! </laughing>"
"""
```

E Supplementary Tables

F Supplementary Figures

Aggr. Tag	Kimi-Audio P / R / F1 (%)	Qwen3-Omni P / R / F1 (%)	Gemini-2.5-Pro P / R / F1 (%)	Gemini-3-Pro P / R / F1 (%)	WESR-Whisper P / R / F1 (%)	WESR-Kimi P / R / F1 (%)	WESR-Qwen P / R / F1 (%)
Breathing	22.6 / 07.8 / 11.6	25.7 / 13.7 / 17.9	30.8 / 30.9 / 30.9	16.6 / 35.2 / 22.5	33.6 / 30.7 / 32.1	37.9 / 27.8 / 32.0	37.0 / 33.8 / 35.3
Cough	51.8 / 44.5 / 47.9	60.7 / 50.8 / 55.3	56.2 / 64.1 / 59.9	70.5 / 62.2 / 66.1	77.2 / 68.8 / 72.7	79.4 / 63.3 / 70.4	71.8 / 61.7 / 66.4
Crying	18.9 / 04.6 / 07.4	68.9 / 33.9 / 45.5	56.4 / 78.8 / 65.8	62.0 / 74.4 / 67.7	61.7 / 87.8 / 72.5	59.2 / 90.1 / 71.5	57.7 / 90.4 / 70.4
Laugh	24.1 / 27.8 / 25.8	11.1 / 24.1 / 15.2	39.1 / 40.3 / 39.7	42.5 / 45.3 / 43.8	50.4 / 43.3 / 46.6	48.9 / 42.8 / 45.6	48.7 / 47.8 / 48.2
Shout	30.2 / 18.0 / 22.6	64.3 / 41.2 / 50.2	77.1 / 52.0 / 62.1	61.0 / 66.3 / 63.5	72.3 / 61.0 / 66.2	65.0 / 54.4 / 59.2	72.2 / 58.8 / 64.8
Singing	57.1 / 00.5 / 00.9	96.4 / 75.6 / 84.8	98.6 / 87.2 / 92.6	97.3 / 81.8 / 88.8	99.8 / 92.6 / 96.1	99.4 / 93.7 / 96.4	98.8 / 92.8 / 95.8
Whispering	0 / 0 / 0	59.5 / 10.8 / 18.3	96.6 / 14.2 / 24.7	94.1 / 37.1 / 53.2	85.6 / 64.5 / 73.6	84.2 / 67.7 / 75.1	85.9 / 70.0 / 77.2
Micro avg.	26.2 / 05.8 / 09.4	58.0 / 32.7 / 41.8	68.2 / 48.3 / 56.6	66.5 / 57.0 / 61.4	75.0 / 71.0 / 73.0	73.6 / 71.9 / 72.8	73.5 / 73.7 / 73.6
Macro avg.	29.3 / 14.7 / 16.6	55.2 / 35.7 / 41.0	65.0 / 52.5 / 53.7	63.4 / 57.5 / 58.0	68.6 / 64.1 / 65.7	67.7 / 62.8 / 64.3	67.5 / 65.0 / 65.4

Table 7: Performance of various models on aggregated vocal event categories in WESR-Bench. The first four columns show results using few-shot prompting (Kimi-Audio, Qwen3-Omni, Gemini-2.5-Pro, Gemini-3-Pro), while the last two columns show results for models fine-tuned on our WESR-Train corpus (WESR-whisper, WESR-Qwen).

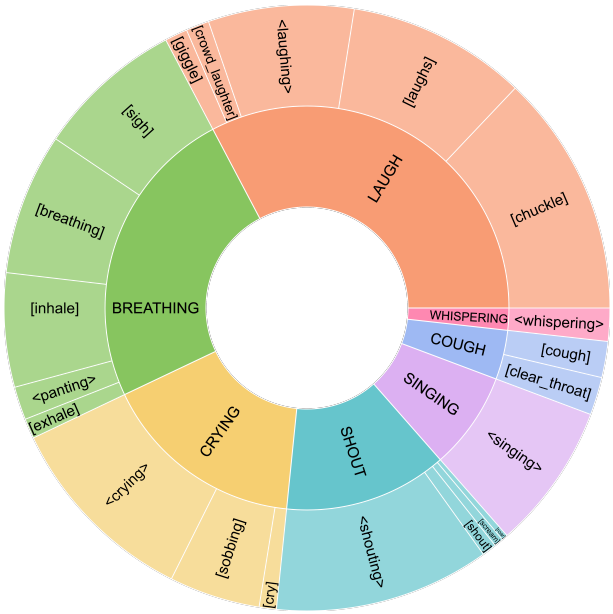

Figure 5: Tag distribution of WESR-Train.

Figure 6: The labeling interface for human annotation.

Tag	Explanation	Annotation Example	Audio Example
[laughs]	Laughter sound	Then [laughs] that was amazing.	(omitted)
[chuckle]	Laugh quietly	I know, right? [chuckle] That's exactly what I thought.	(omitted)
[giggle]	Laugh in a light, high-pitched way from amusement	She told me the story and I just [giggle] couldn't stop laughing.	(omitted)
<laughing>	Speaking while laughing	Later we didn't even need a model; <laughing>this is it.	(omitted)
[crowd_laughter]	Laughter from multiple people	He always hits me... [crowd_laughter] Oh my... [crowd_laughter] ... parent-teacher meetings were the hardest... [crowd_laughter].	(omitted)
[cry]	Audible crying sound	Look at this dress... [cry] It looks awful; I won't wear it.	(omitted)
[sobbing]	Intermittent sobbing sounds	I can't believe it happened... [sobbing] I just can't.	(omitted)
<crying>	Speaking while crying / tearful voice	<crying>I picked it up near Liding Street; that was on their wedding day.	(omitted)
[cough]	Coughing sound	Thanks [cough] for the two days off.	(omitted)
[clear_throat]	Throat clearing sound	[clear_throat] The talk was inspiring; let's all share our thoughts.	(omitted)
[scream]	High-pitched loud vocalization	Help me! [scream] Someone please help!	(omitted)
[roar]	Loud, rumbling vocalization	The contest begins! [roar] Let's go!	(omitted)
[shout]	Loud cry or call	[shout]Who is it? Open the door!	(omitted)
<shouting>	Speaking while shouting	<shouting>Trust me! Please! You have to believe me!	(omitted)
[breathing]	Respiration or panting sound	[breathing] I owe you so much... only you had nothing...	(omitted)
[inhale]	Audible breath intake	I owe you so much [inhale] ... only you [inhale] had nothing...	(omitted)
[exhale]	Audible breath release	I owe you so much... [exhale] only you... [exhale] had nothing...	(omitted)
<panting>	Heavy breathing while speaking	<panting>I can't run anymore... I really can't keep going.	(omitted)
[sigh]	Sighing sound	He resented the hereditary illness... [sigh] it was so unfair.	(omitted)
<whispering>	Speaking in whisper	<whispering>My English is so bad—even simple phrases aren't fluent.	(omitted)
<singing>	Singing voice	<singing>You wrote me into the script and said you must be the lead.	(omitted)

Table 8: Annotation guidelines for nonverbal vocal events. Use angle brackets <...> for interval events and square brackets [...] for point events.

Dataset	Tags	Count
NonverbalTTS	[breath], [grunt], [sniff], [throat clearing], [groan], [sigh], [snore], [cough], [laugh], [sneeze]	10
NVSpeech-170k	[Breathing], [Laughter], [Confirmation-en], [Uhm], [Sigh], [Surprise-ah], [Surprise-oh], [Dissatisfaction-hnn], [Surprise-wa], [Question-yi], [Question-ei], [Cough], [Question-ah], [Question-oh], [Surprise-yo], [Question-en], [Shh], [Crying]	18
NonVerbalSpeech-38K	[snore], [throatclearing], [crying], [breath], [sniff], [laughing], [coughing], [gasp], [yawn], [sigh]	10
SMIIP-NV	[Laughter], [crying], [cough]	3
Synparaspeech	[Sigh], [throat clearing], [laugh], [pause], [tsk], [gasp]	6
MNV-17	[Sneezing], [Clapping], [Hissing], [Whistling], [Clearing Throat], [Coughing], [Lip Smacking], [Exhaling], [Moaning], [Panting], [Sniffling], [Humming], [Laughing], [Applauding], [Inhaling], [Chuckling], [Sighing]	17

Table 9: Summary of non-verbal tags used in prior work.

Aggregated Category	WESR Event Tags
LAUGH	[laughs], <laughing>, [chuckle], [giggle], [crowd laughter]
SHOUT	[scream], [roar], [shout], <shouting>
WHISPERING	<whispering>
SINGING	<singing>
BREATHING	[inhale], [exhale], <panting>, [sigh], [breathing]
COUGH	[cough], [clear_throat]
CRYING	<crying>, [sobbing], [cry]

Table 10: Label aggregation mapping for non-verbal event tags.