From Sound to Symptom: Real-Time Respiratory Signal Understanding for Conversational Healthcare Agents

Tanmay Laud[1], **Herprit Mahal**[1], **Subhabrata Mukherjee**[1]
[1]Hippocratic AI
Correspondence: tanmay@hippocraticai.com

Abstract

Cough events during live spoken conversations carry clinically valuable respiratory signals, yet existing dialogue systems treat them as acoustic noise to be discarded. We present HealthCUES (Clinical Understanding from Embodied Sounds), a streaming pipeline for paralinguistic respiratory monitoring in real-time conversational agents, a capability that, to the best of our knowledge, is absent from all prior systems. HealthCUES processes audio through a rolling buffer aligned with dialogue turn boundaries, enabling sub-second event detection without interrupting conversational flow. Beyond binary cough detection, the system provides fine-grained analytics: (i) differentiation between coughing and throat clearing, (ii) cough subtype classification (dry, wet, barking, whooping) with confidence scores, and (iii) temporal duration estimation with start–end boundaries. To prevent alert fatigue, HealthCUES introduces dialogue-aware gating mechanisms that modulate triggering based on conversational context. The system leverages Qwen3-Omni, a multimodal large language model (MLLM), with constrained structured outputs, decomposing cough analysis into parallel prediction tasks for independent prompt optimization. Evaluation on 847 in-house conversational audio segments demonstrates 93% F1 for cough detection, 0.75 weighted-F1 for wet/dry subtype classification, and average end-to-end latency of 340ms; external validation on the AMI meeting corpus confirms robust cough, throat-clearing, and speech separation in the presence of speech (0.91 macro-F1). A user study with licensed healthcare professionals confirms the clinical relevance of subtype information and the system's utility in telehealth workflows.

1 Introduction

Respiratory sounds, particularly coughs, serve as important biomarkers in healthcare conversations. During telehealth consultations and remote health monitoring, cough characteristics provide actionable clinical insights, distinguishing productive from non-productive cough or identifying patterns such as barking or whooping cough (Porter et al., 2019; Imran et al., 2020). Yet **current conversational AI systems discard these paralinguistic health signals entirely**, routing them to noise suppression rather than clinical analysis. This gap has concrete consequences: audio-based cough assessment is among the most valuable remote diagnostic tools available, yet no existing conversational agent can leverage it (Porter et al., 2019).

Building a real-time cough analysis system for live conversations presents distinct challenges: latency constraints require detection within the natural pause between utterances (<500ms); telephony codecs and background noise substantially degrade signal quality; clinical utility demands subtype differentiation beyond binary detection; and repeated cough mentions become intrusive without dialogue-aware modulation.

We present HealthCUES (Clinical Understanding from Embodied Sounds), a streaming cough analysis pipeline that addresses these challenges. Our contributions are:

1. A **streaming cough detection framework** with turn-aligned audio processing for real-time conversational integration (Section 3).

2. **Fine-grained cough analytics** providing four-way subtype classification, confidence scores, and temporal duration boundaries; to the best of our knowledge, these capabilities are absent from all prior cough detection systems (Section 3).

3. **Dialogue-aware gating** with cooldown, topic suppression, and cough-to-speech ratio monitoring to prevent alert fatigue (Section 3.3).

4. **Systematic evaluation** with baseline comparisons and licensed healthcare professional val-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 559–564
August 2–5, 2026.

idation across three dialogue quality dimensions (Section 4).

2 Related Work

Audio Event Detection and Cough Analysis. General-purpose audio event detection (Kong et al., 2020; Gong et al., 2021) targets broad taxonomies on pre-segmented clips without health-relevant distinctions. Cough analysis spans signal processing (Drugman et al., 2013) and deep learning on dedicated corpora (Orlandic et al., 2021; Sharma et al., 2020), with COVID-19 spurring interest in cough-based screening (Imran et al., 2020; Laguarta et al., 2020). Existing systems share three critical limitations: they analyze isolated recordings rather than streaming audio, perform binary detection without subtype differentiation, and lack any notion of dialogue context for appropriate triggering.

Paralinguistic Signals in Dialogue. Physiological signals have been used to adapt dialogue behavior (Litman and Forbes-Riley, 2014; Katada et al., 2020), and video-based respiratory estimation has been integrated into human-robot dialogue to prevent speech collisions and synchronize robot breathing (Obi and Funakoshi, 2024). Spoken dialogue systems adapting to detected user affect (Litman and Forbes-Riley, 2014) demonstrate that such signal-to-policy integration improves task success and user satisfaction, an evaluation dimension we adopt. LLM-based affect recognition (Feng et al., 2024) further demonstrates community appetite for non-lexical signal integration. In contrast to these antecedents, HealthCUES addresses respiratory health events specifically, providing fine-grained subtype output and dialogue-aware gating designed for clinical triage contexts.

Multimodal Foundation Models. Recent MLLMs demonstrate strong zero-shot audio understanding capabilities (Gong et al., 2023; Chu et al., 2023; Xu et al., 2025). HealthCUES leverages these capabilities, specifically those of Qwen3-Omni (Xu et al., 2025), while adding structured output constraints and dialogue-aware post-processing essential for reliable real-time operation, additions not addressed by the foundation models themselves.

3 System Overview

Figure 1 illustrates the HealthCUES architecture, operating as a module within a conversational agent pipeline.

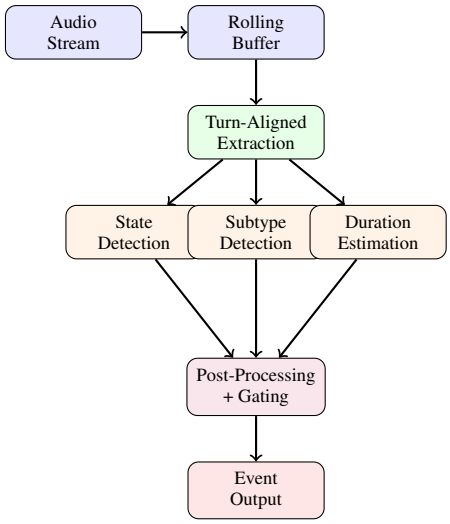

Figure 1: Architecture of HealthCUES. Audio is buffered and extracted at turn boundaries, then processed by parallel inference tasks. Results are combined with dialogue-aware post-processing before output.

3.1 Inputs and Streaming Context

The pipeline receives continuous audio from telephony or microphone input. A **rolling audio buffer** maintains recent audio history, enabling retrospective analysis when a user turn completes. Extraction is **turn-aligned**: when ASR indicates end-of-utterance, the system analyzes a window extending slightly before turn onset (default: 2s look-back) to capture cough events that precede speech. Operating at the turn level matches the granularity at which the dialogue agent plans and acts; because the look-back window spans the preceding utterance, coughs that occur mid-turn or just before speech onset are still captured, so the turn-level trigger does not discard out-of-turn respiratory events.

3.2 Multi-Stage Inference

Rather than a monolithic classifier, HealthCUES decomposes cough analysis into parallel structured prediction tasks via Qwen3-Omni (Xu et al., 2025), a multimodal LLM (MLLM) backend. This decomposition enables independent per-component prompt optimization and graceful degradation if any single task fails.

State Detection performs three-way classification: *coughing*, *throat clearing*, or *none*. Throat clearing is explicitly distinguished, as it carries different clinical significance and warrants different conversational responses.

Subtype Classification categorizes detected

cough events into four commonly used clinical descriptors of cough character (Morice et al., 2007; Chung et al., 2009; Porter et al., 2019): *dry* (non-productive, no mucus expelled), *wet* (productive, with a rattling or gurgling quality from airway secretions), *barking* (seal-like and low-pitched, classically associated with croup), and *whooping* (paroxysmal fits followed by an inspiratory "whoop", associated with pertussis), each with an associated confidence score. These are qualitative descriptors rather than standardized diagnostic categories (Chung et al., 2009), which we revisit when reporting subtype agreement (Section 4).

Duration Estimation outputs start and end timestamps for each event, enabling severity bucketing (short: <2s; medium: 2–6s; long: >6s) as a clinical proxy for severity. All outputs follow a constrained schema: subtype returns code-confidence pairs, e.g. `[wet][0.88]`; duration returns time ranges, e.g. `[1.2-4.3]`.

3.3 Dialogue-Aware Gating

A central contribution of HealthCUES is its gating layer. Without conversational awareness, even accurate detection would be counterproductive: alerting on every cough would disrupt conversation flow, and repeating the same follow-up after each detection would feel robotic. Three mechanisms work together to decide whether a detected event should trigger an agent action.

The pipeline maintains a 60-second sliding window of activity and computes the fraction of recent speaking time occupied by coughing. This *cough-to-speech ratio* (CSR $= \frac{\sum \text{cough\_dur}}{\sum \text{speech\_dur}}$) lets the system treat a brief isolated cough differently from a sustained pattern that warrants more active follow-up. Once an action has been triggered, a configurable quiet period prevents re-triggering while the patient is still coughing through the same episode. Separately, if the agent's recent turns already contain cough-related language, new detections are suppressed, since the system infers the topic is already being addressed.

These mechanisms interact with the detection outputs: clinically urgent subtypes (wet, barking, whooping) and prolonged episodes clear lower thresholds for action than brief dry coughs.

3.4 Output and Dialogue Integration

When a detection clears the gating thresholds, the pipeline produces a natural-language action suggestion passed to the agent's planning module via a lightweight callback API. For example, a 3 second wet cough after two events produces:

> *System action: "Patient has coughed twice. Sounds like a wet cough. Ask about it if not addressed."*
>
> *Agent: "I noticed you've been coughing. Does it feel like you're bringing anything up?"*

Signal	Gating	Action
Dry cough ×2, conf. 0.79, 1.8s	Count threshold met (≥2 events)	"Has that cough been bothering you?"
Barking cough, conf. 0.91, 4.2s	Long duration → immediate trigger	"That cough sounds quite harsh. How long have you had it?"
Wet cough, conf. 0.84; agent recently mentioned "cough"	Topic suppression active	*(suppressed; topic already in dialogue)*

Table 1: Representative HealthCUES outputs illustrating subtype-sensitive responses and dialogue-aware gating. Barking and wet coughs trigger at lower thresholds due to higher clinical significance; topic suppression prevents redundant follow-up.

Table 1 illustrates how subtype and gating interact across three representative scenarios.

4 Evaluation

Setup. We evaluate primarily on 847 held-out conversational audio segments (12.3 hours) spanning telephony-quality recordings (8kHz μ-law) and direct microphone capture (16kHz PCM), with SNR ranging from 5–30dB. These segments were collected in-house to match the conditions of our target telehealth deployment, including its codecs, ambient noise, and overlap between coughs and conversational speech, which public isolated-cough corpora do not reflect. Three U.S.-licensed clinicians labeled event presence, subtype, and timestamps; inter-annotator agreement was substantial for event presence (κ=0.82) and moderate for subtype (κ=0.67), reflecting inherent subjectivity. HealthCUES runs online as a real-time service: it does not require frame-level streaming but instead operates over a rolling audio buffer that is analyzed at turn boundaries during a live call. Our evaluation uses this same rolling-buffer mechanism, without oracle segment boundaries. To probe generalization beyond our own data, we additionally validate on the AMI Meeting Corpus (Carletta et al.,

2006) using publicly released cough annotations,[1] which contain spontaneous coughs, throat clears, and speech recorded in multi-party meetings, a demanding in-the-presence-of-speech setting.

Results. Table 2 shows HealthCUES achieves 93% F1 for cough detection, 0.75 weighted-F1 on the clinically central wet-versus-dry subtype distinction, and 340ms average end-to-end latency, well within the 500ms conversational pause budget. We include generic audio encoders (BEATs (Chen et al., 2023): 59.1% F1) and cough-specialized models (PANNs (Kong et al., 2020): 76.4%; CoughVID fine-tuned (Orlandic et al., 2021): 79.8%) as reference points. These baselines are *not* compute- or backbone-matched: they differ from HealthCUES in model scale, pretraining data, and task formulation, so the comparison is capability-illustrative rather than a controlled head-to-head, and should not be read as a like-for-like accuracy ranking. Its purpose is to situate HealthCUES against the off-the-shelf detectors a practitioner might otherwise adopt, none of which provide the streaming, throat-clearing separation, subtype, duration, and dialogue-aware capabilities that HealthCUES unlocks from a single general-purpose MLLM. The CoughVID fine-tuned model attains the lowest latency (95ms) but is restricted to binary detection; HealthCUES's higher latency (340ms) reflects its richer, multi-task inference. Table 3 confirms HealthCUES is the only evaluated system providing all five capabilities essential for clinical dialogue use.

External Validation (AMI). To test whether detection holds beyond our in-house recordings, we evaluate on the AMI Meeting Corpus (Carletta et al., 2006), where coughs and throat clears co-occur with multi-party conversational speech. With no adaptation to that domain, HealthCUES sustains strong three-way performance: macro-F1 0.91 and accuracy 0.92 (n=719), with per-class F1 of 0.86 (coughing), 0.94 (throat clearing), and 0.92 (none); see Table 4.

Subtype Breakdown. The clinically central distinction is wet versus dry (productive vs. non-productive), which reaches 0.75 weighted-F1 (Table 2): dry coughs, the most common presentation, are detected reliably (F1 0.89) and wet coughs at moderate accuracy (F1 0.55). The rarer barking and whooping subtypes are harder (Table 5): both are acoustically subtler and far less frequent in deployment, so the limited data available for them makes both reliable recognition and reliable evaluation difficult. We therefore treat wet/dry as the dependable operating point today and target the rarer subtypes through ongoing in-domain data collection.

[1] https://github.com/paulleamy/AMI-Cough-Annotations

System	Cough F1	Latency
BEATs[†]	59.1%	180ms
PANNs[‡]	76.4%	210ms
CoughVID FT[§]	79.8%	95ms
HealthCUES	**93.0%**	340ms
Additional HealthCUES metrics		
3-way Macro-F1	0.91	
Wet/Dry Weighted-F1	0.75	
Duration Bucket Acc.	0.91	
Latency p95		520ms

Table 2: Detection comparison and full HealthCUES results on 847 in-house conversational audio segments. Detection and 3-way metrics are macro-averaged; the wet/dry subtype metric is weighted-F1 conditional on cough presence. Baselines are off-the-shelf and *not* compute- or backbone-matched; the comparison is capability-illustrative rather than a like-for-like ranking (see text). [†]Chen et al. (2023); [‡]Kong et al. (2020); [§]Orlandic et al. (2021) fine-tuned.

System	Streaming	Throat Sep.	Subtype	Duration	Dialogue
BEATs[†]	✓	✗	✗	✗	✗
PANNs[‡]	✗	✗	✗	✗	✗
CoughVID FT[§]	✓	✗	✗	✗	✗
Coswara CNN	✗	✗	✗	✗	✗
HealthCUES	✓	✓	✓	✓	✓

Table 3: Capability comparison across the five functions required for conversational integration, each defined operationally in Section 3: *streaming* (turn-aligned rolling-buffer processing), *throat sep.* (three-way state detection), *subtype* (four-way subtype classification), *duration* (start–end estimation), and *dialogue* (dialogue-aware gating, Section 3.3). HealthCUES is the only evaluated system providing all five. See Table 2 for citation keys.

User Study. As a formative feasibility study, three licensed, practicing U.S. registered nurses each interacted with HealthCUES across 6 structured telehealth scenarios, assessing three dialogue quality dimensions: *response appropriateness* (did the agent correctly branch or escalate?), *conversational naturalness* (was the phrasing contextually

Class	P	R	F1
Coughing	0.99	0.76	0.86
Throat clearing	0.97	0.91	0.94
None	0.86	0.99	0.92
Accuracy		0.92	
Macro-F1		0.91	

Table 4: External validation on the AMI Meeting Corpus (Carletta et al., 2006) (n=719 segments with speech present), using Qwen3-Omni with no domain adaptation: three-way state detection (coughing / throat clearing / none).

Subtype	F1
Dry	0.89
Wet	0.55
Barking	0.23
Whooping	0.24

Table 5: Per-class F1 for four-way cough subtype classification (conditional on cough presence). The clinically central wet/dry distinction reaches 0.75 weighted-F1 (Table 2), with reliable dry detection; barking and whooping are harder and limited by data availability.

appropriate?), and *dialogue efficiency* (did cough signals reduce the need for explicit patient questioning?). Evaluators confirmed appropriate branching behavior in scenarios designed to elicit escalation: the agent correctly altered its dialogue strategy based on detected respiratory state. Check-in phrasing was rated natural and contextually appropriate when triggered correctly. Unprompted subtype identification reduced conversational burden by surfacing respiratory status without requiring direct patient self-report, improving information-gathering efficiency within the dialogue flow. Areas for improvement included detection consistency when coughs co-occur with agent speech and gating behavior during prolonged episodes, directly informing ongoing refinements to the mechanisms in Section 3.3.

5 Conclusion

We presented HealthCUES, a streaming cough analysis pipeline providing fine-grained respiratory analytics for real-time conversational agents. By combining MLLM-based audio understanding with dialogue-aware gating, HealthCUES enables conversational systems to reason over paralinguistic health signals, a capability that, to the best of our knowledge, is absent from all prior work. Evaluation demonstrates strong detection accuracy, clinically relevant subtype classification, and sub-500ms latency compatible with natural dialogue flow. Future work includes multilingual validation, expanded paralinguistic signal coverage (wheezing, labored breathing), and larger-scale user studies measuring dialogue success across diverse clinical populations.

Ethical Considerations

Privacy. Audio processed by HealthCUES may contain sensitive health information. In deployment, audio is processed transiently without persistent storage, and no content is retained beyond the active session.

Clinical Limitations. HealthCUES provides informational cough analytics, **not medical diagnosis**. The system triggers follow-up questions rather than clinical recommendations, and users must be clearly informed of these limitations. The system is designed to augment, not replace, clinical judgment.

Bias and Fairness. Cough acoustic characteristics vary across demographics, health conditions, and recording equipment. We acknowledge potential performance disparities and recommend ongoing monitoring across user populations. The MLLM backend's training data composition may introduce biases affecting subtype classification for underrepresented groups.

Limitations

Subtype labels involve inherent clinical judgment; ground truth is imperfect, as reflected in moderate inter-annotator agreement (κ=0.67). Performance may degrade in high-noise environments (SNR <5dB) or with heavily compressed codecs not represented in the evaluation set. Very soft coughs or those co-occurring with loud speech may be missed. The current evaluation is English-only; generalization to other languages and accents requires dedicated validation. HealthCUES targets server-based deployment, so end-to-end latency is bound by server compute rather than on-device resources; the MLLM backend introduces cost and availability constraints, and we do not target edge or fully on-device operation. Our user study (n=3) provides initial validation across response appropriateness, naturalness, and efficiency dimensions; larger-scale studies with diverse populations would further strengthen the evidence base.

References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction (MLMI)*, pages 28–39. Springer.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. Beats: audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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. *arXiv preprint arXiv:2311.07919*.

Kian Fan Chung, Don Bolser, Paul Davenport, Giovanni Fontana, Alyn Morice, and John Widdicombe. 2009. Semantics and types of cough. *Pulmonary Pharmacology & Therapeutics*, 22(2):139–142.

Thomas Drugman, Jérôme Urbain, Nathalie Bauwens, Ricardo Chessini, Carlos Valderrama, Patrick Lebecque, and Thierry Dutoit. 2013. Objective study of sensor relevance for automatic cough detection. *IEEE Journal of Biomedical and Health Informatics*, 17:699–707.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gasic. 2024. Affect recognition in conversations using large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 259–273. Association for Computational Linguistics.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. In *Proceedings of INTERSPEECH*, pages 571–575.

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Interspeech 2023*, pages 2798–2802.

Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, Charles N John, Md Iftikhar Hussain, and Muhammad Nabeel. 2020. Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app. *Informatics in Medicine Unlocked*, 20:100378.

Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is she truly enjoying the conversation?: Analysis of physiological signals toward adaptive dialogue systems. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI)*. 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 Trans. Audio, Speech and Lang. Proc.*, 28:2880–2894.

Jordi Laguarta, Ferran Hueto, and Brian Subirana. 2020. Covid-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:275–281.

Diane J. Litman and Kate Forbes-Riley. 2014. Evaluating a spoken dialogue system that detects and adapts to user affective states. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics.

Alyn H Morice, Giovanni A Fontana, Maria G Belvisi, Surinder S Birring, Kian Fan Chung, Peter V Dicpinigaitis, Jacek A Kastelik, Lorcan P A McGarvey, Jaclyn A Smith, Milos Tatar, and John Widdicombe. 2007. ERS guidelines on the assessment of cough. *European Respiratory Journal*, 29(6):1256–1276.

Chukwuemeka Obi and Kotaro Funakoshi. 2024. Using respiration for enhancing human-robot dialogue. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics.

Lara Orlandic, Tomas Teijeiro, and David Atienza. 2021. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156.

Paul Porter, Udantha Abeyratne, Vinayak Swarnkar, Jamie Tan, Ti-wan Ng, Joanna Brisbane, Deirdre Speldewinde, Jennifer Choveaux, Roneel Sharan, Keegan Kosasih, and Phillip Della. 2019. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. *Respiratory Research*, 20:81.

Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R., Prasanta Kumar Ghosh, and Sriram Ganapathy. 2020. Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. In *Interspeech 2020*, pages 4811–4815.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.