forum acusticum 2026
8-12 September · Graz, Austria

NATURAL LANGUAGE INTERACTION FOR MORPHOLOGY-AWARE AUDIFICATION

Sofia Vallejo Budziszewski[1], Katharina Gross-Vogt[1]

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria*

Abstract

We previously introduced a rule-based framework that classifies time signals into one of 15 morphology classes and assigns class-specific time-scaling factors for strict audification. This paper adds a natural-language interaction layer built on a large language model (LLM) strictly downstream of the pipeline. The LLM receives only structured metadata, never the raw signal, and generates domain-calibrated summaries, supports reclassification and parameter adjustment, and answers open-domain questions. Because the interaction is entirely text-based, it lowers the barrier for non-experts while also opening the tool to blind and low-vision (BLV) users via standard screen readers. Every LLM claim traces back to a number the pipeline already computed, so accessibility is gained without sacrificing analytical grounding.

Keywords: *audification, large language model, morphology classification, blind and low vision, sonification, auditory display*

1. Introduction

Audification, the direct rendering of data as sound through uniform time scaling, offers a compelling approach for exploring scientific data, particularly signals whose temporal structure is difficult to interpret visually [1, 2, 3]. In strict audification, the time-scaling factor is the sole design variable, and selecting it appropriately is the central technical problem [4]. A naive approach applies a single global spectral estimate to derive this factor, but this fails across heterogeneous datasets: a recurrent transient such as an electrocardiogram (ECG) requires mapping into a rhythmic perceptual band, while a tonal signal calls for a range preserving stable pitch relationships [5]. A morphology-first strategy, which classifies signals before estimating the scaling factor, addresses this more robustly; we introduced such a framework at ICAD 2026 [6].

Even with a principled scaling factor, the resulting system remains, from a user's perspective, a label and a number with no explanation attached. At least three groups are excluded by this, not mutually exclusive: people without a signal-processing background, who do not know what a morphology class or scaling factor means; people without expertise in the data's own domain, who cannot judge whether a result is scientifically sensible even once the vocabulary is explained; and people who cannot see the plots, sliders, and dashboards that most sonification tools are built around [7, 8]. This last group is especially notable given that sonification is frequently motivated in the literature as a technology well suited to blind and low-vision (BLV) audiences, since it replaces visual inspection with listening [9, 10]. In practice, however, the tools themselves rarely deliver on that promise: they remain visually mediated interfaces with sound as an add-on, not a self-sufficient channel [7, 11].

This paper lowers all three barriers through the same mechanism: a conversational, natural-language layer on top of the existing rule-based pipeline, which it does not modify. A user, sighted or not, expert or not, uploads a time-series file, provides a short description of what the data are, and the system talks them through the classification, audification, and any detected anomalies, entirely in text a screen reader can speak aloud. They can then reclassify, adjust duration or pitch, or zoom into detected events through follow-up messages in the same conversation. The LLM receives only the pipeline's structured metadata and never the raw signal, so every claim it makes traces back to a number the rule-based classifier already computed.

2. Background

Our prior framework [6] classifies each signal into one of fifteen morphology classes (grouped into four families: oscillatory, transient, stochastic, and trends

***Corresponding author**: vallejo@iem.at.*

& breaks) via a sequential decision pipeline combining stationarity testing (ADF/KPSS [12, 13]), trend detection (HAC-OLS [14]), spectral analysis (Welch PSD/CWT [15, 16]), transient detection (STA/LTA [17]), and structural break detection (CUSUM/Bai–Perron [18, 19]). A class-specific strategy extracts a characteristic frequency and computes a time-scaling factor $k = f_{\text{target}}/f_{\text{char}}$, where the target band is morphology-dependent [4, 5]. Three classes (deterministic trend, random walk with drift, trend break) are excluded from audification and redirected to auditory graphs [20]. Evaluated on 591 real-world signals from eight scientific domains, 92.6% were confidently classified into one of the twelve audifiable classes; on 251 synthetic control signals, family-level accuracy was 86.9%. Full methodological detail is given in [6]; this paper takes that pipeline's output as-is and does not modify it.

This solves the parametric problem of audification but not the communicative one. A domain expert receiving a verdict such as "shock-driven transient, $k = 75\times$" can parse the vocabulary but has no way to judge why that frequency was selected or whether the underlying tests support the call [21]. A non-expert faces a more basic barrier: class names, stretch factors, and test statistics carry no meaning without training [1, 5]. Neither group can push their own knowledge back into the system, and for long recordings with sparse events, a single compressed audio output gives no guidance on where the perceptually relevant moments are.

Sonification has been motivated since its earliest days as a technology that could give BLV users independent access to data [3, 9]; Díaz-Merced's doctoral work demonstrated that sonification enables pattern detection in astrophysical data for blind and sighted researchers alike [9], and tools such as xSonify [11] and SonoUno [10] were explicitly built for BLV astronomers. Yet in our own survey of 57 sonification tools spanning 1990–2025, only 4 active and 5 legacy tools were explicitly built for BLV users, and these tended toward narrow scope, limited sound vocabularies, and, in several cases, discontinuation once initial funding ended [7]. The broader pattern is that sonification tools remain overwhelmingly visual-first, with sound as an add-on rather than a self-sufficient interaction channel [8, 22]. Umwelt [8] addresses this in the data-visualization context by treating sonification and textual description as coequal to visualization; our work applies the same principle to audification.

Large language models have recently shown strong capability in translating structured technical outputs into accessible natural-language explanations and supporting multi-turn interactive dialogue [23, 24]. Their ability to condition responses on structured metadata without requiring access to raw data makes them well suited as a downstream interpretation layer that can lower all three barriers at once: technical vocabulary, domain knowledge, and visual dependence. Since the entire interaction is text-based, a screen reader can speak every part of it aloud with no adaptation layer [22]. In his ICAD 2026 keynote, blind professor Tony Stockman pointed out that many BLV users are specifically skilled at prompting, having long relied on concise, well-formed web search queries, for instance to find a journey connection online, even before LLMs came into practice [25].

3. System Architecture

3.1. LLM Strictly Downstream of Signal Processing

The design decision that makes a natural-language layer possible without compromising analytical rigor is that the LLM operates strictly downstream of the signal processing pipeline and never accesses the raw signal. Once classification and audification adviser have completed, their outputs are serialized into a structured metadata block passed to the LLM as plain text: morphology verdict and confidence, time-scaling factor k, characteristic frequency and target band, original and audified durations, and signal statistics (sample count, sampling rate, mean, standard deviation, IQR, RMS, dominant Welch frequency, outlier count, trend direction). Detected events (transient onsets, structural breaks, rhythm anomalies) are included with timestamps and types when present.

This separation has two consequences. First, the LLM cannot fabricate spectral or statistical properties it was not given: every claim is grounded in numbers the pipeline already computed. Second, the pipeline itself is completely unchanged; the interaction layer can be enabled or disabled without affecting any upstream result. For BLV users, the conversational interface is not a separate accessibility mode bolted on; it consumes the same pipeline output as the visual interface and can replace it entirely.

Figure 1 illustrates the information flow: the pipeline produces a verdict, scaling factor, and detected events, which are packaged into the context block alongside the user-supplied domain description. The LLM then generates a natural-language response, emits a structured action token executed by the system (reclassification, parameter update, zoom), or both.

3.2. Backend Design and Model Selection

The assistant supports two LLM backends, selected automatically by priority based on available credentials and running services, with the rule-based pipeline itself always available as a fallback if neither is reachable. **Mistral API** is the primary option (`mistral-small-latest`), an EU-based provider releasing models under Apache 2.0 with published model cards [23], satisfying reproducibility and transparency requirements [24]; its free tier lowers the barrier to adoption for groups without dedicated compute. **Ollama** is a fully local option requiring no API key or network connection: if a server is reachable at `localhost:11434`, the system selects the first available model from a preference list (Mistral 7B, Mistral Nemo, Llama 3, Phi 3), at the cost of higher hardware requirements (approximately 6 GB RAM for

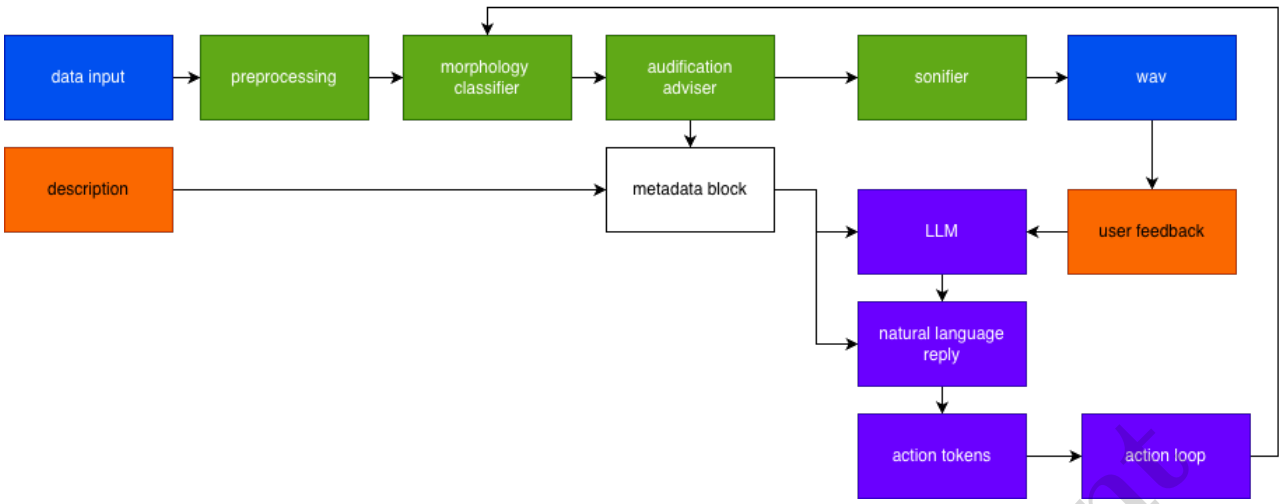

Figure 1: System architecture. The LLM receives only the structured metadata block produced by the pipeline, never the raw signal. Action tokens emitted by the LLM are parsed by the action executor and trigger downstream operations (reclassification, parameter update, or event-focused windowed audification). The entire exchange is plain text, making it natively compatible with screen readers.

Mistral 7B at 4-bit quantization). If neither backend is reachable, the interaction layer is silently skipped and the pipeline runs unchanged. Backend choice requires no code modification and is resolved at runtime.

3.3. Context Construction and Prompting

Each session begins with a system prompt defining the assistant's role, the supported interaction types (summarization, reclassification, duration/pitch adjustment, event zoom, free Q&A), and the JSON action protocol, followed by the context block. The action protocol keeps the LLM out of the signal processing: a reclassification request emits `{"action": "reclassify", "verdict": "<key>"}`, which the action executor uses to re-run the audification adviser and regenerate the WAV file; a zoom request emits `{"action": "zoom", "pre_s": <float>, "post_s": <float>, "events": <list|"all">}`, handled directly by the signal focus module using the existing k. The model is explicitly instructed not to compute k values itself.

An optional user-supplied domain description (e.g., *"ECG with arrhythmias"*) is collected before analysis and prepended to the context block, shaping both the initial summary and all subsequent responses without changing any signal processing behavior. For a BLV user, this is the only setup step.

4. Interaction Model

Immediately after the audification adviser completes, the LLM generates a structured report in four sections: a signal overview, a classification explanation situated in the user's domain, an audification section explaining what the audio will sound like and why, and a notable features section flagging anomalies or events. Because the report is plain text, a screen reader speaks it in full without any visual parsing [22]. Technical depth is calibrated to the domain description: a lay description elicits plain-language explanations, while an absent or technical description prompts a more signal-processing-oriented response. For the three classes excluded from audification, the LLM flags this explicitly and recommends an auditory graph instead.

Once the summary is delivered, the user can continue the conversation to refine the audification in any direction. The LLM interprets the request and, where a concrete pipeline change is needed, emits a structured action token; it never computes signal processing quantities itself. Typical requests include reclassification, duration adjustment, pitch adjustment (where the model first warns that pitch and duration are coupled under uniform time scaling [4]), and event zoom using the same k as the full recording. The user can also ask why a class was chosen, whether the confidence is trustworthy, or anything else about the analysis; the LLM answers from the metadata block it was given.

Because this entire loop is plain text, it works through whatever input method the user relies on, keyboard, voice dictation, or switch-based input, and every response can be spoken aloud by a standard screen reader. This is not a separate accessibility mode or reduced-functionality fallback; it is the primary interaction channel. Even among the small number of tools built for BLV users, such as xSonify [11] and SonoUno [10], the typical design pattern provides an audio rendering but not a verbal explanation of what it encodes or where to listen [7]. The LLM layer supplies exactly this missing verbal context, turning the audification from a sound to listen to into a sound the user has been told what to listen for.

5. Illustrative Example

We briefly illustrate the interaction on a broadband seismogram of the 2004 Sumatra–Andaman earth-

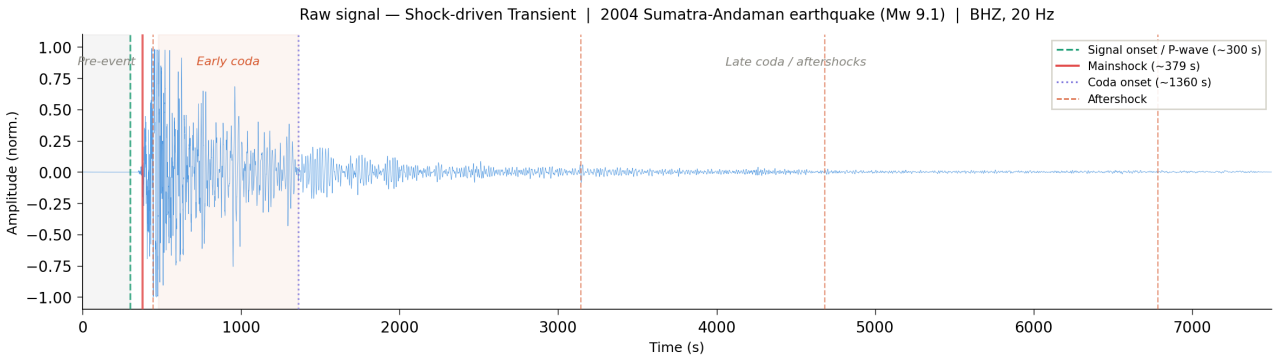

Figure 2: BHZ seismogram of the 2004 Sumatra–Andaman earthquake (20 Hz, 7 500 s). Green dashed line: signal onset at 300 s (P-wave proxy). Red solid line: mainshock at 379 s. Purple dotted line: coda onset. Orange dashed lines: four aftershock onsets detected by the signal focus module. Gray shaded region: pre-event quiet interval.

quake (M_w 9.1), captured at 20 Hz over a 125-minute window, with the user description *"Sumatra earthquake from December 2004"*. The pipeline classified the signal as `shock_driven_transient` with 92% confidence, derived $k = 5\,386$ from a Brune corner-frequency estimate, and compressed the 7 500 s recording to approximately 1.4 s of audio. The resulting LLM summary described a shock-driven transient as a brief, high-energy event followed by a decaying coda, characterized the audio as a low-pitched thud followed by an uneven, fading rumble, and reported that the coda's texture shifts gradually across the compressed recording, giving the user a concrete perceptual anchor before hearing the file. The signal focus module additionally identified five events (the mainshock and four aftershocks; Figure 2), each individually audifiable via a zoom request at the same k; a screen reader user receives this event list as spoken text and can request any subset by name without locating them visually. A second worked example, a resting-elderly-subject ECG segment from the Fantasia database [26] eliciting a plain-language summary calibrated to a lay domain description, is omitted here for space and will be included in an extended version of this paper.

6. Limitations and Ongoing Work

The present work describes a system prototype. As of July 2026, a formal evaluation is underway covering the graphical interface, the LLM interaction layer, and the audification outputs across a range of signal domains and user backgrounds, including expert, non-expert, and, critically, BLV participants evaluating the screen-reader-based workflow specifically. In particular, we aim to assess whether a BLV user can complete a full analysis cycle, upload through classification, summary, refinement, and event zoom, without sighted assistance, something very few existing sonification tools have been formally tested for [7]. Results will be presented in forthcoming work.

A related open question concerns summary quality: the Mistral API and a locally deployed Mistral 7B via Ollama receive identical context blocks but can produce noticeably different summaries in scientific precision, verbosity, and tone; the system prompt was designed manually without formal prompt engineering optimization. The LLM also cannot perceive its own audio output: it can describe what a signal should sound like given the verdict and scaling factor, but cannot verify whether the rendered WAV matches that prediction. Closing this loop, likely via a contrastive language-audio model such as CLAP [27] to embed the rendered WAV and flag mismatches against the LLM's own description, is the main planned technical addition alongside the user evaluation. Furthermore, the system is planned to be extended to further and more flexible sonification techniques, specifically parameter mapping.

7. Conclusion

This paper presented a natural-language interaction layer for a morphology-aware rule-based audification framework, keeping the LLM strictly downstream of the unchanged pipeline while generating domain-calibrated summaries, supporting reclassification and parameter adjustment, and exposing detected events as explicit listening targets. Because the interaction is entirely text-based, it lowers the barrier for people without a signal-processing background, people without domain expertise, and blind and low-vision users, who can operate the full workflow through a standard screen reader with no reduced-functionality fallback. Sonification research has long named the BLV community as one of its most natural audiences [9, 3], while the tools it produces have overwhelmingly remained visual-first [7, 8]. The present system does not solve that problem in general, but demonstrates that an LLM-based conversational layer applied downstream of an existing pipeline can deliver both the audio rendering and its verbal explanation through a single text channel that assistive technology already knows how to speak, without cost to analytical rigor. Automated evaluation confirms the LLM reliably reports pipeline outputs accurately and that parameter computations are arithmetically correct; a formal evaluation with hu-

man participants and a CLAP-based audio verification step are the natural next steps.

References

[1] T. Hermann, A. Hunt, and J. G. Neuhoff, *The Sonification Handbook*. Berlin: Logos Verlag, 2011.

[2] F. Dombois, "Using audification in planetary seismology," in *Proc. of the International Conference on Auditory Display (ICAD)*, Espoo, Finland, 2001.

[3] G. Kramer, Ed., *Auditory Display: Sonification, Audification, and Auditory Interfaces*, ser. SFI Studies in the Sciences of Complexity, Proceedings Volume XVIII. Reading, MA: Addison-Wesley, 1994.

[4] K. Groß-Vogt, M. Frank, and R. Höldrich, "Focused audification and the optimisation of its parameters," *Journal on Multimodal User Interfaces*, vol. 14, pp. 187–198, 2020.

[5] B. N. Walker and M. A. Nees, "Theory of sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin, Germany: Logos Publishing House, 2011, ch. 2, pp. 9–39. [Online]. Available: http://sonification.de/handbook/chapters/chapter2/

[6] S. Vallejo Budziszewski, K. Gross-Vogt, and R. Höldrich, "Rule-based framework for automated audification," in *Proceedings of the 31st International Conference on Auditory Display (ICAD 2026)*. Distributed hubs (Bielefeld, Barcelona, Troy/NY): Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria, July 2026.

[7] S. Vallejo Budziszewski and K. Gross-Vogt, "Legacy, active & future software tools in sonification research," in *Proceedings of the 30th International Conference on Auditory Display (ICAD)*, Coimbra, Portugal, 2025.

[8] J. Zong, I. Pedraza Pineros, M. K. Chen, D. Hajas, and A. Satyanarayan, "Umwelt: Accessible structured editing of multi-modal data representations," in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. New York, NY, USA: ACM, 2024.

[9] W. L. Díaz-Merced, "Sound for the exploration of space physics data," Ph.D. dissertation, University of Glasgow, 2013.

[10] J. Casado, G. de la Vega, and B. García, "SonoUno development: A user-centered sonification software for data analysis," *Journal of Open Source Software*, vol. 9, no. 93, p. 5819, 2024.

[11] B. García, W. Díaz-Merced, J. Casado, and A. Cancio, "Evolving from xSonify: A new digital platform for sonorization," in *EPJ Web of Conferences*, vol. 200, 2019, p. 01013.

[12] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366, pp. 427–431, 1979.

[13] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *Journal of Econometrics*, vol. 54, no. 1–3, pp. 159–178, 1992.

[14] J. D. Hamilton, *Time Series Analysis*. Princeton: Princeton University Press, 1994.

[15] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[16] S. Mallat, *A Wavelet Tour of Signal Processing*, 3rd ed. Burlington: Academic Press, 2009.

[17] R. V. Allen, "Automatic earthquake recognition and timing from single traces," *Bulletin of the Seismological Society of America*, vol. 68, no. 5, pp. 1521–1532, 1978.

[18] R. L. Brown, J. Durbin, and J. M. Evans, "Techniques for testing the constancy of regression relationships over time," *Journal of the Royal Statistical Society: Series B*, vol. 37, no. 2, pp. 149–163, 1975.

[19] J. Bai and P. Perron, "Estimating and testing linear models with multiple structural changes," *Econometrica*, vol. 66, no. 1, pp. 47–78, 1998.

[20] C. Frauenberger, A. de Campo, and G. Eckel, "Analysing time series data," in *Proc. of the International Conference on Auditory Display (ICAD)*, Montreal, Canada, 2007.

[21] Y. Zhao and G. Tzanetakis, "Interactive sonification for health and energy using ChucK and Unity," in *Proceedings of the SoniHED Conference*, 2022. [Online]. Available: https://arxiv.org/abs/2404.08813

[22] J. Zong, C. Lee, A. Lundgard, J. Jang, D. Hajas, and A. Satyanarayan, "Rich screen reader experiences for accessible data visualization," *Computer Graphics Forum*, vol. 41, no. 3, pp. 15–27, 2022.

[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. Renard Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

[24] J. A. Jansen, A. Manukyan, N. Al Khoury, and A. Akalin, "Leveraging large language models for data analysis automation," *PLOS ONE*, vol. 20, no. 2, p. e0317084, 2025.

[25] T. Stockman, "Keynote," Keynote talk, 31st International Conference on Auditory Display (ICAD 2026), Distributed hubs (Bielefeld, Barcelona, Troy/NY), July 2026.

[26] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[27] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.