DOI: 10.61782/fa.2025.0019

FORUM ACUSTICUM
EURONOISE 2025

A MODEL FOR IMPROVING THE PERFORMANCE OF LARGE LANGUAGE MODELS ON SOUNDSCAPE DESCRIPTION TASKS

Zhongju Yuan[1] **Dick Botteldooren**[1*]

[1] Department of Information Technology, Ghent University, Belgium

ABSTRACT

Generative AI is increasingly used in acoustic scene analysis and environmental soundscape description. Audio Language Models (ALMs) are key in this field but have two main drawbacks: they don't provide human-like survey responses and require significant computational resources, making real-time monitoring impractical. These issues stem from differences in how ALMs and humans assess and store sensory experiences. We propose a computationally efficient model with a three-layer architecture. The top layer features an advanced ALM for recognizing, grouping, and remembering audio events. The intermediate layer includes a local salient change detector based on sound feature embedding, which identifies meaningful changes in the auditory environment. When a change is detected, the ALM interprets and remembers the sound. The bottom layer has a memory cell mimicking echoic memory, retaining and reconstructing recent audio inputs. This reconstructed memory is passed to the ALM for recognition when triggered by the salient detector. Tested on the Urban Soundscapes of the World dataset, our model balances computational efficiency with real-time recognition accuracy, paving the way for scalable, intelligent acoustic monitoring solutions.

Keywords: *soundscape, saliency, large language models*

1. INTRODUCTION

Soundscape, the sonic environment as perceived and understood by persons of society within context, is increasingly used to assess and design the (urban) sound environment. With this definition in mind, measuring the soundscape remains a challenge. The ISO/TS 12913-2 standard series suggests surveys, sound walks, and focused listening to assess the soundscape. Artificial intelligence (AI) today allows us to recognize sounds and relate them to emotional attributes, e.g. annoyance [1]. However, such models are not very biologically plausible. Generative AI has recently seen tremendous expansion. Audio Language Models (ALMs), a special branch of large language models, offer the advantage that they can take into account context and common expectations, aspects that were shown to be extremely important in soundscape evaluation [2], [3], [4].

However, a direct use of ALMs on continual sound streams detected in urban public places does not account for the critical role of human attention in urban sound perception, which has often been overlooked in soundscape research. In fact, in daily life, the urban sound environment is seldom the focus of attention of the space user. Therefore in previous work, we have introduced an additional dimension orthogonal to the two-dimensional pleasure - arousal plane, backgrounding, measuring the degree of influence on overall perception of the space [5].

Allowing the sound to be backgrounded, that is gated out from further cognitive processing, also alows to reduce the amount of effort the ALM needs to put in analyses of the sound. Inference of answers from generative AI is a highly energy consuming process [6]. When connected to edge devices for monitoring soundscape, the combined process becomes extremely energy unfriendly and expensive. Nature has solved this problem also encountered by biological brains by allowing to gate out any repeated, continuous or otherwise uninformative sound. To assure that the function of hearing as an alerting and situation monitoring system is kept intact, gating out or in-

**Corresponding author*: dick.botteldooren@ugent.be.

hibition must be accompanied by a saliency detection that triggers attention bottom up [7]. Auditory saliency has been widely studied [8] and has been related to changes in acoustical features [9], extending the classical Kayser-model [10].

In this paper, we propose a hierarchical framework to model the auditory processing system. The bottom layer functions as the echoic memory, responsible for memorizing and retrieving short audio segments. These sliding short segments are passed to the upper short-term memory layer for pattern change detection. When a change in the pattern is detected compared to the previous segment, the higher cognitive layer—referred to as ALM in our context—takes over the processing. At this stage, a longer auditory memory span is analyzed, and responses to several predefined questions are generated.

The proposed framework demonstrates that the model can significantly reduce the workload of the ALM by decreasing query frequency, while maintaining overall performance. By offloading routine pattern recognition and memory operations to the lower layers, the ALM is only engaged when higher-level cognitive processing is required. This hierarchical design not only improves computational efficiency but also better reflects the biological plausibility of human auditory processing. Furthermore, the reduction in unnecessary queries enables the ALM to allocate more resources to complex tasks, thereby enhancing the system's scalability and responsiveness in dynamic auditory environments.

2. MODEL

The proposed model is subdivided in large blocks that roughly correspond to functions in the biological system. The cochlear and brain-stem model conducts the primary spectral filtering. For the purpose of this paper, a simple gammatone filter bank is used for calculation speed, but this block can easily be replaced by a more precise model such as CONNear [11].

The block labeled "echoic memory" stores time traces of the (filtered) sensory input. In line with observations in humans, it allows to go back in time for about 4 seconds. This is essential to allow analysing a salient event that might be detected slightly after it has actually happened. Biologically plausible implementations of sensory, echoic memory might reveal special effects in real-time applications, but for the current paper the stored audio file is simply recalled when needed.

A central element is the predictive model used for detecting changes and saliency. In contrast to earlier work, no attempt is made to extract features that might influence the detection of sensory salient signals. Instead, it is assumed that a convolutional neural network (CNN) trained to detect sounds that are of relevance for humans will identify those features that define a sound object. Thus unpredictable changes of the embedding resulting from encoding a short epoch of sound can be seen as changes in the sound environment that might warrant re-evaluation of the soundscape. If such detections occur, the gate connecting the sensory memory to the large language model is opened for a predefined amount of time. The latter could still be a point of discussion. If the sound is of low interest for the person visiting the environment, top down attention will fade quickly and further processing will stop. However, conversations or longer bird songs may keep attention focussed. In earlier work on notice events we estimated that the decay time of top down attention for uninteresting sounds could be up to 10 seconds [12]. In this paper a fixed time interval of 15 seconds is used.

Any pretrained large generative model can be used as a cognitive model in this framework. We associate these models with human cognition because for soundscape assessment they fulfill roles that are usually associated with cognitive processes: (1) recognizing sounds and associating meaning where meaning is seen as the complex set of associations that are triggered by the sound; (2) putting the sound in a context, e.g. by specifying that a recording is made in an outdoor public place; (3) associating affect (e.g. calming, stimulating). The appropriateness of using the ALM for the latter could be debated as it probably only covers the common sense evaluation. For example, such models may easily link voices of people and traffic to lively environments and birds and water to calming environments, but they do not adequately account for the complexity of sound mixtures [13], where complexity is understood as the effort needed to analyze the auditory scene [14].

Prompt engineering, the craft of prompting a generative model properly, requires some attention. Here, we employ *Qwen-audio* [15] to generate the final responses. After several preliminary experiments, the following prompt was selected for use in the experimental setting:

> **Prompt:** Assume that you are a human in a public place, answer the following questions:
>
> 1. Can you describe what sounds you

hear?

2. What type of public place do you think you are in?
3. In general, how would you categorize the environment you just experienced?
(Options: calming / tranquil, lively / active, or neither)

Provide a structured response for all these questions.

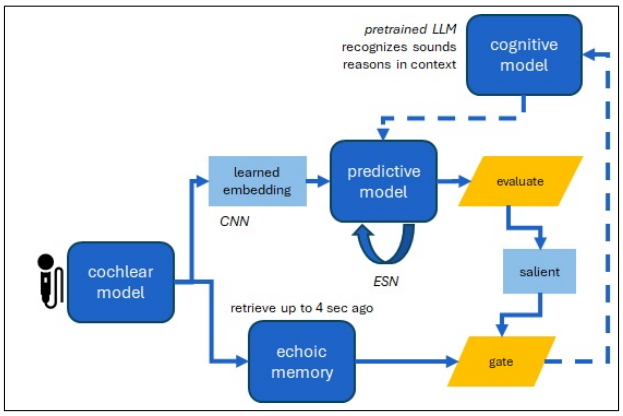

Figure 1. Connection between main components in the proposed model: the predictive model uses a learned embedding of the sound to detect any unpredictable changes in the sound; if such a change occurs it opens the gate to feed the cognitive model with new sound input; the echoic memory allows to retrieve the sound up to 4 seconds before the detection of a change, a salient event.

3. URBAN SOUNDSCAPES OF THE WORLD

As a test case for the approach, we used the Urban Soundscape of the World dataset. This database has been recorded at carefully selected locations in multiple cities across the globe [16]. The database contains high quality 4-channel ambisonics and binaural recordings and is supported by 360 degree video. The original recordings are around 10 minutes long, but 1-minute excerpts have been used in this work [17].

The Urban Soundscapes of the World database has been used in several lab studies. Here, the evaluation of each audiovisual recoding by 20 persons in a carefully constructed virtual reality experiment will be used [5]. In this work, participants are questioned in a much more subtle way than ISO/TS 12913-2. In brief: participants are first ask to rate the environment as a whole on a calming/tranquil to lively/active scale; then they are triggered to reflect on the activities that they could do in this environment in order to set a context for further evaluation; they are then asked whether the sound has drawn their attention during the AV experience (this leads to a rating on the *backgrounded* axis); they are then queried about the possible interruption of the activities they had in mind by the sound (this leads to a rating on the *disruptive* axis); depending on their answer to the first question, they are then asked to rate to what extend the sound environment contributed to that experience (this leads to a rating on the *calm/tranquil* and *lively/active* rating dimensions).

4. RESULTS

4.1 Overall accuracy

The model proposed in this paper is purely auditory, hence it has no visual context or any other stimulus. Thus, it is not possible for the model to rate whether the sound environment would be *backgrounded*. For the same reason *disruption* of potential activities is hard to rate as these potential activities depend largely on the overall knowledge about the place that human participants in an experiment would largely get from visual input. Therefore, this evaluation is restricted to the *calm/tranquil* and *lively/active* dimensions. 2 shows the distribution of average rating by the 20 participants from [5] for two classes predicted by the model: tranquil and lively. In the left figure, all sound fragments are used, in the right figure only those where the people rated that the sound environment had at least a moderate influence on the overall perception. By restricting the evaluation to sound environments that were rated by people, on average, to have an influence on the perception, the two distributions become more disjoint. Yet the model still classifies several sound environments as tranquil while persons would give them a more lively rating.

4.2 Ablation study

The proposed framework improves the ALM's ability to handle real-time scenarios. In this subsection, we demonstrate the performance of the framework. We evaluate its efficiency based on two metrics: the final tranquility level output by Qwen, and the total audio duration processed

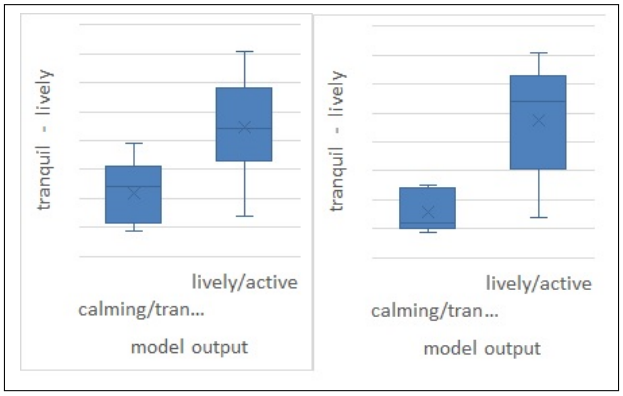

Figure 2. Box plot of the rating by participants of the environments on a *calm/tranquil* and *lively/active* dimension for the group of soundscapes rated as tranquil or lively by the model; left: all recordings, right: only environments where the sound is rated as influential by people.

per sample—since it's not feasible to send audio to the ALM every second for description generation.

First, we compare the tranquil/lively classification produced by Qwen using the complete sound fragment and the excerpt selected by the proposed model (cosine similarity threshold 0.9). We remind the reader that our model only sends a 15-second segment for evaluation when a pattern change is detected. The description from the most recent segment is used to determine tranquil/lively.

A confusion matrix has been calculated comparing the classifications based on analysing the full minute and a classification based on the proposed model. This showed that 54 recordings are classified as *lively/active* by both approaches and 10 recordings are classified as *calm/tranquil* by both approaches. However, 9 recordings that were classified as classified as *lively/active* on the basis of the full minute get classified as *calm/tranquil* by the proposed model and 4 recordings that were classified as *calm/tranquil* on the basis of of the full minute get classified as *lively/active* by the proposed model.

Figure 3 shows the same results as Figure 2 but now using the classification based on an evaluation of the full minute recording by Qwen. Both results are very similar and small differences tend to show that the proposed model slightly better predicts the evaluation by people. The latter observation should be interpreted with caution as no statistical significance of this difference is shown.

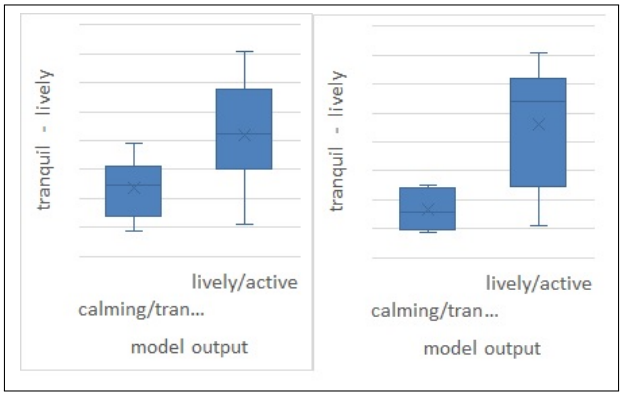

Figure 3. Box plot of the rating by participants of the environments on a *calm/tranquil* and *lively/active* dimension for the group of soundscapes rated as tranquil or lively based on the full one minute recording; left: all recordings, right: only environments where the sound is rated as influential by people.

To clarify that the time length processed by Qwen is significantly shorter, we provide statistical analysis of the processed durations across all audio samples. The results are presented in Fig. 4.

This violin plot shows the distribution of audio segment lengths sent to the Qwen model at a 0.9 threshold, combining density, individual data points, and a box plot. The distribution is bimodal with peaks at 15 and 30 seconds. The mean (22.51s) exceeds the median (15.00s), indicating right skew due to longer segments. Most data range from 15 to 60 seconds, with clusters in the lower range and a few high-value outliers. From Fig. 4, it is evident that approximately half of the audio fragments in the dataset do not exhibit significant pattern shifts. As a result, the framework sends only the initial 15 seconds to Qwen. This shortened processing time does not affect the resulting tranquility level.

5. DISCUSSION

Acoustic Language Models, a special subbranch of Large Language Models that combine sound recognition with the *common sense* knowledge embedded in these pretrained models are becoming notoriously powerful for describing acoustic scenes. However, directly applying such models to continuous sound streams recorded in urban

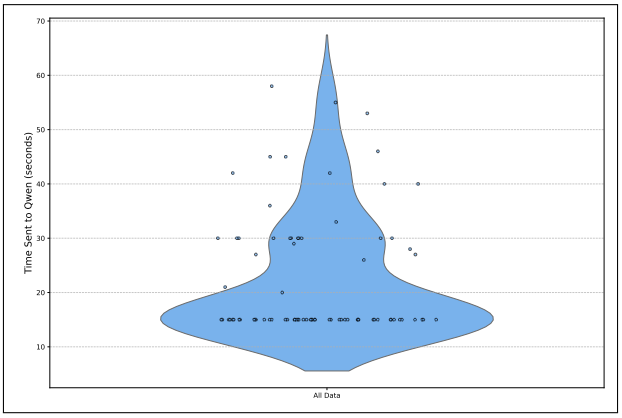

Figure 4. Distribution of Time Length Sent to Qwen at Threshold 0.9. This violin plot illustrates the distribution of audio time lengths processed and sent to the Qwen model at a threshold of 0.9.

sound monitoring for example, is extremely energy consuming and expensive. Inspired by human auditory scene analysis, we propose a model that decides which sound can safely be gated out without losing any essential information.

Environmental sounds often fade into the background for individuals enjoying public spaces. After an initial assessment of the sonic environment, the human mind tends to wander, only reassessing the surroundings when a notable sound event occurs. Numerous models have been proposed to identify these salient sounds, typically grounded in psychoacoustic principles. The proposed model, however, takes a different path. It posits that Artificial Learning Models (ALMs) trained to recognize sounds relevant to humans will develop embeddings that emphasize the features the human auditory system is attuned to. Consequently, a change in these embeddings would signal a shift in sounds significant to human perception, thereby suggesting triggering a release of gating.

This model is used here to assess sound environments available in the urban soundscape of the world database, with soundscape in mind. In particular, the labeling of the sound environment as *lively/active* or *calm/tranquil* is considered. For this, results obtained by human participants evaluation an audiovisual environment in virtual reality [5] can be used as a reference. It is shown that the proposed model predicts human evaluation at least as good as an evaluation based on the whole fragment. At the same time, very often, processing as little as 15 seconds using the ALM is sufficient. Only very seldom more than 30 seconds of sound is needed to reach the decision.

6. ACKNOWLEDGMENTS

This work was supported in part by the Research Foundation - Flanders, Belgium under grant number G0A0220N (FWO WithMe project), and the Flemish Government, Belgium under the "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen"

7. REFERENCES

[1] Y. Hou, Q. Ren, H. Zhang, A. Mitchell, F. Aletta, J. Kang, and D. Botteldooren, "Ai-based soundscape analysis: Jointly identifying sound sources and predicting annoyance," *The Journal of the Acoustical Society of America*, vol. 154, no. 5, pp. 3145–3157, 2023.

[2] K. Filipan, M. Boes, B. De Coensel, C. Lavandier, P. Delaitre, H. Domitrović, and D. Botteldooren, "The personal viewpoint on the meaning of tranquility affects the appraisal of the urban park soundscape," *Applied Sciences*, vol. 7, no. 1, p. 91, 2017.

[3] Z. Kankhuni and C. Ngwira, "Overland tourists' natural soundscape perceptions: influences on experience, satisfaction, and electronic word-of-mouth," *Tourism Recreation Research*, vol. 47, no. 5-6, pp. 591–607, 2022.

[4] H. I. Jo and J. Y. Jeon, "Urban soundscape categorization based on individual recognition, perception, and assessment of sound environments," *Landscape and urban planning*, vol. 216, p. 104241, 2021.

[5] K. Sun, B. De Coensel, K. Filipan, F. Aletta, T. Van Renterghem, T. De Pessemier, W. Joseph, and D. Botteldooren, "Classification of soundscapes of urban public open spaces," *Landscape and urban planning*, vol. 189, pp. 139–155, 2019.

[6] A. A. Chien, L. Lin, H. Nguyen, V. Rao, T. Sharma, and R. Wijayawardana, "Reducing the carbon impact of generative ai inference (today and in 2035)," in *Proceedings of the 2nd workshop on sustainable computer systems*, pp. 1–7, 2023.

[7] C. N. Price and D. Moncrieff, "Defining the role of attention in hierarchical auditory processing," *Audiology Research*, vol. 11, no. 1, pp. 112–128, 2021.

[8] N. Huang and M. Elhilali, "Auditory salience using natural soundscapes," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2163–2176, 2017.

[9] S. R. Kothinti, N. Huang, and M. Elhilali, "Auditory salience using natural scenes: An online study," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2952–2966, 2021.

[10] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[11] D. Baby, A. Van Den Broucke, and S. Verhulst, "A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications," *Nature machine intelligence*, vol. 3, no. 2, pp. 134–143, 2021.

[12] B. De Coensel, D. Botteldooren, T. De Muer, B. Berglund, M. E. Nilsson, and P. Lercher, "A model for the perception of environmental sound based on notice-events," *The Journal of the Acoustical Society of America*, vol. 126, no. 2, pp. 656–665, 2009.

[13] D. Botteldooren, "Urban sound design for all," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 261, pp. 2012–2017, Institute of Noise Control Engineering, 2020.

[14] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[15] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[16] B. De Coensel, K. Sun, and D. Botteldooren, "Urban soundscapes of the world: Selection and reproduction of urban acoustic environments with soundscape in mind," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 255, pp. 5407–5413, Institute of Noise Control Engineering, 2017.

[17] B. De Coensel, D. Botteldooren, S. Kang, and T. Van Renterghem, *Urban soundscapes of the world*. zenodo: https://zenodo.org/records/10106181.