THIS WAY UP: EVALUATING FREQUENCY–ELEVATION MAPPING IN AUDIO–LANGUAGE MODELS

Vincent LOSTANLEN[1] and **Mathieu LAGRANGE**[1]

[1]**Nantes Université, Centrale Nantes, CNRS, LS2N, UMR 6004**, F-44000, Nantes, France

ABSTRACT

In psychoacoustics, frequency-elevation mapping (FEM) refers to the widespread spatial connotation of auditory pitch: e.g., ascending and descending scales. We evaluate whether FEM is an emergent property in two deep learning models which are based on contrastive language-audio pretraining (CLAP): namely, MS-CLAP and LAION-CLAP. With MS-CLAP, the answer seems negative: after multiple attempts at prompt engineering, spatial elevation does not correlate with musical pitch in the TinySOL dataset of isolated notes from 14 instruments. The situation is different with LAION-CLAP, for which the prompt "This sound source is located above my ears" yields a surprisingly good zero-shot fundamental frequency estimator (Spearman's $\rho = 0.673$). However, the same prompt yields chance-level performance in binary classification between environmental sounds which are typically placed above the listener (e.g., fireworks) versus below (e.g., pouring water into a glass). Together, these results suggest that FEM is not robustly or consistently available in the current generation of pretrained audio–language models. We attribute this shortcoming to a misalignment between the two encoders, rather than a limitation of the audio encoder on its own.

1. INTRODUCTION

1.1 Audio–language models and their evaluation

Audio–language models (ALM's) are a new generation of machine listening systems which may indiscriminately respond to an audio signal $\boldsymbol{x}^{\mathrm{a}}$ or a textual prompt $\boldsymbol{x}^{\mathrm{t}}$ [1]. Under its simplest form, an ALM consists of two deep neural networks (DNN's): an audio encoder f^{a} and a text encoder f^{t}, with trainable weights $\boldsymbol{w}^{\mathrm{a}}$ and $\boldsymbol{w}^{\mathrm{t}}$, respectively.

Contrastive language–audio pretraining (CLAP, [2]) optimizes $\boldsymbol{w}^{\mathrm{t}}$ jointly over batches of audio–text pairs $\big((\boldsymbol{x}_i^{\mathrm{a}}, \boldsymbol{x}_i^{\mathrm{t}})\big)_{i=1}^{N}$ so that their output vectors, known as *embeddings*, coincide: intuitively, $f^{\mathrm{a}}(\boldsymbol{x}_i^{\mathrm{a}}, \boldsymbol{w}^{\mathrm{a}}) \approx f^{\mathrm{t}}(\boldsymbol{x}_j^{\mathrm{t}}, \boldsymbol{w}^{\mathrm{t}})$ if and only if $i = j$. As a consequence, a well-trained CLAP embedding performs *cross-modal retrieval*: given a sound $\boldsymbol{x}_0^{\mathrm{a}}$, it is possible to retrieve the most suitable caption $\boldsymbol{x}_{j^*}^{\mathrm{t}}$ among a finite collection of candidates, $(\boldsymbol{x}_j^{\mathrm{t}})_{j=1}^{N}$ as its nearest neighbor in the space of audio–text embeddings:

$$j^* = \arg\min_{1 \leq j \leq N} d\big(f^{\mathrm{a}}(\boldsymbol{x}_0^{\mathrm{a}}, \boldsymbol{w}^{\mathrm{a}}), f^{\mathrm{t}}(\boldsymbol{x}_j^{\mathrm{t}}, \boldsymbol{w}^{\mathrm{t}}\big), \qquad (1)$$

where d is a distance function. Furthermore, the ability for cross-modal retrieval may be harnessed for *zero-shot classification*: by crafting textual prompts $\boldsymbol{x}_j^{\mathrm{t}}$ of the form "the sound of C_j", for C_j a category of interest, Equation 1 becomes a criterion for N-way classification, where the label set $(C_1, \ldots, C_N)$ may be specified at test time.

The invention of CLAP has renewed a long-standing debate in sound and music computing: in what sense, and to what extent, can DNN's be said to "understand" audio? On one hand, the authors of CLAP describe it as "learning audio concepts from natural language supervision" [2]. [3] have praised it as "learning semantically rich audio representations". [4] have shown that conditioning a large language model on f^{t} at training time and f^{a} at test time effectively enables audio captioning abilities. On the other hand, [5] has found "deficiencies in logical reasoning": e.g., a contradiction between $\boldsymbol{x}^{\mathrm{a}}$ being the sound of calm water and $\boldsymbol{x}_*^{\mathrm{t}}$ being "a fast-paced motorboat is speeding through the water". [6] have reported that CLAP fails to retrieve certain timbre qualities such as "cool" and "warm", in the context of music equalization. [7] have shown that a CLAP-like model fails to "really understand [...] contextual concepts such as sequential or concurrent ordering of sound events". [8] have reported that "CLAP-based approaches face notable limitations in terms of speaking styles entanglement". [9] have shown that CLAP "struggle[s] with semantic phenomena such as negation", i.e., contrasting "with vocals" versus "without vocals".

1.2 Frequency–elevation mapping in psychoacoustics

In this article, we aim to contribute to the ongoing debate on the ability of current-generation ALM's to "understand" auditory concepts. Taking inspiration from prior work, above cited, we will craft textual prompts and use them to retrieve pre-recorded audio examples in CLAP feature space. However, we propose to explore *both* literal and figurative uses of a single concept: in this instance, the concept of height. This is in contrast with prior work, which has queried CLAP *either* in the literal sense (e.g., "a fast-paced motorboat" [5]) *or* in the figurative sense (e.g., "warm" as a timbre descriptor [6]). Our study finds its motivation in the "spatial character of high and low tones", as theorized by psychologist Carroll C. Pratt in 1930:

> The tones at one end of the pitch-continuum are called high, and those at the

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

> other end are called low, and qualitative changes along this continuum are spoken of as rising or falling inflections, ascending or descending cadences, downward or upward movements. The conspicuous fitness of such phrases is immediately apparent. [10]

Pratt's effect, known as "frequency–elevation mapping" (FEM), is well established in psychoacoustics. [11] have hypothesized that FEM may result from "an environmental association between lightness, pitch, and size":

> [...] Smaller higher-pitched objects are, statistically-speaking more likely to be located somewhat closer to the source of ambient illumination and hence to be better illuminated and appear visually lighter than larger objects.

To test this hypothesis, [12] have collected multichannel directional recordings at ear's height. For each of them, they estimated elevation via cross-correlation and mapped it to the frequency of the band-pass filter with highest magnitude response. They concluded that

> [...] high-frequency sounds have a tendency to originate from elevated sources in natural auditory scenes. We can only speculate about the origins of this mapping: it could either be that at higher elevations, more energy is generated in high frequencies [...] or it could also be that the absorption of the ground is frequency-dependent in a way that it filters out more of the high-frequency spectrum.

This correlation is reinforced by the anatomy of the outer ear, which acts as an elevation-dependent filter. According to [12], this filtering property reinforces the FEM which is present in natural auditory scenes. More recently, [13] have shown that FEM arises early in human development: newborns are sensitive to the correspondence between auditory pitch (a sliding tone going "up" or "down") and visuospatial elevation (objects moving "up" or "down" on a screen), as measured by their differences in looking behavior. Together, these findings illustrate that FEM is an active research topic in psychoacoustics, with far-reaching implications for the nascent field of human auditory ecology [14]. We refer to [15] for a historical and musicological survey of FEM in European music theory.

1.3 Main contribution and findings

We engineer the following pair of prompts:

$$\boldsymbol{x}^{\mathrm{t}}_{\mathrm{up}} = \text{"This sound source is located } \textit{above} \text{ my ears"},$$
$$\boldsymbol{x}^{\mathrm{t}}_{\mathrm{down}} = \text{"This sound source is located } \textit{below} \text{ my ears"}.$$

Our aim is to discover the acoustical correlates of $\boldsymbol{x}^{\mathrm{t}}_{\mathrm{up}}$ and $\boldsymbol{x}^{\mathrm{t}}_{\mathrm{down}}$ in CLAP space. Thus, we compute text embeddings $\boldsymbol{y}^{\mathrm{t}}_{\mathrm{up}} = f^{\mathrm{t}}(\boldsymbol{x}^{\mathrm{t}}_{\mathrm{up}}, \boldsymbol{w}^{\mathrm{t}})$, $\boldsymbol{y}^{\mathrm{t}}_{\mathrm{down}} = f^{\mathrm{t}}(\boldsymbol{x}^{\mathrm{t}}_{\mathrm{down}}, \boldsymbol{w}^{\mathrm{t}})$ and use them to design features for two zero-shot tasks.

The former task is automatic pitch estimation of isolated musical notes. The feature under consideration is the affine projection of the audio embedding onto the axis going from $\boldsymbol{y}^{\mathrm{t}}_{\mathrm{down}}$ to $\boldsymbol{y}^{\mathrm{t}}_{\mathrm{up}}$, as measured by the dot product

$$z = \left\langle f^{\mathrm{a}}(\boldsymbol{x}^{\mathrm{a}}, \boldsymbol{w}^{a}) - \boldsymbol{y}^{\mathrm{t}}_{\mathrm{down}} \,\middle|\, \boldsymbol{y}^{\mathrm{t}}_{\mathrm{up}} - \boldsymbol{y}^{\mathrm{t}}_{\mathrm{down}} \right\rangle. \qquad (2)$$

The latter task is binary classification between sound sources which are typically above the listener (e.g., fireworks) versus below the listener (e.g., pouring water into a glass). The feature under consideration is the sign of z.

Our results are nuanced and at odds with human psychoacoustics. On one hand, one pretrained instance of the model (MS-CLAP) fares poorly on the former task (Spearman's $\rho = 0.12$) but succeeds moderately well on the latter task (accuracy: 72%). On the other hand, another pretrained instance (LAION-CLAP) succeeds moderately well on the former task (Spearman's $\rho = 0.67$) but performs at chance level on the latter task (accuracy: 50%).

Crucially, the failure of LAION-CLAP to "understand" the literal sense of high and low is not solely attributable to a deficiency of the audio encoder, but to the interaction between the audio encoder and the text encoder. Indeed, there exists support vectors in LAION-CLAP space which achieve perfect binary classification accuracy on the latter task. However, these support vectors are at a large angle with our prompt pair, $(\boldsymbol{y}^{\mathrm{t}}_{\mathrm{up}} - \boldsymbol{y}^{\mathrm{t}}_{\mathrm{down}})$, as encoded by the text encoder f^{t} with the weight vector $\boldsymbol{w}^{t}$. Although it is conceivable to fine-tune $\boldsymbol{w}^{t}$ while leaving $\boldsymbol{w}^{\mathrm{a}}$ unchanged, so as to enhance classification accuracy, doing so might come at the detriment of FEM in terms of Spearman's ρ for the prompt pair above: thus, we regard fine-tuning as off-topic for the present study. We have yet to find an ALM in which frequency-elevation mapping is robust and "emergent", i.e., actionable without any need for fine-tuning. One possible interpretation of our result is that current-generation ALM's such as CLAP lack "common sense", including grounding in the physical world such as gravity and human pose. The recent invention of audio-centric knowledge graphs is a promising step in this direction [16].

The discrepancies between the successes and failures of CLAP models on the two tasks have noteworthy implications for ALM evaluation in sound and music computing. Such implications are discussed in Section 4.

2. MATERIALS AND METHODS

2.1 TinySOL

TinySOL is a dataset of isolated musical notes, originally recorded at Ircam in Paris (France) between 1996 and 1999, as part of a larger project named Studio On Line (SOL). TinySOL contains 2913 audio samples, as played by 14 instruments from the classical instrumentarium [17]. Pitches range from E_1 ($41.2\,\mathrm{Hz}$) to $\mathrm{C}^{\sharp}_8$ ($4.43\,\mathrm{kHz}$) and dynamics range from *pp* to *ff*. All playing techniques are *ordinario*.

2.2 ESC subset

ESC, which stands for environmental sound classification, is a balanced collection of 2000 recordings from 50 different sound categories; i.e., 40 per category [18]. Each recording is five seconds long. We retain eight categories, which are grouped into two supercategories:

Above. *airplane*; *birdsong*; *fireworks*; *helicopter*.

Below. *footsteps*; *keyboard typing*; *mouse clicking*; *pouring water (into a glass)*.

Each of the two supercategories contains four categories, thus $4 \times 40 = 160$ audio examples. In this paper, we denote this collection of 160 sounds by "ESC" for conciseness, although we should acknowledge that it is only a subset of ESC-50, containing $8/50 = 16\%$ of the whole dataset.

2.3 Models: MS-CLAP and LAION-CLAP

MS-CLAP was invented in 2022 at Microsoft [2]. The audio encoder, f^{a}, is a 14-layer convolutional neural network which was pretrained on 2M audio clips from AudioSet, under the name of "pretrained audio neural network" (PANN) [19]. The PANN operates on log-mel-spectrograms (logmelspec). The text encoder, f^{t}, is a bidirectional Transformer (BERT) with 110M parameters, as implemented and pretrained by HuggingFace [20]. Both encoders are fine-tuned jointly on 128k audio–text pairs from four datasets: AudioCaps [21], FSD50k [22], ClothoV2 [23], and MACS [24]. We rely on the 2023 version of the weights [25]. [1]

LAION-CLAP was released in 2023 by the research collective LAION [26]. The audio encoder f^{a} is based on the hierarchical token-semantic audio Transformer (HTSAT), a Vision Transformer (ViT) which is adapted to logmelspec inputs [27]. The text encoder f^{t} is a BERT, specifically RoBERTa [28]. Both encoders are fine-tuned jointly on 634k audio–text pairs from various publicly available sources, as well as 1.9M samples from AudioSet wherein text prompts are generated via a "keyword-to-caption" language model, operating on the original AudioSet keywords. Although [26] have reported some improvements in audio captioning via a dedicated feature fusion technique, we do not use feature fusion in our benchmark because our audio samples are all below ten seconds in duration. We rely on version 0.2.1 of the open-source package. [2]

As a preliminary experiment, we evaluate the audio encoder on its own via a simple nonparametric decision rule: namely, nearest neighbor classification. We consider the ESC subset: sounds $(\boldsymbol{x}_i^{\mathrm{a}})_{i=1}^N$ and their corresponding categories $(\boldsymbol{c}_i)_{i=1}^N \in \{1, \dots 8\}^N$ with $N = 160$. Nearest-neighbor classification associates each sound $\boldsymbol{x}_i^{\mathrm{a}}$ to the category $\hat{\boldsymbol{c}}_i = \boldsymbol{c}_{j^*}$, $1 \leq j^* \leq N$, where the index j^* is chosen such that $\boldsymbol{x}_{j^*}$ is nearest to $\boldsymbol{x}_i$ in CLAP space:

$$j^* = \arg \min_{1 \leq j \leq N} \| f^{\mathrm{a}}(\boldsymbol{x}_i, \boldsymbol{w}^{\mathrm{a}}) - f^{\mathrm{a}}(\boldsymbol{x}_j, \boldsymbol{w}^{\mathrm{a}}) \|_2. \quad (3)$$

Figure 1 summarizes our results. Despite a few confusions, particularly between airplane and helicopter, both MS-CLAP and LAION-CLAP yield nearly perfect classification accuracy.

2.4 Prompt engineering

In this paper, we use parentheses to abbreviate pairs of prompts: returning to the example in Section 1.3, we abbreviate it as "This sound source is located *above (below)*

[1] MS-CLAP: `https://github.com/microsoft/CLAP`
[2] LAION-CLAP: `https://github.com/LAION-AI/CLAP`

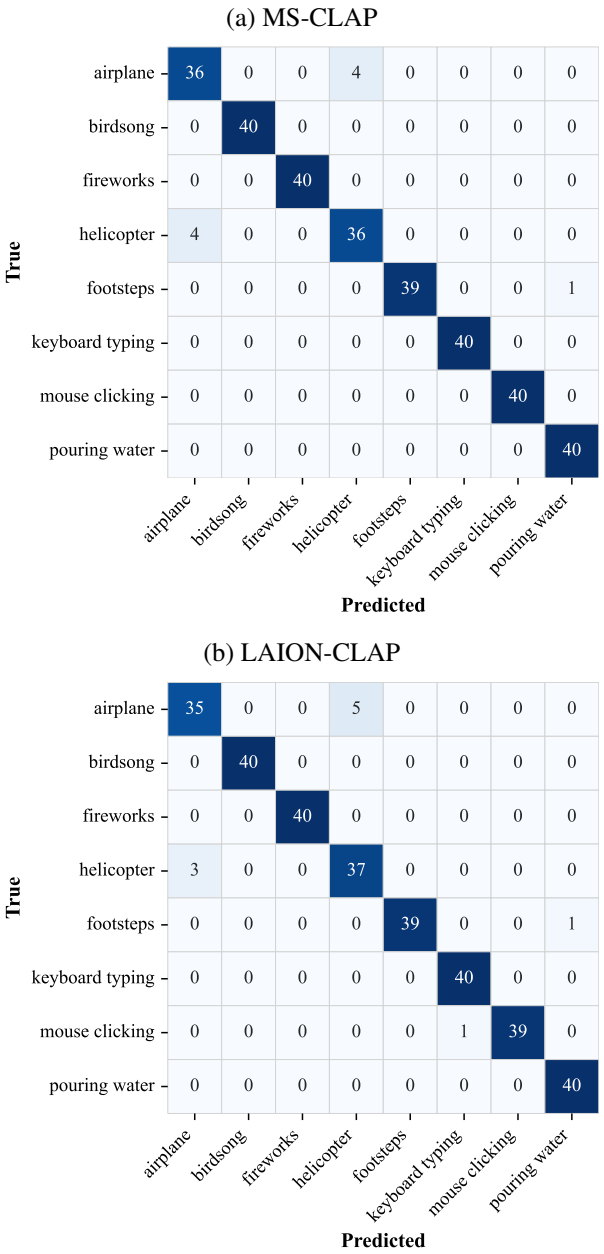


Figure 1. Confusion matrices of nearest neighbor classification with MS-CLAP (a) and LAION-CLAP (b) on our curated subset of ESC. See Section 2 for details.

my ears". In order to evaluate the role of verbal references to oneself and one's body, we engineer three other prompt pairs in which the words "my ears" are replaced by "my head", "me", and "the listener". We also engineer a pair in which the listener is the subject of the sentence: "I am located *below (above)* this sound source"—note the swapping between words *above* and *below* in the pair.

Secondly, as a control experiment, we engineer pairs where FEM is replaced by lateral or ventral/dorsal mappings: "This sound is located to my *left (right)*", "This sound is located *in front of (behind)* my ears", respectively.

Thirdly, we engineer pairs where concepts of high and low are in the figurative sense: "This sound is *high-pitched (low-pitched)*"; "This sound has a *high (low)* pitch"; "This

is a *high-frequency (low-frequency)* sound"; "This sounds *high (low)*". We expect CLAP to perform at least as well with these pairs than with the FEM pairs listed above.

Lastly, we engineer pairs which make indirect references to high and low pitch via specialized musical terms: "This is the sound of a *soprano (tenor)*"; "This sound is *treble (bass)*"; "This note has a *sharp (flat)* accidental". We also include the French sentences, with the premise that this language has dedicated adjectives for high-pitched and low-pitched, which do not hinge on the visuospatial metaphor: "Ce son est *aigu (grave)*" [3] . The role of these pairs is to test to which extent FEM can be enabled in CLAP without making use of polysemic words such as "high" or "above".

3. RESULTS

3.1 Pitch estimation: high versus low

We consider the TinySOL dataset: audio samples $(\boldsymbol{x}_i^a)_{i=1}^N$ and associated pitches $(p_i)_{i=1}^N$ encoded as MIDI note numbers between $(0 \leq p_i < 127)$, $N = 2913$. For every prompt pair $(\boldsymbol{y}_{\mathrm{up}}^{\mathrm{t}}, \boldsymbol{y}_{\mathrm{down}}^{\mathrm{t}})$, we project the audio embedding of all $\boldsymbol{x}_i^a$'s onto the affine line which contains $\boldsymbol{y}_{\mathrm{up}}^{\mathrm{t}}$ and $\boldsymbol{y}_{\mathrm{down}}^{\mathrm{t}}$. Hence the dot products (see Equation 2):

$$z_i = \left\langle f^{\mathrm{a}}(\boldsymbol{x}_i^{\mathrm{a}}, \boldsymbol{w}^a) - \boldsymbol{y}_{\mathrm{down}}^{\mathrm{t}} \,\middle|\, \boldsymbol{y}_{\mathrm{up}}^{\mathrm{t}} - \boldsymbol{y}_{\mathrm{down}}^{\mathrm{t}} \right\rangle, \quad (4)$$

for every i, $1 \leq i \leq N$. We sort the pitches p_i in ascending order, yielding a sequence of ranks $\boldsymbol{r} = (r_i)_{i=1}^N$: for every i, j, $p_i \leq p_j$ implies $r_i \leq r_j$. Following a well-established convention, we represent unisons with fractional-valued ranks, also known as midranks. Likewise for dot products in CLAP space: for every i, j, $z_i \leq z_j$ implies $\hat{r}_i \leq \hat{r}_j$.

Then, we compute Spearman's rank correlation coefficient between rank sequences $\boldsymbol{r}$ and $\hat{\boldsymbol{r}}$:

$$\rho = \frac{\mathrm{cov}(\boldsymbol{r}, \hat{\boldsymbol{r}})}{\mathrm{std}(\boldsymbol{r})\mathrm{std}(\hat{\boldsymbol{r}})}, \quad (5)$$

ranging between -1 and 1: higher values are better.

Figure 2 summarizes our results. For MS-CLAP, the best affine predictor of $(p_i)_{i=1}^N$ from $(f^{\mathrm{a}}(\boldsymbol{x}_i^{\mathrm{a}}, \boldsymbol{w}^{\mathrm{a}}))_{i=1}^N$, obtained via ordinary least squares, achieves a nearly perfect rank correlation coefficient: $\rho = 0.998$, demonstrating the ability of the audio encoder f^{a} to arrange TinySOL sounds by pitch. This is our "topline", i.e., an upper bound on the score that can be achieved by prompt engineering. Prompts such as "This sound has a *high (low)* pitch", "This is the sound of a *soprano (tenor)*", and other variants yield excellent scores, $0.8 < \rho < 0.9$. Despite the moderate score of "This note has a *sharp (flat)* accidental" ($\rho = 0.544$), roughly on par with "This sound is located *in front of (behind)* my ears" ($\rho = 0.579$), these results suggest that MS-CLAP can be effectively prompted with concepts of high and low in the figurative sense; i.e., high-pitched and low-pitched.

In contrast, probing MS-CLAP with concepts of high and low in the literal sense fails to elicit frequency-elevation mapping (FEM). Prompts such as "This sound source is located *above (below)* me", "This sound source is located *above (below)* my ears", and other variants yield weak and inconsistently signed scores, $-0.5 < \rho < 0.2$.

The situation is different with LAION-CLAP. Our prompt pair from Section 1.3 yields a moderate-to-strong correlation: $\rho = 0.673$, see red arrow on Figure 2. To our knowledge, this is the first published example of an ALM in which the textual embedding for elevation (*above* vs. *below* in the literal sense) correlates positively with the audio embedding for frequency (*high* vs. *low* in the figurative sense). That being said, the text encoder of LAION-CLAP appears highly sensitive to variants. Indeed, replacing the words "my ears" by the words "me", "my head", or "the listener" degrades the score: $0.1 < \rho < 0.5$. This caveat aligns with recent prior work, which reported that LAION-CLAP "lack[s] robustness to linguistic variation in natural language inputs" [29].

Figure 3 allows us to take a closer look at the data, focusing on the prompt pair from Section 1.3: "This sound source is located *above (below)* my ears". By plotting z_i versus p_i for all i in TinySOL, we observe a moderate-to-strong positive correlation for LAION-CLAP (bottom) and little or no correlation for MS-CLAP (top).

3.2 Sound classification: above versus below

We now turn to the latter zero-shot task in our benchmark: namely, binary classification between sound sources that are typically above the listener (e.g., fireworks) versus below the listener (e.g., pouring water into a glass). We use sounds from a subset of the ESC dataset, denoted by $(\boldsymbol{x}_i^{\mathrm{a}})_{i=1}^N$, $N = 160$: see Section 2.2. Each $\boldsymbol{x}_i^{\mathrm{a}}$ is paired with reference class labels $c_i = \pm 1$ where $+1$ and -1 stand for "above" and "below" the listener, respectively. We engineer $\boldsymbol{y}_{\mathrm{up}}^{\mathrm{t}}$ and $\boldsymbol{y}_{\mathrm{down}}^{\mathrm{t}}$ as the prompt pair from Section 3.1: "This sound source is located *above (below)* my ears". Then, for each $\boldsymbol{x}_i^{\mathrm{a}}$, we compute the dot product z_i as in Equation 4 and derive a predicted class label $\hat{c}_i = \mathrm{sign}(z_i)$. Lastly, we report classification accuracy as the proportion of samples for which $\hat{c}_i = c_i$. Since our dataset is balanced (with $\frac{1}{2}N = 80$ samples in each class), tossing a fair coin would yield a baseline classification accuracy of 50%.

Figure 4 summarizes our results. For MS-CLAP, the sign of the dot product z_i is a relatively strong predictor of binary elevation (above versus below), leading to a classification accuracy of 70.6%. That being said, we notice large intra-class variability, particularly for *pouring water*, for which the median z_i is nearly zero. For LAION-CLAP, all dot products z_i are positive, regardless of sound category, and so the classification accuracy for binary elevation is at the baseline level, 50%. Remarkably, sounds from the "above" class ($c_i = 1$) tend to have relatively lower positive values for z_i than sounds from the "below" class. Therefore, raising the classification threshold for z_i tends to make performance even worse than the baseline level. Specifically, *airplane* sounds tend to have a lower z_i than *footsteps* sounds.

We stress that this poor performance is not a deficiency of the audio encoder on its own: by training a support vector machine, we have confirmed that the audio samples from classes "above" and "below" are linearly separable in

[3] In French, the word *aigu* means *acute* in the literal sense and *high-pitched* in the figurative sense. The word *grave* means *heavy* in the literal sense and *low-pitched* (also, *serious*) in the figurative sense.

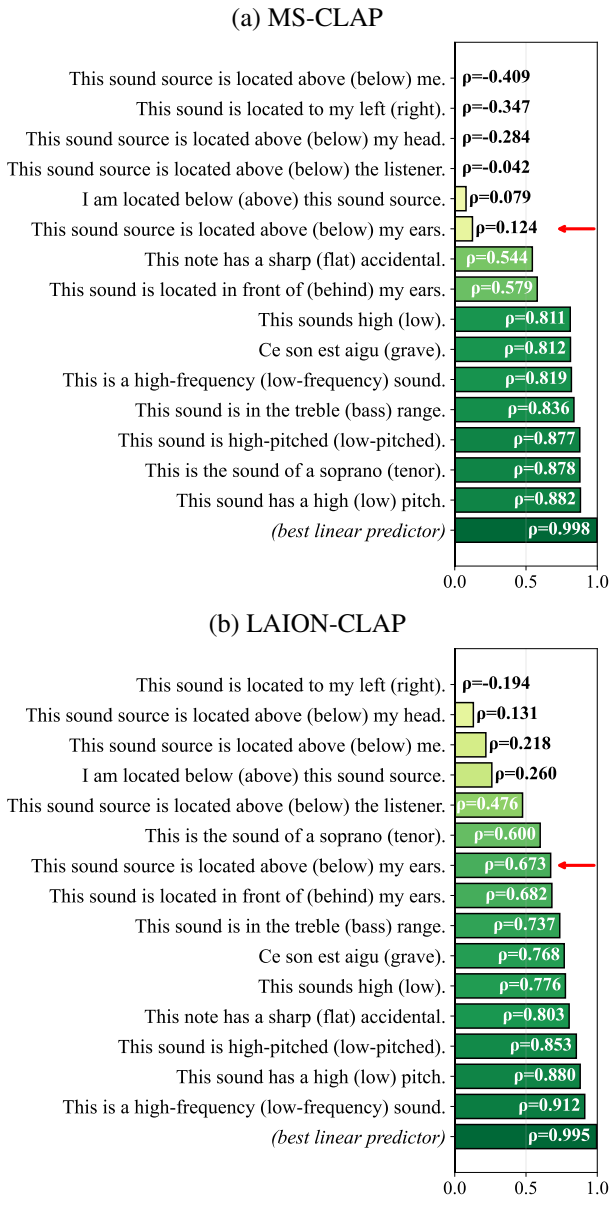


Figure 2. Zero-shot fundamental frequency estimation benchmark on TinySOL with MS-CLAP (a) and LAION-CLAP (b). The red arrow points at the pair of prompts in Section 1.3. See Section 3.1 for details.

terms of LAION-CLAP features. Rather, the issue resides in the interaction between the audio encoder, f^{a}, and the text encoder, f^{t}. Moreover, rephrasing the prompt pair (see Section 2.4 for variants) did not significantly improve binary classification accuracy above chance level.

4. DISCUSSION

Our findings lead us to an unexpected conclusion: when prompting LAION-CLAP, the concept of elevation is operational in the figurative sense but not in the literal sense. Like human listeners, this current-generation ALM tends to describe high-pitched sounds as "located above" and low-pitched sounds as "located below". However, unlike listeners, it lacks common sense for inferring the elevation of sources from monophonic audio: e.g., LAION-CLAP

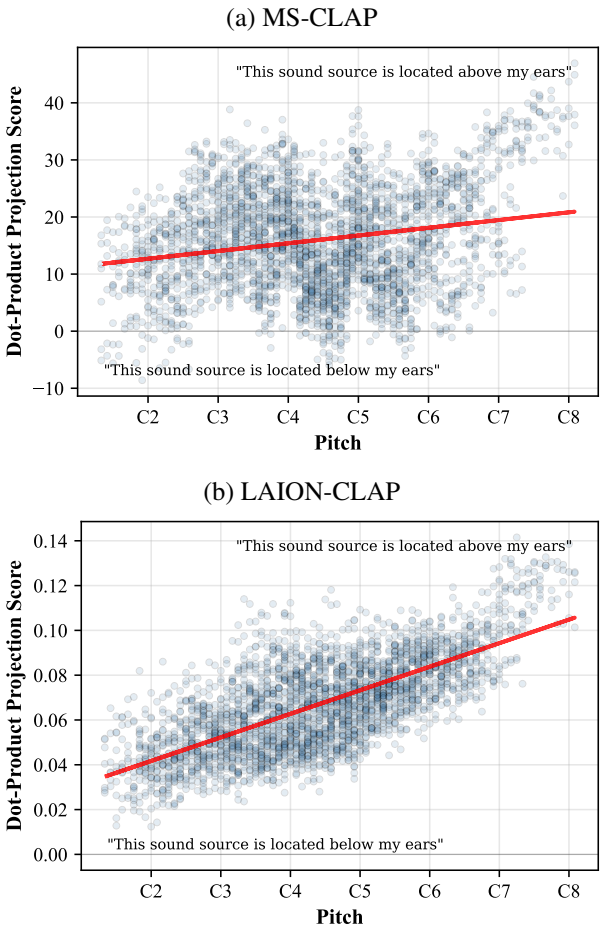


Figure 3. Dot products z_i (see Equation 4) for "This sound is located *above (below)* my ears", encoded with MS-CLAP (top) or LAION-CLAP (bottom). X-axis ticks denote octaves in scientific pitch notation. The red line represents a linear regression. See Section 3.1 for details.

maps the sound of typing on a keyboard as "below" the sound of footsteps and "above" the sound of an airplane. The situation is reversed in the case of MS-CLAP, for which the concept of elevation is operational in the literal sense but not the figurative sense.

Beyond this particular finding, our study points out a potential flaw in the design of evaluation benchmarks for ALM's. The fact that the text encoder f^{t} can be prompted with freely formed English sentences should not mislead us to believe that it necessarily operates on their meaning, defined pragmatically as its whole range of conceivable practical effects. Indeed, the vector $\boldsymbol{y}^{\mathrm{t}} = f^{\mathrm{t}}(\boldsymbol{x}^{\mathrm{t}}, \boldsymbol{w}^{\mathrm{t}})$ may afford all, some, or none of the meaning of $\boldsymbol{x}^{\mathrm{t}}$. Our case study on FEM offers an uncontrived example of audio–language mapping where part of the meaning is retained (e.g., the figurative sense) while another part is lost (the literal sense). To reveal such a nuance, we have taken a step back from cross-modal retrieval at the level of audio–text pairs. Instead, we have conducted comparative surveys on the interaction between a chosen textual prompt and multiple audio corpora, thereby enabling finer technical control on the acoustical features under evaluation.

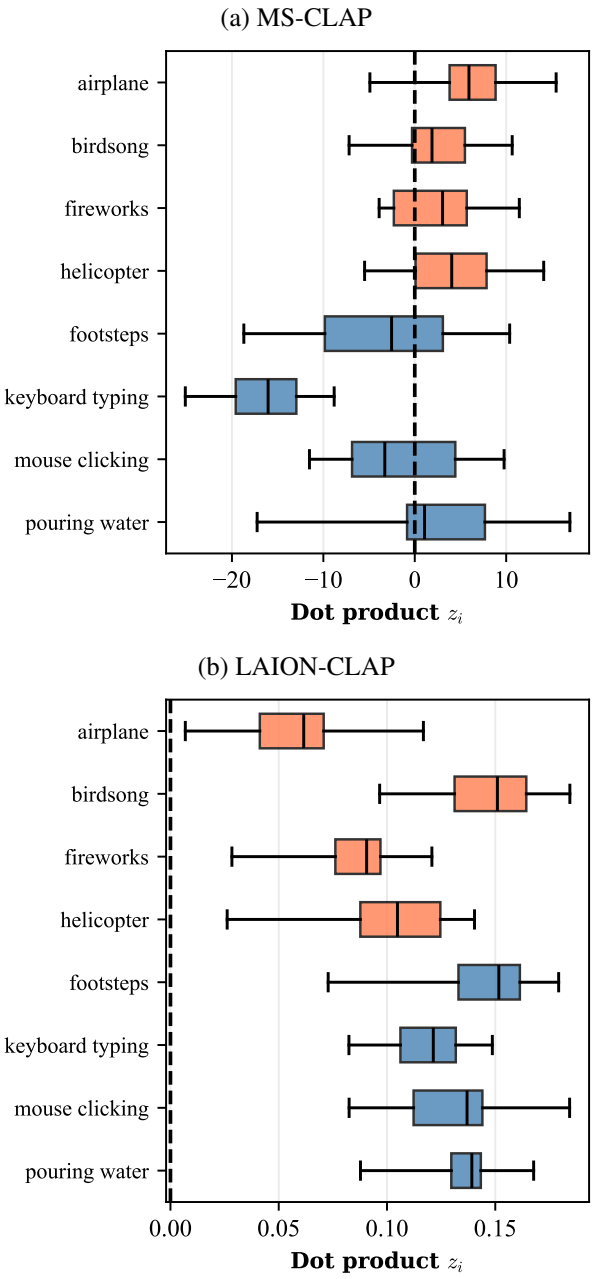


Figure 4. Dot products z_i (see Equation 4) for the ESC subset (see Section 2.2), for MS-CLAP (top) and LAION-CLAP (bottom). Orange and blue denote sound sources which are typically above and below the listener, respectively. Box and whisker ends denote quartiles and extrema respectively. See Section 3.2 for details.

5. CONCLUSIONS

Compared to prior publications on ALM evaluation, we have traded breadth for depth. Starting from a single prompt pair (see Section 1.3) and two instances of the same ALM family (MS-CLAP and LAION-CLAP), we have examined their interaction with audio data in a variety of machine listening settings, both in music and environmental sounds. Moreover, we have analyzed the impact of prompt variants and confirmed that CLAP models are, in some cases, highly sensitive to slight rephrasing of test cases. In this sense, our contribution bears some resemblance with previous research in AI safety, particularly, "red-teaming" language models [30]. Red-teaming can be partially automated and has recently made its way to audio-based question answering systems [31]. Much remains to be done to extend red-teaming methodologies and practices to machine listening at large—this important question is left to future work.

Acknowledgments

We thank Géraldine Carranante for philosophical insights and Zofia Anna Hołubowska for psychoacoustical insights. VL and ML have been funded by the European Union under the MSCA-DN Programme, Grant Agreement No. 101116715 (BioacAI). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible.

6. REFERENCES

[1] Y. Su, J. Bai, Q. Xu, K. Xu, and Y. Dou, "Audio-language models for audio-centric tasks: A survey," *arXiv preprint arXiv:2501.15177*, 2025.

[2] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. IEEE ICASSP*. IEEE, 2023.

[3] X. Mei, G. L. Lan, H. Liu, Z. Ni, V. Nagaraja, Y. Liu, Y. Shi, and V. Chandra, "SLAP: Scalable language–audio pretraining with variable-duration audio and multi-objective training," *Proc. IEEE ICASSP*, 2026.

[4] S. Deshmukh, B. Elizalde, D. Emmanouilidou, B. Raj, R. Singh, and H. Wang, "Training audio captioning models without audio," in *Proc. IEEE ICASSP*. IEEE, 2024.

[5] S. Deshmukh, S. Han, H. Bukhari, B. Elizalde, H. Gamper, R. Singh, and B. Raj, "Audio entailment: Assessing deductive reasoning for audio understanding," in *Proc. AAAI*, vol. 39, no. 22, 2025, pp. 23 769–23 777.

[6] Q. Deng, B. Pardo, and T. N. Pappas, "Do joint language-audio embeddings encode perceptual timbre semantics?" in *Proc. NeurIPS workshop on AI for Music*, 2025.

[7] H.-H. Wu, O. Nieto, J. P. Bello, and J. Salamon, "Audio-text models do not yet leverage natural language," in *Proc. IEEE ICASSP*. IEEE, 2023, pp. 1–5.

[8] H. Lou, H.-y. Paik, W. Hu, and L. Yao, "ParaMETA: Towards learning disentangled paralinguistic speaking styles representations from speech," in *Proc. AAAI*, 2026.

[9] Y. Vasilakis, R. Bittner, and J. Pauwels, "Towards effective negation modeling in joint audio-text models for music," in *Proc. IEEE ICASSP*, 2026.

[10] C. C. Pratt, "The spatial character of high and low tones." *Journal of Experimental Psychology*, vol. 13, no. 3, p. 278, 1930.

[11] C. Spence and O. Deroy, "Crossmodal correspondences: Innate or learned?" *i-Perception*, vol. 3, no. 5, pp. 316–318, 2012.

[12] C. V. Parise, K. Knorre, and M. O. Ernst, "Natural auditory scene statistics shapes human spatial hearing," *Proceedings of the National Academy of Sciences*, vol. 111, no. 16, pp. 6104–6108, 2014.

[13] P. Walker, J. G. Bremner, M. Lunghi, S. Dolscheid, B. D. Barba, and F. Simion, "Newborns are sensitive to the correspondence between auditory pitch and visuospatial elevation," *Developmental Psychobiology*, vol. 60, no. 2, pp. 216–223, 2018.

[14] C. Lorenzi, F. Apoux, E. Grinfeder, B. Krause, N. Miller-Viacava, and J. Sueur, "Human auditory ecology: Extending hearing research to the perception of natural soundscapes by humans in rapidly changing environments," *Trends in Hearing*, vol. 27, p. 23312165231212032, 2023.

[15] M.-E. Duchez, "La repŕesentation spatio-verticale du caractère musical grave-aigu et l'élaboration de la notion de hauteur de son dans la conscience musicale occidentale," *Acta musicologica*, vol. 51, no. Fasc. 1, pp. 54–73, 1979.

[16] M. Olvera, C. Wang, P. Stamatiadis, G. Richard, and S. Essid, "iknow-audio: Integrating knowledge graphs with audio-language models," in *Proc. EMNLP*, 2025, pp. 34 671–34 688.

[17] C. E. Cella, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "Orchideasol: a dataset of extended instrumental techniques for computer-aided orchestration," *arXiv preprint arXiv:2007.00763*, 2020.

[18] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM-MM*, 2015, pp. 1015–1018.

[19] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "HuggingFace's Transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[21] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proc. NAACL-HLT*, 2019.

[22] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," in *Proc. IEEE ICASSP*, 2021.

[23] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE ICASSP*, 2020.

[24] M. Martin *et al.*, "Macs: A multi-annotator captioning dataset for audio and music understanding," in *Proc. IEEE ICASSP*, 2024.

[25] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. IEEE ICASSP*, 2024.

[26] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE ICASSP*. IEEE, 2023.

[27] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. IEEE ICASSP*. IEEE, 2022, pp. 646–650.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[29] R. Selvakumar, S. Kumar, H. K. Giri, N. Anand, A. Seth, S. Ghosh, and D. Manocha, "Do audio-language models understand linguistic variations?" in *Proc. NAACL*, 2025, pp. 899–913.

[30] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," in *Proc. EMNLP*, 2022, pp. 3419–3448.

[31] H. Yang, L. Qu, E. Shareghi, and G. Haffari, "Audio is the achilles' heel: Red teaming audio large multimodal models," in *Proc. ACL*, 2025, pp. 9292–9306.