BiSAID: Bipolar Semantic Adjectives Icons and Earcons Dataset

Zijing Cao (up202001544@fe.up.pt) (0009-0000-0214-4436)[1],
António S. Pinto (asapinto@fe.up.pt) (0000-0003-1629-8385)[1], and
Gilberto Bernardes (gba@fe.up.pt) (0000-0003-3884-2687)[1]

[1]**University of Porto – Faculty of Engineering & INESC TEC**, Porto, Portugal

ABSTRACT

This paper presents BiSAID, a dataset for exploring bipolar semantic adjectives in non-speech auditory cues, including *earcons* and *auditory icons*, i.e., sounds used to signify specific events or relay information in auditory interfaces from recorded or synthetic sources, respectively. In total, our dataset includes 599 non-speech auditory cues with different semantic labels, covering temperature (cold vs. warm), brightness (bright vs. dark), sharpness (sharp vs. dull), shape (curved vs. flat), and accuracy (correct vs. incorrect). Furthermore, we advance a preliminary analysis of brightness and accuracy earcon pairs from the BiSAID dataset to infer idiosyncratic sonic structures of each semantic earcon label from 66 instantaneous low- and mid-level descriptors, covering temporal, spectral, rhythmic, and tonal descriptors. Ultimately, we aim to unveil the relationship between sonic parameters behind earcon design, thus systematizing their structural foundations and shedding light on the metaphorical semantic nature of their description. This exploration revealed that spectral characteristics (e.g. spectral flux and spectral complexity) serve as the most relevant acoustic correlates in differentiating earcons on the dimensions of brightness and accuracy, respectively. The methodology holds great promise for systematizing earcon design and generating hypotheses for in-depth perceptual studies.

1. INTRODUCTION

In sound-related fields, the formal discourse by professional sound designers and musicians often use sensory metaphors as a practical language to describe the attributes of sound. They draw from a myriad fields, such as visual and haptic, encompassing semantic adjectives such as high, low, deep, warm, cold, bright, dark, rough, smooth, big, small, soft, edgy, round, flat, sharp, dull, transparent, translucent, shimmering, sweet, and colorful. These metaphors play a crucial role in shaping the *meaning* of sonic objects and narratives through their crossmodal relationships. For instance, in film sound design, weapons like swords, knives, or daggers are consistently accompanied by high-pitched, sharp sounds, aligning with their semantic characteristics, even if they do not precisely mimic the physical reality [1].

The discourse on the communicative potential of sound has recently gained traction in the field of auditory displays, namely on what concerns non-verbal auditory cues, which can be categorized into two primary types: auditory icons and earcons. Auditory icons consist of recorded non-speech sounds that can be recognized from real-world experiences. For instance, the sound of crumpling paper playing when deleting a digital document [2]. Earcons, on the other hand, are "audio messages used in the user-computer interface to provide information and feedback to the user about computer entities" [3]. Onboard an aircraft, short, structured musical messages can be used as alarms when a passenger requests assistance.

Auditory icons and earcons are instrumental in sound design across mediums like film, games, and human-machine interfaces, effectively conveying specific information to users. In gaming scenarios, for instance, they play a vital role in alerting players to significant events [3–6]. To fully harness the power of non-speech auditory cues, and most notably the synthetically-designed earcons resulting from a greater domain of creative freedom and sonic signifiers, designers must align these cues with their intended meanings. Despite a common understanding and shared recognition of the structural attributes inherent to each semantic adjective among designers across different cultures, there are no established guidelines or systematic analyses of their component structures.

Understanding the sonic attributes inherent to the earcon's design can shed light on human perception, psychology, and communication science by means of cross-modal metaphors, but also in guiding the design principles for optimal communication with sound. For example, the problem of creating an alert sound to signal errors must have a sonic identity that intuitively communicates the information to users.

Research into the structural attributes of semantic adjectives has predominantly utilized music perception methods, focusing on identifying significant attributes through perceptual listening tests. Investigations into the relation between sonic descriptors and the perception of musical instrument sounds have been conducted by various researchers. Disley and Howard [7] explored the timbres of pipe organs, while Bernays and Traube [8] delved into piano tones. Rosi et al. [9] analyzed the meaning of metaphorical sound attributes based on interviews with sound professionals. These studies, despite their differ-

ences, identified common descriptors such as "bright", "warm", and "round", highlighting consistent perception signifiers of musical timbres. A different line of inquiry has examined the impact of audio descriptors on semantic annotations. Studies have shown that the brightness of a sound is associated with its spectral centroid, attack time, and timbral sharpness. Following this approach, Ilkowska and Miskiewicz [10] delved into the effects of spectral variations on the perception of sound. By introducing artificial formants to both noise and music spectra and applying psychophysical scaling to assess their relationship, their research demonstrates that vowel frequency plays a pivotal role in altering perceptions of sharpness and brightness. This influence surpasses that of vowel bandwidth and amplitude, underscoring the critical impact of frequency on the auditory perception of semantic qualities.

Concerning the analysis of earcon-specific attributes, there is a lack of literature on assessing their intrinsic structure. To this end, we present an empirical analysis of bipolar [1] semantic adjectives in earcon design, as a first step towards an in-depth (perceptual) study to unveil the foundational sonic attribute of semantic adjectives. In detail, we infer common attributes from a newly complied dataset of earcons from *weak* user-driven tags, named BiSAID. It includes 599 sound samples for bipolar semantic adjectives covering 10 metaphors in different domains, e.g., temperature (cold vs. warm), and brightness (bright vs. dark). Samples were collected from the crowd-sourced Freesound platform. Samples were annotated with 66 instantaneous and global audio descriptors capturing spectral, temporal, spectrotemporal, and energetic information.

To evaluate the quality of compiled data and initiate the exploration of the sonic attributes of earcon design, we employ statistical and machine learning methods to deduce instantaneous earcon descriptors that effectively differentiate between bipolar adjectives within specific metaphorical domains. Our hypothesis posits that characteristic sonic dimensions, such as pitch, onset density, and spectral bandwidth, among others, are inherent to the sonic structure of earcons corresponding to various metaphors.

We present two detailed case studies. Firstly, we examine which descriptors in the BiSAID dataset are most pertinent for distinguishing between bipolar adjectives such as bright versus dark, and correct versus incorrect. The former represents a well-established baseline case, extensively discussed in existing literature, which should unveil the adequacy of our data in aligning with perceptual results and enlighten the limitations of the current sound collection. The latter introduces a novel metaphor that has received limited attention thus far. Treating each adjective as a distinct sample or class, we statistically infer significant differences between the two sample sets and determine their relative importance through machine learning models.

The reminder of the paper is structured as follows. Section 2 describes the BiSAID dataset structure and contents.

[1] This type of methodology on sound perception was first introduced by von Bismarck [11], based on the semantic differential technique, pioneered by Osgood [12], a method for measuring the meaning of concepts by eliciting ratings on bipolar adjective scales.

Table 1. Dataset composition: earcons, semantic labels, and categories.

Dimension	Adjectives	Category				#
		Earcon	**Icon**	**Other**	Exc.[a]	
Accuracy	correct	125	8	19	248	400
	incorrect	48	13	8	44	113
Brightness	bright	49	0	0	236	285
	dark	47	1	5	202	255
Shape	curved	43	0	30	327	400
	flat	27	0	10	308	345
Sharpness	sharp	46	54	69	231	400
	dull	14	49	32	202	297
Temperature	warm	55	0	59	286	400
	cold	0	20	82	298	400
		454	145	314	2382	3295

[a] Excluded due to erroneous tag.

Section 3.1 provides the description of two case studies on earcon semantic-driven descriptors for brightness (bright versus dark) and accuracy (correct versus incorrect). Section 4 presents preliminary results of the earcon data analysis and discusses the potential and limitation of our data. Finally, Section 5 presents the conclusions of our work and directions for future research.

2. BISAID: A DATASET OF EARCONS AND AUDITORY ICONS WITH SEMANTIC DIFFERENTIAL TAGGING AND ACOUSTIC DESCRIPTION

The BiSAID dataset has been meticulously classified and tagged to encompass the following semantic dimensions: temperature (cold vs. warm), brightness (bright vs. dark), sharpness (sharp vs. dull), shape (curved vs. flat), and accuracy (correct vs. incorrect). Each data point is annotated with semantic differential tags and described through a rich set of acoustic and psychoacoustic descriptors. These earcons and tags are summarized in Table 1, while the descriptors, comprising both global and instantaneous types, are detailed in the remaining of the current section.

The earcons within BiSAID were curated from Freesound [2] [13]. We aimed to collect up to 400 sounds for each tag, classifying them into three categories: *Earcons*, which are short musical sounds designed to convey specific information; *Auditory icons*, brief sounds that mimic familiar non-speech sounds from everyday life; and *Other*, which includes sounds relevant to the tag but not categorized as earcons or auditory icons. In the data cleansing phase, 2382 sounds were excluded due to inaccurate tagging. This step involved listening to each sound to assess if the tags accurately reflected the content, in terms of perception of sound. Despite the crowd-sourced nature of Freesound's tagging system and the resulting variability [14], the objective was to ensure the integrity of tags rather than to standardize their meanings. For example, a sound (id:574006) tagged as "correct" that

[2] `https://freesound.org`, last accessed on 27 February, 2024.

was unrelated to the concept of correctness was removed, as well as a sound (id:195219) tagged as "bright" which descriptors a male voice singing *Monty Python's* "Always Look on the Bright Side of Life"—a clear mismatch with the expected perception of brightness. The final dataset includes 3295 sounds, with 599 specifically identified as earcons or auditory icons.

For the computational analysis of earcons in the BiSAID dataset, we employed the Freesound API extractor, grounded in `Essentia` [15]. Initially, this tool computes a foundational set of 91 [3] primary low- and mid-level descriptors, encompassing a broad spectrum of acoustic and psychoacoustic properties. These foundational descriptors capture diverse acoustic dimensions, including temporal aspects (e.g., zero-crossing rate, effective duration), spectral characteristics (e.g., spectral centroid, spectral spread), spectro-temporal descriptors (e.g., spectral flux, the first derivative of spectral contrast), rhythmic elements (e.g., beats count, onset rate), and tonal attributes (e.g., tuning frequency, chord histogram).

Regarding the temporal analysis scale, we can group these descriptors in global descriptors, which synthesize characteristics over the entire sound duration (e.g., log-attack time, tristimulus), and instantaneous descriptors, calculated for discrete frames within the sound to provide a momentary acoustic snapshot (e.g., spectral rms, pitch salience). Building upon the initial set, the Freesound API further computes statistical aggregators for all instantaneous descriptors, thus expanding the dataset to encompass a total of 452 descriptors. These aggregators include the maximum, minimum, mean, variance, median, and the first and second derivatives, significantly enriching the dataset's descriptive capacity while capturing the acoustic temporal profile of each sound. Notably, approximately 20% of these descriptors are vector-based structures, such as the Equivalent Rectangular Bandwidth (ERB) [16, 17] bands or the Gammatone Frequency Cepstral Coefficients (GFCC), offering multidimensional insights into the sound's acoustic and psychoacoustic profile.

BiSAID is made available in several formats to cater to both manual and computational exploration. This encompasses Excel (`xlsx`) files, which aggregate all sound-related information by dimension, and CSV and JSON files that present audio descriptors alongside semantic labels. The content of these files is curated to include the sound's `id`, `name`, `sound` (a hyperlink for previewing the sound), and `tags` (the full list of tags associated with the sound). After an auditive inspection, additional annotations are made to document the `category` of the sound and any pertinent `note` that offers insight into its categorization. The dataset, distributed under a Creative Commons license, is accessible at `https://figshare.com/articles/media/BISAID/25377589`. To promote reproducibility and transparency, we further provide the scripts utilized in the data collection and analysis phases, ensuring that researchers can readily replicate or extend the work presented herein. The scripts can be found in the paper repository at `https://github.com/ZijingCaoo/sound_analysis`.

[3] For a detailed description of these descriptors, please consult `https://freesound.org/docs/api/analysis_docs.html`.

3. A STUDY ON THE RELATIONSHIP BETWEEN SOUND PARAMETERS OF EARCON AND SEMANTIC DESCRIPTORS

This section presents a preliminary exploration of the BiSAID dataset, emphasizing the semantic dimensions of brightness (bright vs. dark) and accuracy (correct vs. incorrect). Our goal is to uncover the relationship between sound parameters and semantic descriptors, shedding light on the structural underpinnings and metaphorical semantic qualities of earcons. The focus on earcons, in preference to auditory icons, stems from their inherent design process. Unlike auditory icons, which are based on sampled sounds from the environment, earcons are synthesized creations, meticulously crafted by sound designers to convey specific semantic meanings. This deliberate design process allows for a deeper investigation into how certain audio descriptors are intentionally manipulated to represent different semantic labels, offering a more controlled environment to study the precise relationship between sound parameters and their perceived meanings. Moreover, this investigation evaluates the dataset's quality and the data collection process, setting the groundwork for future research and the next phases of our study.

The choice of brightness and accuracy for our study is twofold: the well-documented discussion of brightness in music perception literature [7, 18, 19], and the relatively uncharted territory of the accuracy dimension, critical in sound design for indicating correct or incorrect actions [20, 21].

3.1 Method

To address our objectives, we began with two preliminary steps regarding earcon and descriptor selection. To prevent imbalances, we ensured an equal representation of earcons for each tag, selecting 47 earcons per tag, based on the maximum common count across descriptors for both pairs of correct / incorrect and bright / dark. In terms of descriptors, we aimed for a manageable and interpretable descriptor space, focusing on one-dimensional numerical descriptors and excluding vector-based and non-numeric attributes. Only the mean values of instantaneous descriptors were retained, setting aside other statistical aggregates to streamline our analysis. Additionally, we removed descriptors with constant values across all earcons, as they do not aid in our analysis. An example is the second-peak-spread (`rhythm.second_peak_spread.mean`, according to Freesound original naming) descriptor in the accuracy case study, which showed no variability and was thus excluded.

After establishing the groundwork, our analysis proceeds with three key stages aimed at clarifying how earcon parameters relate to semantic descriptors: statistical analysis of descriptors, descriptor importance ranking, and correlation analysis.

First, we assessed the normality of the distribution of these descriptors using the Shapiro-Wilk test [22]. The

test indicated that approximately 80% of the descriptors deviated from a normal distribution (more precisely, 84% for the accuracy case, and 74% for the brightness case), thus requiring the use of non-parametric tests for further statistical analysis. Consequently, we employed the Mann-Whitney U test [23] [4] to evaluate the statistical significance of differences between groups defined by semantic labels, ensuring the robustness of the descriptor selection process against non-normal distribution patterns. We used a p-value threshold of 0.001 to ensure that the observed differences in medians between the two groups were highly unlikely to have occurred by random chance.

The second phase of our analysis involved the use of two machine learning models, *logistic regression* and *random forests*, to rank the descriptors in terms of the predictive power of the semantic labels. This approach under two alternative machine learning models allowed the identification of both linear and non-linear relationships between the descriptors and the labels, and gave us an idea of the more important descriptors.

The convergence of statistical significance, and model-based predictive power led to a focused subset of descriptors for deeper analysis.

4. SOUND ANALYSIS AND RESULTS

In this section, we examine the correlations between earcon descriptors and semantic labels within two semantic dimensions: *accuracy* (contrasting "correct" vs. "incorrect" bipolar labels) and *brightness* (contrasting "bright" vs. "dark"). Each dimension's analysis follows a structured approach: initially presenting descriptor importance as determined by our computational models, with particular emphasis on descriptors that are statistically significant and hold high importance in both models. Subsequently, we analyze the distributions of these descriptors, providing visual comparisons to elucidate how they vary between semantic contrasts.

4.1 Case I: *Brightness*

Figure 1 presents the ranking of descriptors according to their predictive power for semantic earcon labels of brightness, differentiating between "bright" and "dark" labels. Descriptors that showed statistical significance, as determined by the Mann-Whitney U test with $p < 0.001$, are highlighted with an asterisk next to their labels.

The analysis reveals a notable alignment between the most statistically significant audio descriptors and the random forest method, suggesting the data's potentially non-linear characteristics. However, our focus remains on those descriptors that not only show high statistical significance but also are crucial in the results produced by both predictive models (random forest and logistic regression). Identified as highly relevant in both analyses and marked in dark blue in Figure 1, these descriptors predominantly pertain to the spectral domain

[4] A non-parametric method for comparing differences between two independent groups in instances where the dependent variable is ordinal or continuous yet does not adhere to a normal distribution.

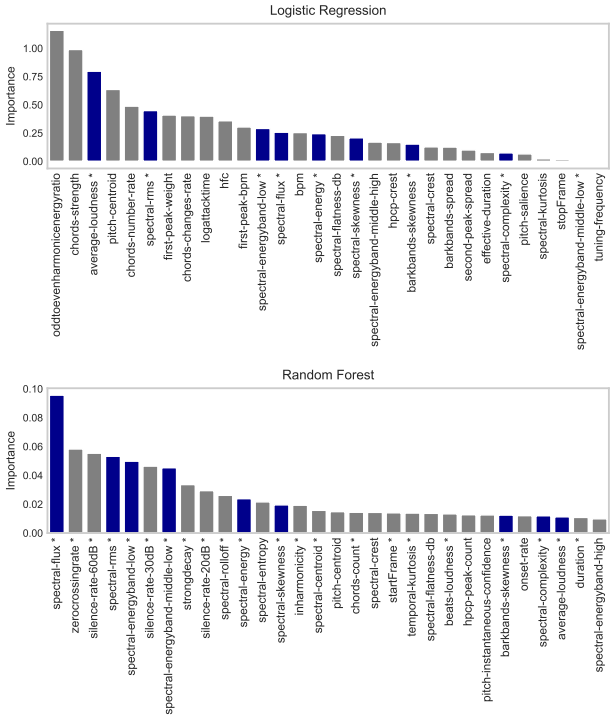

Figure 1. Descriptor importance for semantic label prediction for the *brightness* dimension. Asterisks (*) denote statistically significant descriptors; (*grey*) descriptors highlight high import descriptors in the corresponding model; and (*blue*) descriptors high importance in both models.

(such as *spectral-rms*, *spectral-energyband-low*, *spectral-energyband-middle-low*, *spectral-skewness*, *barkbands-skewness*, *spectral-complexity*, and *spectral-flux*), with the exception of *average-loudness*, which operates in the temporal domain. A significant number of these descriptors are associated with the sound's energy.

Due to the absence of energy normalization in our methodology, we refrain from drawing definitive conclusions about certain descriptors (*spectral-rms*, *spectral-energyband-low*, *spectral-energyband-middle-low*, *spectral-energy*, *average-loudness*). It is noted that "dark" distributions consistently exhibit higher values than "bright" ones in aspects like *average-loudness* and *spectral energy*, particularly at the lower end of the spectrum (*spectral-energyband-low*). This observation aligns with the intuitive association of "darker" sounds, in terms of both luminescence and mood (e.g., "scary"), with greater energy in the lower frequency range and increased loudness.

Given the high correlation between *spectral-skewness* and *barkbands-skewness* (differing only in the scale used for computation), we opt to exclude only the latter from further analysis, as shown in Figure 2, focusing on the former due to its direct relevance. *Spectral-skewness* quantifies the asymmetry of the spectrum's distribution around its mean, indicating more energy on the right-hand side of the distribution at lower values, and conversely, more energy on the left side suggests a higher skewness value.

This descriptor, therefore, is crucial for understanding the distribution of energy across the frequency spectrum.

In addition to *spectral-skewness*, other descriptors under consideration include *spectral-complexity*, which counts the number of peaks in the spectrum between 100Hz and 5KHz, and *spectral-flux*, which gauges the changes in the frequency content over time. The latter uses either the L2- or L1-norm difference between consecutive frames of the magnitude spectrum, offering insights into how the spectral content evolves. These descriptors collectively provide a multifaceted view of the sound's spectral characteristics, illuminating different aspects of how sounds are perceived as either "bright" or "dark".

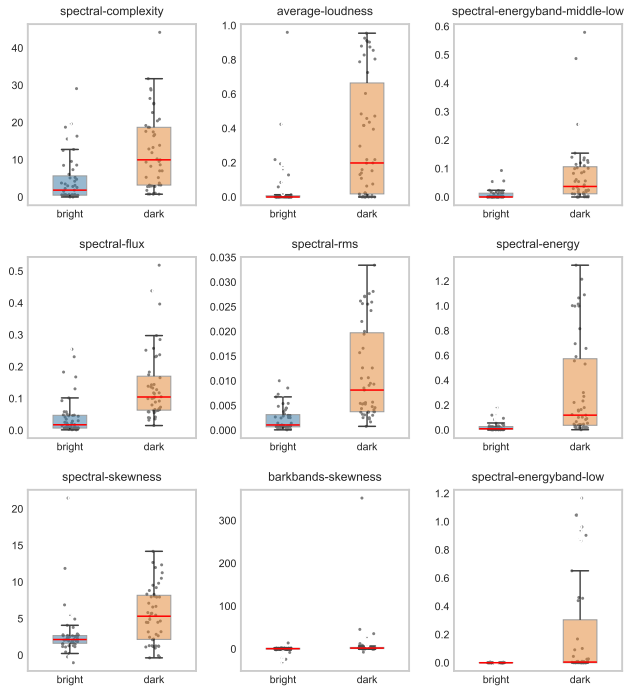

Figure 2. Distribution analysis of key acoustic descriptors for the dimension of *brightness*, contrasting "bright" vs. "dark" semantic labels.

Attempting to interpret the data, the higher skewness values in dark earcons compared to bright ones suggest a trend where the spectral content of "dark" sounds is skewed towards lower frequencies. The broader range of mean values for "dark" instances may imply less consistency in the spectral centroid, hinting at an intuitive connection where "dark" sounds are associated with lower-frequency energy. This observation, while not conclusive, resonates with a common perception that links auditory darkness with a prevalence of lower frequencies.

In the analysis of *spectral-complexity*, dark earcons show a greater complexity than their bright counterparts, particularly in sounds tagged as "scary" or "creepy" on Freesound. This increased complexity might intuitively suggest that "dark" sounds encompass more intricate spectral textures and a higher level of dissonance from overlapping spectral peaks. While this insight does not serve as a definitive explanation, it opens up a pathway for intuitive speculation on how complexity in the spectral domain could influence our perception of sounds as being "dark" or ominous.

Regarding *spectral-flux*, the observation that dark earcons exhibit a higher flux raises speculative thoughts on the nature of these sounds. The higher spectral flux indicates more variability in the spectrum over time, which might intuitively be linked to the presence of elements such as low-frequency rumbles, atonal textures, and noise. This characteristic of "dark" sounds, suggesting a more dynamic spectral evolution, allows for speculative reflection rather than firm conclusions. It hints at how such auditory features could be perceived as "dark" aligning with a common intuition about the complexity and variability in sounds labeled as such.

4.2 Case II: *Accuracy*

Figure 3 delineates the hierarchy of descriptors based on their efficacy in predicting semantic earcon labels concerning accuracy, particularly distinguishing between "correct" and "incorrect" labels. Descriptors achieving statistical significance, as verified through the Mann-Whitney U test with $p < 0.001$, are distinguished by an asterisk beside their names.

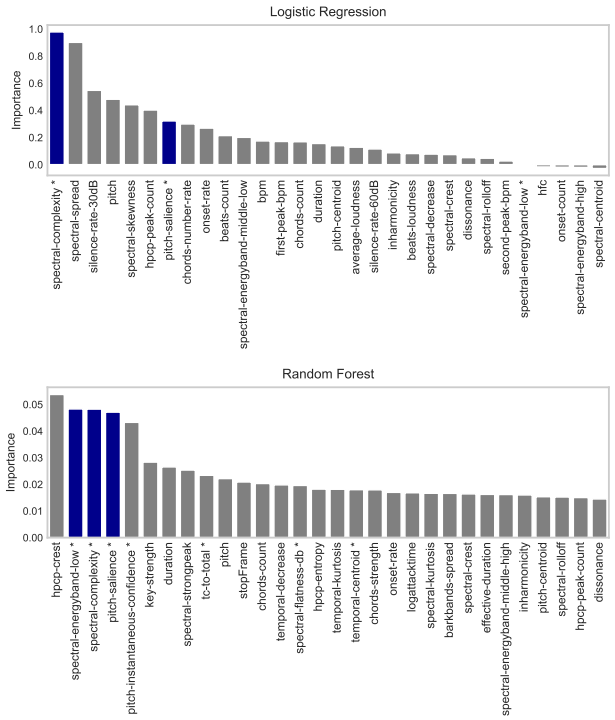

Figure 3. Descriptor importance for semantic label prediction for the *accuracy* dimension. Asterisks (*) denote statistically significant descriptors; (*grey*) descriptors highlight high import descriptors in the corresponding model; and (*blue*) descriptors high importance in both models.

Mirroring the analytical approach of the previous case, we concentrate on the intersection of highly significant descriptors as identified by both logistic regression and

random forest models, illustrated in dark blue in Figure 3. From this convergence, three audio descriptors stand out: *spectral-complexity*, *pitch-salience*, and *spectral-energyband-low*. Descriptive statistics for these salient descriptors are depicted in Figure 4. Preliminary scrutiny of *spectral-energyband-low* indicates a negligible perceptual disparity amidst subtle variations.

The data distribution reveals an uptick in *spectral-complexity* for earcons labeled as incorrect relative to those deemed correct. Closer inspection of the earcons across our dataset exposed clear timbral distinctions between the two groups. Incorrect earcons typically embodied a multitude of percussive sounds characterized by densely populated spectral bands, whereas correct earcons were more likely to manifest as pitched tones within harmonic spectral frameworks. Thus, we may infer that the "incorrect" earcons are typified by a more intricate spectral profile, with a higher density of spectral peaks and a less prominent fundamental frequency.

This narrative of *spectral-complexity* is corroborated by the pitch salience assessments; incorrect earcons are marked by higher *pitch-salience* values. *Pitch-salience* is calculated as the quotient of the spectrum's highest autocorrelation peak over the unshifted autocorrelation baseline, reflecting the prominence of a tone. Sounds defined by pure tones or limited harmonics tend to register lower *pitch-salience* readings, tending towards zero, as noted in correct earcons. In contrast, the presence of multiple harmonics in the spectrum of incorrect earcons results in increased *pitch-salience*, indicative of their intricate harmonic structures.

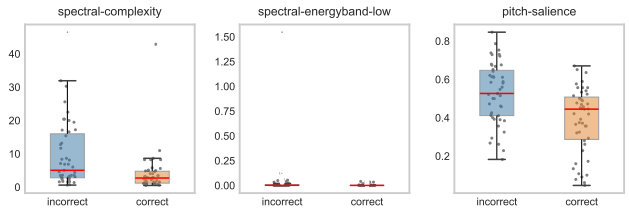

Figure 4. Distribution analysis of key acoustic descriptors for the dimension of *brightness*, contrasting "bright" vs. "dark" semantic labels.

4.3 Discussion

This exploratory study has highlighted several key points regarding the use of audio descriptors for semantic description of sound.

The examination of audio descriptors in relation to semantic bipolar concepts such as "bright" versus "dark" and "correct" versus "incorrect" has demonstrated the potential of descriptors like *spectral-complexity* and *spectral-skewness* to reflect auditory qualities ascribed to these terms. However, the application of these descriptors is not straightforward due to the subjective nature of sound perception. The use of the descriptor "dark" raises issues concerning the alignment between user-generated tags and the a shared perception they are intended to characterize. Literature suggests that subjective tagging can introduce inconsistencies [14], that require further methodological treatment. Bipolar scales can compound this problem by presenting a false dichotomy that may not mirror the auditory experience, as discussed by Kendall and Carterette [18]. An alternative approach using unipolar scales, like the VAME method, avoids such binary oppositions by considering attributes on a continuum, which might better represent the spectrum of auditory qualities.

While the use of crowd-sourced tags offers controlled semantic judgment, it also requires a cautious approach due to its inherent limitations [24]. This necessitates more rigorous data collection methods to ensure the validity of earcon descriptors used in research.

In the context of accuracy, descriptors that might typically indicate sound correctness, such as dissonance or roughness, were not prominent in the dataset, suggesting a gap that warrants further exploration. However, the prominence of *pitch-salience* and *spectral-complexity* could be indirectly linked to these concepts, meriting a future intercorrelation analysis.

The dataset's descriptors showed great deviation from a normal distribution, indicating a complex set of factors at play in computational-based sound perception studies. These include the properties of the sounds or the recording conditions, to mention a few. Such atypical distributions in the descriptors necessitate meticulous outlier management and data transformation in preprocessing to ensure data quality, so not to introduce noise in subsequent analysis. This is especially true when using average-based temporal aggregation methods, which are more vulnerable to outlier effects than those based on median aggregation.

To refine the descriptor extraction process, employing tools like Essentia [5] is recommended. Essentia provides enhanced control over the analysis parameters, allowing for a more robust and controlled extraction process. Utilizing such a tool could improve the precision of sound data analysis, leading to more reliable and significant research findings.

5. CONCLUSIONS AND FUTURE WORK

This study introduces a dataset comprising 1053 non-speech auditory cues, including earcons and auditory icons, each annotated with semantic bipolar adjectives and detailed through 452 global and instantaneous audio descriptors. Our focus was on analyzing a subset of 192 earcons, labeled with contrasting pairs "correct" and "incorrect," and "bright" and "dark," to examine the relationship between earcon parameters and their semantic meanings.

The analysis of semantic earcon labels for brightness differentiation unveils important audio descriptors contributing to perceptual differentiation. Through logistic regression and random forest methods, statistically significant descriptors such as energy-related descriptors, spectral skewness, complexity, and flux are identified. Notably, brighter earcons exhibit lower energy levels, while

[5] https://essentia.upf.edu/

darker ones display higher positive skewness, reflecting timbral characteristics akin to pitched instruments. Spectral complexity analysis reveals intricate frequency components in darker earcons, evident in dense spectral patterns and sensory dissonance. Additionally, spectral flux analysis demonstrates temporal variations, with darker earcons showing greater spectral content variability. These findings offer insights into the nuanced perceptual attributes of earcons, informing future auditory design considerations.

From the analysis of the accuracy study case, featuring correct and incorrect earcons and similar methodology, three key audio descriptors have been highlighted: spectral complexity, pitch salience, and spectral energy band low. Descriptive statistics for these descriptors uncovered higher spectral complexity in incorrect earcons, characterized by dense percussive sounds, in contrast to correct earcons featuring harmonic tones. This finding was reinforced by pitch salience analysis, which indicated variations in harmonic content between correct and incorrect earcons.

Looking ahead, our research paves the way for several avenues in future work. We anticipate the integration of additional models or ensemble methods, such as Gradient Boosting Machines (GBM) [25], to further substantiate and refine the descriptor importance rankings, particularly for capturing complex non-linear relationships. The expansion of our dataset, incorporating a wider variety of sounds selected through a more stringent methodology, stands as a priority. This will enable us to extend our analysis to additional dimensions of earcon design, ultimately extracting actionable insights for sound designers.

Moreover, the true test of our findings lies in their perceptual validation. We propose a perceptual study to confirm the predictive power of the identified descriptors in real-world settings, ensuring that our statistical and machine learning outcomes resonate with actual perceptual differences among listeners. Complementary to this, controlled listening tests with participants will allow us to assess their perception of brightness and accuracy in earcons, aligning subjective assessments with the quantitative data from our analysis.

Through this interconnected approach, we aim to bridge the gap between the theoretical underpinnings of earcon parameters and their practical design applications, fostering a deeper understanding and more intuitive design of auditory signals.

6. ACKNOWLEDGEMENT

The first author would like to thank China Scholarship Council (CSC)[6] for financial support (Grant No.202307920001).

7. REFERENCES

[1] T. Görne, "The emotional impact of sound: A short theory of film sound design," in *EPiC Series in Technology*, 2019, pp. 17–30.

[2] T. Hermann, A. Hunt, and J. G. Neuhoff, *The sonification handbook*. Logos Verlag Berlin, 2011.

[3] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and icons: Their structure and common design principles," *Human-Computer Interaction*, vol. 4, no. 1, pp. 11–44, 1989.

[4] R. D. Patterson, J. Edworthy, and M. C. Lower, *Alarm sounds for medical equipment in intensive care areas and operating theatres*. University of Southampton, Institute of Sound and Vibration Research, 1986.

[5] S. Brewster, V.-P. Raty, and A. Kortekangas, "Earcons as a Method of Providing Navigational Cues in a Menu Hierarchy," *People and Computers XI*, no. August, pp. 169–183, 1996.

[6] G. Leplâtre and S. A. Brewster, "Designing Non-Speech Sounds to Support Navigation in Mobile Phone Menus," in *Proceedings of the International Conference on Auditory Display (ICAD)*. University of Glasgow (United Kingdom), 2000, pp. 190–199.

[7] A. C. Disley and D. M. Howard, "Spectral correlates of timbral semantics relating to the pipe organ," *Speech, Music and Hearing*, vol. 46, pp. 25–39, 2004.

[8] M. Bernays and C. Traube, "Verbal expression of piano timbre: Multidimensional semantic space of adjectival descriptors," in *International Symposium on Performance Science*, no. July, 2011, pp. 299–304.

[9] V. Rosi, O. Houix, N. Misdariis, and P. Susini, "Investigating the Shared Meaning of Metaphorical Sound Attributes," *Music Perception*, vol. 39, no. 5, pp. 468–483, 2022.

[10] M. Ilkowska and A. Miśkiewicz, "Sharpness versus brightness: A comparison of magnitude estimates," *Acta Acustica united with Acustica*, vol. 92, no. 5, pp. 812–819, 2006.

[11] G. von Bismarck, "Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes," *Acustica*, vol. 30, pp. 146–159, 1974.

[12] C. E. Osgood, G. J. Suci, and Percy H. Tannenbaum, *The Measurement of Meaning*. University of Illinois Press, 1957.

[13] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, 2013, pp. 411–412.

[14] F. Font and X. Serra, "Analysis of the folksonomy of Freesound," in *Proceedings of the CompMusic Workshop*, 2012, pp. 48–54.

[15] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: An Audio Analysis Library for Music Information Retrieval," in *International Society for Music Information Retrieval (ISMIR)*, 2013, pp. 2–7.

[6] https: //www.csc.edu.cn.

[16] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The journal of the acoustical society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[17] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[18] R. Kendall and E. Carterette, "Verbal attributes of simultaneous wind instrument timbres: I. von bismarck's adjectives." *Music Perception*, vol. 10, no. 4, pp. 445–467, 1993.

[19] S. McAdams, B. Giordano, P. Susini, G. Peeters, and V. Rioux, "A meta-analysis of acoustic correlates of timbre dimensions," *The Journal of the Acoustical Society of America*, vol. 120, no. 5_Supplement, pp. 3275–3276, 2006.

[20] E. Rovithis, A. Floros, N. Moustakas, K. Vogklis, and L. Kotsira, "Bridging audio and augmented reality towards a new generation of serious audio-only games," *Electronic Journal of e-Learning*, vol. 17, no. 2, pp. 144–156, 2019.

[21] S. K. Adams and L. B. Trucks, "A procedure for evaluating auditory warning signals," in *Proceedings of the Human Factors Society Annual Meeting*, vol. 20, no. 8. SAGE Publications Sage CA: Los Angeles, CA, 1976, pp. 166–172.

[22] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[23] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[24] A. C. Disley, D. M. Howard, and A. D. Hunt, "Timbral description of musical instruments," in *International Conference on Music Perception and Cognition*, 2006, pp. 61–68.

[25] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.