Ontologies of Sound in Neural Network Engineering

Riccardo Ancona
Dipartimento delle Arti
Università di Bologna
Bologna, Italy
riccardo.ancona@unibo.it

Abstract

Contemporary research on AI and sound is predominantly led by engineers and computer scientists, often with non-music-related backgrounds. Their conceptualization of sound and music directly influences the algorithms they design, making it important to understand the ontological frameworks that inform their work. Using mixed-methods linguistic analysis on a corpus of arXiv articles, this study examines word occurrences and collocates in current neural network engineering literature, focusing on key notions like sound, audio, music, listening, nature, originality, and objectivity. The analysis reveals how machine learning practices tend to prioritize quantitative language, performance metrics, and un-situated approaches to auditory phenomena. In contrast to classical research on digital sound technologies, neural network engineering currently appears to be less grounded on phenomenological experience and more detached from artists and audiences.

1 Context

Engineers inscribe their worldview in the artifacts they build. Sociologist of AI Airoldi (2024) argued that viewpoints are codified in the design of algorithms as *deus in machina*, cultural orientations that determine the artifact's behavior. Similarly, Actor-Network Theory believes that all artifacts are inherently sociotechnical and that they 'transcribe and displace the contradictory interests of people and things' (Latour, 1992, p. 226). According to these theories, there must then be a *culture in code* in neural network engineering applied to sound and music. Ontological notions on the nature and constitution of sound, music, and listening are invoked by engineers when designing a model, whether they are aware of them or not. In a period of intensive and extensive development in the field of sonic computation, it is of paramount relevance to inspect what kind of assumptions underlie the research and direct its trajectory. Applications of deep learning technologies are not just producing new forms of sound synthesis, but also reconfiguring the way sound is categorized, interacted with, listened to – and even understood, as psychoacoustics and computational musicology begin to rely extensively on neural networks. The epistemology of audible media and the listening hermeneutics are being reshaped by the way contemporary AI models are designed.[1] This historical passage is also marked by the arrival of new actors in the field of digital sound technology: countless researchers coming from different domains, such as data science, artificial intelligence, and information engineering, from both universities and business-oriented research centers, are contributing to the development of AI audio technologies, bringing with them different interpretations of sound. Tracing down their categorization of the audible is essential to better understand the culture they inscribe in the artifacts and reflect on its aesthetic, as well as ethic, implications. Whether such implications differ significantly from those of traditional digital audio engineering or not is something that needs to be assessed empirically. A typical strategy of analysis borrowed from Science and Technology Studies consists in observing the engineers' enunciations in order to understand the meaning they attribute to their work (see Pinch and Bijker, 1984, p. 414). One of the primary means of communication for researchers is the academic

[1] For a detailed description of the notion of 'listening hermeneutics' and its relationship with technoscience, see (Ihde, 2016).

Proceedings of the 6th Conference on AI Music Creativity (AIMC 2025),
Brussels, Belgium, September 10th-12th

article, a type of document that offers ample analytical opportunities, as it is characterized by the use of carefully-selected language and a descriptive nature. Considering the abundance of papers published every week on the topic of AI and sound, linguistic analysis offers a convenient method to combine quantitative and qualitative approaches to pinpoint which ontologies of sound are mobilized in neural network engineering.

2 Data collection

For this study, a corpus of 131 academic articles published on arXiv between the first of January and the twenty-eighth of February 2025 has been compiled.[2] ArXiv has been chosen as the source because it is the leading open-access archive hosting the most research papers about deep learning technologies. The time constraint of two months in the date of publication of the articles reflects the will to sample the literature in its current state, while also allowing to perform future analyses by using the same criterion. Among the articles tagged with the keyword 'cs.SD', indicating research on sound in the arXiv labeling system, a subset has been selected, ensuring that each paper involved the use of some form of neural networks in the sound and/or music domain. As spoken voices typically entail a specific ontology and distinct listening modalities, articles regarding speech synthesis and analysis have been excluded, aside from those that treated the topic in a multi-modal setting. For the same reason, medical-related papers have also been filtered out. The articles chosen for the corpus cover a wide range of topics, such as music and sound analysis and generation, sound spatialization, music information retrieval, music transcription, and multimodal audio and image generation, including studies on datasets and benchmarks.

A significant aspect to take into account is the affiliation of the articles' authors. The corpus contains contributions from 635 researchers, coming from various fields and institutions: 22.52% of them declare to be working on behalf of private companies, including both start-ups and large media corporations; 21.1% are affiliated with computer science departments in universities; 17.64% come from departments of information technology or technology in general; 10.87% from AI research labs hosted in universities; 7.87% from departments of electrical engineering or engineering in general; 5.35% work for departments of music informatics or sound engineering; 4.09% are receiving state grants or work for state-funded research institutes outside of university; 2.83% come from conservatories or art departments, while 2.52% come from the department of acoustics and/or psychoacoustic and music cognition; 2.05% are researchers from other STEM disciplines, such as maths and economy; 1.73% don't mention any institution but declare to be members of the Institute of Electrical and Electronics Engineers; 1.41% are either independent researchers or do not mention any affiliation. Such proportions seem to confirm the hypothesis that the introduction of deep learning technologies has involved many different actors in the field of digital sound engineering, with the vast majority of researchers coming from university departments or private companies that are not specifically dedicated to sound and music technology. The capability of deep neural networks to be adapted to a variety of data types with minimal adjustments, from text to image and audio, may have significantly contributed to this phenomenon, allowing researchers to work across multiple fields, while using the same language and the same physical facilities – as well as the same set of abstract notions, which may affect the way sound is conceived.

In order to assess if neural network engineering has produced specific ontologies of sound, diverging from the traditional research on digital sound technologies, it is useful to compare the aforementioned corpus with another collection of articles. A corpus containing all the written contributions from the 2015 and 2016 editions of the International Computer Music Conference (ICMC) has been compiled (ICMA, 2015, 2016). It contains approximately the same amount of text as the arXiv corpus, which is convenient for comparing word occurrences. The ICMC is a conference organized by the ICMA, the largest association dedicated to computer music research, reuniting audio engineers, but also composers and practitioners, in a way that manifests the typical intertwining of practice-based research and software design in the field of digital sound technology. The years 2015 and 2016 of the conference have been chosen because they are far enough in time to have only a small minority of the articles containing research on neural networks.

[2]The complete corpus of articles is listed on the references. Differently from the other sources, they include a unique arXiv identifier. Last date of access: 3rd of March, 2025.

3 Word frequency analysis

3.1 Objects, subjects, entities

A first significant indicator in linguistic analysis is word occurrences. The most frequently used words in a corpus may reveal meaningful insights on the type of discourse that is articulated by the enunciators – they reveal the recurrent subjects of interest of the writers, starting from which a worldview is composed. Because the primary aim of this research is to inspect ontological notions, it is appropriate to start by inspecting the lexical frequency of nouns, as it is the grammatical category typically related to subjects, objects, and entities. Table 1 shows a comparison of the fifteen most frequent nouns appearing in the arXiv and the ICMC corpora. Singular and plural occurrences are summed, while terms related with the paper format, such as 'figure' and 'table', have been excluded from the count. Terms that can be both understood as nouns and verbs, like 'control' and 'order', have also been removed, except in cases in which the noun form is highly prevalent, as in 'model'.

Table 1: Top-15 noun occurrences

ArXiv Term	*n*	ICMC Term	*n*
Audio	8534	Music	4804
Model	8340	Sound	3614
Music	5673	System	2221
Dataset	3288	Computer	2176
Speech	2780	Time	1796
Data	2627	Audio	1238
Text	1967	Data	1158
Language	1841	Instrument	1142
System	1679	Model	1138
Sound	1674	Performance	1098
Signal	1651	User	1026
Information	1488	Frequency	987
Input	1448	Score	847
Time	1425	Space	818
Source	1211	Parameter	817

Comparing the two corpora, seven out of fifteen of the most occurring nouns are in common, highlighting a shared ground between them: sound, music, audio, data, system, model, and time are fundamental ontological notions of sound computing. However, their occurrence rate differs significantly. In particular, whereas 'music' has a similar amount of instances among the two corpora, the notion of 'audio' seems to be remarkably more relevant in the arXiv corpus, while the term 'sound' is only at the tenth position. The digital representation of sound as audio is therefore a highly preferred subject than the sonic phenomenon itself. The arXiv corpus appears to dedicate particular attention to entities that can be subjected of manipulation, as indicated by the presence of the words 'dataset', 'signal', 'input', and 'source'. The high frequency of the terms 'text', 'speech', and 'language', despite having filtered out articles about speech synthesis from the corpus, points to the close relationship that audio neural network engineering bears with the field of natural language processing. If we look instead at the terms that are exclusively present in the top fifteen of the ICMC corpus, we can observe that it exhibits a larger variety of entities: there are humans ('user'), devices ('computer', 'instrument'), forms of inscription ('score'), activities ('performance'), and metrics ('parameter'). Most interestingly, the presence of both 'space' and 'time' in the ICMC corpus denotes a stronger phenomenological situatedness than in the arXiv corpus. In the latter, 'space' is at the 239th position in the list of the most occurring words. If one observes the top thousand word frequencies in the arXiv corpus in detail, there is a lack of objects, agents, actions, and phenomena grounded in the experiential, sensory domain, that would be typically associated to musicking and sound research. Table 2 displays the number of occurrences of some of them in the two corpora. Verbs are listed in gerund, but the corresponding value is cumulative of all their forms.

Table 2: Comparison of word frequencies

Term	ICMC	ArXiv
Listening	296	197
Listener	240	117
Hearing	163	161
Perception	187	123
Playing	857	209
Creating	978	336
Performance	1098	2347
Performer	750	29
Composition	830	169
Composer	530	54
Improvisation	213	67
Improviser	54	0
Concert	269	53
Musician	327	89
Instrument	1142	473
Computer	2176	423
User	1026	463
Object	948	108
Interaction	488	131
Interface	814	53
Art	349	386
Artist	166	100
Body	234	36
Ear	88	35
Motion	422	55
Environment	469	203
Place	182	39
Material	434	148
Timbre	376	179
Gesture	862	13
Texture	1167	41

The arXiv corpus appears to be consistently distant from the embodied experience of sound and music, both in the terminology referring to its practice and to that usually describing its fundamental ontologies (objects, bodies, environments). The only exceptions are the words 'art' – although, its occurrence in the arXiv corpus is in the phrasing 'state of the art' 75.4% of the times – and 'performance', whose meaning most often refers to the measurement of the effectiveness of a system. While the articles in the arXiv corpus contain numerous studies on the analysis and generation of sounds and music, their language manifests sheer detachment from the physical conditions in which such procedures are supposed to happen, not to mention the social and creative contexts in which they are supposed to exist. In relation with the theme of this conference, 'The Artist in the Loop', from the data collected in this study it appears that neural network engineers tend not to be particularly concerned with users, artists, musicians, performers, nor listeners. At least in this case, it is notable that research on AI technology in the audio domain demonstrates less interest in human users and artists than traditional audio engineering does. Artists are not in the loop.

3.2 Processes and parameters

The primary concerns of neural network engineering can be distinguished by looking at the impressively rich vocabulary regarding procedures, as shown in Table 3.

Table 3: Thirtyfive most occurring processes in the arXiv corpus. The 'range' column indicates the number of articles in which the word appears.

Term	n	Range
Training	4236	131
Learning	2928	131
Processing	1949	129
Generation	1761	123
Evaluation	1175	130
Classification	1030	99
Detection	946	78
Recognition	834	109
Representation	753	113
Diffusion	667	52
Analysis	661	111
Understanding	621	92
Modeling	539	89
Prediction	441	85
Inference	414	77
Synthesis	378	58
Separation	375	46
Tuning	356	64
Transcription	354	28
Estimation	332	53
Captioning	318	28
Matching	311	56
Extraction	308	63
Capture	296	85
Reasoning	281	25
Transfer	281	57
Query	276	38
Masking	259	30
Encoding	259	62
Augmentation	239	61
Reconstruction	224	32
Generalization	207	62
Identification	198	35
Optimization	195	64
Conditioning	191	38

Interestingly, many of these terms could be applied to any subfield of engineering, ranging from industrial logistics to weapon manifacturing. Other terms are specific to neural architectures. The universe described by audio neural network engineering is constituted by a plurality of types of data and signals undergoing innumerable transformations to distill new types of data and signals. It is a discourse on praxis in which, paradoxically, material contingencies are omitted. Its ontology is characterized by the parcelling out of things into commensurable, divisible parts, as is made apparent by the persistent presence of 'samples' (1453 occurrences), 'tokens' (1259), 'layers' (1198), 'segments' (562), 'components' (409), 'modules' (308), 'parts' (287), 'elements' (279), 'batches' (236), and 'groups' (199). Data and signals thus constitute an intricate mereology in which objects, despite lacking locality and spatial determination, act as modules that can be split or grouped with different degrees of granularity. Processes of data transformation call into account quantitative measurement, as denoted by another class of terms that abounds in the arXiv corpus: 'features' (1933), 'scores' (1315), 'levels' (1080), 'metrics' (705), 'parameters' (667), 'baselines' (641), 'errors' (485), and 'benchmarks' (398). Machine learning engineers seem to be attentive to 'quality' (1069), 'accuracy' (749), 'rate' (618), 'scale' (596), 'size' (562), 'similarity' (407), 'alignment' (368), 'length' (350), 'ability' (332), 'robustness' (328), 'capabilities' (323), 'effectiveness' (314), 'duration' (308), 'distance' (294), 'resolution' (289), 'efficiency' (259), and all other sorts of parametrical judgments. What is described here is a discourse in which functionality appears to be the fundamental property to

inspect in objects and processes. To better understand how the effectiveness of an entity is assessed in this context, it is useful to look at some of the most used adjectives, listed in Table 4.

Table 4: Occurrences of twenty prevalent adjectives in the arXiv corpus. The 'range' column indicates the number of articles in which the word appears.

Term	n	Range
High	1311	121
Different	1191	128
Large	1057	124
First	781	124
Multi	754	109
Low	729	92
Deep	614	118
Real	541	103
Original	463	96
Better	456	113
New	438	111
Long	432	84
Objective	429	88
Available	406	106
Efficient	406	106
Robust	371	84
Open	364	85
Diverse	361	90
Best	356	99
Complex	350	89

Such list is a curious mix of metrical definitions and expressions of primacy. None of them, though, expresses a particular attachment to the sensorial domain that is usually related to judgement. Interestingly, when comparing the frequency of adjectives with their binary opposites, the ratios tend to consistently favour the term referring to the positive or greatest attribute in the dichotomy: for instance, the frequency of 'high' is 1.79 higher than 'low', 'large' occurs 4.02 times more than 'small', 'new' is 39.81 times over 'old', 'better' is 10.13 times over 'worst', and 'long' 2.16 times over 'short'. In this sense, adjectivation is employed as a way to generically emphasize the efficacy of technical procedures, rather than to describe their phenomenological effects.

4 Words in context

4.1 Sound, audio, music, listening

Analyzing the occurrence rate of terms gives an intuitive understanding of which ontologies constitute the discourse of neural network engineering. However, to acknowledge how such ontologies are conceived and employed requires to examine them within their context of use and to evaluate the meanings that are attributed to them. The field of natural language processing offers several effective methods. Collocation extraction algorithms and n-grams are able to retrieve relevant syntagmata, valuable to study how words are grouped together to build larger structures of meaning. Key Word in Context indexing (KWIC) is a practical procedure to align all the sentences containing a given word or group of words and analyze them in their context. For this study, the AntConc corpus analysis toolkit has been employed.

The most important words to dissect are undoubtedly the triad of sound, music, and audio. 'Sound' is the term that generally has the most broadly encompassing meaning among the three, although it is by far the least occurring in the arXiv corpus. It appears in 78% of articles, the majority of the times associated to the syntagma 'sound event(s)', a definition that suggests a temporal attribution. In their context of occurrence, sound events are usually phenomena that need to be detected, recognized, or localized. This suggests that they are something out-there, already present in the world, to which

neural network engineers relate by means of classification and detection. In fact, the word 'sound' often occurs in the syntagma "sound source(s)". In its singular form, sound is sometimes associated to the concepts of localization, positioning, separation, and distancing, suggesting a sense of locality and situatedness of the phenomenon. However, the occurrence of such syntagmata is limited to the subset of articles in the corpus dedicated to sound spatialization. In the other papers, this concept of locality is usually not manifested. Much more frequent than 'sound' is the term 'audio', which appears in 96% of the articles. The contexts of usage of the term encompass the entire semantic field of the corpus, from 'audio generation' and 'captioning', to 'audiovisual' and other forms of multimodality. As sound, audio is also associated to 'events', but it is more often thought in terms of 'audio data', 'signals', 'samples', or 'recordings'. It has, therefore, a distinct objectual nature. A vast vocabulary of operations are documented in the corpus: audio is sampled, used, converted, divided, transformed, created, encoded, extracted, fused, generated, identified, mixed, observed, degraded, sliced, standardized, trained, and transcribed. Judging from such phraseology, audio is an intrinsically passive entity, understood and experienced by means of manipulations it is subjected to. Apparently, audio has no agency on its own. Thus, in the ontology described in the arXiv corpus, sound and audio are distinct in the sense that sound is a phenomenon that has a spatiotemporal determination, whose nature has to be captured by means of information retrieval, whereas audio is a malleable material, the recipient and the object that has to be transformed and refined.
Differently from sound and audio, the meaning of 'music' is often made explicit by the researchers: it is 'an art form centered on sound' (Tian et al., 2025b, p. 1), whose enjoyment is 'a highly personal experience' (Harada et al., 2025, p. 1); music is 'an extremely specific modality' (Zhu et al., 2025a, p. 1), 'a universal medium [...] crossing cultural and linguistic boundaries' (Wu et al., 2025a, p. 2). Such definitions converge in attributing to music a strong cultural complexity, which, by the way researchers describe it, seems to exceed the representational and hermeneutical capabilities proposed by the engineers' models. Nevertheless, some of the most occurring syntagmata suggest the attempt to overcome these difficulties, engaging in 'music generation', 'retrieval', 'tagging', 'transcription', 'representation', 'analysis', and even 'understanding'. The phrasing 'music understanding' occurs 134 times and it is generally intended as a subfield of music AI devoted to a variety of tasks, such as 'genre classification, emotion prediction, and key detection' (Zhu et al., 2025a, p. 1). The use of 'understanding' in this context is quite peculiar, as in common language this verb is generally associated to a more nuanced intellectual and emotional grasping of reality than numerical tasks such as detection and classification. Neural network engineers in the arXiv corpus often seem to not recognize such nuance, as it is also attested by the usage of the verb 'reasoning' associated to algorithmic computation. This type of terminology appears to be in contradiction with the aforementioned emphasis on the cultural and experiential complexity of music. In relation to sound, audio, and music, another term would be useful to analyze. The word 'listening' is present in 43% of the articles in the arXiv corpus, with roughly 40% of them mentioning it only once. With the exclusion of a handful of outliers, 'listening' is hardly a relevant category in the corpus, revealing the weak integration of phenomenological, subjective experience into the discourse of neural network engineering.

4.2 Natural, real, original

Other notions worth of examining are those related to the oppostion between nature and artificiality. Following the tradition of Actor-Network Theory, the focus is on observing how these notions are constructed and employed in the agents' enunciations, without giving them for granted. In the analyzed articles, 'nature' is never nature in itself but the nature of something – like, for instance, 'the nature of the training process'. Nature is therefore not a subject but often an attribute, as the adjective 'natural' is present in 69.5% of the papers. While often it is phrased in the syntagma "natural language processing", it also assumes other meanings. 'Natural' is a quality attributed to sounds and signals coming from the preexisting world, either anthropic or not – as exemplified by the phrasing 'natural characteristics of unaltered audio' (Farooq et al., 2025, p. 4). 'Natural' stands for intact, not yet subjected to manipulation. At the same time, it is an attribute that the neural network models need to achieve: engineers try to 'generate high-resolution, natural-sounding audio' (Hong and Choi, 2025, p. 2), 'natural and high quality voices' (Guo et al., 2025, p. 1), 'convincing levels of natural sounding vocals' (Sharma and Gupta, 2025, p. 1), 'speech that is natural and understandable' (Xue et al., 2025, p. 26), 'achieving natural and expressive audio' (Chen et al., 2025a, p. 3), and 'more natural sound reconstruction' (Hong and Choi, 2025, p. 5). The resemblance of naturality is thus perceived as an intrinsically positive characteristic. Another recurring adjective is 'real', present in 78% of the

articles, often in the syntagma 'real world'. There are 'real world applications', 'real world scenario', 'real world audio', 'real world data' – a recurrent referencing to a world outside the engineering practice, to which, apparently, neural networks do not fully belong, but which need to imitate in the most 'realistic' fashion. The same kind of meaning is given to the adjective 'original', appearing in 73% of the texts, in relation with entities that have to be either transformed into something else or reconstructed by models. The dichotomy between the natural and artificial world is therefore considered an a priori condition, but at the same time the authors of the papers strive to bridge the gap between the two worlds by means of technical enhancement. This, in turn, could mean that such a priori distinction is not understood as essential, as mere improvement can dissolve it – or, alternatively, that engineers believe that technoscience is capable of overcoming essential ontological differences.

4.3 Objective and subjective metrics

A dichotomy also manifests in the opposition of 'objective' and 'subjective' judgements, this time in a more epistemological rather than ontological sense. The word 'objective' occurs 1.87 times more than 'subjective', although it is often intended as a noun, synonym of 'target'. Very often, 'subjective' and 'objective' occur as collocates as two parallel forms of assessment of the models' results. The dialectic between the two methodologies is best exemplified by the phrase in Liu et al. (2025a, p. 1): objective metrics 'offer quick and convenient evaluations, yet they often show a weak correlation with human-perceived quality. In contrast, subjective evaluations directly reflect human judgement, but are time-consuming, labour-intensive, and lack reproducibility, making it difficult to compare results across different works'. The limitations of objective metrics are often pinpointed, 'given the challenge of establishing objective metrics that fully capture musicality' (Wang et al., 2025b, p. 6). Engineers seem to be fully aware of such limitations, yet the results of their work is way more often assessed by means of objective metrics than subjective ones. 'Objective', in this context, although it might mean that the measurement is universally valid, has an ambiguous epistemological status, because objective metrics are constantly contested and redefined in the discourse of machine learning engineering. Reading the articles in the arXiv corpus, a proliferation of new units of measurement and benchmarks are constantly negotiated, all of them equally defined 'objective', as they all descend from statistical inference. As in the case of bridging the gap between the natural and the artificial, the gap between what quantitative measures are able to capture and the actual qualities that need to be grasped is understood as a distance that can be progressively reduced to zero by gradual technical refinements. This, in tandem with the aforementioned notions of machine 'reasoning' and 'understanding', might suggest that some neural network engineers might be convinced that the capability of grasping the qualitative side of sound is merely a task that can be fulfilled by a statistical algorithm, once it reaches a certain perfectioning. Yet, at the current state of the art, such ability is pretended to be reached only with many caveats on the apparently irreducible complexity of sound and music.

5 Discussion

The data collected and analyzed in the previous sections point to a peculiar technoscientific culture, whose core values and methodologies diverge to the preexisting computer music culture. However, it must be kept in mind that due to the type of analysis method employed and the chosen dataset, the conclusions of this research are not necessarily generalizable to the entirety of neural audio literature. The purpose of the analysis, in fact, is not to assert that all neural network engineering would align with one and the same way of ontologizing the world, but rather to identify a specific subculture, whose contribution to the discipline seems to be becoming, if perhaps only quantitatively, increasingly relevant. While it is true that choosing other repositories of academic articles other than arXiv might have revealed other perspectives on sound, it is impossible to ignore the amount of scholarly material found on that platform. It is also difficult to overlook that a significant subset of these articles come directly from AI labs belonging to some of the most important big tech companies, as well as from some of the world's most distinguished academic research centres. Hence, while this study does not capture the totality of research in this area, it certainly highlights a way of thinking and manipulating sound that seems to have a growing influence in the contemporary industrial-academic apparatus. Neither did this study attempt to trace the causes of this transformation in sound ontologies. In continuity with the Social Construction of Technology (SCOT), an attempt was made to describe a specific social group – the audio neural network engineers who publish on arXiv – in order to

understand the way in which they discuss their work. Other methodologies coming from other disciplines can certainly try to trace back the genealogy of these ideas and how they are reinforced by social and economic dynamics. Here, the aim was simply to draw attention to the language of those working in this field, so as to instigate a conversation on the subject and invite other researchers to engage with it, putting their expertise into practice. Notwithstanding all these caveats, it is significant to observe some salient features of the articles examined. Firstly, it is noted that the type of neural network engineering studied in this article manifests a culture of sound without a culture of listening. The ontological category of sound is essentially subsumed by that of audio, which is seen as data rather than cultural artefact. The mathematical inscription is conflated with the thing itself. From this perspective, the direct sensory experience of the phenomenon is of no particular relevance to these engineers: all that matters is latently encapsulated in the information, in the form of a hidden statistical pattern. One consequence of this way of thinking is that the traditional actors of music are less and less present. It is no coincidence that there is a growing interest in the literature for the design of automatic sound generation systems in non-real time, according to a paradigm that is increasingly detached from the spatial, temporal, and cultural contingencies within which music generally emerges. This culture focuses on producing content in the most industrially optimal way, rather than asking how deep learning can expand the potential of artists and practitioners. Not all the articles in the corpus converge on this perspective: there are relevant exceptions, such as (Lee and Pasquier, 2025) and (Pasquier et al., 2025). Nevertheless, the trend seems to follow a process that philosopher Bernard Stiegler has called 'industrialization of the spirit' (*industralisation de l'esprit*) (Stiegler, 2018).

6 Conclusions

To shed light on the way machine learning engineers might conceive sound, music, and audio, a linguistic analysis of a corpus of academic articles has been conducted. The data has been analyzed in comparison with a corpus of older papers on classical audio technology, with the aim of noticing differences among the fields. Compared to traditional research on digital sound, neural network engineering seems to give larger priority to the notion of audio and its manipulation. Its ontology of audio is consistently detached from phenomenological experience and it largely relies on purely quantitative procedures and measurements, outlining an empirical science without perception as its fundamental activity. The experience of listening is apparently absent from the practice of neural audio design, as are absent the traditional agents of musicmaking, such as musicians and musical instruments. However, the considerations formulated in the analysis of the corpus might not necessarily generalize to the entirety of the research field – as is the case with every inductive study, there might be significant outliers. Data has been curated and analyzed in a way that is replicable, with the aim of encouraging future examinations of different corpora. Additionally, to improve the depth of the research, the study of the engineers' enunciations could be accompanied in the future by other types of studies, including dataset analysis,[3] post-phenomenological investigation of the neural models, as well as participant observation of the practices of neural network engineering.

References

Afchar, D., Meseguer-Brocal, G., and Hennequin, R. (2025). AI-Generated Music Detection and its Challenges. arXiv:2501.10111 [cs].

Agarwal, M., Wang, C., and Richard, G. (2025). F-StrIPE: Fast Structure-Informed Positional Encoding for Symbolic Music Generation. arXiv:2502.10491 [cs].

Airoldi, M. (2024). *Machine Habitus. Sociologia degli algoritmi*. LUISS.

Akram, M. W., Dettori, S., Colla, V., and Buttazzo, G. C. (2025). ChordFormer: A Conformer-Based Architecture for Large-Vocabulary Audio Chord Recognition. arXiv:2502.11840 [cs].

Ancona, R. (2024). Una prospettiva critica sui dataset per la sintesi text-to-audio. In *Projecting Memories. 24th CIM – Colloquium on Music Informatics Conference Proceedings*, pages 107–114. AIMI.

[3] Dataset analysis has been explored in Ancona (2024).

Barański, M., Jasiński, J., Bartolewska, J., Kacprzak, S., Witkowski, M., and Kowalczyk, K. (2025). Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio. arXiv:2501.11378 [cs].

Barusco, M., Borsatti, F., Pezze, D. D., Paissan, F., Farella, E., and Susto, G. A. (2025). From Vision to Sound: Advancing Audio Anomaly Detection with Vision-Based Algorithms. arXiv:2502.18328 [cs].

Basu, A., Chaudhari, P., and Caterina, G. D. (2025). Fundamental Survey on Neuromorphic Based Audio Classification. arXiv:2502.15056 [cs].

Berger, C., Badeau, R., and Essid, S. (2025). Perceptual Noise-Masking with Music through Deep Spectral Envelope Shaping. arXiv:2502.17527 [cs].

Bhandari, K., Chang, S., Lu, T., Enus, F. R., Bradshaw, L. B., Herremans, D., and Colton, S. (2025a). ImprovNet: Generating Controllable Musical Improvisations with Iterative Corruption Refinement. arXiv:2502.04522 [cs].

Bhandari, K., Wiggins, G. A., and Colton, S. (2025b). Yin-Yang: Developing Motifs With Long-Term Structure And Controllability. arXiv:2501.17759 [cs].

Bjare, M. R., Cantisani, G., Lattner, S., and Widmer, G. (2025). Estimating Musical Surprisal in Audio. arXiv:2501.07474 [cs].

Carson, A., Välimäki, V., Wright, A., and Bilbao, S. (2025). Resampling Filter Design for Multirate Neural Audio Effect Processing. arXiv:2501.18470 [eess].

Chen, W., Sha, B., Yang, J., Wang, Z., Fan, F., and Wu, Z. (2025a). Singing Voice Conversion with Accompaniment Using Self-Supervised Representation-Based Melody Features. arXiv:2502.04722 [cs].

Chen, Y., Shimada, K., Simon, C., Ikemiya, Y., Shibuya, T., and Mitsufuji, Y. (2025b). CC-Stereo: Audio-Visual Contextual and Contrastive Learning for Binaural Audio Generation. arXiv:2501.02786 [cs].

Cheng, K. J., Li, T., and Anumanchipalli, G. (2025). Audio Texture Manipulation by Exemplar-Based Analogy. arXiv:2501.12385 [cs].

Cheston, H., Balen, J. V., and Durand, S. (2025). Automatic Identification of Samples in Hip-Hop Music via Multi-Loss Training and an Artificial Dataset. arXiv:2502.06364 [cs].

Choi, S., Kim, K. W., and Kang, M. (2025). MMVA: Multimodal Matching Based on Valence and Arousal across Images, Music, and Musical Captions. arXiv:2501.01094 [cs].

Chou, B. S.-H., Jajal, P., Eliopoulos, N. J., Nadolsky, T., Yang, C.-Y., Ravi, N., Davis, J. C., Yun, K. Y.-J., and Lu, Y.-H. (2025). Detecting Music Performance Errors with Transformers. arXiv:2501.02030 [cs].

Chung, Y., Eu, P., Lee, J., Choi, K., Nam, J., and Chon, B. S. (2025). KAD: No More FAD! An Effective and Efficient Evaluation Metric for Audio Generation. arXiv:2502.15602 [cs].

Clarke, C. J. (2025). LACTOSE: Linear Array of Conditions, TOpologies with Separated Error-backpropagation – The Differentiable "IF" Conditional for Differentiable Digital Signal Processing. arXiv:2502.15829 [cs].

Comunità, M., Steinmetz, C. J., and Reiss, J. D. (2025). Differentiable Black-box and Gray-box Modeling of Nonlinear Audio Effects. arXiv:2502.14405 [cs].

Dai, S., Wang, Y., Dannenberg, R. B., and Jin, Z. (2025a). Everyone-Can-Sing: Zero-Shot Singing Voice Synthesis and Conversion with Speech Reference. arXiv:2501.13870 [cs].

Dai, Y., Wang, C., Li, C., Wang, C., Du, J., Li, K., Wang, R., Ma, J., Sun, L., and Gao, J. (2025b). Latent Swap Joint Diffusion for Long-Form Audio Generation. arXiv:2502.05130 [cs].

Deshmukh, S., Han, S., Singh, R., and Raj, B. (2025). ADIFF: Explaining audio difference using natural language. arXiv:2502.04476 [cs].

Diao, X., Zhang, C., Wu, T., Cheng, M., Ouyang, Z., Wu, W., and Gui, J. (2025). Learning Musical Representations for Music Performance Question Answering. arXiv:2502.06710 [cs].

Doh, S., Choi, K., and Nam, J. (2025). TALKPLAY: Multimodal Music Recommendation with Large Language Models. arXiv:2502.13713 [cs].

Donahue, C., Wu, S.-L., Kim, Y., Carlton, D., Miyakawa, R., and Thickstun, J. (2025). Hookpad Aria: A Copilot for Songwriters. arXiv:2502.08122 [cs].

Dong, Y., Wang, Q., Hong, H., Jiang, Y., and Cheng, S. (2025). An Experimental Study on Joint Modeling for Sound Event Localization and Detection with Source Distance Estimation. arXiv:2501.10755 [cs].

Emmerich, L., Aste, P., Brandão, E., Nolan, M., Cuenca, J., Svensson, U. P., Maeder, M., Marburg, S., and Zea, E. (2025). A data-driven two-microphone method for in-situ sound absorption measurements. arXiv:2502.04143 [cs].

Fang, K., Wang, Z., Xia, G., and Fujinaga, I. (2025). Exploring GPT's Ability as a Judge in Music Understanding. arXiv:2501.13261 [cs].

Farooq, M. U., Khan, A., Uddin, K., and Malik, K. M. (2025). Transferable Adversarial Attacks on Audio Deepfake Detection. arXiv:2501.11902 [cs].

Genova, D., Esling, P., and Hurlin, T. (2025). Keep what you need : extracting efficient subnetworks from large audio representation models. arXiv:2502.12925 [cs].

Grinberg, P., Kumar, A., Koppisetti, S., and Bharaj, G. (2025). What Does an Audio Deepfake Detector Focus on? A Study in the Time Domain. arXiv:2501.13887 [cs].

Gröger, F., Baumann, P., Amruthalingam, L., Simon, L., Giurda, R., and Lionetti, S. (2025). Evaluation of Deep Audio Representations for Hearables. arXiv:2502.06664 [cs].

Guo, W., Zhang, Y., Pan, C., Huang, R., Tang, L., Li, R., Hong, Z., Wang, Y., and Zhao, Z. (2025). TechSinger: Technique Controllable Multilingual Singing Voice Synthesis via Flow Matching. arXiv:2502.12572 [cs].

Hakala, A., Kincy, T., and Virtanen, T. (2024). Automatic Live Music Song Identification Using Multi-level Deep Sequence Similarity Learning. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 31–35. arXiv:2501.08129 [eess].

Harada, T., Motomitsu, T., Hayashi, K., Sakai, Y., and Kamigaito, H. (2025). Can Impressions of Music be Extracted from Thumbnail Images? arXiv:2501.02511 [cs].

Hasumi, T., Komatsu, T., and Fujita, Y. (2025). Music Tagging with Classifier Group Chains. arXiv:2501.05050 [cs].

Heikkinen, M., Politis, A., Drossos, K., and Virtanen, T. (2025). Gen-A: Generalizing Ambisonics Neural Encoding to Unseen Microphone Arrays. arXiv:2501.08047 [eess].

Hexeberg, S., Chitre, M., Hoffmann-Kuhnt, M., and Low, B. W. (2025). Semi-supervised classification of bird vocalizations. arXiv:2502.13440 [cs].

Hong, S. and Choi, Y.-H. (2025). RingFormer: A Neural Vocoder with Ring Attention and Convolution-Augmented Transformer. arXiv:2501.01182 [cs].

Hu, H., Siniscalchi, S. M., Yang, C.-H. H., and Lee, C.-H. (2025). Variational Bayesian Adaptive Learning of Deep Latent Variables for Acoustic Knowledge Transfer. arXiv:2501.15496 [eess].

Huang, T., Li, A., Li, X., Zhang, J., Xian, S., Zhang, Q., Lu, M., Chen, G., Xiong, L., and Hu, X. (2025a). Unsupervised CP-UNet Framework for Denoising DAS Data with Decay Noise. arXiv:2502.13395 [cs].

Huang, W., Gu, Y., Wang, Z., Zhu, H., and Qian, Y. (2025b). Generalizable Audio Deepfake Detection via Latent Space Refinement and Augmentation. arXiv:2501.14240 [eess].

ICMA (2015). *Proceedings of the 41st International Computer Music Conference, ICMC 2015, Denton, TX, USA, September 25 - October 1, 2015*. Michigan Publishing.

ICMA (2016). *Proceedings of the 2016 International Computer Music Conference, ICMC 2016, Utrecht, The Netherlands, September 12-16, 2016*. Michigan Publishing.

Ihde, D. (2016). *Acoustic Technics*. Postphenomenology and the philosophy of technology. Lexington Books, Lanham, [Maryland] Boulder, [Colorado] New York, [New York].

Im, J. and Nam, J. (2025). FlashSR: One-step Versatile Audio Super-resolution via Diffusion Distillation. arXiv:2501.10807 [eess].

Jiang, X., Dindar, S. S., Choudhari, V., Bickel, S., Mehta, A., McKhann, G. M., Flinker, A., Friedman, D., and Mesgarani, N. (2025a). AAD-LLM: Neural Attention-Driven Auditory Scene Understanding. arXiv:2502.16794 [cs].

Jiang, Y., Chen, Q., Ji, S., Xi, Y., Wang, W., Zhang, C., Yue, X., Zhang, S., and Li, H. (2025b). UniCodec: Unified Audio Codec with Single Domain-Adaptive Codebook. arXiv:2502.20067 [eess].

Justus, A. A. (2025). Music Generation using Human-In-The-Loop Reinforcement Learning. arXiv:2501.15304 [cs].

Kang, H., Han, G., Jeong, Y., and Park, H. (2025). AudioGenX: Explainability on Text-to-Audio Generative Models. arXiv:2502.00459 [cs].

Kang, J. and Herremans, D. (2025). Towards Unified Music Emotion Recognition across Dimensional and Categorical Models. arXiv:2502.03979 [cs].

Kheir, Y. E., Samih, Y., Maharjan, S., Polzehl, T., and Möller, S. (2025). Comprehensive Layer-wise Analysis of SSL Models for Audio Deepfake Detection. arXiv:2502.03559 [eess].

Kim, H., Kwon, T., and Nam, J. (2025a). D3RM: A Discrete Denoising Diffusion Refinement Model for Piano Transcription. arXiv:2501.05068 [cs].

Kim, K., Koo, J., Lee, S., Joung, H., and Lee, K. (2025b). TokenSynth: A Token-based Neural Synthesizer for Instrument Cloning and Text-to-Instrument. arXiv:2502.08939 [cs].

Kim, L., Jeon, S., Heo, W., and Park, J. (2025c). Note-Level Singing Melody Transcription for Time-Aligned Musical Score Generation. arXiv:2502.12438 [cs].

Kong, Y., Meseguer-Brocal, G., Lostanlen, V., Lagrange, M., and Hennequin, R. (2025a). S-KEY: Self-supervised Learning of Major and Minor Keys from Audio. arXiv:2501.12907 [cs].

Kong, Z., Shih, K. J., Nie, W., Vahdat, A., Lee, S.-g., Santos, J. F., Jukic, A., Valle, R., and Catanzaro, B. (2025b). A2SB: Audio-to-Audio Schrodinger Bridges. arXiv:2501.11311 [cs].

Korgialas, C., Tsingalis, I., Tzolopoulos, G., and Kotropoulos, C. (2025). InterGridNet: An Electric Network Frequency Approach for Audio Source Location Classification Using Convolutional Neural Networks. arXiv:2502.10011 [cs].

Koudounas, A., Quatra, M. L., Siniscalchi, M. S., and Baralis, E. (2025). voc2vec: A Foundation Model for Non-Verbal Vocalization. arXiv:2502.16298 [eess].

Kow, P.-Y. and Kow, P.-Z. (2025). An efficient light-weighted signal reconstruction method consists of Fast Fourier Transform and Convolutional-based Autoencoder. arXiv:2501.01650 [cs].

Krishnan, V. V., Alben, N., Nair, A., and Condit-Schultz, N. (2025). Sanidha: A Studio Quality Multi-Modal Dataset for Carnatic Music. arXiv:2501.06959 [cs].

Lanzendörfer, L. A., Grötschla, F., Ungersböck, M., and Wattenhofer, R. (2025). High-Fidelity Music Vocoder using Neural Audio Codecs. arXiv:2502.12759 [cs].

Latour, B. (1992). Where are the missing masses, sociology of a few mundane artefacts. In Bijker, W. E. and Law, J., editors, *Shaping Technology-Building Society. Studies in Sociotechnical Change*, pages 225–259. MIT Press.

Le, D.-V.-T., Bigo, L., and Keller, M. (2025). Evaluating Interval-based Tokenization for Pitch Representation in Symbolic Music Analysis. arXiv:2501.04630 [cs].

Lee, K. J. M., Ens, J., Adkins, S., Sarmento, P., Barthet, M., and Pasquier, P. (2025). The GigaMIDI Dataset with Features for Expressive Music Performance Detection. *Transactions of the International Society for Music Information Retrieval*, 8(1). arXiv:2502.17726 [cs].

Lee, K. J. M. and Pasquier, P. (2025). Musical Agent Systems: MACAT and MACataRT. arXiv:2502.00023 [cs].

Li, D., Zang, Y., and Kong, Q. (2025a). Piano Transcription by Hierarchical Language Modeling with Pretrained Roll-based Encoders. arXiv:2501.03038 [cs].

Li, J., Zhao, W., Huang, Z., Guo, Y., and Tian, Y. (2025b). Do Audio-Visual Segmentation Models Truly Segment Sounding Objects? arXiv:2502.00358 [cs].

Li, Y., Liu, J., Zhang, T., Zhang, T., Chen, S., Li, T., Li, Z., Liu, L., Ming, L., Dong, G., Pan, D., Li, C., Fang, Y., Kuang, D., Wang, M., Zhu, C., Zhang, Y., Guo, H., Zhang, F., Wang, Y., Ding, B., Song, W., Li, X., Huo, Y., Liang, Z., Zhang, S., Wu, X., Zhao, S., Xiong, L., Wu, Y., Ye, J., Lu, W., Li, B., Zhang, Y., Zhou, Y., Chen, X., Su, L., Zhang, H., Chen, F., Dong, X., Nie, N., Wu, Z., Xiao, B., Li, T., Dang, S., Zhang, P., Sun, Y., Wu, J., Yang, J., Lin, X., Ma, Z., Wu, K., li, J., Yang, A., Liu, H., Zhang, J., Chen, X., Ai, G., Zhang, W., Chen, Y., Huang, X., Li, K., Luo, W., Duan, Y., Zhu, L., Xiao, R., Su, Z., Pu, J., Wang, D., Jia, X., Zhang, T., Ai, M., Wang, M., Qiao, Y., Zhang, L., Shen, Y., Yang, F., Zhen, M., Zhou, Y., Chen, M., Li, F., Zhu, C., Lu, K., Zhao, Y., Liang, H., Li, Y., Qin, Y., Sun, L., Xu, J., Sun, H., Lin, M., Zhou, Z., and Chen, W. (2025c). Baichuan-Omni-1.5 Technical Report. arXiv:2501.15368 [cs].

Liu, C., Wang, H., Zhao, J., Zhao, S., Bu, H., Xu, X., Zhou, J., Sun, H., and Qin, Y. (2025a). MusicEval: A Generative Music Corpus with Expert Ratings for Automatic Text-to-Music Evaluation. arXiv:2501.10811 [cs].

Liu, X., Feng, Z., Kanojia, D., and Wang, W. (2025b). DGFM: Full Body Dance Generation Driven by Music Foundation Models. arXiv:2502.20176 [cs].

Liu, Y., Lu, L., Jin, J., Sun, L., and Fanelli, A. (2025c). XAttnMark: Learning Robust Audio Watermarking with Cross-Attention. arXiv:2502.04230 [cs].

Liu, Z., Ding, S., Zhang, Z., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., and Wang, J. (2025d). SongGen: A Single Stage Auto-regressive Transformer for Text-to-Song Generation. arXiv:2502.13128 [cs].

Lu, S.-C., Yeh, C.-C., Cho, H.-L., Hsu, C.-C., Hsu, T.-L., Wu, C.-H., Shih, T. K., and Lin, Y.-C. (2025). YNote: A Novel Music Notation for Fine-Tuning LLMs in Music Generation. arXiv:2502.10467 [cs].

Ma, Z., Chen, Z., Wang, Y., Chng, E. S., and Chen, X. (2025). Audio-CoT: Exploring Chain-of-Thought Reasoning in Large Audio Language Model. arXiv:2501.07246 [cs].

Mao, Z., Zhao, M., Wu, Q., Wakaki, H., and Mitsufuji, Y. (2025). DeepResonance: Enhancing Multimodal Music Understanding via Music-centric Multi-way Instruction Tuning. arXiv:2502.12623 [cs].

Masuyama, Y., Ueno, N., and Ono, N. (2025a). Mel-Spectrogram Inversion via Alternating Direction Method of Multipliers. arXiv:2501.05557 [eess].

Masuyama, Y., Wichern, G., Germain, F. G., Ick, C., and Roux, J. L. (2025b). Retrieval-Augmented Neural Field for HRTF Upsampling and Personalization. arXiv:2501.13017 [eess].

Mehta, A., Chauhan, S., Djanibekov, A., Kulkarni, A., Xia, G., and Choudhury, M. (2025). Music for All: Exploring Multicultural Representations in Music Generation Models. arXiv:2502.07328 [cs].

Müller, N., Kawa, P., Stan, A., Doan, T.-P., Jung, S., Choong, W. H., Sperl, P., and Böttinger, K. (2025). DeePen: Penetration Testing for Audio Deepfake Detection. arXiv:2502.20427 [cs].

Nam, H. and Park, Y.-H. (2025). JiTTER: Jigsaw Temporal Transformer for Event Reconstruction for Self-Supervised Sound Event Detection. arXiv:2502.20857 [eess].

Naman, A. and Zhang, G. (2025). FAST: Fast Audio Spectrogram Transformer. arXiv:2501.01104 [cs].

Nihal, R. A., Yen, B., Shi, R., and Nakadai, K. (2025). Weakly Supervised Multiple Instance Learning for Whale Call Detection and Localization in Long-Duration Passive Acoustic Monitoring. arXiv:2502.20838 [cs].

Pasini, M., Lattner, S., and Fazekas, G. (2025). Music2Latent2: Audio Compression with Summary Embeddings and Autoregressive Decoding. arXiv:2501.17578 [cs].

Pasquier, P., Ens, J., Fradet, N., Triana, P., Rizzotti, D., Rolland, J.-B., and Safi, M. (2025). MIDI-GPT: A Controllable Generative Model for Computer-Assisted Multitrack Music Composition. arXiv:2501.17011 [cs].

Petermann, D. and Kalayeh, M. M. (2025). Seeing Sound: Assembling Sounds from Visuals for Audio-to-Image Generation. arXiv:2501.05413 [cs].

Pinch, T. J. and Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science*, 14(3):399–441.

Pourmehrani, H., Hosseini, R., and Moradi, H. (2025). Water Flow Detection Device Based on Sound Data Analysis and Machine Learning to Detect Water Leakage. arXiv:2501.11151 [cs].

Puhle, C. (2025). Deep learning-based filtering of cross-spectral matrices using generative adversarial networks. arXiv:2502.21097 [cs].

Quan, M. K., Wijayasundara, M., Setunge, S., and Pathirana, P. N. (2025). Quantum-Enhanced Transformers for Robust Acoustic Scene Classification in IoT Environments. arXiv:2501.09394 [eess].

Quelennec, A., Chouteau, P., Peeters, G., and Essid, S. (2025). Masked Latent Prediction and Classification for Self-Supervised Audio Representation Learning. arXiv:2502.12031 [cs].

Rahimi, A., Afouras, T., and Zisserman, A. (2025). Reading to Listen at the Cocktail Party: Multi-Modal Speech Separation. arXiv:2501.01518 [eess].

Rho, K., Lee, H., Iverson, V., and Chung, J. S. (2025). LAVCap: LLM-based Audio-Visual Captioning using Optimal Transport. arXiv:2501.09291 [cs].

Rouard, S., Roman, R. S., Adi, Y., and Roebel, A. (2025). MusicGen-Stem: Multi-stem music generation and edition through autoregressive modeling. arXiv:2501.01757 [cs].

Roy, A., Liu, R., Lu, T., and Herremans, D. (2025). JamendoMaxCaps: A Large Scale Music-caption Dataset with Imputed Metadata. arXiv:2502.07461 [cs].

Seidel, E., Enzner, G., Mowlaee, P., and Fingscheidt, T. (2024). Neural Kalman Filters for Acoustic Echo Cancellation. *IEEE Signal Processing Magazine*, 41(6):24–38. arXiv:2501.16367 [eess].

Serrà, J., Araz, R. O., Bogdanov, D., and Mitsufuji, Y. (2025). Supervised contrastive learning from weakly-labeled audio segments for musical version matching. arXiv:2502.16936 [cs].

Sharma, F. and Gupta, P. (2025). Deepfake Detection of Singing Voices With Whisper Encodings. arXiv:2501.18919 [cs].

Shaulov, A., Shaharabany, T., Shaar, E., Chechik, G., and Wolf, L. (2025). Classifier-Guided Captioning Across Modalities. arXiv:2501.03183 [cs].

Shetty, M. V. and Kumar, Y. D. S. (2025). Audio-Based Classification of Insect Species Using Machine Learning Models: Cicada, Beetle, Termite, and Cricket. arXiv:2502.13893 [cs].

Singh, S., Benetos, E., Phan, H., and Stowell, D. (2025). LHGNN: Local-Higher Order Graph Neural Networks For Audio Classification and Tagging. arXiv:2501.03464 [cs].

Stiegler, B. (2018). *La Technique et le Temps*. Fayard.

Su, Y., Bai, J., Xu, Q., Xu, K., and Dou, Y. (2025). Audio-Language Models for Audio-Centric Tasks: A survey. arXiv:2501.15177 [cs].

Sulun, S., Viana, P., and Davies, M. E. P. (2025). Video Soundtrack Generation by Aligning Emotions and Temporal Boundaries. arXiv:2502.10154 [cs].

Sánchez, J. C. A., Comanducci, L., Pezzoli, M., and Antonacci, F. (2025). Towards HRTF Personalization using Denoising Diffusion Models. arXiv:2501.02871 [cs].

Tang, J., Cooper, E., Wang, X., Yamagishi, J., and Fazekas, G. (2025a). Towards An Integrated Approach for Expressive Piano Performance Synthesis from Music Scores. arXiv:2501.10222 [cs].

Tang, M., Li, J., Yang, L., Zhang, Z., Tian, J., Li, Z., Zhang, L., and Wang, P. (2025b). NOTA: Multimodal Music Notation Understanding for Visual Large Language Model. arXiv:2502.14893 [cs].

Tavares, T. F. and Ayres, F. J. (2025). Multi-label Cross-lingual automatic music genre classification from lyrics with Sentence BERT. arXiv:2501.03769 [cs].

Tian, H., Lattner, S., McFee, B., and Saitis, C. (2025a). Hybrid Losses for Hierarchical Embedding Learning. arXiv:2501.12796 [cs].

Tian, S., Zhang, C., Yuan, W., Tan, W., and Zhu, W. (2025b). XMusic: Towards a Generalized and Controllable Symbolic Music Generation Framework. arXiv:2501.08809 [cs].

Tjandra, A., Wu, Y.-C., Guo, B., Hoffman, J., Ellis, B., Vyas, A., Shi, B., Chen, S., Le, M., Zacharov, N., Wood, C., Lee, A., and Hsu, W.-N. (2025). Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound. arXiv:2502.05139 [cs].

Vuilliomenet, A., Balvanera, S. M., Aodha, O. M., Jones, K. E., and Wilson, D. (2025). acoupi: An Open-Source Python Framework for Deploying Bioacoustic AI Models on Edge Devices. arXiv:2501.17841 [cs].

Wang, W., Hou, R., Chang, H., Shan, S., and Chen, X. (2025a). MATS: An Audio Language Model under Text-only Supervision. arXiv:2502.13433 [cs].

Wang, Y., Wu, S., Hu, J., Du, X., Peng, Y., Huang, Y., Fan, S., Li, X., Yu, F., and Sun, M. (2025b). NotaGen: Advancing Musicality in Symbolic Music Generation with Large Language Model Training Paradigms. arXiv:2502.18008 [cs].

Watcharasupat, K. N., Ding, Y., Ma, T. A., Seshadri, P., and Lerch, A. (2025). Uncertainty Estimation in the Real World: A Study on Music Emotion Recognition. arXiv:2501.11570 [cs].

Watcharasupat, K. N. and Lerch, A. (2025). Separate This, and All of these Things Around It: Music Source Separation via Hyperellipsoidal Queries. arXiv:2501.16171 [eess].

Wei, H., Yuan, J., Zhang, R., Dai, Q., and Chen, Y. (2025). MAJL: A Model-Agnostic Joint Learning Framework for Music Source Separation and Pitch Estimation. arXiv:2501.03689 [cs].

Wen, C., Torfs, G., and Verhulst, S. (2025). Artifact-free Sound Quality in DNN-based Closed-loop Systems for Audio Processing. arXiv:2501.04116 [eess].

Wu, S., Guo, Z., Yuan, R., Jiang, J., Doh, S., Xia, G., Nam, J., Li, X., Yu, F., and Sun, M. (2025a). CLaMP 3: Universal Music Information Retrieval Across Unaligned Modalities and Unseen Languages. arXiv:2502.10362 [cs].

Wu, Y.-C., Marković, D., Krenn, S., Gebru, I. D., and Richard, A. (2025b). ComplexDec: A Domain-robust High-fidelity Neural Audio Codec with Complex Spectrum Modeling. arXiv:2502.02019 [eess].

Xiao, E., Cheng, H., Shao, J., Duan, J., Xu, K., Yang, L., Gu, J., and Xu, R. (2025). Tune In, Act Up: Exploring the Impact of Audio Modality-Specific Edits on Large Audio Language Models in Jailbreak. arXiv:2501.13772 [cs].

Xue, L., Zhou, Z., Pan, J., Li, Z., Fan, S., Ma, Y., Cheng, S., Yang, D., Guo, H., Xiao, Y., Wang, X., Shen, Z., Zhu, C., Zhang, X., Liu, T., Yuan, R., Tian, Z., Liu, H., Benetos, E., Zhang, G., Guo, Y., and Xue, W. (2025). Audio-FLAN: A Preliminary Release. arXiv:2502.16584 [cs].

Yao, S., Dai, J., Qin, X., Wang, S., Wang, S., Niu, K., and Zhang, P. (2025). SoundSpring: Loss-Resilient Audio Transceiver with Dual-Functional Masked Language Modeling. *IEEE Journal on Selected Areas in Communications*, pages 1–1. arXiv:2501.12696 [eess].

Yonay, O., Hammond, T., and Yang, T. (2025). Myna: Masking-Based Contrastive Learning of Musical Representations. arXiv:2502.12511 [cs].

Yun, J.-H., Kim, S.-B., and Lee, S.-W. (2025a). FLowHigh: Towards Efficient and High-Quality Audio Super-Resolution with Single-Step Flow Matching. arXiv:2501.04926 [eess].

Yun, S., Masukawa, R., Chen, H., Jeong, S., Huang, W., Rezvani, A., Na, M., Yamaguchi, Y., and Imani, M. (2025b). Hyperdimensional Intelligent Sensing for Efficient Real-Time Audio Processing on Extreme Edge. arXiv:2502.10718 [cs].

Zeng, D. and Ikeda, K. (2025). Metric Learning with Progressive Self-Distillation for Audio-Visual Embedding Learning. arXiv:2501.09608 [cs].

Zhang, J., Dinkel, H., Song, Q., Wang, H., Niu, Y., Cheng, S., Xin, X., Li, K., Wang, W., Wang, Y., and Luan, J. (2025a). The ICME 2025 Audio Encoder Capability Challenge. arXiv:2501.15302 [cs].

Zhang, J., Parada, P. P., Jalal, M. A., and Saravanan, K. (2025b). Diffusion based Text-to-Music Generation with Global and Local Text based Conditioning. arXiv:2501.14680 [eess].

Zhang, L., Zhao, Q., Wang, R., Bian, S., Gungor, O., Ponzina, F., and Rosing, T. (2025c). Offload Rethinking by Cloud Assistance for Efficient Environmental Sound Recognition on LPWANs. arXiv:2502.15285 [cs].

Zhang, Z., Liang, S., Shimada, D., and Xu, C. (2025d). Rethinking Audio-Visual Adversarial Vulnerability from Temporal and Modality Perspectives. arXiv:2502.11858 [cs].

Zhao, L., Chen, S., Feng, L., Zhang, X.-L., and Li, X. (2025). DualSpec: Text-to-spatial-audio Generation via Dual-Spectrogram Guided Diffusion Model. arXiv:2502.18952 [cs].

Zhou, Y., Wang, W., Ding, H., Xu, J., Zhu, J., Gao, X., and Li, S. (2025). SYKI-SVC: Advancing Singing Voice Conversion with Post-Processing Innovations and an Open-Source Professional Testset. arXiv:2501.02953 [cs].

Zhu, H., Zhou, Y., Chen, H., Yu, J., Ma, Z., Gu, R., Luo, Y., Tan, W., and Chen, X. (2025a). MuQ: Self-Supervised Music Representation Learning with Mel Residual Vector Quantization. arXiv:2501.01108 [cs].

Zhu, X., Tian, W., Wang, X., He, L., Wang, X., Zhao, S., and Xie, L. (2025b). CosyAudio: Improving Audio Generation with Confidence Scores and Synthetic Captions. arXiv:2501.16761 [eess].

Zuo, H., You, W., Wu, J., Ren, S., Chen, P., Zhou, M., Lu, Y., and Sun, L. (2025). GVMGen: A General Video-to-Music Generation Model with Hierarchical Attentions. arXiv:2501.09972 [cs].