SCHAEFFER: A DATASET OF HUMAN-ANNOTATED SOUND OBJECTS FOR MACHINE LEARNING APPLICATIONS

Maurizio Berta

Conservatorio 'Giuseppe Verdi' di Torino, Italy
KTH Royal Institute of Technology Stockholm, Sweden
mberta@kth.se

Daniele Ghisi

Conservatorio 'Giuseppe Verdi' di Torino, Italy
daniele.ghisi@conservatoriotorino.eu

ABSTRACT

Machine learning for sound generation is rapidly expanding within the computer music community. However, most datasets used to train models are built from field recordings, foley sounds, instrumental notes, or commercial music. This presents a significant limitation for composers working in acousmatic and electroacoustic music, who require datasets tailored to their creative processes. To address this gap, we introduce the SCHAEFFER Dataset (Spectromorphological Corpus of Human-annotated Audio with Electroacoustic Features For Experimental Research), a curated collection of 1000 sound objects designed and annotated by composers and students of electroacoustic composition. The dataset, distributed under Creative Commons licenses, features annotations combining technical and poetic descriptions, alongside classifications based on pre-defined spectromorphological categories.

1. INTRODUCTION

The application of machine learning techniques to sound generation is one of the fastest-growing areas of research within the computer music community. In particular, the recent surge of text-to-image generative models has sparked a parallel interest in text-to-audio models capable of translating a description into a fully formed audio file. Relevant examples of this type of model are Google's MusicLM [1] and AudioLM [2], Meta's MusicGen [3] and AudioGen [4], or commercial products such as Stable Audio [5], Suno [6], and Udio [7].

Most current applications focus on generating commercially usable audio - typically songs or sound effects. By contrast, little attention has been given to generating atomic units of sound rather than complete audio files. While this is admittedly a niche application, this type of 'atomic' sound generation is precisely what experimental electroacoustic composers need most.

A key theoretical framework for addressing this challenge is Pierre Schaeffer's concept of the 'sound object' [8]. There is a substantial body of research on how electronic music can be understood, decomposed, and recomposed through sound objects and their generalizations. [9, 10, 11, 12].

And yet, despite their analytical significance, these ideas have remained peripheral to the machine learning community — for some good reasons. While datasets of songs and sound effects are relatively easy to compile (leveraging existing commercial collections), assembling a dataset of musically significant sound objects is far more challenging. Extracting sound objects from existing compositions is difficult and, in many cases, impossible, as they are often layered or intertwined.

This work aims to establish a foundation for tackling this problem. We propose an innovative approach: constructing a dataset of annotated sound objects, from scratch, in collaboration with pedagogical institutions.

The notion that these objects are semantically meaningful units of sound – "equivalent to a unit of breath or articulation, a unit of instrumental gesture," as Pierre Schaeffer describes them [8] – is central to the philosophy of this project. In this sense, the database aligns with the Schaefferian tradition. To underscore this foundational idea, we named the dataset SCHAEFFER – an acronym for Spectromorphological Corpus of Human-annotated Audio with Electroacoustic Features For Experimental Research. SCHAEFFER comprises 1000 spectromorphologically annotated sound objects, released under Creative Commons licenses, and is primarily designed for use with state-of-the-art classification and regression techniques.

The structure of this article is as follows: Section 2 reviews publicly available datasets related to our work; Section 3 outlines the SCHAEFFER dataset, including its ontology, statistics, and data collection methodologies; Section 4 provides the necessary pointers to obtain the data; Section 5 presents some present and future use cases; Section 6 discusses the dataset's limitations and outlines future research directions.

2. RELATED WORK

Currently, there are several datasets of labelled audio available, each with their own scope, size, data quality, labelling methodology, and audio file format.

- Urbansound8k [13] is a dataset containing 8732 typologically labeled sound excerpts of urban sounds from 10 classes drawn from the urban sound taxonomy. All excerpts are taken from field recordings uploaded to Freesound and the target is sound event detection.
- Freesound Dataset 50k [14] (or FSD50K for short) is a dataset of typologically labeled sound events containing 51197 clips from Freesound unequally distributed in 200 classes drawn from the AudioSet Ontology. It is targeted at sound event detection.
- IRMAS [15] consists of 9579 audio files of 3 seconds from more than 2000 distinct recordings. It contains musical examples from different genres and various decades, and it was created with instrument recognition in mind.
- OrchideaSOL [16] is a dataset of 13265 samples, each containing a single musical note from one of 14 different instruments, containing many combinations of mutes

and extended playing techniques. It can be employed as a dataset for computer-aided orchestration as well as for instrument or playing technique recognition, and fundamental frequency estimation.

- AudioSet [17] consists of a collection of 2084320 human-verified automatic labels of 10-second sound clips drawn from YouTube videos. They are classified according to an ontology in which the event categories are hierarchically distributed.
- MusicCaps [1] is a subset of 10-second music clips from AudioSet, containing 5521 music examples, each of which is labeled with an English *aspect list* and a *free text caption* written by musicians. The text is solely focused on describing *how* the music sounds, not metadata like the artist's name.

None of these datasets is tailored to work on sound objects or to classify sounds according to spectromorphological characteristics. Furthermore, a number of these datasets (such as Urbansound8k, IRMAS, and portions of FSD50K) do not allow commercial usage, which also prevents many types of creative applications, such as the recording of a commercial album.

3. THE SCHAEFFER DATASET

To address these deficiencies, we have created the SCHAEFFER dataset. SCHAEFFER is a crowdsourced dataset of 1000 sound objects, described both with predefined labels and free-form text, and designed to be used primarily for experimental music practices.

Although a database of only 1000 sounds is admittedly small for today's standards, our focus was on the quality and experimental nature of the sounds, on the openness of the license, and the extensibility of the structure, in the hope that future additions may extend the existing base.

3.1. Framework and format

The framework of SCHAEFFER is inspired by Pierre Schaeffer's concept of a 'sound object' [8], in turn connected to the practice of 'reduced listening'. Reduced listening is achieved by repeated listening, which enables to focus on the intrinsic properties of a sound, disconnecting it from its context.

While accounting for many post-Schaefferian developments (such as Smalley's spectromorphology [12], the *unités sémiotique temporelles* [10], and Lasse Thoresen's Aural Sonology [11]), we have tried to adapt categories and labels to modern terminology and usage (including harmonizing them with popular online sample libraries).

Although it is impossible to impose theoretical limits on the duration of a sound object (for instance, short clicks or long drones can very well refer to the paradigm), for practical reasons (included, but not limited to, the ability to be ready for use with already developed machine learning algorithms), we had to introduce 'soft' constraints.

We decided to follow the line of work of MusicCaps, encouraging contributors to upload sounds between 5 and 10 seconds, while allowing them to also contribute with shorter or longer sounds.

Table 1: *Spectromorphological labels available in SCHAEFFER*

Property	Categories
Type	Soundscape, Drone, Chop, Sub, Glitch, Impact, Stab (Attack-Resonance), Synthesis, Vocal, Scratch, Crackle, Noise, Textural, Instrumental, Chirp, Percussive, Honk, Choral
Mass Type	Sinusoidal Sound, Harmonic Sound, Inharmonic Sound, Cluster Sound, Breathlike Sound, Noisy Sound, Composite or Stratified Sound, Combination of Harmonic Sounds
Complexity	Very Simple Element, Relatively Simple Element, Moderately Complex Element, Very Complex Element, Simple Emergence from Complex Details
Onset	Sharp Onset, Marked Onset, Flat Onset, Swelled Onset, Fade In
Sustain	Flat Sustain, Vacillating Sustain, Ostinato, Decaying Sustain, Uplifting Sustain, Iteration, Accumulation, No Sustain, Chaotic
Offset	Sharp ending, Sudden Stop, Flat ending, Soft ending, Laissez Vibrer, Interrupted Resonance, Fade Out
Pulse	Impulse, Regular Pulse Train, Irregular Pulse Train, Irregular Sporadic Pulses, No Pulse
Processes	Layering, Chorus, Tremolo, Distortion, Fuzzy, Granular, Loop, Bit-reduction, Reverb, Filtered, Resonators, Flanger, Pitch-shift, Stretched, Delay, Echo, Vibrato, Filter Modulation, Feedback
Direction	Fulfilled Forward Push, Evaded Forward Push, Suspended Forward Push, Backward Push, Neutral, Glissando Up, Glissando Down, Glissando Complex, Evanescent Appearance

3.2. Sound description

The sound objects in SCHAEFFER are described in two different ways:

Via predefined categories. We have identified a set of properties, and for each provided a number of predefined labels to choose from. The choice of a predefined label was made to ensure a coherent analysis, avoiding the use of synonyms or out-of-scope terms. The full taxonomy is displayed in Table 1. Each sound only allowed a single label per class (e.g. a sound object could have only one type of onset). Exceptions to this criterion were the "Type" and "Processes" classes, where multiple labels at once were permitted. Contributors were also allowed to include a set of additional, user-defined labels.

Via free-form captions. The second type of annotation of the sound objects in SCHAEFFER is free-form captions, in a similar style to MusicCaps. Contributors were, however, encouraged to provide descriptions relating both to a technical level (detailing the low-level electroacoustic features of the sound) and to a poetical level (describing the sound in more musical, and even metaphorical, terms).

An example of annotation is displayed in Listing 1.

```
{
"object" :          {
    "username" : "*",
    "filename" : "lo-mid_simpleDrone.wav",
    "description" : "A simple drone starting from
    low frequencies and expanding in the mid-upper
    range with a marked beating resonance. Subtly
    moving this sound object resembles what a sea
    diver feels when slowly moving towards
    the deep.",
    "labels" :                  {
        "type" : [ "Drone", "Synthesis" ],
        "mass-type" : "Inharmonic sound",
        "complexity" : "Relatively simple element",
        "onset" : "Swelled onset",
        "sustain" : "Vacillating sustain",
        "offset" : "Sudden stop",
        "pulse-typology" : "Irregular sporadic
        pulses",
        "processes" : "Resonators",
        "directionality" : "Evaded forward push"
    }
,
    "userlabels" : "*",
    "filelength(seconds)" : 9.659501133786847
}
}
```


Listing 1: *Example of JSON file containing the metadata for one of the sound objects in the SCHAEFFER dataset.*

3.3. Contributors

One of the fundamental principles of SCHAEFFER is its close integration with pedagogical objectives. Most of its contributors are students from Bachelor's and Master's programs at Italian conservatories (primarily the Turin Conservatory, as visible in Figure 1).

Students of the classes of Electroacoustic Composition and Analysis were tasked with submitting approximately 20 sound objects each — a requirement that served dual purposes. From a pedagogical perspective, crafting annotations and captions for one's own sound object is beneficial for self-analysis, and more challenging than it might appear, offering a valuable learning opportunity for aspiring musicians. Moreover, each student's unique background and artistic sensibility contributed to the rich diversity of the dataset.

At the same time, the contributions remained anchored to a shared reference paradigm — the post-Schaefferian framework — which students were already familiar with through their earlier composition and analysis courses.

3.4. Data collection

The data collection was carried out via a Max patch [18], which provided a simplified user interface for tagging, labeling, and uploading (see Figure 2).

A video tutorial was provided to contributors, and the necessary steps to follow were numbered in the Max patch for simplicity. As a result of the analysis, the patch produced a JSON file containing the relevant information, which was then sent to a Google Cloud Bucket for storage, using Node.js.

3.5. License

All sound files are released under a CC-BY license to both ensure rightful attribution to each contributor and allow for commercial and creative uses. The license choice was guided by Freesound guidelines for licensing audio. [19]

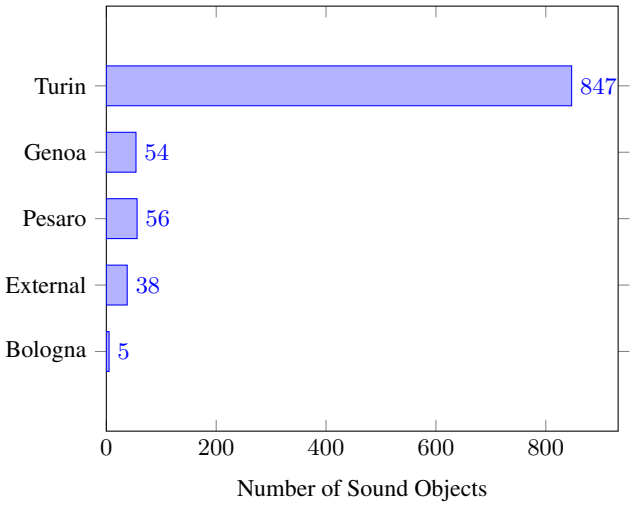

Figure 1: *Provenance of sound objects (distribution across Conservatories)*

If the dataset is used in its integrity, it is only necessary to attribute the dataset. Otherwise, if portions of SCHAEFER or single sounds are used, it is required to give credit to each author. The attribution should include the author's name (or username), a link to the original file, the CC-By license and a copyright notice.

3.6. Data Cleaning

The crowdsourced collection process allowed us to diversify the dataset and distribute the workload among several people. However, crowdsourcing carried out several problems like mislabelling and careless captions. To overcome this issue and ensure the consistency as well as the quality of the labels throughout the dataset, the sound objects' analyses were manually revised and corrected.

This revision process was practiced in three phases. In the first phase, all audio files were converted to the WAV format. In the next phase, duration issues were addressed; trailing silences were removed with a Python script, and files longer than 30 seconds were trimmed. Finally, in the third phase, all file labels were manually revised using a Max patch similar to the one used for uploading the files. As visible in Table 2, a large number of files were partially lacking categorical descriptions. To resolve this, we manually added missing labels, accepting empty categorical values only in specific cases. The most evident errors were also corrected, for example, *fade-in* marked as *sharp onset* or multiple selections of unique labels.

3.7. Data distribution

Figure 3 shows the distribution of sound objects according to their length. The mean duration is 7.86 seconds, but it is interesting to notice the bump between 9 and 10 seconds. This was clearly influenced by our guidelines, leading to the cropping of many textural or iterative objects, which in principle might have been infinite in length, to this duration. It is in any case interesting to notice, for future expansions of the project, that, ignoring textural and iterative sounds, a sort of 'natural' range for sound objects turns out to be between 0 and 15 seconds.

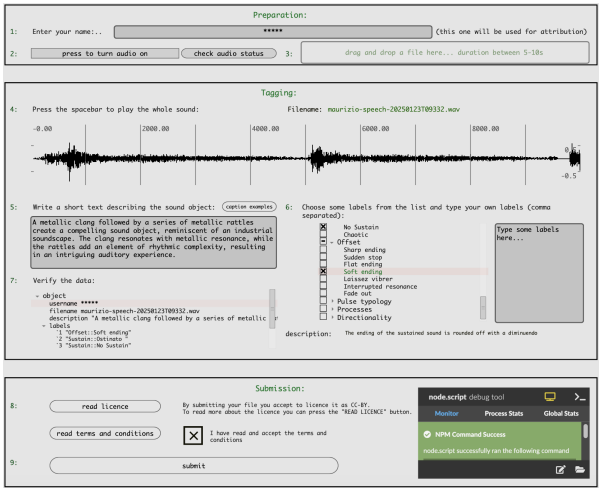

Figure 2: *The Max patch used to analyze and upload the sound objects.*

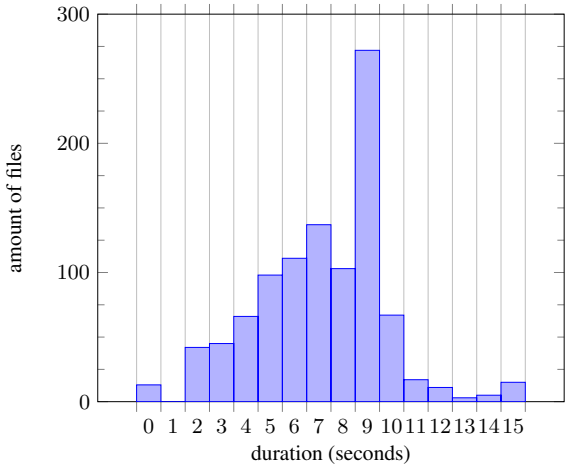

Figure 3: *Distribution of length of sound objects in SCHAEFFER.*

Table 2: *Comparison of missing data attributes before and after revision*

	Before correction	After correction
Type	55	4
Processes	252	131
Mass Type	165	0
Directionality	398	0
Onset	266	0
Sustain	211	0
Offset	263	0
Pulse Typology	291	1
Complexity	111	0

Table 3 presents the distribution of categories in SCHAEFFER, emphasizing the class imbalance - some categories are well-represented, while others do not appear frequently.

Figure 4 shows a word cloud with the most relevant terms used in the free-form description of the audio files.

4. DATA AVAILABILITY

The SCHAEFFER dataset is accessible on Kaggle[1] and Huggingface[2]. A GitHub page was also created, including the uploading app, the correction app, and the Python scripts. (The page also contains interactive notebooks that tackle text-to-audio starting from the SCHAEFFER dataset)[3]

5. USE CASES

We tested the dataset by training a text-to-audio machine learning model. We started from the Riffusion's model checkpoint [20], a

Figure 4: *Most recurrent free-form description terms in SCHAEFFER.*

[1] https://www.kaggle.com/datasets/maurizioberta/test-schaeffer/

[2] https://huggingface.co/datasets/dbschaeffer/SCHAEFFER

[3] https://github.com/mauriziobrt/SCHAEFFER

Table 3: *Sound Element Categorization and Distribution (long names are abbreviated)*

Typology		**Processes**		**Mass Type**		**Pulse**		**Complexity**	
Synthesis	468	Filtered	307	Composite	352	Irreg. pulse	286	Rel. simple	429
Textural	257	Granular	269	Inharmonic	253	No pulse	232	Mod. complex	317
Noise	238	Layering	242	Noisy	160	Reg. pulse	192	Very simple	145
Glitch	226	Distort	217	Harmonic Comb.	106	Irreg. sporadic	149	Very complex	84
Instrument	198	Reverb	216	Harmonic	91	Impulse	141	Simple emerg.	25
Percussive	187	Pitch-shift	167	Cluster	19				
Drone	126	Stretched	145	Sinusoidal	16				
Soundscape	119	Filter Mod.	81	Breathlike	13				
Impact	97	Delay	78						
Crackle	96	Resonators	65						
Sub	73	Fuzzy	63						
Vocal	68	Loop	60						
Stab	47	Tremolo	52						
Chop	29	Chorus	33						
Choral	29	Feedback	31						
Scratch	29	Bit reduction	28						
Chirp	28	Echo	20						
Honk	4	Vibrato	20						
		Flanger	7						
Onset		**Sustain**		**Offset**		**Directionality**			
Marked	336	Iteration	328	Soft ending	270	Neutral	425		
Swelled	222	Vacillating	202	Fade out	220	Forward	165		
Sharp	198	Decaying	157	Sudden stop	185	Backward	134		
Fade in	135	Flat	90	Flat ending	136	Suspended	93		
Flat	109	Chaotic	76	Laissez vib.	105	Evaded	76		
		Uplifting	59	Sharp end	74	Gliss. Complex	33		
		Ostinato	40	Interrupted	12	Evanescent	31		
		Accumulation	39			Gliss. Up	22		
		No Sustain	11			Gliss. Down	21		

fine-tuning of Stable Diffusion specifically made for music generation, and further fine-tuned the model on our dataset using LoRA [21]. This model was trained solely on pairs of audio and captions. Training the model took around 14 hours on a local NVIDIA RTX 3060 using CUDA acceleration.

During inference, the model generated typologically coherent material if presented with words that are common in the dataset, for example "pulsar synthesis". Training using LoRA didn't overwrite most of the pre-learned material. This caused the model to hallucinate often, producing sounds similar to a standard Riffusion generation. The model didn't manage to produce morphologically coherent sounds from a semantic description. Cherry-picked audio examples are available on GitHub, as well as training and inference Python notebooks.

This model may serve as a baseline for exploring audio generation from spectromorphological descriptions using the SCHAEFFER dataset. While the results were influenced by limitations in dataset size, training time, and computational resources, we see significant potential for future applications. These include semantically conditioned audio sculpting and automatic audio classification.

6. CONCLUSIONS AND FUTURE WORK

While 1000 annotated sounds provide a useful resource for fine-tuning, they are insufficient for large-scale applications. The limited size of the dataset resulted in class imbalance, with certain labels being underrepresented. Future expansions could address these limitations by increasing both the quantity and diversity of sounds. Broadening the dataset's scope beyond Italian conservatories to include contributions from music schools and conservatories worldwide would enhance its representativeness and usability. This should be the primary focus for a future second phase of the project. At the same time, specific terminology issues also need refinement. Some labels were ambiguous or difficult to interpret without auditory examples. This is particularly evident in the dynamic morphology categories (*onset, sustain, offset*), where participants often labeled a *'fade-in' as a 'swelled onset'*. Similarly, in the offset category, terms like *'fade out'*, *'soft ending'*, *'laissez vibrer'*, and, in rare cases, *'flat ending'* were used interchangeably, highlighting inconsistencies in classification.

There are two possible solutions to address label ambiguity. The first is to merge similar labels, simplifying machine learning training by reducing data complexity. However, this approach may also diminish the descriptive richness of sound objects, making them less meaningful for analysis. The second solution is to enhance the interface by incorporating graphical representations of morphological labels. While this could improve clarity, an overly dense interface might hinder usability and make the tagging process less intuitive. Striking a balance between these approaches could lead to a more efficient tagging process and a more consistent dataset for analysis.

Another issue observed after the collection phase is the lack of certain labels. For example, whistling or audio-mangling labels were lacking in the typology category. A thoughtful consideration regarding the enhancement of labels should be the baseline for future iterations of the dataset.

7. ACKNOWLEDGMENTS

The authors sincerely thank the students and composers who participated in the creation of the dataset, as well as their teachers and supervisors for their invaluable support. Maurizio Berta's work has been partially funder by a grant from the Swedish Research Council (Grant VR 2023-04496) - "Non-verbal communication of invisible information: New methods for communicating complex

and inaccessible information through sound".

8. REFERENCES

[1] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "Musiclm: Generating music from text," 2023.

[2] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour, "Audiolm: a language modeling approach to audio generation," 2022.

[3] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," 2023.

[4] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "Audiogen: Textually guided audio generation," 2023.

[5] "Stable audio," https://stability.ai/stable-audio.

[6] "Suno," https://suno.com/.

[7] "Udio | AI Music Generator - Official Website," https://www.udio.com.

[8] Pierre Schaeffer, *Treatise on Musical Objects: An Essay across Disciplines*, University of California Press, Apr. 2019.

[9] Michel Chion, "Guide des objets sonores: Pierre schaeffer et la recherche musicale," 1983.

[10] François Delalande, *Analyser la musique, pourquoi, comment ?*, chapter 10, pp. 170–181, Ina Expert. Ina Editions, 2013.

[11] Lasse Thoresen, Andreas Hedman, James Grier, University of Western Ontario Department of Music Research, and Composition Don Wright Faculty of Music, *Emergent musical forms : aural explorations*, Department of Music Research and Composition, Don Wright Faculty of Music, University of Western Ontario London, Ontario, London, Ontario, 2015.

[12] Denis Smalley, "Spectromorphology: explaining sound-shapes," *Organised Sound*, vol. 2, no. 2, pp. 107–126, Aug. 1997.

[13] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[14] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra, "Freesound datasets: A platform for the creation of open audio datasets," in *International Society for Music Information Retrieval Conference*, 2017.

[15] Juan J. Bosch, Ferdinand Fuhrmann, and Perfecto Herrera, "Irmas: a dataset for instrument recognition in musical audio signals," June 2018.

[16] Carmine Emanuele Cella, Daniele Ghisi, Vincent Lostanlen, Fabien Lévy, Joshua Fineberg, and Yan Maresz, "Orchideasol: a dataset of extended instrumental techniques for computer-aided orchestration," 2020.

[17] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. mar 2017, IEEE.

[18] Miller Puckette, "Max at seventeen," *Computer Music Journal*, vol. 26, no. 4, pp. 31–43, Dec. 2002.

[19] "Freesound - Help - Frequently Asked Questions," https://freesound.org/help/faq/#licenses.

[20] Seth* Forsgren and Hayk* Martiros, "Riffusion - Stable diffusion for real-time music generation," 2022.

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 2021.