MACHINE LISTENING IN A NEONATAL INTENSIVE CARE UNIT

Modan Tailleur[1*], *Vincent Lostanlen*[1†], *Jean-Philippe Rivière*[1], *Pierre Aumond*[2]

[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[2] Université Gustave Eiffel, CEREMA, UMRAE, F-44344 Bouguenais, France

ABSTRACT

Oxygenators, alarm devices, and footsteps are some of the most common sound sources in a hospital. Detecting them has scientific value for environmental psychology but comes with challenges of its own: namely, privacy preservation and limited labeled data. In this paper, we address these two challenges via a combination of edge computing and cloud computing. For privacy preservation, we have designed an acoustic sensor which computes third-octave spectrograms on the fly instead of recording audio waveforms. For sample-efficient machine learning, we have repurposed a pretrained audio neural network (PANN) via spectral transcoding and label space adaptation. A small-scale study in a neonatological intensive care unit (NICU) confirms that the time series of detected events align with another modality of measurement: i.e., electronic badges for parents and healthcare professionals. Hence, this paper demonstrates the feasibility of polyphonic machine listening in a hospital ward while guaranteeing privacy by design.

Index Terms— Computational environmental audio analysis, edge computing, machine learning methods, privacy.

1. INTRODUCTION

Sound is a reliable and non-invasive carrier of information about human health [1]. Yet, historically, the subfield of medical acoustics has mainly focused on analyzing sounds as *produced* by patients: stutter [2], crackles [3], cough [4], and so on. Much less is known about the sounds as *heard* by patients in a clinical setting: as experimental psychologists have pointed out, the detailed description of acoustic events in intensive care units (ICU's) is typically overlooked in favor of sound pressure level measurements (SPL) [5]. Meanwhile, exposure to anthropogenic noise at unsafe SPL levels is known to induce stress, cognitive impairment and sleep disorders in children [6] and adults [7], thus calling for urgent remediation.

The case of neonatal intensive care units (NICU's), where premature babies receive special care to grow and survive, presents an even greater gap in research than adult ICU's [8]. During their time in the NICU, preterm infants are exposed to unpredictable sensory stimuli while undergoing a protracted period of rapid brain growth, causing lasting effects on cognitive ability [9]. Unfortunately, the auditory physiology and cognition of neonates have received insufficient attention from scientists until recently [10].

What is known with certainty is that parents have an essential role to play in the development of their newborn babies [11]. Indeed, an approach sometimes described as "kangaroo care" involves prolonged periods of skin-to-skin contact between the baby and either of the two parents, in addition to incubator placement. Pro-
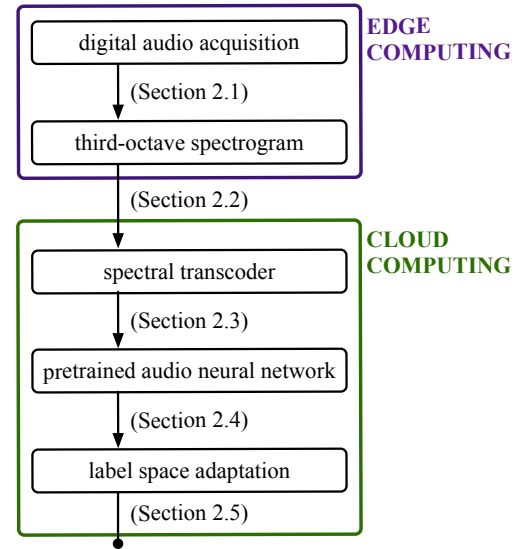
Figure 1: Flowchart of stages in the proposed approach. The first two stages are performed "on the edge". The last three stages are performed "on the cloud", i.e., on a central server.

moting this approach requires to take the well-being of parents into consideration so that they feel included into collective care work.

For this purpose, we have launched a project on "listening to family experiences in the neonatological ward", or LIFEWARD for short. Here, the word "listening" is understood as both qualitative and quantitative: i.e., as enacted by interviews with parents as well as autonomous acoustic sensors. Although there is scientific consensus around the value of semi-structured interviews for neonatology—see, for example, [12]—the same cannot be said about machine listening. This is for at least three reasons. First, the deployment of acoustic sensors in a hospital raises pressing concerns about privacy preservation and cybersecurity. Secondly, the application of machine learning to the NICU is not straightforward, for lack of annotated training data. Thirdly, and perhaps most fundamentally, machine listening systems have not yet demonstrated their ability to reconstruct objective information about social bonds in the NICU. Addressing these three challenges is necessary before envisioning the integration of machine listening instruments within the toolkit of patient experience research.

In this article, we present a proof of feasibility of machine listening for neonatology. Prior work in this domain has focused on a single class of sound event—namely, the spontaneous cries of preterm newborns [13]. Meanwhile, our system is a multilabel

sound event detector for adult voices, footsteps, oxygenators, and alarm devices. Furthermore, the originality of our approach is that it integrates all aspects of machine listening, from digital audio acquisition to sound event detection, into a mixed pipeline that involves both edge computing and cloud computing.

Figure 1 decomposes our approach into five stages, described in Sections 2.1 through 2.5. To comply with standards of privacy and security, the LIFEWARD sensor does not store the acquired waveforms. Rather, it extracts a third-octave spectrogram on the fly; i.e., a coarse estimate of power spectral density over windows of duration 125 ms. This, in turn, brings its own challenges for sound event detection, which typically requires finer spectrotemporal information. We address this challenge via a pretrained "spectral transcoder"; i.e., a deep neural network for nonuniform resampling the time–frequency domain. We pass the output of the spectral transcoder to a pretrained audio neural network (PANN), with AudioSet as its label space. Lastly, we use domain-specific knowledge to narrow down this taxonomy for the NICU.

Section 3 presents the result of an in-progress study at an NICU, and provides tentative evidence for the feasibility of the proposed approach. Indeed, neural network predictions appear to coincide with isolated sound events of interest (Section 3.1) as well as times-tamps from a non-audio modality of human presence (Section 3.2).

2. METHODS

2.1. Acoustic sensor

Our acoustic sensor is a Raspberry Pi, inspired by previous work on urban noise monitoring [14]. It acquires audio from an external USB microphone, specifically, the micW i436. The i436 is an omnidirectional electret microphone with a capsule diameter of approximately 7 mm, in compliance with NF EN 61672 Class-2 standards. Its sensitivity and frequency response has been calibrated manually by the manufacturer. After digital–analog conversion, the sample rate is 32 kHz. Our sensor is powered by the grid and "airgapped", i.e., physically isolated from the public Internet and from each other. This is to reduce the risk of malicious data access.

2.2. Third-octave spectrogram

We use fast Fourier transforms (FFT) to design a third-octave filterbank with bands ranging from 20 Hz to 12.5 kHz, in compliance with the ANSI S1.1-1986 and IEC 61260-1:2014 standards [15]. We extract the magnitude response of each filter over non-overlapping subbands of duration 125 ms. These operations are implemented in the C language, compiled for the Raspberry Pi, and executed in real time. The result is stored incrementally on a non-volatile memory ("SD card").

A perceptual evaluation on twelve subjects has shown that the third-octave spectrogram does not contain sufficient information to recover intelligible speech, at least via classical signal processing techniques—namely, Moore-Penrose pseudoinverse and Griffin-Lim algorithm for phase retrieval [15]. Thus, the third-octave spectrogram representation can be said to be privacy-aware, in the sense it mitigates the severity of a security breach should the SD card were to be lost or stolen in the healthcare facility.

Another advantage of computing third-octave spectrograms on the edge resides in its bitrate: around 3.71 kilobytes per second (kbps). This is lower than MP3 (128–320 kbps) and lossless audio (around 1 Mbps). The bitrate of third-octave spectrograms translates to around 320 megabytes per day, or 117 gigabytes per year.

Thus, a single SD card suffices to contain all the spectrogram data over a longitudinal survey spanning the full length of stay of the preterm infant at the NICU.

2.3. Spectral transcoder

Previous work in urban environments has shown the potential of the third-octave spectrogram as a feature for sound event classification, both in supervised and self-supervised scenarios [16]. Yet, this previous work is unapplicable in the context of the NICU, for lack of annotated training data. Furthermore, note that it would not be possible to launch our own annotation campaign because, as explained before, our sensors do not record audio. We propose to circumvent this problem by relying on a pretrained audio neural network (PANN) for multilabel sound event detection and classification [17].

Here, a second issue arises: PANN does not operate upon the third-octave spectrogram but on a mel-frequency spectrogram, which has a finer temporal resolution (hop size of 10 ms) and a finer spectral resolution (64 bins on the mel scale). In principle, the required change of resolution could be achieved by a linear non-uniform resampler. Yet, in practice, this produces a blurry time–frequency representation which is not recognized by PANN as containing any events of interest. Against this issue, a deep neural network was developed by Tailleur et al. [18], which we call *spectral transcoder*, so as to recover a plausible mel-frequency spectrogram from a third-octave spectrogram measurement.

The spectral transcoder is a convnet with six layers. It is trained on TAU Urban Acoustic Scenes 2020 Mobile dataset [19] in a "teacher–student" scenario. The teacher is the composition of mel-frequency spectrogram and PANN whereas the student is the composition of third-octave spectrogram, spectral transcoder, and PANN. In other words, the spectral transcoder is not trained to minimize its mean square error with the mel-frequency ground truth (as a linear model would) but to generate a mel-frequency spectrogram whose spectrotemporal content has the same distribution of sound events as the ground truth. The training process involves minimizing a binary cross-entropy loss, computed between the PANN output of the student and that of the teacher, by updating solely the transcoder's parameters. This is a kind of super-resolution procedure in which the implicit knowledge about the spectrotemporal characteristics of natural audio sounds is distilled from PANN into the spectral transcoder under the form of convnet weights. We refer to [18] for more details on the spectral transcoder.

2.4. AudioSet classification with PANN

Our PANN of choice is a residual network with 38 layers, or ResNet38 for short. It contains around 74M parameters. To this day, it is regarded as one of the most accurate general-purpose multilabel audio classifier among those which take the mel-frequency spectrogram as input. The PANN is trained on AudioSet, a dataset which contains over 2M 10-second audio clips which were extracted from YouTube videos. In the next sections, we refer to the composition of pretrained spectral transcoder and PANN as "PANN-1/3oct" model. We refer to [17] for further details on PANN.

Since PANN is a multilabel classifier, its output vector is unnormalized. For the sake of visualization, we have found it beneficial to rank predictions in decreasing order, normalize rank by the number of classes, and apply an inverse power transform. Given predictions $x[k]$ for each class k, this procedure yields the α-compressed

reciprocal rank

$$\boldsymbol{y}[k] = \left(\frac{K}{\sigma^{-1}[k]} \right)^{\alpha}, \qquad (1)$$

where k is the class index, K is the total number of classes, σ is the sorting permutation such that $(\boldsymbol{x} \circ \sigma)[1] > \ldots > (\boldsymbol{x} \circ \sigma)[K]$, and $\alpha < 1$ is a constant exponent. We set $\alpha = 0.5$ in this paper.

2.5. Label space adaptation

The PANN-1/3oct model analyzes a third-octave spectrogram snippet of duration equal to 10 seconds and returns a vector of dimension 527, corresponding to the classes in AudioSet dataset. These classes are a subset of the 623-class AudioSet ontology[1], which has been defined by a Google Research team after scraping web-scale text data for "Hearst patterns", i.e., of either of these forms [20]:

[...] sounds such as X or Y [...]

[...] X, Y, and other sounds [...]

This approach has proven fruitful for general-purpose audio classification: we refer to [21] for a review. Yet, it is unsuitable for the ICU, whose distribution of sound events is inadequately represented by textual mentions of sound events on the web. At the same time, training a classifier from scratch on a new taxonomy is out of the question for reasons of privacy preservation, as explained earlier.

Instead, we simply run the PANN-1/3oct model on third-octave spectrogram data from the NICU and look for some frequently occuring AudioSet classes. We find four activity patterns of interest: "conversation", "walk, footsteps", "train", and "electronic music". Although the former two sound events are plausible, the latter two are clearly not. Yet, after interviewing NICU employees, we may hypothesize that they yield indirect information: i.e., that "train" actually corresponds to the rumble of the oxygenator while "electronic music" corresponds to the ringtone of the hospital phone. We summarize this correspondence in Table 1.

Neonatal Intensive Care Unit (NICU)	AudioSet
Conversation	Conversation
Footsteps	Walk, footsteps
Oxygenator	Train
Hospital phone	Electronic music

Table 1: Mapping of sound event labels from the neonatal intensive care unit (NICU) to AudioSet.

3. APPLICATION

3.1. Deployment in a neonatal intensive care unit

Since 2018, a design company[2] have been partnering with Nantes University hospital and a nonprofit organization[3] to enhance the inclusion of parents in the NICU. The nonprofit organization collaborated with designers to refurnish a care room so as to facilitate the presence of parents alongside their newborn. In this context, the LIFEWARD sensor has offered the necessary guarantees for a safe and privacy-aware deployment in the NICU. We have obtained the

---

[1]Link to complete list of classes in the AudioSet dataset:
https://research.google.com/audioset/dataset/index.html

[2]Sensipode

[3]the B.E.R.S.E association

approval of an ethical review board to deploy this sensor[4]. Six families have given their informed consent to participate in the LIFEWARD study: three in the aforementioned redesigned room and three in a standard room. The length of stay is approximately 90 days for each family. Thus, we have collected third-octave spectrogram data over 18 cumulated months.

3.2. Visualization of sound events

We now collect a few waveform-domain samples from the sound events of interest in a real NICU environment. This data collection stage is carried out with a handheld device, over short durations, and with the collaboration of NICU professionals. Specifically, we ring various kinds of alarms, activate oxygenators and other pumps, stomp our feet, and so forth. Admittedly, these sounds are too few to offer an independent quantitative evaluation of PANN-1/3oct: we refer to [18] for that matter. Still, they may serve as suggestive evidence for the fact that the correspondences which we hypothesized in Table 1 are adequate and useful in practice.

Figure 2 illustrates our findings for each of the four classes of interest. For example, we notice vertical patterns of high energy in the recording of footsteps, versus horizontal patterns in the recordings of oxygenator. These simple observations corroborate the prediction of the PANN-1/3oct model with label space adaptation: see Table 1. Those examples demonstrate one of the advantages of using the transcoder: one can double-check model predictions by displaying the transcoded spectrogram, despite not being able to listen to the underlying audio waveform.

3.3. Proof of feasibility for continuous monitoring

The previous section has confirmed the interest of the PANN-1/3oct model in the context of isolated sounds from the NICU, as acquired by a handheld device. It remains to be seen if this model remains informative in a real-world polyphonic context, as acquired by the LIFEWARD sensor. For this purpose, we propose to compare the detected events with another modality of measurement: i.e., electronic badges worn by parents and healthcare professionals. Via near-field communication (NFC), these badges yield information about who is present in the care room at any given time. Hence, they offer indirect confirmation for the feasibility of machine listening in the NICU, while remaining non-invasive and privacy-aware.

Figure 3 shows an example of PANN-1/3oct predictions from our real-world NICU dataset, together with timestamps from electronic badges. We notice that segments during which two adults are present in the room coincide with a rise in the presence of conversation—and, to a lesser degree, of footsteps. Meanwhile, the lowest values for the "conversation" class correspond to segments in which only one adult is present in the room. Yet, we recognize that these are only anecdotal observations. Future research is needed to expand the comparison of acoustical and non-acoustical information to a larger scale; i.e., multiple days and multiple rooms.

4. CONCLUSION

The DCASE community has a key role to play at the intersection between sound design and healthcare. Yet, fulfilling this role comes with challenge of its own, such as: privacy, cybersecurity, and limited labeled data. In this article, we have presented a first prototype

---

[4]Groupe Nantais d'Éthique dans le Domaine de la Santé (GNEDS) - n°22-09-090
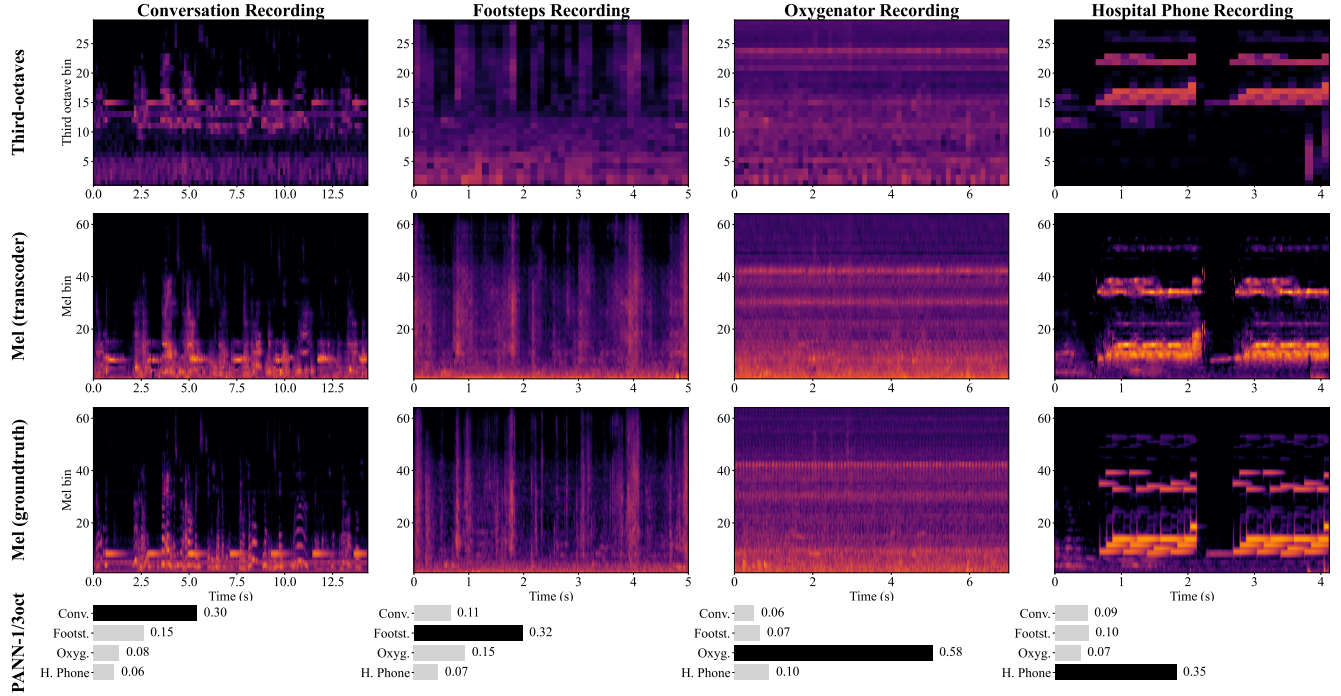
Figure 2: Spectrograms from audio recordings in the Neonatal Intensive Care Unit (NICU). First row corresponds to audio recordings transformed into fast third-octaves spectrograms. Second row corresponds to Mel spectrograms transcoded with the transcoder. Third row corresponds to groundtruth Mel spectrograms, obtained with Mel transformation on the waveform. PANN-1/3oct predictions, using the mapping between Audioset classes and NICU classes, are shown in fourth row.
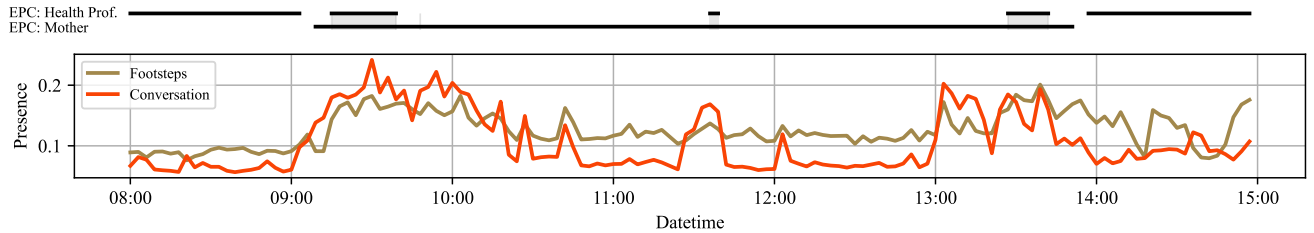

Figure 3: Presence of conversation and footsteps on a day of April 2023 in one room, as averaged over three-minute intervals. The badge of the health professional (EPC) and of the mother are also shown during the period. The shaded areas denote intervals in which more than one adult is present in the room.

of acoustic sensor which demonstrates the feasibility of sound event detection in a neonatal intensive care unit (NICU). The main limitation of our study is that, because our sensor does not record audio as waveforms, it is impossible to establish a "ground truth" by expert annotation. We have circumvented this limitation in two way: first, by evaluating the system on well-controlled isolated sounds; and second, by matching the sequence of detected sound events with non-acoustical information. In the future, we plan to refine the integration of multiple data modalities towards a more comprehensive understanding of patients and their lived experiences.

5. ACKNOWLEDGMENT

6. REFERENCES

[1] K. W. Beach and B. Dunmire, "Medical acoustics," in *Springer Handbook of Acoustics*. Springer, 2014, pp. 877–937.

[2] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "Machine learning for stuttering identification: Review, challenges and future directions," *Neurocomputing*, vol. 514, pp. 385–402, 2022.

[3] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PLOS ONE*, vol. 12, no. 5, p. e0177926, 2017.

[4] K. S. Alqudaihi, N. Aslam, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. A. R. Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith, H. M. Alharthi, W. M. Alghamdi, and M. S. Alshahrani, "Cough Sound Detection and Diagnosis Using Artificial Intelligence Techniques: Challenges and Opportunities," *IEEE Access*, vol. 9, pp. 102 327–102 344, 2021.

[5] J. Mackrill, "Experiencing the hospital ward soundscape: Towards a model," *Journal of Environmental Psychology*, vol. 36, pp. 1–8, 2013.

[6] A. L. Bronzaft, "Supporting healthier urban environments with a sound and noise curriculum for students," *Cities & health*, vol. 5, no. 1-2, pp. 118–121, 2021.

[7] E. Murphy and E. A. King, *Environmental noise pollution: Noise mapping, public health, and policy*. Elsevier, 2022.

[8] S. Lenzi, S. Spagnol, and E. Özcan, "Improving the quality of the acoustic environment in neonatal intensive care units: A review of scientific literature and technological solutions," *Frontiers in Computer Science*, vol. 5, 2023.

[9] L. Gray and M. K. Philbin, "Effects of the neonatal intensive care unit on auditory attention and distraction," *Clinics in perinatology*, vol. 31, no. 2, pp. 243–260, 2004.

[10] M. K. Philbin, "The sound environments and auditory perceptions of the fetus and preterm newborn," *Early vocal contact and preterm infant brain development: Bridging the gaps between research and practice*, pp. 91–111, 2017.

[11] J. Baley, C. on Fetus, Newborn, K. Watterberg, J. Cummings, E. Eichenwald, B. Poindexter, D. L. Stewart, S. W. Aucott, K. M. Puopolo, and J. P. Goldsmith, "Skin-to-skin care for term and preterm infants in the neonatal icu," *Pediatrics*, vol. 136, no. 3, pp. 596–599, 2015.

[12] I. M. F. Medina, J. Granero-Molina, C. Fernández-Sola, J. M. Hernández-Padilla, M. C. Ávila, and M. d. M. L. Rodríguez, "Bonding in neonatal intensive care units: Experiences of extremely preterm infants' mothers," *Women and Birth*, vol. 31, no. 4, pp. 325–330, 2018.

[13] S. Cabon, B. Met-Montot, F. Porée, O. Rosec, A. Simon, and G. Carrault, "Automatic extraction of spontaneous cries of preterm newborns in neonatal intensive care units," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1200–1204.

[14] J. Ardouin, L. Charpentier, M. Lagrange, F. Gontier, N. Fortin, D. Ecotière, J. Picaut, and C. Mietlicky, "An innovative low-cost sensor for urban sound monitoring," in *Proceedings of Inter-Noise*, vol. 258, no. 5. Institute of Noise Control Engineering, 2018, pp. 2226–2237.

[15] F. Gontier, M. Lagrange, P. Aumond, A. Can, and C. Lavandier, "An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach," *Sensors*, vol. 17, no. 12, p. 2758, 2017.

[16] F. Gontier, V. Lostanlen, M. Lagrange, N. Fortin, C. Lavandier, and J.-F. Petiot, "Polyphonic training set synthesis improves self-supervised urban sound classification," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 4309–4326, 2021.

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

[18] M. Tailleur, M. Lagrange, P. Aumond, and V. Tourre, "Spectral trancoder: Using pretrained urban sound classifiers on undersampled spectral representations," in *Proceedings of the International Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023.

[19] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the International Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2018.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, 2017.

[21] T. Pellegrini, I. Khalfaoui-Hassani, E. Labbé, and T. Masquelier, "Adapting a ConvNeXt model to audio classification on AudioSet," in *Proceedings of the International Speech Association Conference (INTERSPEECH)*, 2023.