electronics

MDPI

Article

Audio Recognition of the Percussion Sounds Generated by a 3D Auto-Drum Machine System via Machine Learning

Spyros Brezas [1,2,*], Alexandros Skoulakis [3,4], Maximos Kaliakatsos-Papakostas [1], Antonis Sarantis-Karamesinis [1,4], Yannis Orphanos [1,2,4], Michael Tatarakis [3,4], Nektarios A. Papadogiannis [1,2,4], Makis Bakarezos [1,2,4], Evaggelos Kaselouris [1,2,4] and Vasilis Dimitriou [1,2,4]

1 Department of Music Technology and Acoustics, Hellenic Mediterranean University, Perivolia, 74133 Rethymnon, Greece; maximoskp@hmu.gr (M.K.-P.); mta52@edu.hmu.gr (A.S.-K.); yorphanos@hmu.gr (Y.O.); npapadogiannis@hmu.gr (N.A.P.); bakarezos@hmu.gr (M.B.); vagfem@hmu.gr (E.K.); dimvasi@hmu.gr (V.D.)
2 Physical Acoustics and Optoacoustics Laboratory, Hellenic Mediterranean University, Perivolia, 74133 Rethymnon, Greece
3 Department of Electronic Engineering, Hellenic Mediterranean University, 73133 Chania, Greece; skoulakis@hmu.gr (A.S.); mictat@hmu.gr (M.T.)
4 Institute of Plasma Physics and Lasers-IPPL, University Research and Innovation Centre, Hellenic Mediterranean University, 74100 Rethymno, Greece
* Correspondence: sbrezas@hmu.gr

Abstract: A novel 3D auto-drum machine system for the generation and recording of percussion sounds is developed and presented. The capabilities of the machine, along with a calibration, sound production, and collection protocol are demonstrated. The sounds are generated by a drumstick at pre-defined positions and by known impact forces from the programmable 3D auto-drum machine. The generated percussion sounds are accompanied by the spatial excitation coordinates and the correspondent impact forces, allowing for large databases to be built, which are required by machine learning models. The recordings of the radiated sound by a microphone are analyzed using a pre-trained deep learning model, evaluating the consistency of the physical sample generation method. The results demonstrate the ability to perform regression and classification tasks when fine tuning the deep learning model with the gathered data. The produced databases can properly train machine learning models, aiding in the investigation of alternative and cost-effective materials and geometries with relevant sound characteristics and in the development of accurate vibroacoustic numerical models for studying percussion instruments sound synthesis.

Keywords: cymbal sounds; sound database; audio recognition; machine learning

Citation: Brezas, S.; Skoulakis, A.; Kaliakatsos-Papakostas, M.; Sarantis-Karamesinis, A.; Orphanos, Y.; Tatarakis, M.; Papadogiannis, N.A.; Bakarezos, M.; Kaselouris, E.; Dimitriou, V. Audio Recognition of the Percussion Sounds Generated by a 3D Auto-Drum Machine System via Machine Learning. *Electronics* **2024**, *13*, 1787. https://doi.org/10.3390/electronics13091787

Academic Editor: Chang Wook Ahn

Received: 28 March 2024
Revised: 29 April 2024
Accepted: 2 May 2024
Published: 6 May 2024

1. Introduction

Percussion instruments are musical instruments that produce sound when struck by a percussion mallet, beater, or hand, or when scraped, rubbed, or struck against another similar instrument [1]. The percussion section of an orchestra usually consists of instruments such as the timpani, snare drum, bass drum, tambourine (membranophones), cymbals, and triangle (idiophones). Beyond the orchestral setting, percussion instruments play a crucial role in various musical genres, from the pulsating beats of jazz and rock to the complex rhythms of world music, and the percussive family also includes a vast array of exotic and culturally specific instruments like djembes, congas, tabla, and marimbas. Although they are the most widely played instruments, they have not been as extensively studied as wind or string instruments [2,3].

For the study of percussion musical instruments, vibroacoustic analysis is of major importance [4–6]. The literature contains research studies on the vibroacoustic behavior of percussion instruments. For example, Skrodzka et al. [7] performed a modal analysis

of a batter head of a snare drum and measurements of the instrument sound spectrum. Modal analysis and the acoustic radiation measurements of the kettledrum were performed by Tronchin [8]. Sunohara et al. [9] experimentally measured the sound spectrum, the vibration of the body, drumstick driving force, directivity, and the sound intensity vector of Japanese wooden drums (mokugyo). More recently, detailed numerical simulations were performed to study the vibroacoustic behavior of crash and splash cymbals [10,11], while nonlinear sound syntheses of cymbals were studied in the work presented in [12,13]. Moreover, ongoing research focuses on robotic-based experimental measurements [14–18].

The existing research on the automated generation of databases (DBs) of percussion sounds and their classification and recognition via machine learning (ML) is limited. Recently, Boratto et al. [19–21] recorded 276 audio samples corresponding to four drum cymbals made of three different bronze alloys. There were environmental and microphone variations during the recording procedure. Chhabra et al. [22] implemented drum instrument classification using machine learning. Recently, Li et al. [23] performed an audio recognition of Chinese traditional instruments, including percussion instruments, based on machine learning.

In this research work, we introduce an integrated method for the development of large percussion sound DBs, generated by known impact force spatiotemporal conditions, which are mapped and classified by audio recognition via machine learning. A novel 3D Auto-Drum Machine (3D-ADM) system capable of generating and collecting cymbal impact drum sounds is developed and presented. The capabilities of the 3D-ADM, along with initialization and calibration, sound production, and recording protocol for a reliable and repeatable measurement system, are demonstrated. The 3D-ADM excites the percussion instruments with a wooden drumstick at various points, sequentially, along a radial path at programmable spatial intervals, with known impact forces. The audio signal data produced at each excitation point of the vibrating object under study, including the metadata details of spatial coordinates of the excitation point and the impact force value, are recorded. The data collected during the calibration and initialization of the 3D-ADM are fed into a transformer-based audio neural network (ML model) that has been pretrained on speech signal. The internal representations of the ML model are visualized in 2D by a dimensionality reduction technique, which verifies the assumption that the collected data are separable (almost linearly) based on material and geometry. Additionally, based on the visualization process, it is expected that when enough data become available through the process described, models that perform geometry and material classification, as well as excitation-point estimation through regression, will be trained. On one hand, the ML-based analysis offers confirmation that the collection process is robust and consistent; on the other hand, it acts as a proof of principle, indicating that further development of large databases of 3D-ADM data will allow for the implementation of innovative research pathways relating to vibroacoustic characteristics, playing (excitation) positions, and generated sounds.

The generated and recorded percussion sounds are accompanied by the spatial excitation coordinates and the correspondent impact forces, allowing for the development of large and detailed DBs required for machine learning models. Thus, a high repetition rate for data production may be achieved and gathered in the future within sound DBs capable of training ML models. These DBs may serve as a reference for investigating the alternative and cost-effective materials and geometries with relevant sound characteristics [24]. Using an ML classification algorithm, it becomes possible to determine how much an alternative or cheaper material and/or manufacturing processes may alter the resulting sound. For future studies, this could be complemented by hearing tests. Moreover, apart from percussion instruments, this study could be applied to other musical instruments, also involving impulse testing [2,3,25]. Furthermore, the generation of sound DBs will constitute a foundation to properly train ML models, aiding in the development of accurate vibroacoustic numerical models and the exploration of percussion instrument sound synthesis [10–13].

The contributions of this paper can be summarized as follows: (i) a system that can automate the process of collecting percussive audio data, including the excitation force and

spatial position is presented; (ii) a qualitative analysis based on a pretrained ML model is developed that further evaluates the consistency of the data collection process; and (iii) the presented results demonstrate the feasibility of large sound database development. Such databases can properly train ML models for the possible use of alternative and cost-effective materials and geometries with similar sound characteristics and the development of accurate vibroacoustic numerical models used in percussion instruments sound synthesis.

Two 8-inch splash cymbals a classic medium thin, and a bell-shaped, made of a B8 Bronze and MS63 alloy, respectively, along with an 8-inch circular aluminum sheet used as a reference, are chosen for this research study. In Section 2, the methodology is presented. The experimental setup, the 3D-ADM system initialization and calibration, and the ML model are described in detail. In Section 3, the results of the data capturing, as well as the results of the exploratory analysis using ML are presented and discussed. The conclusions are described in Section 4.

2. Methodology

2.1. Experimental Setup

The 3D-ADM system developed for the generation and recording of the impact sounds is presented in Figure 1a. The 3D-ADM system is developed to excite the object under study by a known impact force. The circular aluminum sheet under study in Figure 1 is used as a reference and has an 8-inch diameter and 3 mm thickness. As shown in Figure 1c, the excitation mechanism can excite (strike) the sheet at any point along the radial path over the X-axis. The generated vibration can be detected and recorded by the 3D-ADM system, which is fully automated and Computer Numerical Controlled (CNC). This process is repeatable and accurate since the excitation conditions remain constant until the predefined and programmed number of points are excited. The corresponding CAD model presents the 3D-ADM in detail, shown in Figure 1b. A permanent metal stud is welded at the center of the front edge of the stage on the *X*-axis, to host and hold the object under study. The reference metal sheet, with a radius R1, is held at height H1 by bolts, whose torque is measured by a gauge while fastened, and is maintained the same for all the samples studied. The stepper motor SM1-X translates the stage along the *X*-axis, and the stepper motor SM2-Z translates the excitation mechanism along the *Z*-axis.

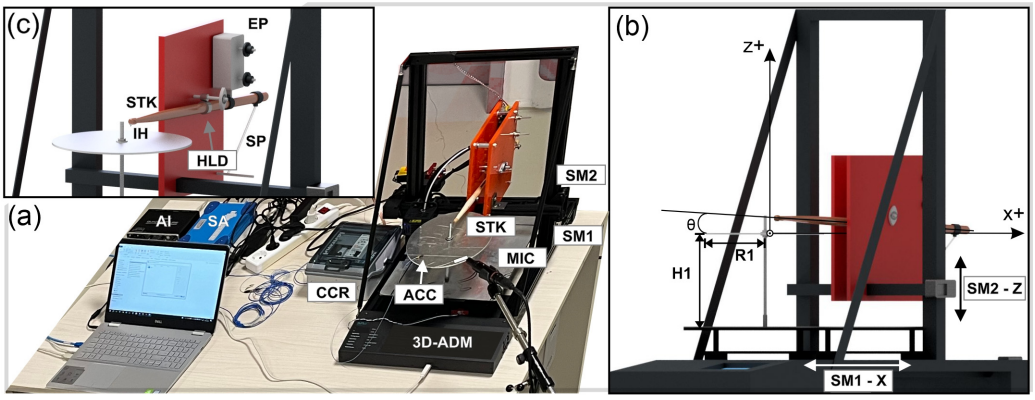

Figure 1. The 3D-ADM system generating and collecting the impact sounds: (**a**) Experimental set-up of the system, (**b**) CNC spatial adjustment, and (**c**) Excitation mechanism.

The excitation mechanism is rigidly and permanently attached to the center of the horizontal beam of 3D-ADM, as shown in Figure 1c. For the excitation, any type of drumstick can be used. Herein, a modified drumstick (STK) incorporating an impact hammer (IH, Model 086E80, PCB, Depew, NY, USA) on one side of its tip is used. The STK is held by a double ring holder (HLD), which allows for movement along a circular path due to a central shaft and two pillow block bearings. The movement of the STK is

induced by an electromagnetic piston (EP), driven by a circuit control system, denoted as the current controller (CCR) in Figure 1a, and results in the impact force applied. The STK is restored to its equilibrium position by a spring (SP), which is attached to the STK by a ring holder. The G-code synchronizes the spatial motion of the CNC 3D-ADM on the XZ plane with the motion of the STK.

The emitted sound is recorded by a measurement microphone (MIC, Model MiniSPL, NTI, Schaan, Liechtenstein), which is connected to an audio interface (AI, Model QUAD Capture, Roland, Los Angeles, CA, USA). Additionally, a miniature accelerometer (ACC, Model TLD352A56, PCB, Depew, NY, USA) can be used for the detection of the vibration. Thus, the system allows for the simultaneous recording of two signals, one by ACC (converted to wav format), and one by MIC, providing additional capabilities for vibroacoustic measurements to the 3D-ADM system.

The measurements are performed in the recording studio of the Department of Music Technology and Acoustics of the Hellenic Mediterranean University to eliminate the influence of background noise and reverberation on the recorded signals. The emitted sounds are recorded within sequential time windows, from excitation point to point, thus any noise produced by the machine and the motors is avoided. The development of a large training sound data base will further follow by implementing recording studio directional microphones.

2.2. System Initialization & Calibration

The initialization and calibration of the 3D-ADM system is performed using the circular 8-inch Al sheet set at a height H1, as presented in Figure 1b. The modified drumstick with the hammer tip is used for the spatial zero-set of the X and Z axis, with reference to the excitation point ExP-1, the first ExP to be measured. Therefore, angle θ is set to 0° by adjusting the spring accordingly, and thus securing the perpendicular hit of the hammer tip on the flat metal sheet. Figure 2a shows the coordinate system to be used for the G-code with an origin on ExP-1, and the next equidistant points along the radial path of the metal sheet, to be excited from X = 0 to R1, while the Z coordinates are set to zero since the object under study is flat. The coordinates of the 16 ExPs are imported in the G-code, setting a sequence of 16 equidistant points along the radial path of the metal sheet to be struck. The total number of the excitation points is defined by the user, requiring the corresponding changes to the G-code without any limitation. The spring is adjusted to its initial position to set the angle θ between the axis of drumstick and the plane of the Al sheet to θ = 3°. The microphone is fixed with a direction to the center of the circular Al sheet under study at a distance of 20 cm, and the accelerometer is attached 30 mm far from the center on the circular sheet, without affecting its vibroacoustic behavior.

The calibration of the 3D-ADM system is based on the CCR, which is equipped by a potentiometer that enables the control of the impact force via the variation of the current intensity that drives the EP. Given the initialization parameters, the angle θ is set to 3° by the help of spring and the CCR is set to deliver an impact force of 40 N, which is validated by the IH sensor. To automate the measuring and recording procedure, the time delay (td) for vibration relaxation before the excitation of the next ExP is determined. A trial excitation is performed with a nominal force of 40 N and the accelerometer records the duration of the vibrational signal to be ~2 s. Therefore, a td of 10 s is introduced in the G-code to allow the synchronization of the EP movement with the CCR between each ExP excitation, and to provide sufficient relaxation time for the vibrating Al sheet.

After initialization and calibration, the efficiency of the 3D-ADM system in terms of repeatability and accuracy is determined. The G-code developed for the sequential excitation and the recording of the 16 ExPs is executed seven more times. All the signals of the accelerometer, the force measured by the hammer, and the sound captured by the microphone are recorded and stored. Figure 2b shows the mean force value for each excitation point, which reveals a maximum deviation of 1.3 N at ExP-16 from the impact force of 40 N. The deviation of the mean value is expressed as a percentage in Figure 3. The

mean value of the absolute deviation percentage for all 16 ExPs is 1.53%. The demonstrated results reveal that the smaller deviation corresponds to points from ExP-5 to ExP-12. Such force value deviations result only in small changes in the sound energy, which is recorded by the microphone, and thus may be neglected. The same process was also repeated for the impact forces of 20 N, 30 N, and 50 N, resulting in similar deviation values, securing the independency of the measurements accuracy from the excitation parameters. It should be noted that the contact duration between the drumstick and the metal sheet was measured to be ~15 ms for all cases.

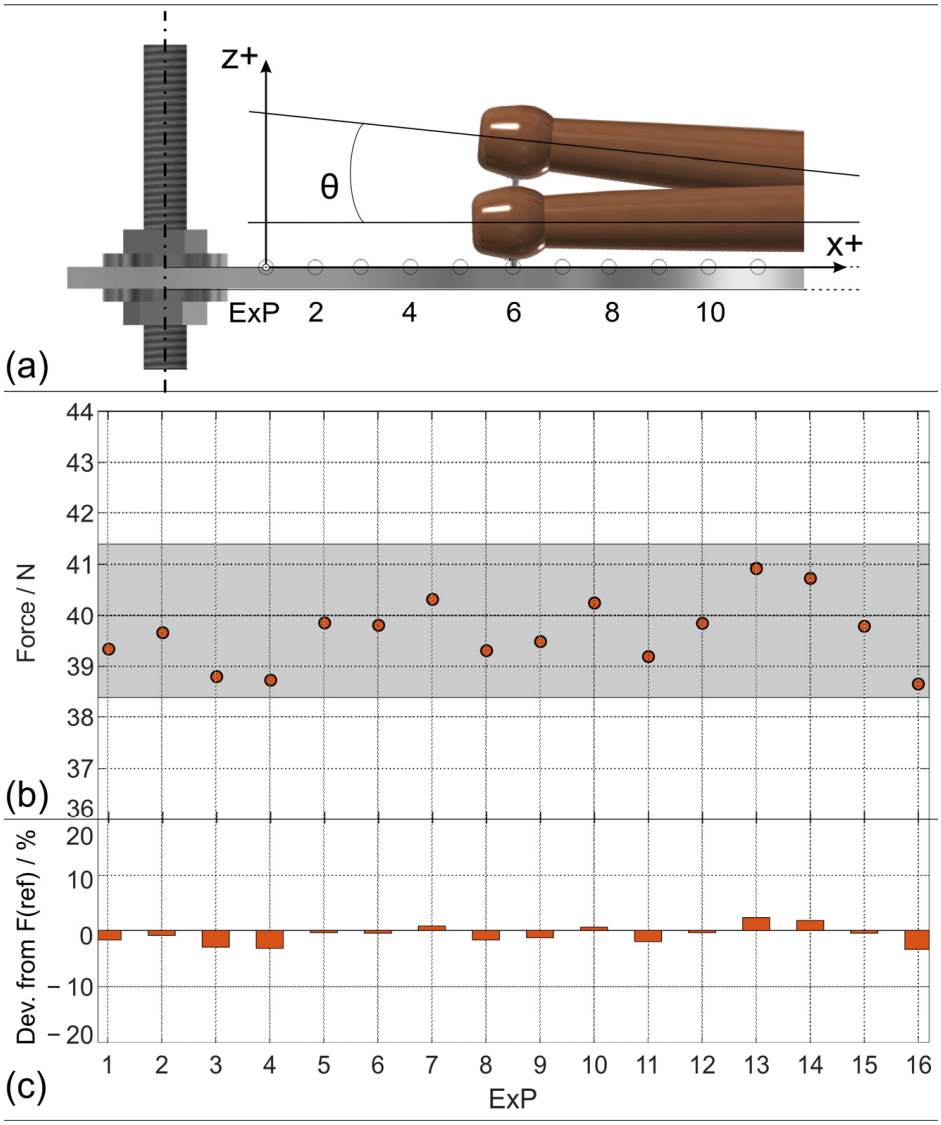

Figure 2. (**a**) The modified drumstick with the hammer tip for θ = 0° and 3° at ExP-6 and the axis origin set on ExP-1, (**b**) mean force for the eight repeated sets of measurements on the 16 ExPs along R1 and (**c**) deviation % from the reference force F = 40 N.

The described protocol secures the measurement process and their accuracy with the help of the modified drumstick. The drumstick is rotated by 180° along its axis within the HLD without any other system alterations, ensuring the predefined value of the same impact force on the same ExPs. The proposed measuring methodology can be directly applied to any type of sample utilizing impact excitations under different force values.

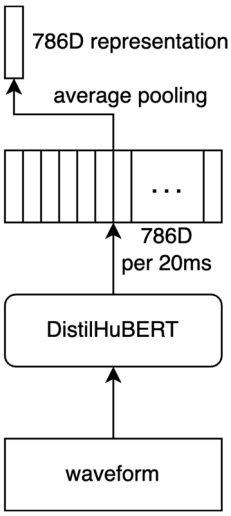

Figure 3. Transformation of a waveform in a 768-dimensional representation.

2.3. Machine Learning Model

The data collected by the 3D-ADM system can produce a sufficiently large dataset to study many aspects regarding the relations between the sound and physical attributes of objects (e.g., geometry, material, force, and the position of impact among others). To verify the efficiency and usability of such a database, the consistency of this early developed data collection process must be investigated (i.e., the pairs of same intended ExPs and forces produce similar waveforms). Additionally, a proof-of-concept overview regarding the prediction capabilities on the newly developed and currently small dataset collected from two splash cymbals and the Al sheet of different materials and geometries must be explored.

A pretrained ML model on speech data through self-supervised learning is employed for the exploratory analysis. Such models are pretrained on large datasets of speech audio data and they can be readily employed with, or even without, fine-tuning on music-related tasks [26]. The specific model used in this study is DistilHuBERT [27], which is a light-weight version of the HuBERT model [28], which is a "discretized" version of the wave2vec 2.0 model [29]. This model takes an audio waveform as the input with a 16 kHz sample rate and converts it into a contextualized representation of 20 ms frames. This representation encompasses information for each 20 ms-long frame of audio into 768-dimensional vectors that capture the context of the entirety of frames in the sequence through the transformer attention mechanism [30]; vectors that belong to the same context are more similar. Since we are not fine-tuning the system, we use the "frozen" version as given in the work presented in [27], where the details for the system hyper-parameters can be found.

The motivation behind using DistilHuBERT and not any other network trained with self-supervision (e.g., MERT [31]) is two-fold. First, to use a reliable system, which has been tested in several tasks that are not limited to music. Second, to test our results on a system that was built to identify speech (not music), since speech models are sensitive to noise-related spectra because they can identify fine details in the spectra of fricatives. Although other models should be tested in the future, in the context of this work it suffices determining the prospects of such pretrained models, which are robust and "noise-sensitive" enough, capable of processing the collected acoustic data.

To account for different waveform lengths, average pooling is applied, i.e., the average of all 768-dimensional vectors in the sequence are averaged out for each dimension, leading to a single average 768-dimensional representation for each audio file. The process for extracting the 768-dimensional representation for a recorded waveform is depicted in Figure 3. This model is not fine-tuned to any downstream task, but it is rather used in its readily available pretrained state.

3. Results and Discussion

3.1. Data Capturing

The drumstick is rotated by 180° along its axis to allow for the excitation of the same 16 ExPs of the Al sheet by the wooden tip. Typical sound spectra recorded by the microphone (excitation at the points 6, 12, and 14) are shown in Figure 4, along with the spectra recorded during calibration using the hammer tip. The main aspect revealed for 3D-ADM by observing the results in Figure 4 is that both excitations by the hammer and the wooden tip produce sounds with similar frequency characteristics, verifying again that the modified drumstick allows for excitation with the same initial conditions. As expected, the sound spectra resulting from the wooden drumstick tip and the hammer tip are not identical, even if the impact force is the same, because the hammer excitation induces a point force due to its metal conical tip, while the wooden tip of the drumstick induces dynamic pressure. These measurements correspond to real percussion sounds generated by drummers during performance with common wooden drumsticks but preserve two advantages: (i) the known impact conditions, and (ii) the elimination of human interaction influence. Thus, a high repetition rate for data production may be achieved and gathered within sound DBs, capable of both training ML models in the future and being used as a reference for the development of accurate vibroacoustic numerical models and studying percussion instruments sound synthesis.

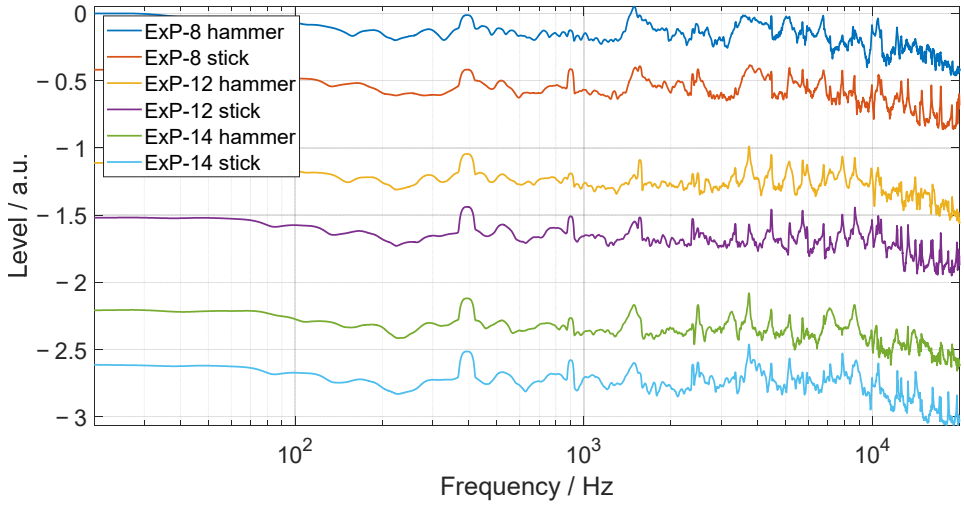

Figure 4. Spectra of the emitted sound excited by the modified drumstick with impact force F = 40 N at the representative points ExP-6, -12, and -14 by the hammer vs. the wooden tip (level in a.u.).

The recorded sound time series and spectra, including the meta-data details of spatial coordinates of the excitation point and the impact force value, are collected and seeded to the sound DB. Additionally, the frequency response function using a kinematic quantity can also be determined and further used in FEM models for vibroacoustic analysis [10–13]. Extra development and the enrichment of the DB can be achieved by varying the applied force via the alterations of the current intensity that drives the EP of CCR. In this context, two measurement sets are additionally performed with a mean force value of 20 N and 50 N. The differences in the sound spectrum after exciting the Al sheet at the same point with the hammer by applying different forces are shown in Figure 5. The increase in the input energy to the sheet is evident in the spectrum by the level increase. It must be noted that the frequency spectra characteristics for both forces, even if their values differ by 30 N, preserve the same behavior, as expected.

The 8-inch Al reference sheet is replaced by the 8-inch bell-shaped MS63 alloy splash cymbal, supported at height H1 by bolts fastened again with the same torque measured by the gauge. The G-code is modified and adjusted at the new Z-coordinates for each of the sixteen ExPs, according to the CAD axisymmetric cymbal profile. The accelerometer

and the microphone are kept at the same positions during the initialization measurements. The sound delay is measured again to determine the time interval between consecutive measurements. The same methodology is also applied to the measurements of a splash medium-thin 8-inch B8 cymbal. The acoustic measurement sequence for sixteen ExPs and for the impact forces of 20, 30, 40, and 50 N are performed. The sound data and the measurements results are recorded and stored, including the corresponding meta-data and are added to the DB under development. Figure 6a shows the microphone recorded sound spectra of the MS63 and the B8 cymbals, after excitation at the ExP-7 within the bow area by an impact force of 30 N by the wooden tip of drumstick, with reference to the 8-inch Al sheet. It must be noted that the ExPs of interest regarding cymbals are in the bow area ranging from ExP-5 to ExP-14, and these ExPs belong to the range where the minimum force deviation values was measured and presented in Figure 2. The representative spectral information in Figure 6a on a given frame, disregarding spectral evolution over time, demonstrates that the detailed spectral characteristics may provide the indications about the geometry and the material of the cymbals required for regression and recognition. Figure 6b presents the distribution of the ExPs for each 8-inch percussion instrument under study. The CAD axisymmetric profile used for the G-codes developed is highlighted and the characteristic changes in the curvature are also obvious and will be used in the future as classes for the ML models.

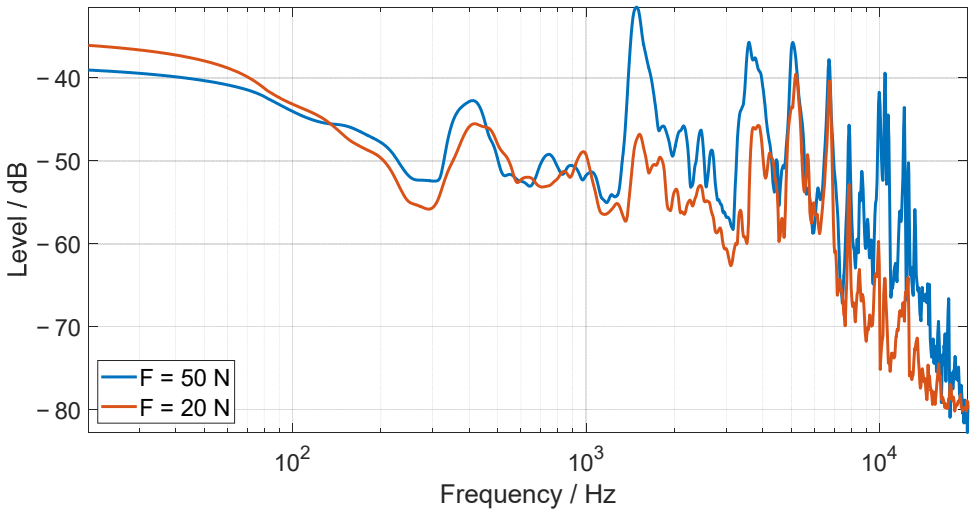

Figure 5. ExP-7 emitted sound spectra excited by 20 N (red line) and 50 N (blue line).

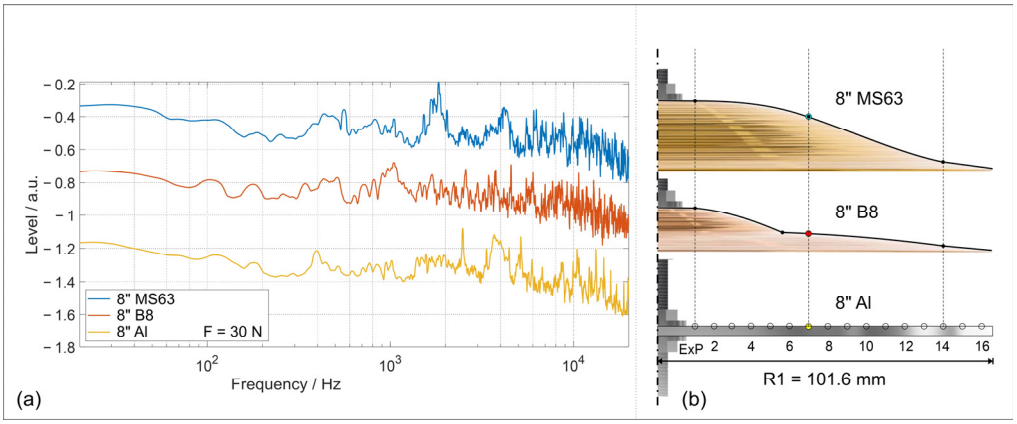

Figure 6. (**a**) Microphone recorded sound spectra of MS63, B8 cymbals and Al 8-inch, excited by 30 N on the ExP-7. (**b**) The geometrical characteristics of cymbals with reference to the 8-inch Al sheet.

3.2. Exploratory Analysis

For examining the goals at hand, recordings produced by the microphone are employed. Specifically, two repetitions of the same force and same position were collected for 16 positions (32 audio waveforms in total), for three different percussion instruments as already described (different geometry and material), leading to a total of 96 audio waveforms. The t-SNE [32] dimensionality reduction method was applied to transform the 768-dimensional representations of all 96 recordings to 2D representations. This dimensionality reduction method disentangles several assumed low-dimensional manifolds (in this case, 2D manifolds are assumed) that are embedded in high-dimensional manifolds (i.e., 768D), based on linear approximations of neighboring data points within low-dimensional manifolds.

Figure 7 shows the resulting mapping for all 96 recordings, where the colors of each point represent the material/geometry of each cymbal. The results are clearly separable, almost linearly, a fact that indicates that speech-pretrained model performs well on the basic categorization of the examined data. Figure 8 includes a zoomed-in region of Figure 7; this region includes numbers for each point that indicate the corresponding ExP. Namely, numbers 1–16 indicate the 16 ExPs with increasing radius and numbers 17–32 indicate the repetitions, where the repetition of excitation "i" is depicted as "i + 16". In Figure 8 the repetitions appear close to each other (9–25, 10–26, etc.). This fact indicates that the collected data is consistent, verifying that repeated attempts lead to neighboring results. Furthermore, the placement of consecutive ExPs on consecutive positions on the graph shows that it is possible to do ML-based regression for estimating the ExP of a given sound on a given geometry/material. Similar results can be obtained for other parts of the graph presented in Figure 7.

Furthermore, there is an almost monotonic trajectory for consecutive hits in both graphs. For instance, in the zoomed-in region presented in Figure 8, where the effect is clearer, the excitations of the ExPs 9, 10, 11, 12, etc., are on a path that moves from left to right. This indicates that it is possible to fine tune an ML system to predict the ExP coordinates, given the waveform and the geometry/material.

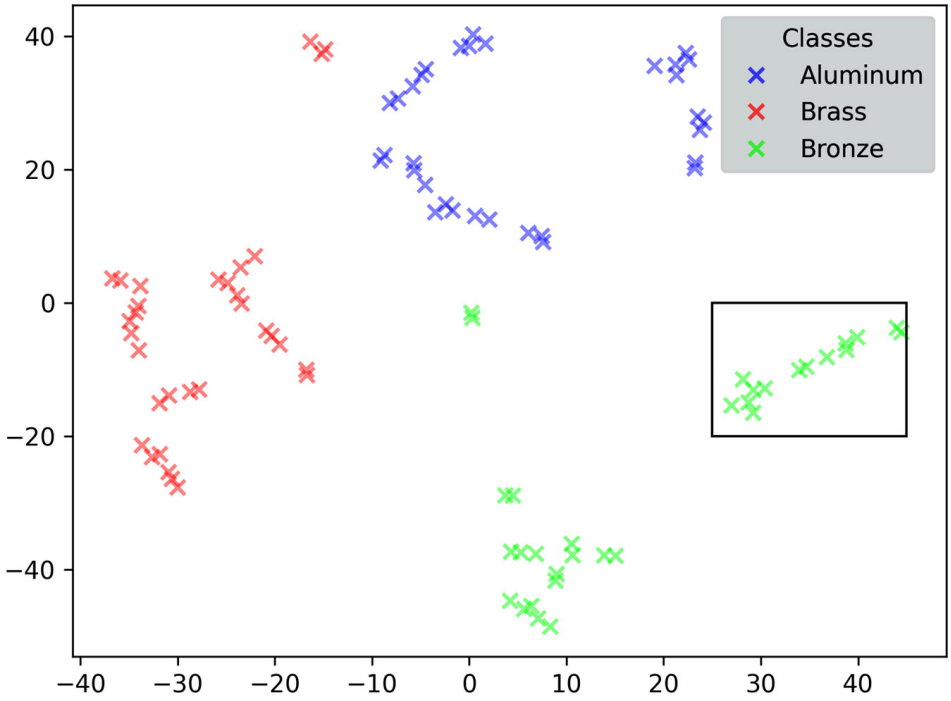

Figure 7. 2D layout derived from t-SNE on the 768-dimensional representation for 96 recordings. The 32 recordings for each geometry material are displayed in different color and are almost linearly separable. The area delineated by the rectangle is further analyzed in Figure 8.

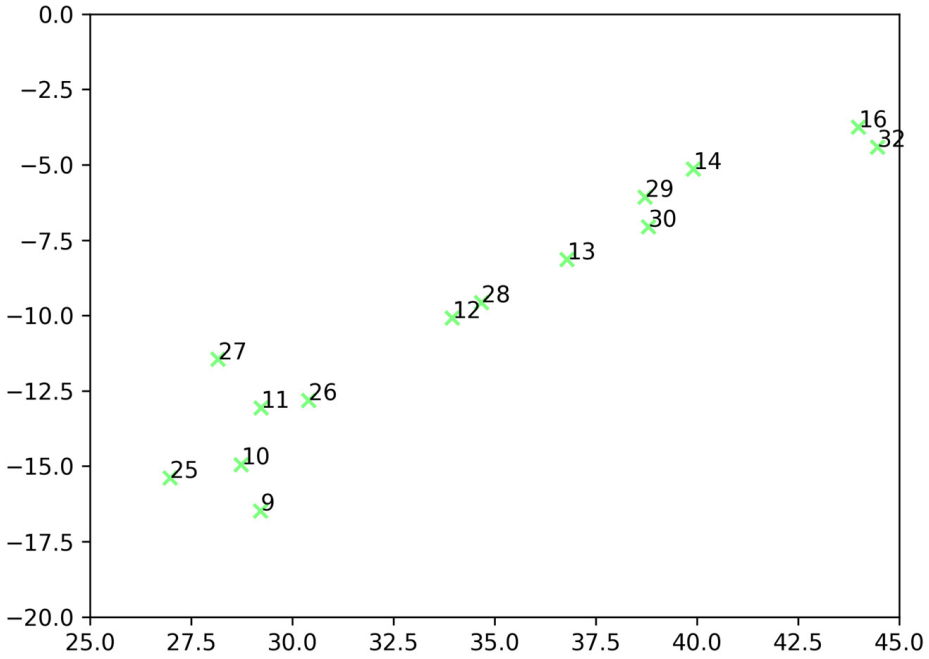

Figure 8. The zoomed-in region of Figure 7, demonstrating the repetition and ExP spatial position consistency.

4. Conclusions

A novel automated process for the generation, collection, classification, and recognition of audio files corresponding to percussion sounds is presented. The sounds are produced by known impact force excitation from the developed 3D-ADM, free from human interference. The proposed excitation and measurement ADM system includes a microphone and a miniature accelerometer, used to record the sound and vibration, resulting after excitation. The machine is CNC programmable, allowing for variations of the excitation points and the impact force. The repeatability and the efficiency of the measurement procedure is explored and validated.

After initialization and calibration, the 3D-ADM system is used to excite two cymbals and a flat circular plate, which differ in material and geometry. The recorded sounds are processed by a ML model pretrained on speech signals. The visualizations of the internal representations of the ML model validate the consistency of the process followed for data measurements and collection. The results presented demonstrate the capability of generation of large sound databases and provides pointers for future work regarding ML-based modeling to relate materials, geometries, playing positions, and generated sounds. This work might include fine-tuning ML models to classify the material, geometry, force, and position of impact either separately (i.e., give a probability for each attribute independently), conditionally (e.g., given a material and force/position of impact, estimate the geometry), or jointly (i.e., give a single probability for a combination of attributes).

Author Contributions: Conceptualization, E.K., M.K.-P., S.B. and V.D.; methodology, A.S., A.S.-K., M.B., M.K.-P., S.B., V.D. and Y.O.; investigation, M.K.-P., M.T., N.A.P., S.B. and V.D.; writing—original draft preparation, E.K., M.K.-P., S.B. and V.D.; writing—review and editing, E.K., M.B., M.K.-P., S.B. and V.D.; supervision, V.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available within the article.

Acknowledgments: This research was funded by the Hellenic Mediterranean University, within the project «Recording and metrological analysis of the vibro-acoustic characteristics of musical instruments for the investigation of alternative and low-cost materials and geometries with relevant sound characteristics».

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bucur, V. *Handbook of Materials for Percussion Musical Instruments*, 1st ed.; Springer: Cham, Switzerland, 2022.
2. Wolfe, J.; Fletcher, N.H.; Smith, J. The interactions between wind instruments and their players. *Acta Acust. United Acust.* **2015**, *101*, 211–223. [CrossRef]
3. Kaselouris, E.; Bakarezos, M.; Tatarakis, M.; Papadogiannis, N.A.; Dimitriou, V. A review of finite element studies in string musical instruments. *Acoustics* **2022**, *4*, 183–202. [CrossRef]
4. Rossing, T.D. Acoustics of percussion instruments: Recent progress. *Acoust. Sci. Technol.* **2001**, *22*, 3. [CrossRef]
5. Rossing, T.D. Acoustics of percussion instruments: An update. *Acoust. Sci. Technol.* **2004**, *25*, 6. [CrossRef]
6. Morrison, A.C.; Rossing, T.D. Percussion Musical Instruments. In *Springer Handbook of Systematic Musicology*, 1st ed.; Bader, R., Ed.; Springer: Heidelberg, Germany, 2018; pp. 157–170.
7. Skrdodzka, E.B.; Hojan, E.; Proksza, R. Vibroacoustic investigation of a batter head of a snare drum. *Arch. Acoust.* **2006**, *31*, 289–297.
8. Tronchin, L. Modal analysis and intensity of acoustic radiation of the kettledrum. *J. Acoust. Soc. Am.* **2005**, *117*, 926–933. [CrossRef] [PubMed]
9. Sunohara, M.; Furihata, K.; Asano, D.K.; Yanagisawa, T.; Yuasa, A. The acoustics of Japanese wooden drums called "mokugyo". *J. Acoust. Soc. Am.* **2005**, *117*, 2247–2258. [CrossRef] [PubMed]
10. Kaselouris, E.; Alexandraki, C.; Bakarezos, M.; Tatarakis, M.; Papadogiannis, N.A.; Dimitriou, V. A detailed FEM Study on the Vibro-acoustic Behaviour of Crash and Splash Musical Cymbals. *Int. J. Circuits Syst. Signal Process* **2022**, *16*, 948–955. [CrossRef]
11. Kaselouris, E.; Paschalidou, S.; Alexandraki, C.; Dimitriou, V. FEM-BEM Vibroacoustic Simulations of Motion Driven Cymbal-Drumstick Interactions. *Acoustics* **2023**, *5*, 165–176. [CrossRef]
12. Nguyen, Q.B.; Touzé, C. Nonlinear vibrations of thin plates with variable thickness: Application to sound synthesis of cymbals. *J. Acoust. Soc. Am.* **2019**, *145*, 977–988. [CrossRef]
13. Samejima, T. Nonlinear physical modeling sound synthesis of cymbals involving dynamics of washers and sticks/mallets. *Acoust. Sci. Technol.* **2021**, *42*, 314–325. [CrossRef]
14. Ness, S.; Trail, S.; Driessen, P.; Schloss, A.; Tzanetakis, G. Music Information Robotics: Coping Strategies for Musically Challenged Robots. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), Miami, FL, USA, 24–28 October 2011.
15. Kapur, A.; Singer, E.; Benning, M.S.; Tzanetakis, G. Trimpin, Integrating hyperinstruments, musical robots & machine musicianship for North Indian classical music. In Proceedings of the 2007 Conference on New Interfaces for Musical Expression (NIME07), New York, NY, USA, 6–10 June 2007.
16. Wu, X.; Liu, T.; Deng, Y.; Wu, X.; Luo, D. Developing Robot Drumming Skill with Listening-Playing Loop. In *Advances in Swarm Intelligence*; Tan, Y., Takagi, H., Shi, Y., Niu, B., Eds.; ICSI 2017. Part of Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10386, pp. 559–566.
17. Sui, L.; Su, Y.; Yi, Y.; Li, Z.; Zhu, J. Intelligent Drumming Robot for Human interaction. In Proceedings of the 2020 International Symposium on Autonomous Systems, ISAS 2020, Guangzhou, China, 6–8 December 2020.
18. Long, J.; Murphy, J.W.; Carnegie, D.A.; Kapur, A. A closed-loop control system for robotic hi-hats. In Proceedings of the International Conference on New Interfaces for Musical Expression, Copenhagen, Denmark, 15–19 May 2017.
19. Boratto, T.H.A.; Cury, A.A.; Goliatt, L. A Fuzzy Approach to Drum Cymbals Classification. *IEEE Lat. Am. Trans.* **2022**, *20*, 2172–2180. [CrossRef]
20. Boratto, T.H.A.; Saporetti, C.M.; Basilio, S.C.A.; Cury, A.A.; Goliatt, L. Data-driven cymbal bronze alloy identification via evolutionary machine learning with automatic feature selection. *J. Intell. Manuf.* **2022**, *35*, 257–273. [CrossRef]
21. Boratto, T.H.A.; Cury, A.A.; Goliatt, L. Machine learning-based classification of bronze alloy cymbals from microphone captured data enhanced with feature selection approaches. *Expert Syst. Appl.* **2023**, *215*, 119378. [CrossRef]
22. Chhabra, A.; Singh, A.V.; Srivastava, R.; Mittal, V. Drum Instrument Classification Using Machine Learning. In Proceedings of the 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 18–19 December 2020.
23. Li, R.; Zhang, Q. Audio recognition of Chinese traditional instruments based on machine learning. *Cogn. Comput. Syst.* **2022**, *4*, 108–115. [CrossRef]
24. Brezas, S.; Katsipis, M.; Kaleris, K.; Papadaki, H.; Katerelos, D.T.G.; Papadogiannis, N.A.; Bakarezos, M.; Dimitriou, V.; Kaselouris, E. Review of Manufacturing Processes and Vibro-Acoustic Assessments of Composite and Alternative Materials for Musical Instruments. *Appl. Sci.* **2024**, *14*, 2293. [CrossRef]
25. French, M.; Bissinger, G. Testing of Acoustic Stringed Musical Instruments-an Introduction. *Exp. Tech.* **2001**, *25*, 40–43. [CrossRef]

26. Ma, Y.; Yuan, R.; Li, Y.; Zhang, G.; Chen, X.; Yin, H.; Lin, C.; Benetos, E.; Ragni, A.; Gyenge, R.; et al. On the effectiveness of speech self-supervised learning for music. *arXiv* **2023**, arXiv:2307.05161.
27. Chang, H.-J.; Yang, S.-W.; Lee, H.-Y. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 200), Singapore, 22–27 May 2022.
28. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
29. Baevski, A.; Yuhao, Z.; Abdelrahman, M.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
30. Ashish, V.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 17), Long Beach, CA, USA, 4–9 December 2017.
31. Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Xiao, C.; Lin, C.; Ragni, A.; Benetos, E.; et al. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv* **2023**, arXiv:2306.00107.
32. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.