Machines that listen: towards a machine listening model based on perceptual descriptors

Marco Buongiorno Nardelli, Mitsuko Aramaki, Sølvi Ystad, Richard Kronland-Martinet

Machines that listen: towards a machine listening model based on perceptual descriptors.

Marco Buongiorno Nardelli[1,2,3,4,5][0000−0003−0793−5055], Mitsuko Aramaki[5][0000−0001−6518−374X], Sølvi Ystad[5][0000−0001−9022−9690], and Richard Kronland-Martinet[5][0000−0002−7325−4920]

[1] CEMI, Center for Experimental Music and Intermedia, University of North Texas, Denton, TX 76203, USA
[2] iARTA, Initiative for Advanced Research in Technology and the Arts, University of North Texas, Denton, TX 76203, USA
[3] Department of Physics, University of North Texas, Denton, TX 76203, USA
[4] IMéRA - Institut d'études avancées d'Aix-Marseille Université, Marseille 13004, France
[5] CNRS, Aix Marseille University, PRISM (Perception, Representations, Image, Sound, Music), Marseille, France
mbn@unt.edu
http://www.musicntwrk.com

Abstract. Understanding how humans use auditory cues to interpret their surroundings is a challenge in various fields, such as music information retrieval, computational musicology and sound modeling. The most common ways of exploring the links between signal properties and human perception are through different kinds of listening tests, such as categorization or dissimilarity evaluations. Although such tests have made it possible to point out perceptually relevant signal structures linked to specific sound categories, rather small sound corpora (100-200 sounds in a categorization protocol) can be tested this way. The number of subjects generally do not exceed 20-30, since it is also very time consuming for an experimenter to include too many subjects. In this study we wanted to test whether it is possible to evaluate larger sound corpora through machine learning models for automatic timbre characterization. A selection of 1800 sounds produced by either wooden or metallic objects were analyzed by a deep learning model that was either trained on a perceptually salient acoustic descriptor or on a signal descriptor based on the energy contents of the signal. A random selection of 180 sounds from the same corpus was tested perceptually and used to compare sound categories obtained from human evaluations with those obtained from the deep learning model. Results revealed that when the model was trained on the perceptually relevant acoustic descriptors it performed a classification that was very close to the results obtained in the listening test, which is a promising result suggesting that such models can be trained to perform perceptually coherent evaluations of sounds.

Keywords: Sound descriptors · Sound perception · Machine learning · Networks · Machine Listening

1 Introduction

Analyzing our surroundings through the many sounds that are continuously produced by our environment is a trivial task that humans do more or less automatically. Both natural sounds from the environment such as waves, rain, wind, or sounds from humans, machines or animals can be recognized and localized without any practicing. It is however much more complicated to tell a machine how to recognize such events through sounds. For that purpose we need to identify perceptually relevant sound structures for each source that somehow can be considered as the signature that characterizes an aspect of the sound, such as the action that caused the sound or the object, such as its shape, size or material of the sound source. Several previous studies have tempted to identify such characteristics and have identified sound structures linked to the perceived size [11] and the material of which it is composed [12,8,3]. In the case of more complex situations reflecting for instance interactions between sound sources, the listener perceives properties related to the event as a whole. Warren and Verbrugge [17] showed that objects that bounce and break can be distinguished by listeners with a high degree of accuracy, while Repp [14] revealed that subjects were able to recognize their own recorded clapping and the hand position from recordings when someone else is clapping. More recently, Thoret [16] showed that subjects were able to recognize biological motions and certain shapes from friction sounds produced when a person is drawing on a paper.

The present study focuses on how two different material categories, wood and metal, can be distinguished. Previous studies on the identification of material categories [4,3], have shown that both temporal aspects such as the damping and frequency related aspects such as the spectral bandwidth or the roughness are perceptually salient signal structures for such sounds. These approaches enabled us to design evocative sound synthesis models that enable to control sounds from verbal labels (material, size, shape etc). However, for more general uses, such as the identification of sound categories within large databases, the previous approaches are less adapted, since they rely on a combination of several acoustic descriptors obtained from a rather small set of sounds and therefore might not be adapted to general models that characterize environmental sounds. In the present study we therefore propose to focus on the log Mel energy of the sound by using a more global descriptor, namely the MFCC that has been commonly used in automatic classification tasks [9]. Although the MFCCs were initially designed for speech recognition based on source filter models [7], they integrate perceptual properties and mainly discard the source part making them rather pitch independent. It is therefore interesting to use such descriptors on large sets of sounds that cannot be easily pre-treated and equalized in pitch and intensity. Their ability to capture global spectral envelope properties is also an important advantage from a perceptual point of view [15].

The objective of this study is to take advantage of network-based modeling, analysis, and visualization techniques to perform automatic categorization tasks that mimic human perception. Similarly to social networks, gene interaction networks and other well-known real-world complex networks, the data-set of

sounds can be treated as a network structure, where each individual sound is represented by a node in a network, and a pair of nodes is connected by a link if the respective two sounds exhibit a certain level of similarity according to a specified quantitative measure. In this approach, one can see a lot of conceptual similarities between sound networks and social networks, where individual nodes are connected if they share a certain property or characteristic (i.e., sounds can be connected according to shared physical or perceptual properties, and people are connected according to their acquaintances, collaborations, common interests, etc.) Clearly, different properties of interest can determine whether a pair of nodes is connected; therefore, different networks connecting the same set of nodes can be generated.

In this paper we take a further step in exploring this promising direction of research. Specifically, many complex systems can be better analyzed via network representations as networks provide a nice mathematical tool to explore these systems. Uncovering the topological structures of networks may help to understand the organizing principles of underlying complex systems. Furthermore, the knowledge acquired via this approach has motivated us to explore machine learning models for automatic timbre characterization and classification and the effectiveness of such approaches compared to results from human perception tests. Automatic classification with deep learning approaches have previously been applied to large datasets of both speech and music [9,10], but fewer studies have investigated automatic timbre classification of environmental sounds.

The paper is organized as follows: in Sec. 2 we discuss the various methodologies employed in this research, from the definition of acoustical descriptors, to network metrics and machine learning models; in Sec. 3 we discuss the results of our analysis; we end with a few concluding remarks and a look towards future applications of this study.

2 Methodology

All the computational results in this paper have been obtained with the `MUSICNTWRK` package. `MUSICNTWRK` is a python library for pitch class set and rhythmic sequences classification and manipulation, the generation of networks in generalized music and sound spaces, deep learning algorithms for timbre recognition, and the sonification of arbitrary data. The software authored by one of the coauthors (MBN) and it is freely available under GPL 3.0 at www.musicntwrk.com [6][5].

2.1 Impact sound data.

We have compiled a database of ca. 1800 impact sounds produced by metal or wood objects from Splice Sounds collections. The length of each recording was equalized to 22050 sample points (5.0 sec. at a sample rate of 44100 Hz) by either zero-padding or truncation. Sounds span a broad palette of timbre and

provide a good data-set for statistical analysis. Of these 1800 sounds, we extract a random sub-set of 180 sounds that we used for the human perception tests and analyzed with network techniques. The remaining 1620 sounds have been used to train the machine learning models with a 80-20% split between training and validation sets.

2.2 Audio descriptors and metrics in the generalized timbre space.

Power Cepstrum and PSCC The Power Cepstrum of a signal gives the rate of change of the envelope of different spectrum bands and is defined as the squared magnitude of the inverse Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal:

$$\text{PSCC} = \left| FT^{-1} \left\{ \log(|FT\{f(t)\}|^2) \right| \right|^2 \tag{1}$$

In this work We always considered the first 13 cepstrum coefficients (PSCC), where the 0-th coefficient corresponds to the power distribution of the sound over time.

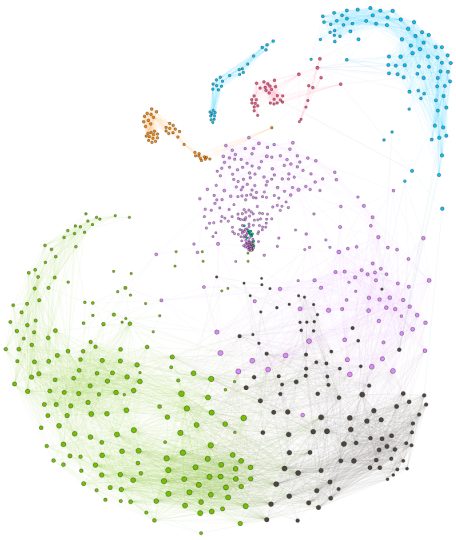

Fig. 1. Section of the network of the MFCCs built from the 1620 sounds that have been used to train the machine learning models. Colors indicate the classification based on their modularity class: Green, mostly high frequency tones from wood; Gray, mostly high to mid-frequency tones of wood; Purple, mostly deep frequency tones of wood; Cyan, mostly dry metal tones; Pink, mostly metal; and Orange, mostly "choked" metal sounds.

Mel Frequency Cepstrum and MFCC. The Mel Frequency Cepstrum of a signal is obtained as in Eq. 1 and differ from the power cepstrum by the choice of the spectrum bands that are mapped over the Mel scale using triangular overlapping windows. The mapping of the frequency bands on the Mel scale better approximates the human auditory system's response than the linearly-spaced frequency bands used in the normal cepstrum. We use a 16 bands Mel filter, and we considered the first 13 cepstrum coefficients (MFCC) as in the previous case. As for the PSCC, the 0-th coefficient corresponds to the power distribution of the sound over time.

Both PSCC and MFCC are obtained using 64 bins in the short time Fourier transform.

Networks and metric in timbre space. Network analysis methods exploit the use of graphs or networks as convenient tools for modeling relations in large data sets. If the elements of a data set are thought of as "nodes", then the emergence of pairwise relations between them, "edges", yields a network representation of the underlying set. Similarly to social networks, biological networks and other well-known real-world complex networks, entire data-set of sound structures can be treated as a network, where each individual descriptor (PSCC, MFCC) is represented by a node, and a pair of nodes is connected by a link if the respective two objects exhibit a certain level of similarity according to a specified quantitative metric. Pairwise similarity relations between nodes are thus defined through the introduction of a measure of "distance" in the network: a "metric". In this study we use the Euclidean norm (generalized Pythagoras theorem in N-dimensions) to quantify similarity between sound descriptors:

$$\text{distance}(I, J) = \sqrt{\sum_i \left(x_i^I - x_i^J\right)^2}, \tag{2}$$

where $\mathbf{x}$ is the chosen sound descriptor for sound I and J. In Figure 1 we show the network of the MFCC built from the 1620 sounds that have been used to train the machine learning models. In the figure we display the principal component for which less than 2% of all possible edges are built, that is, we allow an edge only if two MFCCs are at a distance that is less than 3% of the maximum diameter of the network. This representation reveals that the classification is coherent with material categories, as it can be observed from the emergence of clusters of sounds belonging to the same material (see for instance the green cluster of wood sounds on the lower left part of the network and the cyan, pink and orange clusters of metallic sounds in the upper part). This result validates the corpus with respect to the sound quality. For a general review on networks and graph theory the reader is referred to Albert and Barabási [2].

2.3 Machine learning model.

We implemented a deep learning model based on convolutional neural network (CNN) architecture inspired by similar approaches used in image and sound

recognition [13]. The CNN is built using the Keras kernel of Tensorflow [1] and it is trained on the full PSCC or MFCC data, after proper scaling and normalization. We retained only models with validation accuracy higher than 90%. After an appropriate model is chosen, it is tested on the set initially chosen for the human perception experiment. Each model chosen retains a similar accuracy on this set. A typical result of a training session on 30 epochs is shown in Fig. 2.

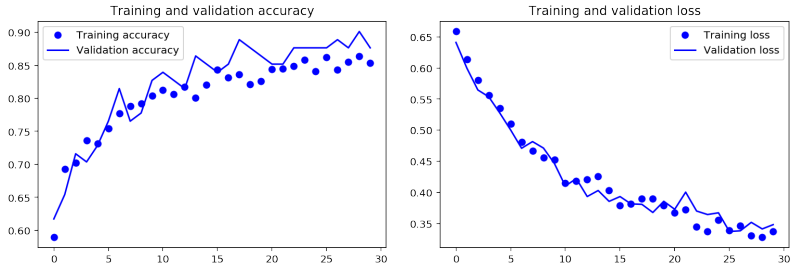

Fig. 2. Accuracy and loss in a typical model training run: (left) accuracy, (right) loss.

2.4 Experimental setup

The sounds were presented randomly to the participants thought headphones. The participants were asked to categorize each sound in either the Metal or Wood category by selecting the label shown on a graphical interface developed with Matlab. They could listen to each sound as often as desired.
Participants: Twenty-seven volunteers (21 males, mean age: 37 years-old) participated in the experiment. They declared no hearing nor cognitive problems.

3 Results and Discussion

A set of 180 impact sounds randomly extracted from the initial database (section 2.1) was used in a perceptual listening test. In Figure 3 and Table 1, we summarize the data of the perceptual test compared with four scenarios based on the machine learning models. In Figure 3 the scores represent the percentage of classification in the metal category. From this figure certain sounds (157) were clearly classified without ambiguity. 60 sounds were classified by 100% of the subjects in the labeled material category and 23 sounds that were less clearly classified were defined as "ambiguous", i.e. classified by less than 50% of the subjects in the labeled material category (for more details see Table 1). The above sounds were fed to our ML models in four different fashions: 1. with the model trained on the MFCCs to classify the sound with its MFCC (perceptual measure with perceptual model); 2. with the model trained on the MFCCs to classify the sound with its PSCC (physical measure with perceptual model); 3. with the

model trained on the PSCCs to classify the sound with its PSCC (physical measure on physical model); and 4. with the model trained on the PSCCs to classify the sound with its MFCC (perceptual measure on physical model). It should be noted that the ML models 1. and 3. have an accuracy that exceeds 90% in both cases. In the first column of Table 1 we report the percentage of classification in the metal category in the perceptual test. The successive columns list the sounds that are common with sounds characterized incorrectly by the four ML models (marked as X). Interestingly, we find a large degree of overlap between the perceptual scores and the ML scenario n. 4, physical model and perceptual measure. Moreover, a closer inspection to the data, shows that the majority of sounds, even if identified correctly by the model, retain a certain degree of uncertainty, as demonstrated by the probability of that sound to be characterized as metal listed in the last column. Indeed the majority of the scores are within $50\pm10\%$ and they could vary depending on the specific training of the machine learning model. To further support this observation, we have built the network of MFCCs for the full set of 180 sounds used in the perceptive test, shown in Fig. 4. The figure displays the first few giant components built with the shortest distances, less than 1% of all possible edges, that is, we allow an edge only if two MFCCs are at a distance that is less than 0.5% of the maximum diameter of the network. From the figure clusters of both unambiguous and ambiguous sounds can be observed confirming the observations made on the original network shown in Figure 1. It is evident that the "ambiguous" sounds are all clustered together and belong to the same modularity class demonstrating the robustness of the MFCCs as a relevant descriptor from a perceptual point of view.

4 Conclusion

By testing sounds using a perceptive measure (MFCC) on a machine learning model trained on physical parameters (PSCC), we reproduced a distribution of ambiguity in the classification of the origin of the sound that is coherent with the results of a human listening tests. In this way we can obtain scores that are coherent with perceptual tests. This is a first step towards a more general machine listening methodology that, if associated with perceptually salient acoustic descriptors that characterize the acoustic information used by the auditory system, might replace time-consuming listening tests and enable perceptual evaluation of huge databases of sounds. It could also be a valuable tool for cognitive studies to point out relevant sound structures (invariants) associated to perceptual categories based on very large data sets. Such sound structures open new possibilities to design evocative synthesis models that enable to control sounds in a perceptually consistent way.

Acknowledgments MBN wishes to acknowledge useful discussions with Alexander Veremyer and the financial support of IMéRA - Institut d'études avancées d'Aix-Marseille Université during his residency in Marseille in the Spring 2019.

Table 1. Table of sounds that are classified by less than 50% of subjects in the labeled material category in the perceptual test. We compare the perceptual scores with four different machine perception scenarios and list the sounds that overlap between the perceptual scores and the selected ML model. HM, percentage of classification in Metal category by human subjects (perceptual scores); MM, sounds in common with Model MFCC - MFCC; MP, sounds in common with model MFCC - PSCC; PP, sounds in common with model PSCC - PSCC; PM, sounds in common with model PSCC - MFCC; and ML, metal score (%) for model PSCC-MFCC. See test for a complete discussion

Ambiguous sounds	HM (%)	MM	MP	PP	PM	ML (%)
metal_092.wav	26				X	16
metal_106.wav	15				X	27
metal_119.wav	48	X				100
metal_1_04.wav	18					59
metal_1_12.wav	15					51
metal_2_06.wav	37				X	45
metal_2_10.wav	33					59
metal_3_10.wav	30					58
metal_4_03.wav	15					78
metal_4_10.wav	19					66
metal_1_06.wav	19					53
metal_3_03.wav	19				X	31
metal_4_05.wav	37				X	36
metal_4_11.wav	37				X	34
metal_1_02.wav	19				X	11
metal_1_03.wav	26				X	8
metal_2_04.wav	30				X	12
metal_3_07.wav	48				X	9
metal_3_12.wav	48				X	9
metal_1_11.wav	44				X	39
wood_019.wav	52		X		X	99
wood_005.wav	52					8
wood_16.wav	52		X		X	67

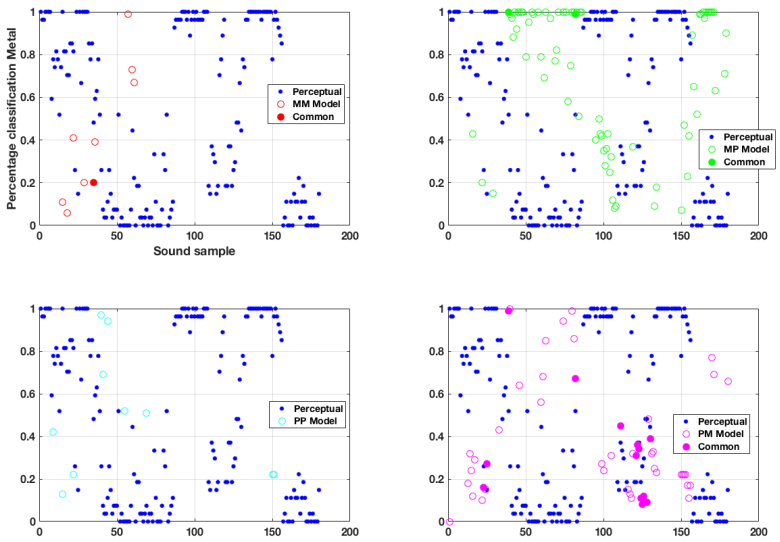

Fig. 3. In blue: Mean scores (corresponding to the percentage of classification in the Metal category) obtained from the perceptual test. These scores are compared with the scores of sounds characterized incorrectly by the four models (MM, MP, PM and MM models). The sounds that are in common are represented with filled markers.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (Jan 2002), https://link.aps.org/doi/10.1103/RevModPhys.74.47
3. Aramaki, M., Besson, M., Kronland-Martinet, R., Ystad, S.: Controlling the perceived material in an impact sound synthesizer. IEEE Transactions on Audio, Speech, and Language Processing 19(2), 301–314 (2011)
4. Aramaki, M., BRANCHERIAU, L., Kronland-Martinet, R., Ystad, S.: Perception of impacted materials: sound retrieval and synthesis control perspectives. In: Ystad, Kronland-Martinet, Jensen (eds.) Computer music modeling and retrieval: genesis of meaning in sound and music, pp. 134–146. Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2009), https://hal.archives-ouvertes.fr/hal-00462245
5. Buongiorno Nardelli, M.: `MUSICNTWRK`: data tools for music theory, analysis and composition. Proceedings of CMMR 2019 *in press* (2019), also at https://arxiv.org/abs/1906.01453

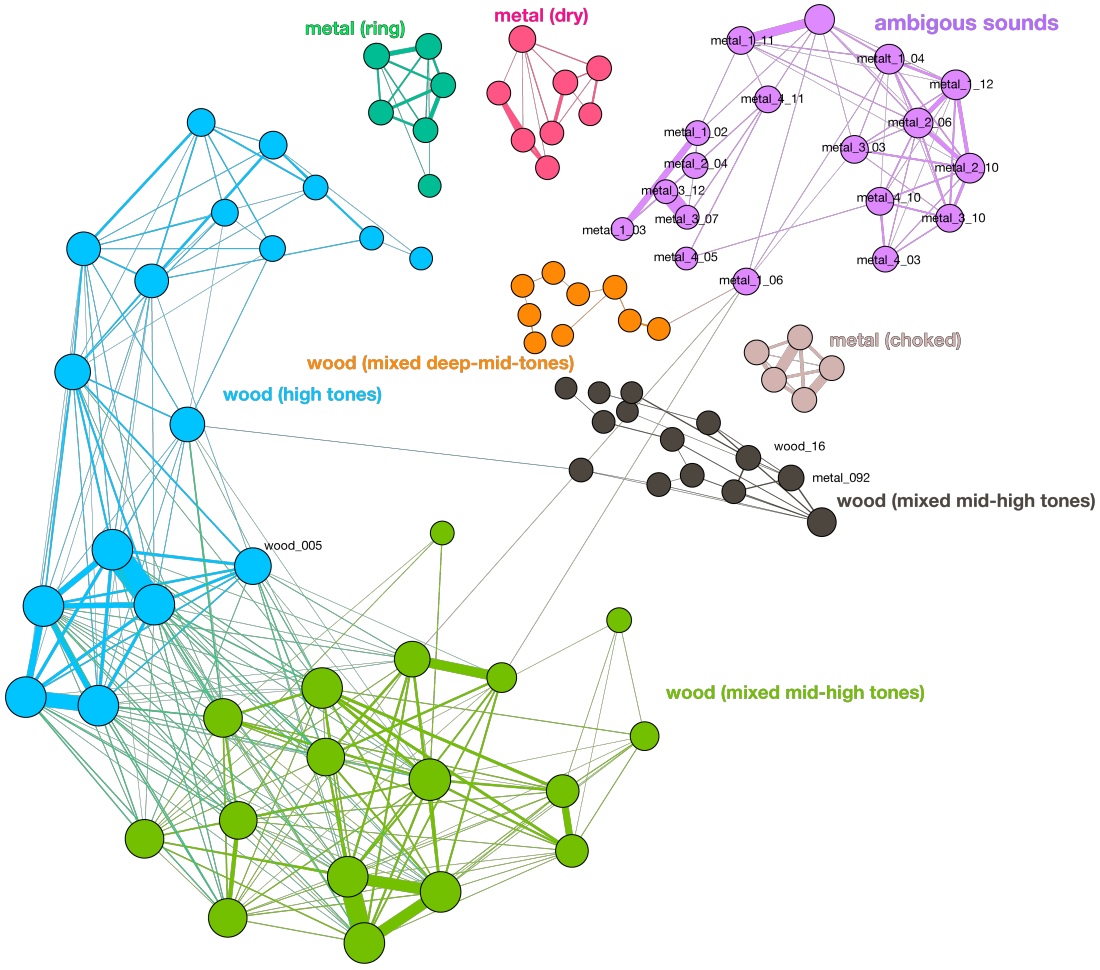

Fig. 4. Network of the MFCC of the sounds from the set used in the perceptual test.

6. Buongiorno Nardelli, M.: Topology of networks in generalized musical spaces. Leonardo Music Journal *in press* (2020), also at http://arxiv.org/abs/1905.01842
7. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing 28(4), 357–366 (August 1980)
8. Giordano, B.L., McAdams, S.: Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. The Journal of the Acoustical Society of America 119(2), 1171–1181 (2006)
9. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.W.: Cnn architectures for large-scale audio classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 131–135 (2017)
10. Kell, A.J., Yamins, D.L., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H.: A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 93(3), 630–644 (2018)
11. Lakatos, S., McAdams, S., Chaigne, A.: The representation of auditory source characteristics : simple geometric form. Perception and Psychophysics 59, 1180–1190 (1997)
12. McAdams, S., Chaigne, A., Roussarie, V.: Psychomechanics of simulated sound sources. Journal of Acoustical Society of America 115(3), 1306–1320 (2004)
13. Piczak, K.J.: Environmental sound classification with convolutional neural networks. 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP) pp. 1–6 (2015)
14. Repp, B.H.: The sound of two hands clapping: An exploratory study. The Journal of the Acoustical Society of America 81(4), 1100–1109 (1987)
15. Serizel, R., Bisot, V., Essid, S., Richard, G.: Acoustic Features for Environmental Sound Analysis. In: Virtanen, T., Plumbley, M.D., Ellis, D. (eds.) Computational Analysis of Sound Scenes and Events, pp. 71–101. Springer (2017), https://hal.archives-ouvertes.fr/hal-01575619
16. Thoret, E., Aramaki, M., Kronland-Martinet, R., Velay, J.L., Ystad, S.: From sound to shape: auditory perception of drawing movements. Journal of Experimental Psychology: Human Perception and Performance 40(3), 983–994 (Jan 2014), https://hal.archives-ouvertes.fr/hal-00939025
17. Warren, W.H., Verbrugge, R.R.: Auditory perception of breaking and bouncing events: a case study in ecological acoustics. Journal of Experimental Psychology : Human Perception and Performance 10(5), 704–712 (1984)