Spectral and Temporal Timbral Cues of Vocal Imitations of Drum Sounds

Alejandro Delgado[1,2†], Charalampos Saitis[1], and Mark Sandler[1]
[1] Centre for Digital Music, Queen Mary University of London, London, United Kingdom
[2] Research and Development Team, Roli Ltd., London, United Kingdom
[†] Corresponding author: a.delgadoluezas@qmul.ac.uk

Introduction

The imitation of non-vocal sounds using the human voice is a resource we sometimes rely on when communicating sound concepts to other people. Query by Vocal Percussion (QVP) is a subfield in Music Information Retrieval (MIR) that explores techniques to query percussive sounds using vocal imitations as input, usually plosive consonant sounds. The goal of this work was to investigate timbral relationships between real drum sounds and their vocal imitations. We believe these insights could shed light on how to select timbre descriptors for extraction when designing offline and online QVP systems. In particular, we studied a dataset composed of 30 acoustic and electronic drum sound recordings and vocal imitations of each sound performed by 14 musicians [1]. Our approach was to study the correlation of audio content descriptors of timbre [2] extracted from the drum samples with the same descriptors taken from vocal imitations. Three timbral descriptors were selected: the *Log Attack Time* (LAT), the *Spectral Centroid* (SC), and the *Derivative After Maximum* of the sound envelope (DAM). LAT and SC have been shown to represent salient dimensions of timbre across different types of sounds including percussion [2]. In this sense, one intriguing question would be to what extent listeners can communicate these salient timbral cues in vocal imitations. The third descriptor, DAM, was selected for its role in describing the sound's tail, which we considered to be a relevant part of percussive utterances.

Method

After computing the three descriptors using the Essentia library, we constructed two distance matrices for each of them: one for the acoustic space of drum samples and another one for the acoustic space of vocal imitations. The first one was built by taking the euclidean distance between a single descriptor taken from all drum samples. It was, therefore, a symmetric matrix of dimensions 30x30. For the second distance matrix, we first measured the euclidean distances between the descriptors taken from the vocal imitations of individual participants and we then averaged the 14 resulting matrices into a single one of size 30x30. Once the two distance matrices were built, we ran the Mantel test, which measures the degree of statistical correlation between two symmetric matrices. Lastly, we picked the two most correlated descriptors and applied a hierarchical clustering algorithm to group imitators based on the distances between their Mantel test scores (Fig. 1b). This was done to see if there were any interpersonal differences in the way these descriptors were imitated. For reference, we also plotted the values of these two descriptors against each other for both drum sounds and vocal imitations, using different colors for the five imitated instruments (Fig. 1a). We did this to assess whether the descriptors from same-category instruments were close to their imitations.

Results

The Mantel test scores of the DAM descriptor for all 14 imitators gave a mean result of $\bar{r}=.561$, a standard deviation of $\sigma_r=.114$, and a maximum p-value of $p<.001$. For the SC, the results were $\bar{r}=.428$,$\sigma_r=.153$, $p<.024$ for a subset of 13 participants, and $\bar{r}=.102$, $p=.411$ for one participant alone. For the LAT descriptor, the scores were $\bar{r}=.011$,$\sigma_r=.130$, $p<.997$.

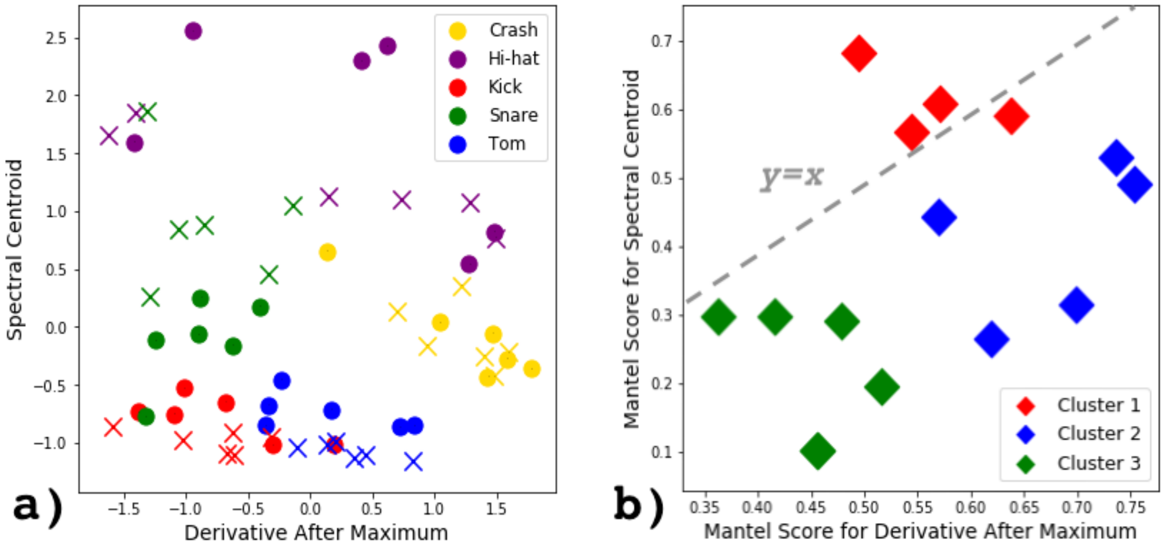

Figure 1: (1a) Normalised mean values of SC plotted against DAM for all sounds (circle markers for drum sounds and crosses for vocal imitations averaged across imitators). (1b) Mantel scores of SC plotted against those of DAM for all imitators (applied hierarchical clustering to group participants).

Discussion

The fact that the LAT descriptor failed to be a good predictor for one acoustic space given the other indicates that, despite playing an important role in timbre perception, LAT might be difficult to reproduce vocally with enough precision, at least in the case of percussive utterances (quick attack). Instead, it appears that listeners can better imitate a temporal envelope cue related to the length of the sound's tail (DAM), an observation that inspires further research from a timbre perception perspective. Interestingly, participants appear to imitate the SC cue dexterously, which seems to reinforce its role in describing one of the most salient dimensions of timbre. Considering QVP systems, the irrelevance of the LAT could make online QVP more challenging, needing more complex descriptors to quickly classify utterances. Meanwhile, offline QVP systems can safely rely on the SC and the DAM to link imitations and samples. Fig. 1a shows how these two descriptors can help cluster the different instruments and their imitations. The emergence of imitator clusters in Fig. 1b warrant further investigation into interpersonal differences in vocal imitation of non-vocal sounds, for example, differences between expert and naive listeners.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068

References

Mehrabi, A., Choi, K., Dixon, S., & Sandler, M. (2018, April). Similarity measures for vocal-based drum sample retrieval using deep convolutional auto-encoders. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 356-360). Calgary, Canada.

Caetano, M., Saitis, C., & Siedenburg, K. (2019). Audio content descriptors of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. Popper and R. Fay (eds) *Timbre: Acoustics, Perception, and Cognition* (pp. 297-333). Springer Handbook of Auditory Research, vol 69. Springer, Cham.