Bonfire happiness or a scratchy affair?: Patterns in free labelling and categorization of percussive stimuli

Lauren Fink[1,2], Madita Hörster[3,4], David Poeppel[2,4,5,6], Melanie Wald-Fuhrmann[1,2], Pauline Larrouy-Maestri[1,2,4]

[1] Music Dept., Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany [2] Max-Planck-NYU Center for Language, Music, and Emotion (CLaME), New York, USA & Frankfurt am Main, Germany [3] Department of Psychology, Ludwig-Maximilians-University, Munich, Germany [4] Neuroscience Department, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany [5] Psychology Department, New York University, New York, USA [6] Ernst Struengmann Institute for Neuroscience, Frankfurt am Main, Germany

Background

Categorizing sounds is a crucial and often automatically performed task, typically studied using short stimuli (e.g., phonemes [1], screams [2]).

Our previous work on sound sequences [3] used:

- 6-10 sec recordings of Nigerian dùndún talking drum performances that were intended to be speech or music
- a categorization task: is the sequence music- or speech-like?

We found: familiarity and acoustic features shape listeners' categorizations. Even participants unfamiliar with the dùndún could categorize above chance whether the drum was talking or playing music.

BUT the labels "speech" and "music" were given to participants, whereas categorization of our auditory environment is usually label-free.

HERE we asked how listeners group the same percussive stimuli, when given no labels.

Methods

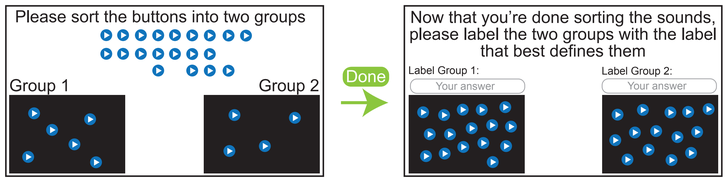
Online participants. N = 180, mean age 26.2 (+/- 8) years

Material. Cleaned versions (removed background noise, clipping, etc.) of the recordings used in [3].

Feature extraction. Pitch, spectral entropy (timbre), amplitude envelope (intensity), inter-onset-intervals (IOI), ratio of IOIs, amplitude modulation spectrum (AMS) peak, and pulse clarity, were calculated using custom scripts and third-party toolboxes in MATLAB.

Procedure.

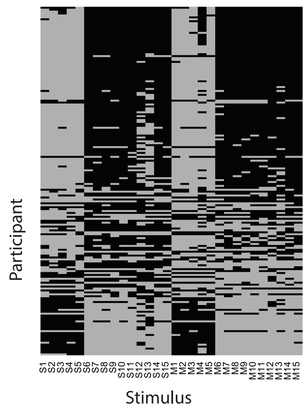
- Participants click to play each sound displayed on the screen, and drag and drop each into one of two separate boxes (see below).
- When the task was completed, participants were asked to label the two boxes.



Left: 30 stimuli were represented in a randomized order as blue play buttons. Participants were asked to group them into two black boxes labelled "Group 1" and "Group 2." Right: subsequently participants labelled their two groups.

Results

Raw grouping data

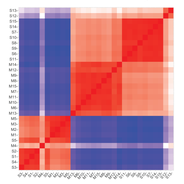


Rows: participant groupings for all 30 stimuli (columns), sorted by # stimuli put into left (gray) vs. right (black) box. Columns are not sorted (S = speech, M = music)

Left box
Right box

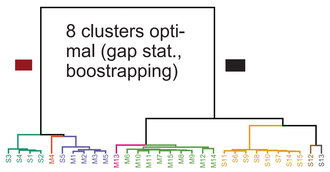
Analyze similarity of participants' response patterns

Dissimilarity matrix

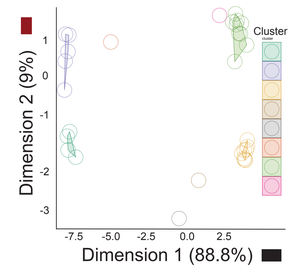


Cluster stimuli based on their distances

Hierarchical clustering

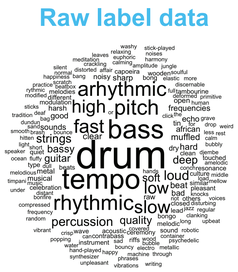


Principal components analysis



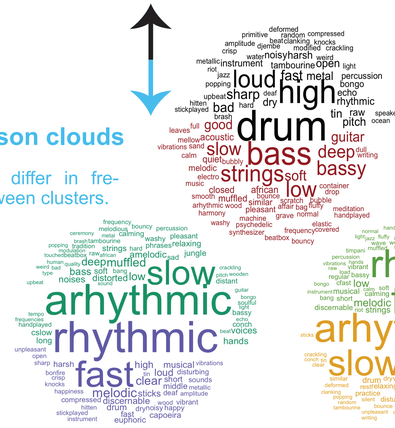
Predict stimulus position in dimensions 1 and 2, respectively, from acoustic features.

Raw label data

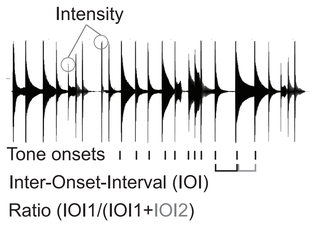


Comparison clouds

Words that differ in frequency between clusters.

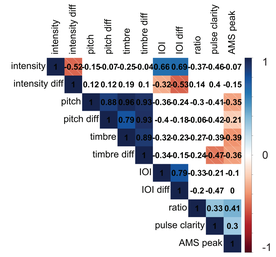


Acoustic features



Check for high correlations among features

Correlation matrix



Proceed with reduced set of features, all $r < |.7|$

Linear regression modelling

	Dim. 1 F(7,22) = 13.95, $p < .001$			**Dim. 2** F(7,22) = 14.22, $p < .001$		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	100.64	-81.76 – 283.04	0.265	30.2	-27.28 – 87.68	0.288
intensity (mean)	**-10.72**	**-20.83 – -0.62**	**0.038**	-2.91	-6.09 – 0.28	0.072
intensity (difference)	2.92	-5.18 – 11.02	0.462	**2.59**	**0.04 – 5.14**	**0.047**
timbre (mean)	**-74.31**	**-96.42 – -52.20**	**<0.001**	-1.97	-8.94 – 4.99	0.563
IOI (mean)	0.02	-0.00 – 0.04	0.103	**-0.01**	**-0.02 – -0.01**	**<0.001**
ratio (mean)	-130.21	-482.87 – 222.45	0.452	-55.05	-166.19 – 56.09	0.316
pulse clarity	-0.01	-15.38 – 15.36	0.999	2.79	-2.06 – 7.63	0.246
amp. mod. spectrum peak	-0.17	-0.68 – 0.34	0.492	0.05	-0.11 – 0.21	0.497
	R^2 / R^2 adj. 0.816 / 0.758			R^2 / R^2 adj. 0.819 / 0.761		

Discussion

- Despite being given no labels, participants show a high degree of consistency in the way they freely categorize dùndún stimuli.
- The highest level of distinction maps onto "loud, high, drums" vs. "bass, strings, low" labels, consists of a mix of speech-like and music-like stimuli, and is predicted by the timbre and intensity of the stimuli.
- The next subdivisions of groupings consist almost exclusively of speech-like vs. music-like stimuli, which map onto the labels of "slow, arhythmic" and "fast, rhythmic," respectively, and similarly to intensity differences between notes and the mean IOI of the stimuli (in line with our previous results).

Ultimately, these results show hierarchical categorization of unfamiliar percussive stimuli presented without context.

Contact Lauren Fink
lauren.fink@ae.mpg.de
website: lkfink.github.io
pdf available (use QR)

References [1] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review, 74*(6), 431.
[2] Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2021). The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific reports, 11*(1), 1-10.
[3] Durojaye*, C., Fink*, L., Roeske, T., Wald-Fuhrmann, M., & Larrouy-Maestri, P. (2021). Perception of Nigerian dùndún talking drum performances as speech-like vs. music-like: The role of familiarity and acoustic cues. *Frontiers in psychology, 12*, 1760.