Perceptual ratio scales of timbre-related audio descriptors

Savvas Kazazis[1†], Philippe Depalle[1], and Stephen McAdams[1]
[1] Schulich School of Music, McGill University, Montreal, Quebec, Canada
[†] Corresponding author: savvas.kazazis@mail.mcgill.ca

Introduction

Most of the past research on timbre psychophysics has focused on determining acoustic correlates of perceptual dimensions derived from multidimensional scaling of dissimilarity ratings, in order to quantify the ways in which we perceive sounds to differ. However, there is little empirical evidence to date demonstrating that acoustic features derived from correlational analysis causally correspond to psychological dimensions. Most importantly, even for cases in which the causality has been verified, there is almost no research on understanding how the sensation magnitudes of such acoustic features are apprehended. To the best of our knowledge, there have been no previous attempts in psychophysical scaling of timbre-related audio descriptors other than perhaps the preliminary results of Almeida et al. (2017) who attempted to derive a ratio scale of timbral brightness as a function of spectral centroid.

In order to investigate whether listeners perceive timbre-related descriptors on perceptual ratio scales, we conducted a ratio scaling experiment in which we tested the following descriptors: spectral centroid, spectral spread, spectral skewness, odd-to-even harmonic ratio, spectral deviation, and spectral slope (measured in dB/octave) (Peeters et al., 2011).

Method

Twenty participants, 6 female and 14 male, with a median age of 25 years (range: 18–41) were recruited from the Schulich School of Music, McGill University. All of them were self-reported amateur or professional musicians with formal training in various disciplines such as performance, composition, music theory, and sound engineering. Participants were compensated for their time.

The stimulus sets of each audio feature were synthesized and (wherever possible) independently controlled through additive synthesis with appropriate spectral amplitude distributions. For spectral centroid, spread, and skewness, the stimuli had an f0 of 120 Hz and the initial spectrum (prior to filtering according to Gaussian distributions) contained harmonics up to Nyquist frequency (i.e., 22.05 kHz). In addition, for spectral spread and skewness, three different stimulus sets were constructed with centroids at 1640, 5600, and 7800 Hz. For odd-to-even ratio, deviation, and slope, three different sound sets were constructed with f0's at 120, 300, and 720 Hz and with 9 harmonics except for the sets of spectral deviation in which 16 harmonics were used.

The listener's task was to equisect a continuum of a particular audio feature. Each equisection was performed using the *progressive solution* according to which listeners progressively partition the continuum formed by the stimuli into a number of equal-sounding intervals. In order to create a continuum within a range of a particular feature, several stimuli were constructed with multiple imperceptible successive differences. The total number of sounds used for each stimulus set and the ranges of feature values for a particular set are indicated in Table 1. In a first step, listeners bisected the continuum of an audio feature into two equal-sounding intervals, by triggering each stimulus with a cursor along a horizontal bar that contained the stimuli, and by placing a marker over the stimulus-bar. Each resulting section was then bisected in the next step. In total there were three bisections: the first one was made between the stimuli of the total range, and the other two between the lower and upper bisected ranges. In a final step, listeners were presented with all their bisections and were instructed to make further fine adjustments so that all four intervals they had created in the previous steps sounded equal. The equality of sensory intervals implies that the intervals themselves have ratio properties (Marks & Gescheider, 2002) and thus, the results of this experiment led to ratio scale measurements.

The equisection scales were then derived by fitting well-behaving functions to the listeners' ratings (Figure 1, left panel)[1]. The criteria used for choosing the form of the function were monotonicity, maximum explained variance, and good continuation (i.e., no oscillations) outside the tested range, which was useful for extrapolating the fitting function (Figure 1, right panel). The reliability of the derived scales across listeners was evaluated according to *Cronbach's alpha* (α).

Table 1: The ranges of feature values within designated stimulus set are shown in bold. The number of sounds on which the feature values were computed are shown in parentheses. The reported ranges for the spectral spread and skewness stimulus sets were computed on stimuli with 5600-Hz spectral centroid. Linear regression over normally distributed spectral amplitudes is futile.

Stimulus Sets (# sounds)	Feature Ranges					
	Centroid (Hz)	Spread (Hz)	Skewness	Odd-to-Even Ratio	Deviation	Slope (dB/octave)
Centroid (505)	**[1642, 9560]**	[479, 480]	[0.00, 0.02]	[1.00, 1.00]	[0.00, 0.00]	-
Spread (100)	[5600, 5600]	**[181, 1439]**	[0.00, 0.00]	[1.00, 1.00]	[0.00, 0.00]	-
Skewness (97)	[5600, 5600]	[1079, 1080]	**[–0.88, 0.96]**	[1.00, 1.00]	[0.00, 0.00]	-
Odd-to-Even Ratio (349)	[1260, 1500]	[768, 848]	[0.00, 0.21]	**[0.25, 1250.00]**	[0.00, 0.11]	[–11.67, –2.62]
Deviation (265)	[1723, 2550]	[1292, 1396]	[0.00, 0.28]	[1.00, 1.19]	**[0.00, 0.06]**	[0.00, –5.04]
Slope (349)	[332, 2082]	[134, 785]	[–1.04, 6.68]	[1.25, 15.05]	[0.00, 0.03]	**[–24.00, 5.44]**

Results

The results indicate that listeners can produce ratios of descriptor values, which in turn enabled the construction of perceptual ratio scales of each descriptor tested. As evidenced by *Cronbach's alpha*, the reliabilities of the scales were overall excellent, with the scales of spectral centroid and spectral skewness having the highest reliability ($\alpha = 0.96$). The lowest reliability was observed for the equisections of spectral spread ($\alpha = 0.89$), followed by the odd-to-even ratio ($\alpha = 0.78$). With the exception of spectral skewness, for which the best fitting function on the median ratings was a third-order polynomial, the best fitting functions for the rest of the descriptors were all power functions albeit exhibiting significantly different shapes, which indicates that each descriptor is perceived on a different psychophysical scale. After identifying the form of the function fitted on listeners' equisections, the psychophysical scales were constructed by defining a *unit* for each scale and its *zero point*. With the exception of spectral centroid, for which the zero point of the scale was assigned to 20 Hz, which marks the lower limit of pitch perception, the rest of the scales were assigned a zero point that has a physical meaning (e.g., zero skewness). The units of the scales were defined by empirically assigning specific numerals to the points of the equisection scale (e.g., by assigning the numeral 10 to the 1-kHz spectral centroid), so as to facilitate comparisons between the derived perceptual ratio scales of all descriptors (Figure 1).

Discussion

The stimuli used in each of the presented experiments were constructed through specifically designed additive-synthesis algorithms that enabled control of each audio descriptor independently of the rest and that therefore isolated as much as was feasible the effect of each descriptor on listeners' perceptions. However, spectral slope is the descriptor that is the most difficult to control independently of centroid, spread and skewness descriptors, because these two sets of descriptors are physically intercorrelated (Table 1).

The construction of psychophysical scales based on such univariate stimuli allowed for the establishment of *cause and effect* relations between audio features and perceptual dimensions, contrary to past research that has relied on multivariate stimuli and has only examined the correlations between the two. The derived

[1] The upper limit of the centroid range is high (10 kHz), essentially to serve as an anchor point, and accommodate specific high-pass sound signals such as a hi-hat.

scales along with their respective units designate a *perceptual coordinate system of audio features* in which sounds can be grouped and ordered according to their perceived sound qualities that relate to each descriptor. The perceptual coordinate system of descriptor values could potentially be used to define a "control surface" for applications that include computer-aided orchestration, perceptually motivated sound effects, and synthesis algorithms.

In addition, audio descriptors have been widely used as predictor variables in statistical regression models for interpreting and predicting listeners' responses on a variety of tasks that relate to timbre. However, the physical values of these predictors may lead to false-positive interpretations about their perceptual significance on a particular task. The derived scales allow timbre researchers to use perceptually informed values of spectral descriptors as predictors in their statistical models that may lead to more sustainable conclusions and accurate interpretations in terms of perception.

Nonetheless, it should be mentioned that this study does not imply that all the descriptors tested here constitute perceptual dimensions because, it only provides evidence that individual descriptors can be perceived on perceptual ratio scales when the rest of them remain relatively constant. However, this study does not test the extent to which each descriptor is independently perceived when multiple descriptors covary. Verifying or rejecting that hypothesis is left for future work.

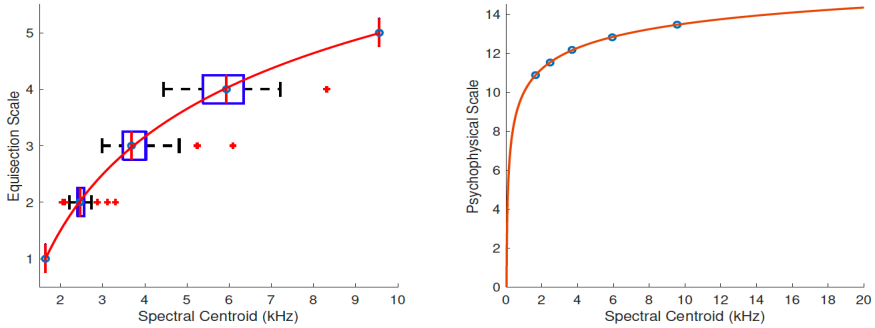

Figure 1: Equisection and psychophysical scales of spectral centroid. Whiskers extend to 2.7 the Standard Deviation.

Acknowledgments

We would like to thank Bennett K. Smith for programming the user interface of the experiment, and Erica Huynh for her help in recruiting and running participants.

References

Almeida, A., Schubert, E., Smith, J., & Wolfe, J. (2017). Brightness scaling of periodic tones. *Attention, Perception & Psychophysics*, *79*, 1892–1896.

Marks, L. E., & Gescheider, G. (2002). Psychophysical scaling. In H. Pashler & J. Wixted (eds.), *Stevens' handbook of experimental psychology: Perception and motivation; learning and cognition* (pp. 91–138). Hoboken, NJ, US: John Wiley & Sons Inc.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, *130*(5), 2902–2916.