Timbre Discrimination for Brief Instrument Sounds

Bigoni, Francesco; Dahl, Sofia

Published in:
Proceedings of the 19th International Society for Music Information Retrieval Conference

Creative Commons License
CC BY 4.0

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Bigoni, F., & Dahl, S. (2018). Timbre Discrimination for Brief Instrument Sounds. In *Proceedings of the 19th International Society for Music Information Retrieval Conference* (pp. 128-134). dblp Computer Science Bibliography. http://ismir2018.ircam.fr/doc/pdfs/249_Paper.pdf

TIMBRE DISCRIMINATION FOR BRIEF INSTRUMENT SOUNDS

Francesco Bigoni
Sound and Music Computing
Aalborg University – Copenhagen
fbigon17@student.aau.dk

Sofia Dahl
Dept. of Architecture, Design and Media Technology
Aalborg University – Copenhagen
sof@create.aau.dk

ABSTRACT

Timbre discrimination, even for very brief sounds, allows identification and separation of different sound sources. The existing literature on the effect of duration on timbre recognition shows high performance for remarkably short time window lengths, but does not address the possible effect of musical training. In this study, we applied an adaptive procedure to investigate the effect of musical training on individual thresholds for instrument identification. A timbre discrimination task consisting of a 4-alternative forced choice (4AFC) of brief instrument sounds with varying duration was assigned to 16 test subjects using an adaptive staircase method. The effect of musical training has been investigated by dividing the participants into two groups: musicians and non-musicians. The experiment showed lowest thresholds for the guitar sound and highest for the violin sound, with a high overall performance level, but no significant difference between the two groups. It is suggested that the test subjects adjust the weightings of the perceptual dimensions of timbre according to different degrees of acoustic degradation of the stimuli, which are evaluated both by plotting extracted audio features in a feature space and by considering the timbral specificities of the four instruments.

1. INTRODUCTION

Timbre is a primary vehicle for sound source recognition and, from a cognitive perspective, sound identity [10]. The auditory system is designed to identify sound sources: this enables us to discern a melody in a complex soundscape, follow what is being said by a speaker, or step aside when something fast and dangerous appears to be approaching. As an example which is more related to music consumption, listeners are able to identify musical genres better than chance in a fraction of a second (the shortest duration tested is 125 ms [9]).

Although so important to our auditory system, timbre is often defined in a negative manner — as, for instance, in Plomp's (1970) operational definition: "Timbre is that attribute of sensation in terms of which a listener can judge that two steady complex tones having the same loudness, pitch and duration are dissimilar" (quoted in [12]). Timbre can be described as a set of perceptual attributes which are either continuously varying (timbral semantics such as attack sharpness, brightness, richness) or discrete (perceptual features such as the pinched offset of a harpsichord sound) [10]. For the former category of attributes, a number of objective acoustic correlates can generally be found among spectro-temporal audio features, e.g. spectral centroid, attack time and spectral envelope; for the latter, the objective correlates are harder to identify [10].

Being complex and multidimensional, timbre is usually modelled employing a so-called multidimensional scaling (MDS), i.e. fitting the dissimilarity ratings given by a group of listeners on a set of sounds to a *timbre space* of perceptual attributes and respective acoustic correlates [10]. While the basic MDS model assumes the same perceptual dimensions for all listeners, more recent models (e.g. CLASCAL by McAdams et al. [11]) account for different weightings of the perceptual dimensions (by individual listeners or classes of listeners) and for the effect of the features that are specific to an individual timbre, called "specificities" (basically related to the aforementioned discrete features).

The studies that evaluate the effect of brief duration on timbre perception exhibit a decidedly different approach from MDS research: quoting Suied et al., MDS models aim at finding the most *prominent* perceptual dimensions of specific sounds through dissimilarity ratings, whereas the problem of timbre recognition for brief sounds asks to identify the most *informative* ones [21]. Inside this field, the prevalent area of interest is speech: in a seminal paper from 1942, Gray investigated phoneme cues for short vowel sounds, and coined the term "phonemic microtomy" [5]. More recently, Robinson and Patterson found that timbral cues for brief vowel stimuli are not pitch-assisted [18]. Generally, the measured performance is above chance for durations as short as a single glottal pulse cycle, on the order of 3 ms.

Only a few studies deal with non-speech sounds: Clark et al. asked their test subjects to identify orchestral instruments for varying window length and position for gated stimuli [3]; Robinson and Patterson replicated their previous study using synthesized instrument sounds, achieving slightly lower performance levels than for voice stimuli [17]. In later articles, the sound recognition problem has been stated in different terms, by taking the subjective reaction times rather than the temporal thresholds of the stimuli into account [1,22].

In 2014, Suied et al. published what we consider by far the most exhaustive contribuition on the topic, as well as the most relevant reference for our paper [21]. In a series of timbre discrimination experiments, participants were asked to identify whether a sound belonged to a target category (e.g. strings, percussion, voice) or a distractor category. Very short duration thresholds were found, on the order of 8-16 ms. The best performance was for voice, followed by percussion (marimba and vibraphone). Subsequent experiments rejected the effect of feedback and training on the performance for the voice stimuli; and, finally, demonstrated that source recognition based on timbral cues is fast and robust to stimulus degradation, with a clear advantage for voice signals.

While it may not be surprising that humans are highly trained to identify sounds as belonging to the "voice" category, one could expect more variability in the exposure to instrumental sounds. Suied et al. [21] did not report any information regarding the musical training of their participants. We would expect that listeners who are trained as musicians would exhibit lower threshold values compared to non-musicians.

Previous studies [17, 21] have used constant stimuli lengths, with durations that are doubled. The increasing differences in durations help to reduce the test time, but also make it difficult to pinpoint where and how thresholds differ between individuals or groups of listeners. We expect musically trained and untrained listeners to differ in the overall threshold of instrument discrimination, but there may be an interaction with the instrument type. Suied et al. [21] found a lower performance for the "strings" category compared to "percussion". In order to efficiently target the listeners' individual thresholds, an adaptive approach is an attractive alternative.

In this paper, we investigate whether musical training has an effect on the perceptual interaction between timbre and duration through a timbre discrimination task, using brief sounds of varying length. Our goals were threefold: 1) applying an adaptive staircase method to estimate the temporal thresholds of timbre discrimination for a small sound set (four instruments: guitar, clarinet, trumpet and violin); 2) determining if musical training has an effect on the task; 3) relating the degradation of timbre descriptors (caused by the length shortening) to the perceptual adaptation strategies of the participants.

2. METHOD

We anticipated the range of thresholds to vary between participants, and therefore opted for an adaptive test procedure. Adaptive methods are designed to be time-efficient and focus the presentation of stimuli around the perceptual threshold of interest by adapting the level of presentation according to the past responses of the participant (in our case, the indication of the heard instrument). Compared to the method of constant stimuli, the adaptive procedure can quickly move from presenting clearly audible and distinguishable stimuli to a range where performance is more difficult. By gradually decreasing the *step size* after a change in test subject performance, the method allows to zoom in rapidly on the threshold level. Depending on the criteria for changes in presentation level and step size, the adaptive procedure can be designed to target different performance levels on the psychometric function (see [7] for an overview). For this study, we chose the simple up-down staircase method [8], as this does not require assumptions on the shape of the psychometric function.

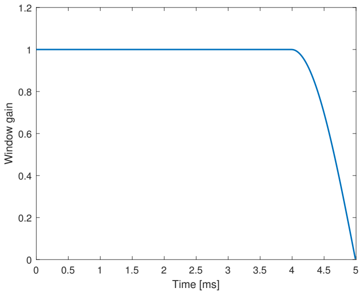

Figure 1. Time window employed for stimulus gating (in this case, the window length is 5 ms).

2.1 Stimuli

The four stimuli (guitar, clarinet, trumpet and violin) were picked from an existing database of anechoic recordings of acoustic instruments [20] [23]. The audio files were recorded at a sample rate of 48 kHz and a resolution of 24 bits, using a 32-channel microphone array. The audio editing was performed in the digital audio workstation Reaper. Four source files were created by mixing down the respective 32 channels to a mono track (with no instrument-specific mix), bounced at 16-bit/44.1 kHz. In the source files, the instrumentalists are playing a C4 at a *ff* dynamic. The pitch of the source files was already normalized, as the instruments were all tuned at A4 = 443 Hz [1] . Sounds were loudness-normalized to -18 LUFS using the SWS extension in Reaper. The sound snippets were prepared on the fly in MATLAB between the presentations, by applying a suitable window (i.e. a rectangular window with 4 samples of silence at the start and a 1 ms equal-power fade-out at the end) of the required duration, starting from time 0: an example is shown in Figure 1. Thus, onset information has been included in each snippet.

2.2 Participants

A convenience sample of 16 participants was tested, consisting of 13 males and 3 females with ages ranging from 22 to 50 ($\mu_{age} = 32$, $\sigma_{age} = 9$) recruited through author Bigoni's personal network. Participants indicated their age and sex (if willing to disclose) and whether they had normal hearing (no testing was made to assess this); they

[1] This gives a fundamental frequency of $443/2^{(9/12)} = 263.41$ Hz at C4.

were informed that their personal data would be anonymized. Each test subject was assigned to one of two groups (*musicians* or *non-musicians*) by asking if he/she had 5 or more years of formal musical training and/or performance experience. This left the border between the two groups somewhat flexible, giving the option to music students and amateur musicians to choose their group according to their confidence level. The groups were fairly balanced with respect to sample size: 9 musicians and 7 non-musicians. Of the 9 musicians, 5 are primarily performing on wind instruments, whereas the other 4 play different combinations of guitar, piano, drums and electronics. Despite this inter-group difference, we do not assume that any of the subjects were biased towards a specific instrument type.

2.3 Setup

The playback system consisted of the laptop internal sound card, driven with ASIO4ALL drivers for Windows, and a pair of Beyerdynamic DT 990 Pro, 250 Ohm headphones. Even though the DT 990 Pro do not have a flat frequency response, we assume that the sound coloration introduced by the headphones did not alter the relative timbre perception.

2.4 Procedure

The experimental setup was implemented in MATLAB. It features a simple GUI and consists of three steps: 1) creation of a test subject entry in a database; 2) soundcheck: the subject can click on four buttons to play the source files (guitar, clarinet, trumpet, violin) while adjusting the headphones volume to a comfortable level. The soundcheck also constitutes a small training session on the four timbres, to avoid confusion at the beginning of the discrimination task; 3) timbre discrimination task: an adaptive staircase method (simple up-down) with four interleaved tracks (the four instrument timbres). For each sound presentation, the participants made a 4-alternative forced choice (4AFC) test. For each track, the following procedure applied: when a participant correctly identified the instrument, the duration of the next presentation (of the same instrument) would be reduced by the step size (initially, 40 ms); on the other hand, the duration of the next presentation would be augmented by the step size after a wrong answer. In the literature, right and wrong answers usually get represented by positive (+) and a negative (-) signs respectively. In the light of this notation, a *run* consists of a sequence of presentations that get answers of the same sign, and a *reversal* occurs at each change of sign. Thus, after the first misidentification, the first reversal would occur, the first run would end, the step size would be halved and the duration would be lengthened (by 20 ms) for each wrong answer; at the next right answer, another reversal would occur, the second run would would end and the next presentation would, again, be shortened by the step size. For each track, the initial snippet length was set to 500 ms, the step sizes (halved at the end of each odd run) to 40/20/10/5/2 ms and the stop criterion to 8 reversals.

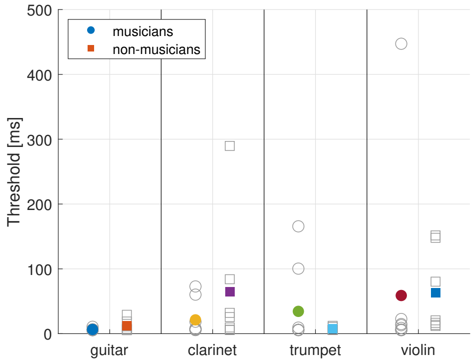

Figure 2. Thresholds for each group and instrument sound, both individual (grey) and group-based means (colour-filled symbol) for musicians (circle) and non-musicians (square). The different variability in thresholds between instruments is clearly seen.

The order of stimuli presentations was made by interleaving the four tracks using a random permutation of a 4x4 integer sequence of indices. This technique allows the same timbre to be replayed before a sequence of 4 is completed, removing a potential bias by avoiding the possibility of the subject anticipating the next sound. The typical test time was 15-20 minutes (setup + 150-200 presentations), with a shortest played duration of 1 ms.

2.5 Analysis

The typical shortest durations played during tests ranged between 1 and 10 ms across all participants. The four thresholds (one per instrument) were computed as the mean of the thresholds at reversals. The simple up-down estimates point $p = 0.50$ on the psychometric function, which is well above chance performance for 4AFC ($p = 0.25$).

The performance difference between the two groups (musicians and non-musicians) was estimated by performing a mixed ANOVA (between-subjects variable: 2 levels of musical training, within-subjects variable: 4 instrument sounds).

Moreover, eight sound snippets were created using the found thresholds and two audio descriptors (spectral centroid and spectral irregularity) were computed in MATLAB using *MIRtoolbox 1.7* [6].

3. RESULTS

The outcome of the experiment is shown in Figure 2, with the threshold means and standard deviations re-stated in Table 1. It can be seen that the threshold values vary considerably across groups and instruments, with very low mean values for guitar and trumpet (for non-musicians only), and mean values almost 10 times higher for violin. Furthermore, variability is large for all thresholds, except guitar. A Q-Q plot showed that the data violates

	Mean (std) [ms]	
Stimulus	**Mus**	**Non-mus**
Guitar	6.4 (1.6)	12.2 (8.8)
Clarinet	21.3 (26.1)	64.7 (102.9)
Trumpet	34.4 (58.2)	7.4 (2.7)
Violin	58.9 (145.7)	63.2 (63.7)

Table 1. Mean and standard deviation of thresholds of timbre discrimination. Mus = values from 9 musicians, Non-mus = values from 7 non-musicians.

the normality assumption, while a Levene's test indicated that group variances are homogeneous. After improving normality with a 10-log transformation, we proceeded with a mixed ANOVA (α = 0.05) on the transformed data, looking for statistically significant effects of musical training and stimulus. The between-subjects factor was group (musicians/non-musicians) and the within-subjects factor was target (instrument). The Q-Q plot of the ANOVA residuals is approximately linear, so we assume that this analysis is robust with respect to our dataset. While the stimulus effect was statistically significant ($F(3, 42) = 5.035$, $p = 0.005$), the musical training effect on timbre discrimination of brief sounds was not ($F(1, 14) = 1.134$, $p = 0.305$). The interaction effect was not significant either ($F(3, 56) = 2.416$, $p = 0.080$).

Post-hoc pairwise t-tests (two-sided, Holm-Bonferroni correction) on the instrument thresholds showed that the guitar mean threshold was significantly different from violin ($t(15) = -1.833$, $p = 0.024$), but not from clarinet ($t(15) = 4.873$, $p = 0.095$) or trumpet ($t(15) = -1.187$, $p = 0.871$). No other contrasts were significant — trumpet vs violin ($t(15) = -1.780$, $p = 0.081$), clarinet vs trumpet ($t(15) = -1.913$, $p = 0.472$), clarinet vs violin ($t(15) = -2.097$, $p = 0.871$).

As a rough approximation of an MDS model (with equal perceptual weightings and no specificities), we created a feature space using two calculated audio descriptors: spectral centroid and spectral irregularity. Spectral irregularity is a measure of the amplitude deviation between successive peaks of the spectrum (implemented in *MIRToolbox 1.7* [6]), a feature analogous to spectral deviation. The two descriptors were chosen for two reasons: 1) they are informative as a set, as they are not strongly correlated (see e.g. [15]); 2) they can be computed as single-number features, and are thereby easy to visualize and more robust to the short snippet durations than other descriptors which require frame-based analysis, e.g. spectral flux.

The feature space is seen in Figures 3 (mean thresholds) and 4 (individual thresholds), with colours denoting the four instruments: guitar (black), clarinet (blue), trumpet (red), and violin (green). The values for the 2 s source files are not plotted in Figure 4, thereby the different x-axis scale. As a general trend, the spectral centroid gets lowered for reduced duration. On the other hand, the spectral irregularity seem to either stay constant (clarinet), increase (guitar and violin), or fluctuate (trumpet) depending on the instrument sound.

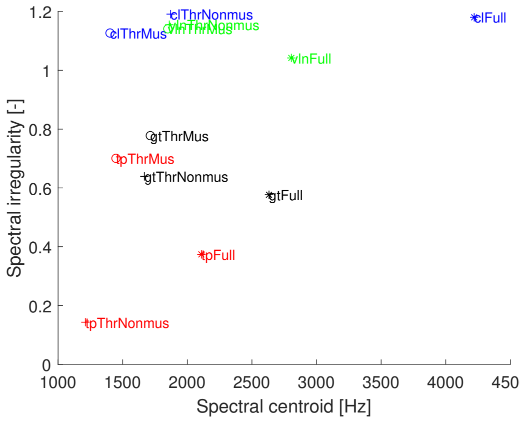

Figure 3. Mean thresholds for the four instruments in a feature space spanned by two audio descriptors: spectral centroid and spectral irregularity (computed using [6]) for the four instruments. Labels of the format xyz, with x defining the instrument (gt = guitar (black), cl = clarinet (blue), tp = trumpet (red), vln = violin (green)), y defining the duration of the audio file (Thr = audio snippet cut at threshold length, $Full$ = 2 s long source file) and z defining the test group (Mus = musicians, $Nonmus$ = non-musicians).

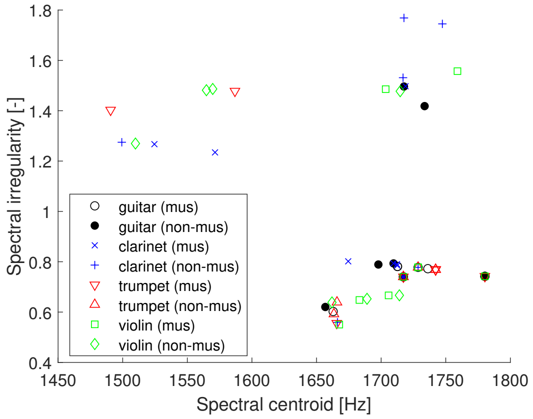

Figure 4. Thresholds for all participants and the four instruments in a feature space spanned by two audio descriptors: spectral centroid and spectral irregularity (computed using [6]) for the four instruments. The 2 s long source files are labelled $xFull$, with x defining the instrument (gt = guitar (black), cl = clarinet (blue), tp = trumpet (red), vln = violin (green).

Additionally, we computed the log attack times of the four source files ($LAT_{guitar} = -1.921$, $LAT_{clarinet} = -0.930$, $LAT_{trumpet} = -0.506$, $LAT_{violin} = -1.092$). However, since the attack phase is incomplete for the thresshold snippets, this feature was less informative in relation to the perceptual result.

4. DISCUSSION

Using an adaptive procedure, we investigated the temporal thresholds for timbre discrimination of different sounds for musically trained and untrained listeners. Our findings agree with the existing literature with respect to the overall high performance of both groups. The overall performance was best for guitar, both with respect to duration thresholds and variability, as confirmed by post-hoc tests.

While we measured an average violin threshold of about 60 ms, Suied et al. [21] report window lengths corresponding to above chance performance as small as 8 ms for string sounds in a first experiment, then doubled to 16 ms in a subsequent trial (with other instruments as distractors). Suied et al. offer two plausible interpretations of the high performance levels: a successful adjustment of the auditory representation of the stimuli, which is specific to the signal gating setup – which is described as a computationally challenging form of unsupervised learning – or an efficient activation of spectral cues even for deteriorated stimuli. Both interpretations could apply to our experiment.

Some differences between our method and that used by Suied et al. [21] are worth mentioning. While we asked our participants to indicate the instrument heard in a 4AFC-task that got more challenging over time, Suied et al. asked their subjects to indicate whether the sound was a target sound (50% of presentations) or not. The range of possible sounds was also wider in their study, with seven other instruments beside those belonging to the target. Before their test, however, the participants listened to the targets repeatedly for all stimuli durations. It is therefore difficult to judge whether this would result in a harder task for the participants compared to the one we chose. The task of categorizing (target vs distractor) or identifying (4AFC) sound are different: although the thresholds are still in the same range, the peculiarity of the tasks might explain the discrepancy between string instrument thresholds.

Our results showed no effect of musical training, possibly as a result of the moderate sample size and our operational definition of musician. Rather than dividing the participants in two groups (musicians and non-musicians), using an index of musical sophistication (e.g. [13, 14]) could provide a more sensitive measure and allow for a regression analysis. Moreover, differences between musicians and non-musicians have been shown in a combined instrument/voice discrimination task [2] as well as in brain activity [4], so group differences in terms of cognitive strategies should not be dismissed. On the other hand, our result agrees with the thesis of MDS researchers, meaning that the inter-individual differences in perceptual weighting of different timbral dimensions are independent of musical training [10].

The very low thresholds and low variability for guitar (both groups) and trumpet (non-musicians) seem to indicate the presence of early acoustical markers that could be identified by listeners. Even though it is commonly assumed that onset is highly significant for sound recognition (see e.g. [16] and [19]), this premise is not universally accepted by timbre/duration studies. It has been doubted by Clark et al. [3] and then strongly disputed by Suied et al., who argue that onset information might even be misinformative for the discrimination of string instruments (due to the noisy transients caused by the initial contact between bow and string) [21]. However, the results shown by Suied et al. seem to indicate that the performance difference between the two window constraints (random and onset) is both stimulus-specific and inconsistent across window lengths. The gating used in their experiment applied a raised-cosine window, while we applied a rectangular window with a fixed fade-out length (1 ms) for all durations, as shown in Figure 1. Thus, our approach would be more likely to preserve the original amplitude for longer time (but with a sharper fade-out), while the stimuli prepared by Suied et al. would decrease in amplitude in a quicker and smoother manner. As for the acoustic analysis of the stimuli, Suied et al. explain the effect of gating in terms of “spectral splatter” (the smearing of spectral features when short time windows are applied) and refute the assumption that trivial spectral features are relevant to the timbre discrimination task, based on a simulation of auditory excitation patterns derived by the employed stimuli [21].

As a direct investigation of the stimuli, we placed the source files and threshold sound snippets in a feature space (Figures 3 and 4). Without perceptual weightings, this representation lacks the depth of MDS models, but it is useful to trace the deterioration of a set of audio descriptors (spectral centroid and spectral irregularity) for reduced duration. Even though the full set of thresholds forms three clusters (Figure 4) and most of the guitar data points are located in one of the clusters (lower right), it is hard to conclude that the guitar advantage is due to the fact that the stimulus retains specific audio features for brief durations. The threshold differences could instead be explained by the different placement of discrete timbral features (specificities), which are hard to correlate to the audio descriptors. The guitar advantage might by explained by our choice of the onset condition, which preserves the characteristic ”twang” even for very brief window lengths. A more systematical investigation of the evolution of a set of audio descriptors for different stimuli and progressively decreasing duration would be worth considering for future work.

5. CONCLUSION

In this paper, we have investigated the temporal thresholds of timbre discrimination for four instrument sounds. Although the thresholds from the staircase method varied significantly between stimuli, with means ranging from < 15 ms (guitar) to $\approx$ 60 ms (violin), there was no significant effect of musical training on timbre discrimination. The guitar advantage can be explained by considering our choice of window position (always including the sound onset) and the timbral specificities of the guitar sound. The overall low thresholds agree with the findings of the existing literature, and the adaptive staircase method seems to constitute a viable alternative to the method of constant stimuli for the chosen task. As a result, participants were able to adjust the weights of the perceptual dimensions of timbre to the acoustic degradation of the stimuli as the durations were reduced. Future research could use this investigation as a point for departure for a further examination of the duration thresholds, using larger sound sets and multiple window conditions.

6. ACKNOWLEDGEMENTS

The authors would like to thank all the participants and two anonymous reviewers for helpful comments on an earlier version of the manuscript.

Author Bigoni performed the main part of this work as part of a course in Music Perception and Cognition at the MSc program Sound and Music Computing, Aalborg University – Copenhagen. Author Dahl supervised the work and participated in writing the manuscript.

7. REFERENCES

[1] Trevor R. Agus, Clara Suied, Simon J. Thorpe, and Daniel Pressnitzer. Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, 131(5):4124–4133, May 2012.

[2] Jean-Pierre Chartrand and Pascal Belin. Superior voice timbre processing in musicians. *Neuroscience Letters*, 405(3):164–167, Sep 2006.

[3] Jr Clark, David Luce, Robert Abrams, Howard Schlossberg, and James Rome. Preliminary Experiments on the Aural Significance of Parts of Tones of Orchestral Instruments and on Choral Tones. *Journal of the Audio Engineering Society*, 11(1):45–54, January 1963.

[4] Garry C. Crummer, Joseph P. Walton, John W. Wayman, Edwin C. Hantz, and Robert D. Frisina. Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch. *The Journal of the Acoustical Society of America*, 95(5):2720–2727, May 1994.

[5] Giles Wilkeson Gray. Phonemic microtomy: The minimum duration of perceptible speech sounds. *Communications Monographs*, 9(1):75–90, 1942.

[6] Olivier Lartillot and Petri Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[7] Marjorie R Leek. Adaptive procedures in psychophysical research. *Perception & psychophysics*, 63(8):1279–1292, 2001.

[8] H. Levitt. Transformed Up–Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2B):467–477, Feb 1971.

[9] Sandra T Mace, Cynthia L Wagoner, David J Teachout, and Donald A Hodges. Genre identification of very brief musical excerpts. *Psychology of Music*, 40(1):112–128, 2012.

[10] Stephen McAdams. Musical Timbre Perception. In Diana Deutsch, editor, *The Psychology of Music*, pages 35–67. Elsevier, 3rd edition edition, 2013.

[11] Stephen McAdams, Suzanne Winsberg, Sophie Donnadieu, Geert De Soete, and Jochen Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, Dec 1995.

[12] Brian Moore. *An Introduction to the Psychology of Hearing: Sixth Edition*, pages 284–286. BRILL, 2013.

[13] Daniel Müllensiefen, Bruno Gingras, Jason Musil, and Lauren Stewart. The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one*, 9(2):e89642, 2014.

[14] Joy E Ollen. *A criterion-related validity test of selected indicators of musical sophistication using expert ratings*. PhD thesis, The Ohio State University, 2006.

[15] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, 2011.

[16] Jean-Claude Risset and David L Wessel. Exploration of timbre by analysis and synthesis. In Diana Deutsch, editor, *The Psychology of Music*, pages 113–169. Elsevier, 1999.

[17] Ken Robinson and Roy D. Patterson. The Duration Required to Identify the Instrument, the Octave, or the Pitch Chroma of a Musical Note. *Music Perception: An Interdisciplinary Journal*, 13(1):1–15, October 1995.

[18] Ken Robinson and Roy D. Patterson. The stimulus duration required to identify vowels, their octave, and their pitch chroma. *The Journal of the Acoustical Society of America*, 98(4):1858–1865, October 1995.

[19] E. L. Saldanha and John F. Corso. Timbre Cues and the Identification of Musical Instruments. *The Journal of the Acoustical Society of America*, 36(11):2021–2026, Nov 1964.

[20] Noam R. Shabtai, Gottfried Behler, Michael Vorländer, and Stefan Weinzierl. Generation and analysis of an acoustic radiation pattern database for forty-one musical instruments. *The Journal of the Acoustical Society of America*, 141(2):1246–1256, Feb 2017.

[21] Clara Suied, Trevor R. Agus, Simon J. Thorpe, Nima Mesgarani, and Daniel Pressnitzer. Auditory gist: Recognition of very short sounds from timbre cues.

The Journal of the Acoustical Society of America, 135(3):1380–1391, March 2014.

[22] Clara Suied, Patrick Susini, Stephen McAdams, and Roy D. Patterson. Why are natural sounds detected faster than pips? *The Journal of the Acoustical Society of America*, 127(3):EL105–EL110, March 2010.

[23] Stefan Weinzierl, Michael Vorländer, Gottfried Behler, Fabian Brinkmann, Henrik von Coler, Erik Detzner, Johannes Krämer, Alexander Lindau, Martin Pollow, Frank Schulz, and Noam R. Shabtai. A Database of Anechoic Microphone Array Measurements of Musical Instruments, Apr 2017. Available at http://dx.doi.org/10.14279/depositonce-5861.2.