Modeling the onset advantage in musical instrument recognition

Kai Siedenburg,[1] Marc René Schädler,[1] and David Hülsmeier[1]

Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all*, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany*

(Dated: 26 November 2019)

Sound onsets provide particularly valuable cues for musical instrument identification by human listeners. It has yet remained unclear whether this *onset advantage* is due to enhanced perceptual encoding or the richness of acoustical information during onsets. Here this issue was approached by modeling a recent study on instrument identification from tone excerpts [Siedenburg 2019, JASA, 145(2), 1078-1087]. A simple Hidden Markov Model classifier with separable Gabor filterbank features simulated human performance and replicated the onset advantage observed previously for human listeners. These results provide evidence that the onset advantage may be driven by the distinct acoustic qualities of onsets.

The following article appeared in K. Siedenburg, M.R. Schädler, D. Hülsmeier (in press). Modeling the onset advantage in musical instru- ment identification The Journal of the Acoustical Society of America, 146(6), EL523–EL529 and may be found at https://doi.org/10.1121/1.5141369

I. INTRODUCTION

The identification of musical instruments is a central task in music perception (e.g., Rentfrow and Levitin, 2019). Research on the acoustical underpinnings of instrument identification still constitutes rough terrain, mainly because the candidate acoustical feature sets are high-dimensional and redundant (Handel, 1995; McAdams, 2019). A landmark effect concerns sound onsets, which are suspected to provide particularly valuable cues for instrument identification: if presented with sound excerpts, human listeners more easily identify instrument sounds from onset portions compared to other portions of the sound (Saldanha and Corso, 1964; Schaeffer, 2017)—a behavioral effect that we refer to as *onset advantage*. Importantly, this effect does not imply that all instrumental sounds become unidentifiable without onsets, because informative cues may be extracted across the full sound duration and the degree to which this is possible may depend on the specific instrument at hand (Agus *et al.*, 2019). However, the psychoacoustical factors playing into the onset advantage largely remain unclear. Auditory modeling approaches are in a position to provide valuable insights into this issue.

Previous research has started to characterize the onset advantage, although results have not been unequivocal. Among the more recent studies, Suied *et al.* (2014) had listeners categorize sounds into broad categories such as sung voices, percussion sounds, or string instruments, using gated excerpts of musical sounds. Categorization performance was above chance for very short gates: 4 ms for voices and 8 ms for instruments whereas identification scores were at ceiling at 64 ms gate duration or more. The authors obtained mixed results

regarding the importance of onsets: instrumental sounds, but not vocal sounds benefited from gates being positioned at sound onsets. This means for vocal sounds there was sufficient redundancy of cues across the whole duration for listeners to achieve robust categorization, which may be due to the general robustness of voice recognition (Agus *et al.*, 2012). Ogg *et al.* (2017) measured the durations required for human listeners to discriminate between musical instrument sounds, human speech, and human environmental sounds. Results suggested that listeners required 25 ms for robust discrimination and that the presence of onsets was beneficial, in particular for instrument sounds. In the present study, only musical instrument sounds were considered.

For a refined discussion, it is important to distinguish between the notions of onset and the so-called *transient*. Here, transients are understood as short-lived and stochastic sound bursts that are measurable in the sound signal (e.g., the hammer hitting the piano string without the quasi-stationary sound waves that emanate from the harmonically vibrating string). Therefore, transients should not be confused with the onset as a whole: all sounds have onsets but not necessarily pronounced transients (e.g., the smooth onset of a clarinet sound). Siedenburg (2019) then quantified the individual contribution of transient components to instrument identification. Stationary and transient components were extracted from the audio signal and instrument identification was tested for gated excerpts containing stationary plus transient components, or stationary components alone. Results indicated that the omission of transient components at the onset impaired identification accuracy only by 6 percentage points. A much stronger effect was obtained by shifting the position of the gate from the onset to the middle portion of the tone, impairing overall identification accuracy

by 25 percentage points. These results portrayed short-lived transient as of relatively minor importance in instrument identification compared to the importance of retaining the onset.

Nonetheless, in the experiment by Siedenburg (2019) the important question regarding the origin of the onset advantage was left open: Are listeners focussing on informative acoustic features that are only available during the onset? Or do equally informative acoustic features exist throughout the full duration of instrumental sounds that listeners ignore, either because of their redundancy or because of the particular salience of onsets in auditory neural processing? Based on the analysis of timbre dissimilarity ratings, Grey (1977) suggested that the buildup of sinusoidal components acts as a perceptual dimension of musical instrument sound. Because of the various differences between dissimilarity rating and identification tasks (cf., Siedenburg and McAdams, 2017), and because of a lack of replication this finding has not been very conclusive. Alternatively, one may suppose that neural coding in the auditory system is tuned to onsets. It is known that already the cochlear nucleus exhibits specialized onset units (Rhode and Greenberg, 1992) and neurons all along the auditory pathway exhibit particularly strong responses to onsets (Heil, 1997). In psychophysics, onset dominance in binaural processing has been thoroughly documented (Houtgast and Aoki, 1994) and the adaptation mechanisms implemented in models of auditory processing yield a pronounced response overshoot at onsets (Jepsen *et al.*, 2008), even for simple signals such as ramped sinusoids. All together, these factors suggest a more elaborate encoding of onsets which in turn could imply that acoustic onset features are taken as more reliable for sound identification, whether they are acoustically more informative or not. Thus, the degree to which acoustical and neural factors play into the onset advantage remains unclear.

Here, a modeling approach is used to further disentangle these issues. We utilize a Hidden Markov Model (HMM) classifier in conjunction with separable Gabor filterbank (SGBFB) features, an approach that has proven valuable in the domain of speech recognition and psychoacoustic modelling for normal-hearing listeners as part of the Framework for Auditory Discrimination Experiments (FADE) (Schädler *et al.*, 2016). Originally conceived as speech recognizer, FADE has quite successfully modelled human performance on a variety of psychoacoustic tasks without requiring any internal calibration data beyond training on the specific task at hand. As additional baseline features, we use Mel-frequency cepstral coefficients (MFCCs) and log-Mel spectra. The simulations are set up in an analogous way to the main experiment from Siedenburg (2019), and the results are interpreted in terms of their implications on the role of acoustical factors in the onset advantage.

II. METHODS

A. Previous experiment used for modeling

The main experiment by Siedenburg (2019) serves as the starting point for the present modeling study. That experiment was divided into a training and a test phase. In the training phase, musician participants were presented with sounds from ten test instruments: piano, guitar, harp, vibraphone, marimba, trumpet, clarinet, flute, violin, and cello. Sounds were of 250 ms duration and for each instrument there were sounds with twelve different pitch levels (C4/262Hz to B4/494Hz). Subsequently, listeners were trained in the identification task and obtained feedback on 60 trials (6 per instrument). Notably, before the start of each

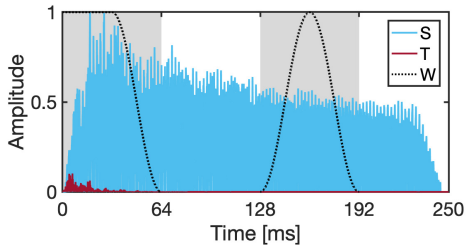

FIG. 1. (Color online). A rectified waveform of a piano tone, decomposed into stationary (S) and transient (T) components. The dashed line corresponds to the gating window (W). In the experimental conditions S+T@0ms and S@0ms, the down-ramped window starting at 0 ms was used. In the condition S+T@128ms, the window extending from 128 to 192 ms was used.

experimental block, participants again went through a passive exposure phase, listening to all the original 250 ms sounds, as in the very first part of the training. This means, participants had extensive exposure to the full 250 ms sounds before being tested on short excerpts. In the subsequent test phase, listeners identified instruments from 64 ms segments of sounds composed of stationary and transient components ($S+T$) or stationary components alone (S), taken either from the onset (*@0ms*), or from the middle portion of the sounds (*@128ms*). Stationary and transient extraction was achieved by applying a specifically developed algorithm (Siedenburg and Doclo, 2017). Figure 1 shows the example of a piano tone, including the stationary and transient components and the gating window at the *@0ms* and *@128ms* positions.

B. Modeling rationale

The modeling was set up in such a way as to create an analogous scenario compared to the instrument identification experiment with human listeners: a classifier was trained on

the set of full 250 ms sounds and tested on the short 64 ms excerpts. Hence, the scenario required the classifier to generalize to sounds with durations of around one fourth of the training sounds.

As back-end, we used a classic HMM classifier with one Gaussian component per state. Several points speak for using HMM for the present modeling task: HMM perform well with small training sets, explicitly encode the temporal dimension of auditory stimuli, and are not overly powerful (in comparison to more recent architectures such as deep neural networks, which may even learn high-level tasks from raw data), hence allowing us to differentiate the quality of the acoustic input features. Recent research has demonstrated that this modeling approach is well suited for various aspects of auditory modeling, including speech intelligibility, elementary psychoacoustics, and hearing loss (Kollmeier *et al.*, 2016; Schädler *et al.*, 2016, 2018). Here, we tested HMM with a variable number of 1-6 states. For every instrument and number of states, separate classifiers were trained for stimuli adjusted to the input levels 65 dB SPL $\pm$ 0, 3, 6, 9, 12, 15 dB. This range was selected to cover short- and long-term level changes that usually occur in music recordings. Each classifier was then tested in the 65 dB condition and we report average recognition performance. A more detailed description of the modeling architecture is provided by Schädler *et al.* (2016).

C. Acoustic features

Three sets of acoustic features were used in the simulations: log-Mel Spectrograms, Mel-frequency cepstral coefficients (MFCC), and separable Gabor Filterbank features (SGBFB). Log-Mel spectra were computed with a window length of 25 ms and 10 ms hop size, and

the linear frequency axis was subsequently warped to 36 bins with Mel spacing, that is, with frequency centers between 64 and 11874 Hz. MFCCs were computed by applying a discrete cosine transform in the spectral dimension, and the first 21 coefficients were used. These were concatenated with the first and second order derivatives along the time axis (the so-called *delta* and *double-delta* coefficients, respectively). SGBFB features were computed by using a temporal and a spectral modulation filterbank operating on the log-Mel spectrogram. These covered spectral modulations from 0.03 to 0.25 cycles per channel, and temporal modulations from 6.2 to 25 Hz. The combination of temporal and spectral filters resulted in a feature set of 570 coefficients (as compared to 63 MFCC and 36 Log-Mel features). The SGBFB and MFCC features were used with mean and variance normalization, the log-Mel spectrum was used without normalization. For more information on the feature sets and details of their implementation, the reader is referred to Schädler *et al.* (2012, 2016) or the model code[1].

III. RESULTS

A. Performance comparison

Accuracies (i.e., proportion correct classifications) for classifiers trained on the full 250 ms sounds are depicted in Fig. 2, together with experimental results from human listeners. The SGBFB features performed best and yielded accuracies of .99, .62, .55, and .41 (averaged across the different number of states of the classifier) for the four conditions 0-250ms, S+T@0ms, S@0ms, and S+T@128ms, respectively. MFCCs obtained weaker results with

averages of .99, .31, .27, .20 for the four conditions. This means that both feature sets could be easily fitted to the training set 0-250ms, but generalized considerably worse to the shorter test excerpts. With accuracies of .61, .36, .35, and .31 across the four conditions, log-Mel spectra did not yield a similarly good fit to the training set, but slightly better results compared to the MFCC on the short test excerpts. Importantly, both the SGBFB and MFCC features still provided a pattern of results that qualitatively resembled that of human listeners.

As further visible in Fig. 2, the number of states of the HMM was a critical factor for the classification performance: SGBFB yields highest performance for an HMM with three states, and notably, this classifier was the only one that reaches human performance in the test conditions up to error tolerance (95% confidence intervals of human performance). With MFCCs, the best classifier contained five states and performed worse than human performance by around 20 percentage points, but showed a very similar decay of performance across the test conditions. Surprisingly, log-Mel spectra did not at all resemble the human pattern of performance and the best performing classifier had only one state, that is, it only encoded static spectral information. Note that log-Mel spectra are the only features that did not explicitly encode spectro-temporal information or modulations (SGBFB are tailored to do so; MFCC do so by virtue of its discrete cosine transform and the delta-coefficients). This supports the view that robust generalization in instrument identification relies on the explicit encoding of spectro-temporal information (cf., Patil *et al.*, 2012).

None of the classifiers performed better than human listeners in the test conditions, which leaves open the possibility that this was due to general difficulties in classifying the short

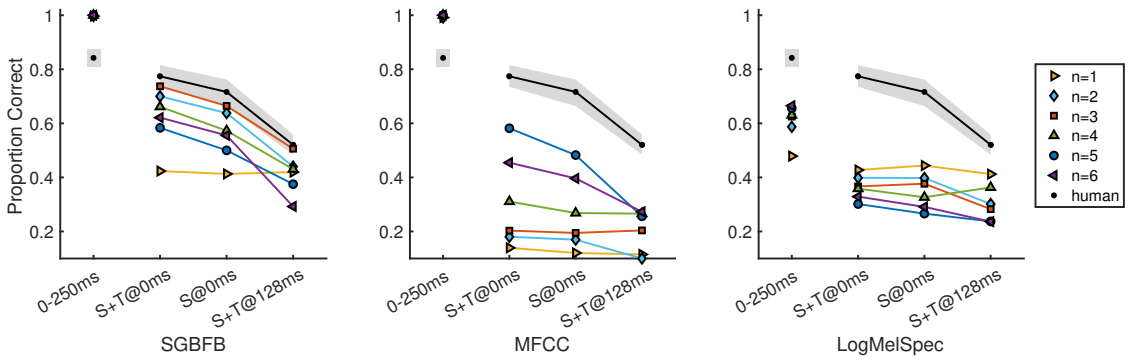

FIG. 2. (Color online). Identification results for the three different feature sets. Test conditions on x-axis: condition 0-250 ms refers to the unaltered sounds with 250 ms duration (the reference condition for human listeners and the model). The other conditions are indexed by whether stationary or transient components were used: for instance, $S+T$@0ms indicates that test signals consisted of stationary (S) and transient (T) components with the gating window of 64 ms length starting at 0 ms. Performance for the three different feature sets is provided in the three different panels. The number of states of the HMM is indicated in the legend. The shaded area indicates 95% confidence intervals of human scores.

64 ms excerpts. Table I provides accuracies for classifiers that were trained on a merged set of sounds, containing both the full 250 ms sounds and the 64 ms excerpts from the three conditions S+T@0ms, S@0ms, and S+T@128ms. With this merged training set, performance was much better for the short excerpts with averages of 75% correct classifications or more for the SGBFB and MFCC models. That is, the recognition performance for this merged training set was on par with or even exceed human performance. Therefore, this

TABLE I. Proportion correct classification for models trained on a merged set of both 250 ms and 64 ms sounds. Columns index the test sets, rows the feature sets. Table entries correspond to mean and range (square brackets) across the number of states (1-6).

	0-250ms	S+T@0ms	S@0ms	S+T@128ms
SGBFB	.60 [.54, .70]	.83 [.79, .86]	.76 [.72, .79]	.80 [.79, .82]
MFCC	.27 [.21, .34]	.83 [.72, .88]	.75 [.63, .79]	.75 [.69, .78]
Log-Mel Spec	.56 [.45, .68]	.48 [.40, .55]	.49 [.44, .53]	.40 [.37, .43]

latter simulation suggests that the primary difficulty for the present classifiers was not to classify short excerpts per se, but to generalize from 250 ms sounds to 64 ms excerpts.

B. Recognition of excerpts over time

In order to more detailedly probe the distribution of acoustic information over time, the classifiers trained on the full 250 ms sounds were tested on 64 ms excerpts that were obtained through gating with a raised cosine window starting at different temporal positions (0, 32, 64, ..., 192 ms). Note that the last excerpt only had a duration of 58 ms (extending from 192-250 ms). An additional test condition extending from 0-64 ms was included, which did not use the full cosine window but only the fade out part, hence preserving the original attack and closely resembling the experiment condition S+T@0ms (although it also contained the residual noise that was missing in S+T@0ms). Here, classifiers with three or five states

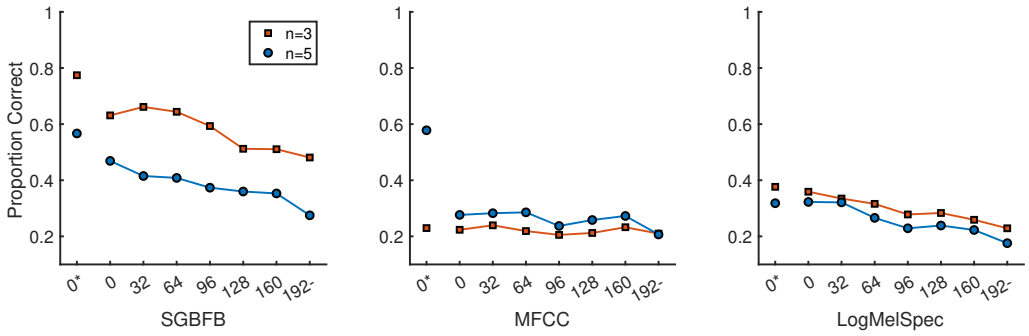

FIG. 3. (Color online). Identification results for models with different feature sets and number of states tested on different excerpts. The x-axis indicates the starting point (in ms) of the 64 ms gating window. The condition 0* denotes the first 64 ms of the sound without fade in, but fade out; all other sounds contain both fade in and out. The condition 192^{-} corresponds to the gate extending from 192-250 ms.

were evaluated because these models had performed best on the test set in conjunction with SGBFB and MFCC features.

Figure 3 depicts the results for the three different feature sets. Performance dropped by more than 10 percentage points from the 0-64ms* to the 0-64ms condition for the classifier using SGBFB features with three states. The SGBFB features further exhibited a gradual decline of performance over time after the onset, which suggests that the distinctiveness of acoustical cues gradually worsens over time for classifiers trained on the full 250 ms sounds. Even more drastic was the drop of more than 30 percentage points for the classifier using MFCC features with five states, which continued to perform poorly for the consecutive excerpts. This finding indicates that classifiers that best performed in these simulations are

those that rely on the encoding of precise onset information, providing further evidence for acoustic information to play an important role in the onset advantage.

IV. CONCLUSION

To investigate the importance of acoustic cues for musical instrument identification, we trained HMM classifiers using SGBFB, MFCC, and log-Mel spectrum features on 250 ms sounds and tested generalization to short 64 ms excerpts that contained stationary and transient sound components. Classifiers using SGBFB with three states generalized best to the test excerpts and showed very similar results compared to human listeners. Testing the classifiers on excerpts gated at different time points of the sound indicated that the best performing classifiers rely on precise onset information. More specifically, performance dropped drastically when the initial onset was manipulated through gating and performance degraded gradually the more the excerpts stemmed from later portions of the sound. These results provide converging evidence that the acoustical richness of sound onsets itself could be exploited by listeners as a cue, which then may give rise to the onset advantage.

It is important to bear in mind that the current results stem from a scenario which trained the classifiers on the full 250 ms sounds, that is, that the full sounds acted as a reference. Although this seems to be the best possible analogy to the main experiment of Siedenburg (2019) where listeners were trained and heavily (re-)exposed to the full sounds, future research could contextualize this scenario by exposing listeners and machine classifiers to sounds of varying durations in a training phase. These pursuits could also attempt to

account for the long-term knowledge about more diverse classes of instrument sounds that the musician participants may have utilized in the experiment.

These simulations demonstrate that a simple Hidden Markov Model classifier with separable Gabor filterbank features replicates the onset advantage in instrument identification. The classifier essentially implemented an elaborate encoding of acoustic information and no component of the classifier architecture was dedicated to onsets per se (e.g. much in contrast to Newton and Smith, 2012). On the basis of these results, one does not need to assume a specialized neural encoding of onsets to explain the onset advantage in instrument identification. In the convoluted reality of auditory processing, however, it may well be the case that the acoustical properties and the neural encoding of sounds act in concert and jointly contribute to the onset advantage. In fact, neural sound coding could even have evolved to optimally exploit the acoustic richness of sound onsets (Młynarski and McDermott, 2018; Theunissen and Elie, 2014).

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for insightful comments. KS has received funding from the European Unions Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Skodowska-Curie Grant Agreement No. 747124. This work was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 390895286 EXC 2177/1 and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 352015383 SFB 1330 A 3.

[1]Model code is available at https://github.com/m-r-s/fade/

Agus, T. R., Suied, C., and Pressnitzer, D. (**2019**). "Timbre recognition and sound source identification," in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Springer), pp. 59–85.

Agus, T. R., Suied, C., Thorpe, S. J., and Pressnitzer, D. (**2012**). "Fast recognition of musical sounds based on timbre," The Journal of the Acoustical Society of America **131**(5), 4124–4133.

Grey, J. M. (**1977**). "Multidimensional perceptual scaling of musical timbres," The Journal of the Acoustical Society of America **61**(5), 1270–1277.

Handel, S. (**1995**). "Timbre perception and auditory object identification," in *Hearing*, edited by B. C. Moore, Handbook of Perception and Cognition (Academic Press, San Diego, CA), pp. 425–461.

Heil, P. (**1997**). "Auditory cortical onset responses revisited. I. First spike timing," Journal of Neurophysiology **77**(5), 2616–2641.

Houtgast, T., and Aoki, S. (**1994**). "Stimulus-onset dominance in the perception of binaural information," Hearing research **72**(1-2), 29–36.

Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception.," The Journal of the Acoustical Society of America **124**(1), 422–438.

Kollmeier, B., Schädler, M. R., Warzybok, A., Meyer, B. T., and Brand, T. (**2016**). "Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation

noise with fade: Empowering the attenuation and distortion concept by plomp with a quantitative processing model," Trends in Hearing **20**, doi: 10.1177/2331216516655795.

McAdams, S. (**2019**). "The perceptual representation of timbre," in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Springer), pp. 23–57.

Młynarski, W., and McDermott, J. H. (**2018**). "Learning midlevel auditory codes from natural sound statistics," Neural Computation **30**(3), 631–669.

Newton, M. J., and Smith, L. S. (**2012**). "A neurally inspired musical instrument classification system based upon the sound onset," The Journal of the Acoustical Society of America **131**(6), 4785–4798.

Ogg, M., Slevc, L. R., and Idsardi, W. J. (**2017**). "The time course of sound category identification: Insights from acoustic features," The Journal of the Acoustical Society of America **142**(6), 3459–3473.

Patil, K., Pressnitzer, D., Shamma, S. A., and Elhilali, M. (**2012**). "Music in our ears: The biological bases of musical timbre perception," PLOS Computational Biology **8**(11), e1002759.

Rentfrow, P. J., and Levitin, D. J. (**2019**). *Foundations in music psychology: Theory and research* (MIT Press, Cambridge, MA).

Rhode, W. S., and Greenberg, S. (**1992**). "Physiology of the cochlear nuclei," in *The mammalian auditory pathway: Neurophysiology*, edited by A. N. Popper and R. R. Fay, Springer Handbook of Auditory Research (Springer, Heidelberg, Germany), pp. 94–152.

Saldanha, E., and Corso, J. F. (**1964**). "Timbre cues and the identification of musical instruments," The Journal of the Acoustical Society of America **36**(11), 2021–2026.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (**2012**). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," The Journal of the Acoustical Society of America **131**(5), 4134–4151.

Schädler, M. R., Warzybok, A., Ewert, S. D., and Kollmeier, B. (**2016**). "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception," The Journal of the Acoustical Society of America **139**(5), 2708–2722.

Schädler, M. R., Warzybok, A., and Kollmeier, B. (**2018**). "Objective prediction of hearing aid benefit across listener groups using machine learning: Speech recognition performance with binaural noise-reduction algorithms," Trends in Hearing **22**, doi: 10.1177/2331216518768954.

Schaeffer, P. (**2017**). *Treatise on Musical Objects: An Essay Across Disciplines*, **20** (Univ of California Press).

Siedenburg, K. (**2019**). "Specifying the perceptual relevance of onset transients for musical instrument identification," The journal of the Acoustical Society of America **145**(2), 1078–1087.

Siedenburg, K., and Doclo, S. (**2017**). "Iterative structured shrinkage algorithms for stationary/transient audio separation," in *Proc. of the 20th Int. Conf. on Digital Audio Effects (DAFx-20), Edinburgh, Sep 5–8.*

303 Siedenburg, K., and McAdams, S. (**2017**). "Four distinctions for the auditory "wastebasket"
304 of timbre," Frontiers in Psychology **8**, 1747.
305 Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., and Pressnitzer, D. (**2014**). "Auditory
306 gist: Recognition of very short sounds from timbre cues," Journal of the Acoustical Society
307 of America **135**(3), 1380–1391.
308 Theunissen, F. E., and Elie, J. E. (**2014**). "Neural processing of natural sounds," Nature
309 Reviews Neuroscience **15**(6), 355–366.