Specifying the perceptual relevance of onset transients for musical instrument identification

Kai Siedenburg[a)]
Department of Medical Physics and Acoustics, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

(Received 12 September 2018; revised 24 January 2019; accepted 4 February 2019; published online 26 February 2019)

Sound onsets are commonly considered to play a privileged role in the identification of musical instruments, but the underlying acoustic features remain unclear. By using sounds resynthesized with and without rapidly varying transients (not to be confused with the onset as a whole), this study set out to specify precisely the role of transients and quasi-stationary components in the perception of musical instrument sounds. In experiment 1, listeners were trained to identify ten instruments from 250 ms sounds. In a subsequent test phase, listeners identified instruments from 64 ms segments of sounds presented with or without transient components, either taken from the onset, or from the middle portion of the sounds. The omission of transient components at the onset impaired overall identification accuracy only by 6%, even though experiment 2 suggested that their omission was discriminable. Shifting the position of the gate from the onset to the middle portion of the tone impaired overall identification accuracy by 25%. Taken together, these findings confirm the prominent status of onsets in musical instrument identification, but suggest that rapidly varying transients are less indicative of instrument identity compared to the relatively slow buildup of sinusoidal components during onsets. © *2019 Acoustical Society of America*. https://doi.org/10.1121/1.5091778

[JJL] Pages: 1078–1087

I. INTRODUCTION

It is a common idea in music psychoacoustics that timbre cues at sound onsets are of central importance for the identification of musical instruments by human listeners. Acoustical explorations of this idea may date back as far as the 1940s, when the advent of tape recording technology allowed sounds to be systematically manipulated by means of cutting and splicing. The radio engineer and musician Pierre Schaeffer pioneered in testing the perceptual implications of different temporal gatings of sounds (cf. Schaeffer, 2017) and made the observation that sounds such as piano tones lose aspects of their identity if presented bare of onsets. This has led to the idea that onset information is perceptually more valuable compared to other sound components that are present in the so-called *steady state*, the portion of a tone where its waveform (or short-time spectrum) is relatively constant. As of today, however, surprisingly little is known about the specific acoustic ingredients that give rise to this effect.

A component of specific importance to onsets is the so-called *transient*. Here, transients are defined as short-lived and chaotic bursts of acoustical energy, such as the sound of the hammer hitting the piano string (without the sound from the harmonically vibrating string). It is important to note that according to this definition, transients should not be confused with the full onset: all sounds have onsets but not necessarily pronounced transients—think of a clarinet tone with a smooth attack. Neither do transients exclusively occur at the onset—think of the return of the hopper of the harpsichord at the release of the key (usually accompanied by sustained harmonic resonance in the soundboard).

Regarding the perceptual identification of instruments, rapidly varying onset transients are often claimed to be of prime importance, particularly in the audio processing literature (Daudet, 2005; Zaunschirm *et al.*, 2012), although no definitive proof has been provided to date. There yet exist alternative acoustic properties of sound onsets that could bear diagnostic information about sound identity, such as the comparatively slow buildup of sinusoids which could be particularly informative at sound onsets (Grey, 1977). The primary goal of the present study was to better understand the relevance of transients and more slowly varying sinusoidal components for the identification of musical sounds.

A. Previous research

Rigorous empirical research on instrument identification has emerged in the 1960s. Early studies used tape recordings of musical instrument tones that were manipulated by means of cutting and splicing for experimental purposes. In a well-known study, Saldanha and Corso (1964) suggested that several factors contribute to the identification of orchestral instruments: pitch, the presence of vibrato, the experimental session (test/re-test), and the presented excerpt (onset, steady state, offset). Although identification accuracy was generally poor (around 40% correct identifications), offsets did not bear perceptually useful information and shortening the steady state from 9 to 3 s did not negatively affect the results. On the contrary, discarding onsets decreased identification accuracy by 15 percentage points, although performance remained above chance. Unfortunately, no clear criterion

a)Electronic mail: kai.siedenburg@uni-oldenburg.de

was provided as to how the endings of onsets were determined and hence the durations of the segments that were used as onsets remained unclear.

Other research from around that time came to similar conclusions regarding the role of onsets. Clark *et al*. (1963) presented excerpts from the onset or steady part of recorded instrument tones to listeners with durations varying from 60 to 600 ms. The authors observed that even short portions such as the first 60 ms of tones contained sufficient information for musicians to discriminate instruments. Using recorded tones of 6 s duration, Elliott (1975) observed that discarding the first and last half second from sustained instrument tones with an overall duration of 6 s significantly impaired identification performance of several orchestral instruments.

Exploring timbre dissimilarity perception, Grey (1977) used musical instrument tones emulated by additive synthesis and observed that the ordering of sounds along one dimension of a timbre space obtained from dissimilarity ratings corresponded to the synchronicity of the onsets of sounds' sinusoidal components. Also studying dissimilarity ratings, Iverson and Krumhansl (1993) tested the role of onsets by using three sets of tones: full tones (duration: 2–3.3 s), onsets (first 80 ms), and the remainder (first 80 ms removed). They found strong commonalities between the multidimensional scaling solutions of all three sets, which were interpreted as reflecting a form of acoustical invariance across segments. However, today it is known that an excerpt of 80 ms can be more than enough for instrument identification (Suied *et al*., 2014), making it likely that listeners also relied on instrument identity or sound source properties in their dissimilarity judgments (cf. Siedenburg *et al*., 2016). Unfortunately, it thus seems hard to differentiate whether the supposed invariance in Iverson and Krumhansl (1993) arose from invariance of aspects of the sensory representations or from invariance in the inferred sound source mechanism (which in turn may have affected dissimilarity judgments) or a combination of both aspects.

Subsequent research has shown that relatively short durations are necessary to discriminate instruments. Robinson and Patterson (1995) presented listeners with short sound excerpts, excised from synthetic emulations of brass, flute, harpsichord, and string sounds. For the identification of isolated sounds, it was observed that even for single cycles of periodic tones (corresponding to 2.9–30.5 s depending on pitch), musicians and nonmusicians achieved an impressive performance of around 75% and 50% of correct responses, respectively. Note that because cycles were presented repeatedly, no temporal cues (onset, offset) were present in the sounds, which highlights the importance of spectral cues for instrument identification. In a similar vein, Suied *et al*. (2014) tested the minimal duration required for the correct recognition of sound source categories. Listeners heard cosine-shaped gated segments of musical sounds and were required to respond to target categories (sung voices, percussion sounds, string instrument sounds). Categorization performance was above chance for surprisingly short gates, 4 ms for voices and 8 ms for instruments, and scores were at ceiling at 64 ms gate duration. Mixed results were obtained for the effect of onset information: instrumental, but not vocal sounds benefited from gates being positioned at sound onsets.

Most recently, Thoret *et al*. (2016, 2017) showed that instrument identification is determined by specific instrument-specific spectrotemporal modulations, although their approach did not allow them to draw specific conclusions about the role of onsets. Ogg *et al*. (2017) studied the minimal duration required to discriminate between musical instrument sounds, human speech, and human environmental sounds. They found that listeners required 25 ms for robust discrimination and that the presence of onsets was beneficial, even for vocal sounds.

Two conclusions may be drawn from this review regarding the role of onsets in instrument identification. First, the presence of the onset portion appears to improve sound identification but does not seem to be strictly necessary for correct identification. The relative importance of onsets appears to depend on the specific instrument at hand. Second, and more generally, whether implemented by digital gating or by excised tape, the experimental approach of presenting temporal segments has conceptually remained identical throughout the last 60 years (even though the analog scalpel may be less precise than today's digital means). This approach assumes that sounds can be meaningfully separated into discrete temporal states. However, as will be demonstrated in Sec. II, short-lived transients and quasi-stationary sinusoidal components cannot be strictly separated in time because both regimes overlap and one dynamically transforms into the other (Levine and Smith, 2007; Reuter, 1995). Therefore, the studies outlined above can only, to a limited degree, allow for conclusions about the importance of specific acoustical components such as transients—more flexible tools for separating signal components (sharpened acoustical scalpels) are needed.

B. The present study

The goal of this study was to use a novel transient/stationary separation algorithm to circumvent some of the methodological limitations of the literature. This algorithm is described in Sec. II. In the main experiment described in Sec. III, listeners identified short segments extracted from the sounds of ten musical instruments. These segments were processed by the separation algorithm and contained stationary and transient information, or only stationary information. Segments were extracted from the onset or from the middle portion of the sound. The goal of an additional control experiment described in Sec. IV was to assess whether the transient components were generally discriminable.

II. TRANSIENT SEPARATION

A. Description of the algorithm

Developments in audio signal processing have made it possible to separate overlapping stationary and transient components from mixtures (for a general review, see Müller, 2015, Chap. 8). A classical approach to this problem was provided by Serra and Smith (1990), approximating

transients in a global manner by time-varying filtered noise. Recently, the present author presented a more fine-grained algorithm to estimate transients by using an iterative multi-resolution analysis (Siedenburg and Doclo, 2017). The algorithm exploits the orthogonal orientation of components in the time-frequency plane: Whereas the quasi-stationary (S) components are sparse in frequency and persistent over time, rapidly varying transient (T) are sparsely distributed in time and persistent across frequency. Both types of components are extracted iteratively from Short-Term Fourier Transform (STFT) representations, using long window lengths (46 ms) for stationary components, yielding spectral precision, and short window lengths (3 ms) for transient components, yielding temporal precision. In technical terms, the separation process is based on a shrinkage operation of STFT coefficients that specifically extracts coefficients which are part of groups of relatively strong coefficients that extend over time or frequency (so-called *neighborhoods*, see Siedenburg and Doclo, 2017; Siedenburg and Dörfler, 2011). The result is an approximation of the original signal y in terms of three components, $y = S + T + e$, where e denotes the residual signal. The residual signal usually is of rather low energy and captures reverberation and microphone noise, but also faint phase-distorted versions of the stationary and transient components. The algorithm accurately separates stationary and transient components in synthetic examples and provides plausible separation results for recorded audio signals from acoustic musical instruments (although by definition there is no ground truth in this case). In the following experiment, S and S + T were used to study instrument identification. Consequently, if there was unintended distortion from the signal processing, it would have appeared not only in S but also in S + T.

B. Acoustic analyses

Figure 1 depicts the example of an A4 (440 Hz) piano sound of 250 ms duration. Throughout this study, the same settings of the algorithm were used as described in the original publication (Siedenburg and Doclo, 2017). The algorithm separates the impulsive sound of the hammer from the vibrating string (sound examples are provided as part of the supplementary material[1]). Figure 1(A) depicts the spectrogram (using a window length of 25 ms) of the original sound and a zoom into the onset is shown in Fig. 1(B). Figure 1 illustrates that beyond harmonic components, there is transient energy present in the onset portion of the sound. Moreover, the more detailed visualization in Fig. 1(B) suggests that the partial tones do not all start at the same time, but that lower components precede higher ones. Figures 1(C) and 1(D) depict the waveform of the separated stationary and transient components. The extracted time-frequency coefficients are shown in Figs. 1(E) and 1(F). Stationary components are sparse in frequency (although some subharmonic energy seems to be captured by the stationary estimate because of its relatively long extension in time). Transients have impulsive characteristics. Notably, the extracted transients are short-lived but overlap in time with the stationary components. This example hence demonstrates the limitations of considering musical sounds as a sequence of discrete states that can be neatly spliced apart in the time domain. To the contrary, components overlap and are continuously transformed over time, and thus transient components should not be confused with onsets as a whole.

The residual signal is depicted in Figs. 1(G) and 1(H). It is visible that the residual contains residual traces of both the harmonic stationary components and the impulsive transient of this piano tone.

In the perceptual experiment reported below, ten instruments at 12 different pitch levels were used (see Sec. III B 2 for details). Analyses indicated that these sounds had transients of much lower overall energy compared to the stationary components. Specifically, the stationary-to-transient energy level ratios averaged across pitch was highest for the vibraphone (mean 18 dB), followed by the marimba (24 dB),

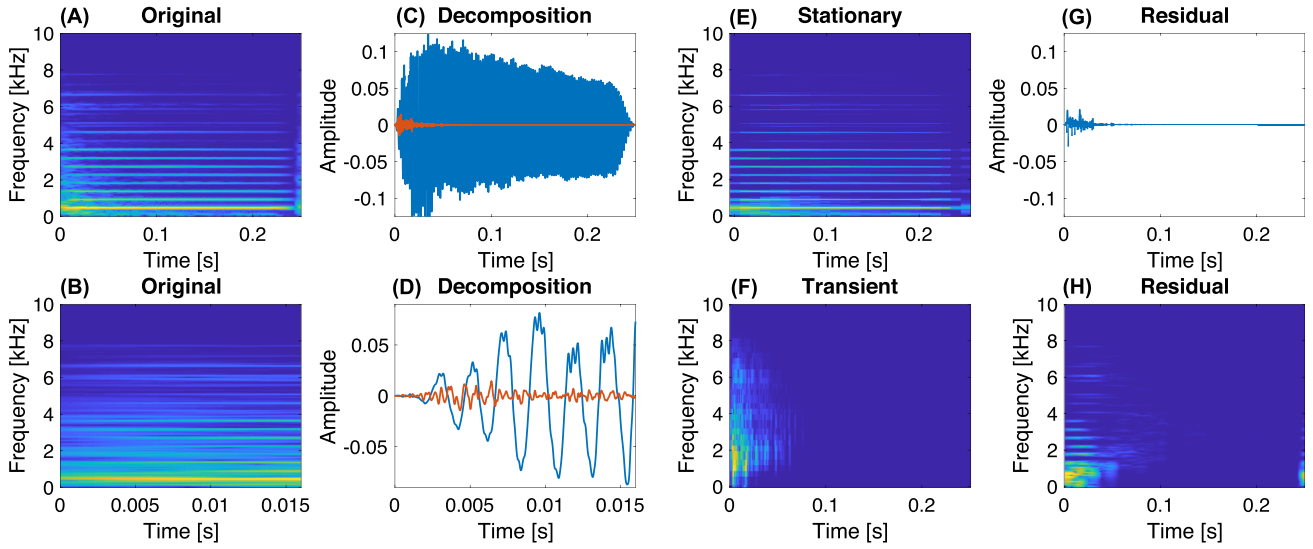

FIG. 1. (Color online) Example of a piano sound A4 (440 Hz) of 250 ms separated into stationary and transient components. (A) Spectrogram of original sound (window length 25 ms). (B) Zoom into first 16 ms of the original sound's spectrogram. (C) Waveform of separated stationary components (dark blue) and transients (light red). (D) Zoom into first 16 ms of the separated components' waveform. (E) Estimated stationary coefficients. (F) Estimated transient coefficients. (G) Waveform of residual. (H) Spectrogram of residual.

trumpet (28 dB), guitar (30 dB), piano (31 dB), cello (36 dB), harp (40 dB), violin (42 dB), flute (44 dB), and finally the clarinet (52 dB) with the weakest transient. Somewhat surprisingly, these ratios indicate that it is not generally possible to infer the sound excitation mechanisms of instruments by virtue of the relative transient energies, because the harp (an impulsive instrument) had lower relative transient energy compared to the trumpet (a sustained instrument).

The temporal evolution of transient and stationary energy is depicted in Fig. 2 (rows 1–2). Figure 2 shows the average temporal and spectral envelopes of the stationary and transient signal components (for temporal envelopes, gray background indicates the positioning of the gates in experiment 1). Here, temporal envelopes were extracted by computing the magnitude of the analytic signal, filtered with a third-order Butterworth lowpass-filter at a cutoff frequency of 50 Hz. The levels plotted in Fig. 2 correspond to signal intensities taken to the power of 0.3 (following Steven's law to approximate loudness). Figure 2 shows that the extracted transients do not extend much further than 64 ms into the tone and exhibit exponential decay characteristics for the impulsive instruments. This also holds, albeit to a much smaller degree, for the trumpet, violin, and cello. For the flute and clarinet, however, transients are of very low intensity, potentially more reflecting continuous blowing noise. Regarding the stationary component, Fig. 2 further indicates marked differences in envelope slope of impulsively excited instruments (top row) compared to sustained instruments (bottom row), the latter only reaching their energy peak in the middle portion of the tone.

The two bottom rows of Fig. 2 shows the average spectral power for the original signal, and the stationary and transient components (as for temporal envelopes raised to the power of 0.3 to reflect loudness). Spectral envelopes were obtained by smoothing the computed magnitude spectra by using a first-order Butterworth lowpass filter with a cutoff frequency of 1000 Hz. Figure 2 illustrates that the extracted transients had energy at relatively high frequencies, with spectral peaks at frequencies around or higher than 1 kHz. Figure 2 also highlights the distinct spectral shapes of the

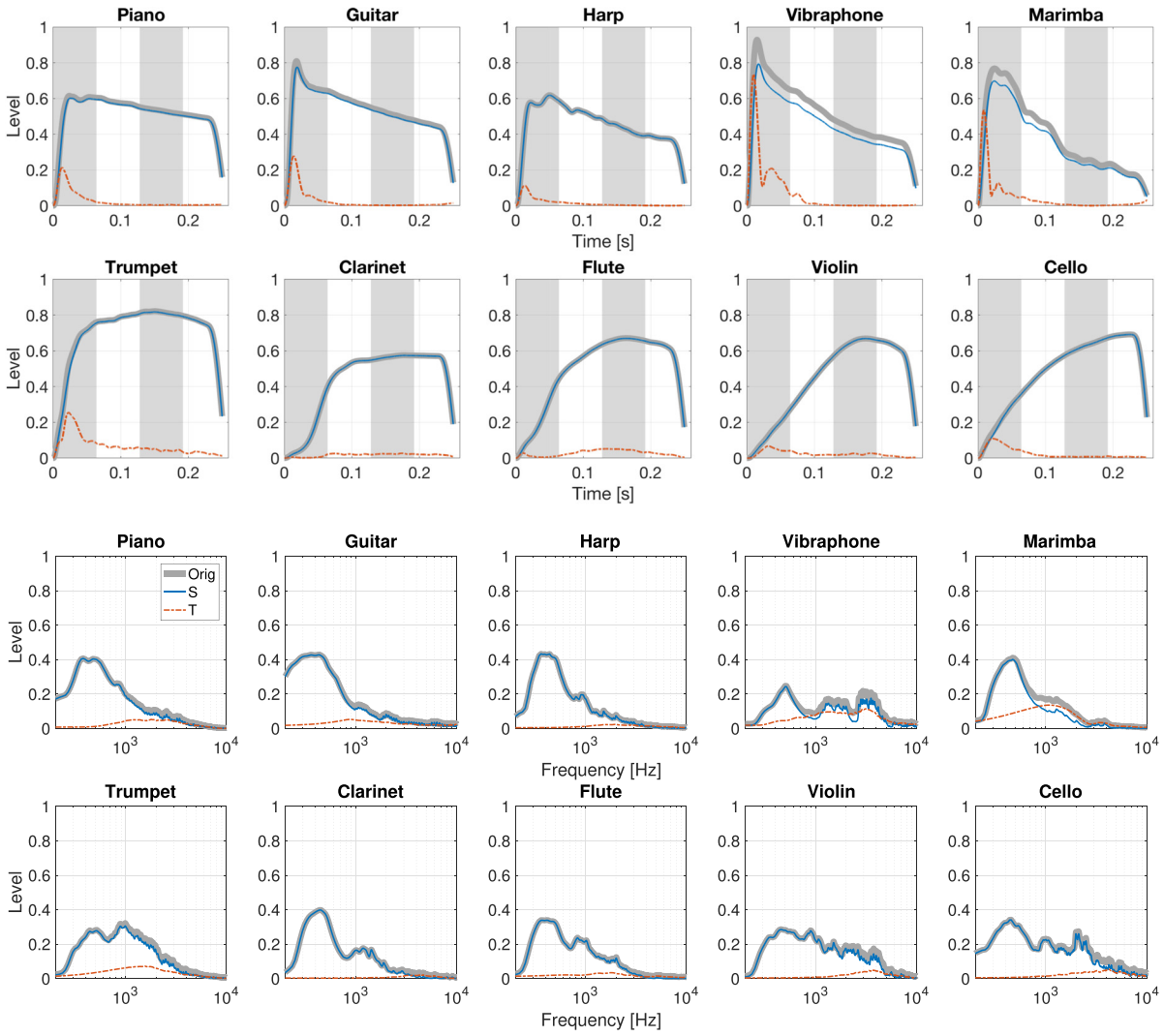

FIG. 2. (Color online) Temporal amplitude envelopes (rows 1–2) and spectral envelopes (rows 3–4). Level corresponds to signal intensity raised by 0.3 to approximate loudness according to Steven's law. Original sounds: gray, separated stationary components: blue, transient components: red, dashed-dotted. Lines depict averages across all 12 pitch levels. For temporal amplitude envelopes (rows 1–2), shaded areas correspond to the position of the gating used in experiment 1.

instruments' stationary components compared to the relatively similar spectral shapes of transient components. Experiment 1 tested the perceptual relevance of these components.

III. EXPERIMENT 1: INSTRUMENT IDENTIFICATION

A. Rationale

The present experiment compared instrument identification for harmonic instrument sounds resynthesized with and without transient components. In order to avoid ceiling performance and to be able to account for the importance of the onset position, sounds were gated with short gates of 64 ms duration. The resulting segments were taken from the onset of the original sounds and presented with and without transient components. In order to obtain an estimate about the general relevance of the onset, a third signal condition was tested that presented segments obtained from the middle portion of sounds (128–196 ms) with stationary and transient components. Note, however, that in the present sound set, the energy of transients was very small for the middle portion (see Fig. 2). Therefore, excerpts from the middle portion with only stationary components were not included in the experiment.

B. Methods

1. Participants

Eighteen listeners (13 female, 4 male, 1 other) with self-reported normal hearing and a mean age of $M = 26.1$ yr [standard deviation (SD) = 6.7, range: 21–48] participated in this experiment. Participants had played their primary musical instrument for an average of $M = 9.3$ yr (SD = 6.6, range: 1–22) and were dedicating $M = 10.5$ h per week to musical activities (SD = 11.0, range: 1–35). Participants were recruited via advertisements at the University of Oldenburg online job board and received a compensation with 10 EUR per hour.

2. Stimuli and apparatus

Stimuli were derived from orchestral instrument samples, obtained from the Vienna Symphonic Library (http://vsl.co.at). The following instruments were used in this study: piano, guitar, harp, vibraphone, marimba, trumpet, clarinet, flute, violin, and cello. Guitar samples were obtained from a Yamaha P155 synthesizer. Each instrument was played at 12 pitch levels: C4 (262 Hz) to B4 (494 Hz). From the stereo samples, only the left channels were used. Tones were played at *forte* dynamics and conceived as 8th-notes at a tempo of 120 quarter notes per minute, corresponding to a duration of 250 ms. The actual recordings were longer than this and of varying duration, so a 25 ms raised cosine function was applied as fade out to obtain a consistent duration of 250 ms.

In the experiments by Suied *et al.* (2014), instrument categorization performance levelled off at a duration of 64 ms. Hence, this gate duration was chosen for the current experiment in order to ensure that participants would be able to perform the task. Furthermore, this gate duration was short enough to meaningfully compare different placings of the gate within sounds. When the gate started at the beginning of the sound (0–64 ms: *@0 ms*), the original onset was preserved and a raised-cosine fade-out was used (cf. Suied *et al.*, 2014). When the gate was positioned in the middle of the sound (128–192 ms: *@128 ms*), both a raised-cosine fade-in and fade-out was used. Gated sounds were normalized in root-mean-square energy. The decomposition algorithm described above was used to extract the stationary and transient signal components from the gated sounds. Overall, there were three signal conditions: (1) stationary (S) and transient (T) components gated at the onset ($S + T@0\,ms$), (2) stationary components at the onset ($S@0\,ms$), and (3) stationary and transient components in the middle of the tone ($S + T@128\,ms$). Figure 3 shows the gating function and the temporal envelopes of the individual components for an exemplary piano tone.

The experiment was run with MATLAB and sounds were converted with an RME Fireface audio interface at an audio sampling frequency of 44.1 kHz and 24 bit resolution. Sounds were presented diotically over Sennheiser HDA 200 headphones at an average level of 65 dBA sound pressure level, as calibrated by a Norsonic Nor140 sound-level meter with a G.R.A.S. IEC 60711 artificial ear to which the headphones were coupled. Listeners were tested individually in a soundproof lab and provided responses on a computer mouse.

3. Procedure

The experiment comprised a training and test phase. The training phase was conducted to ensure that participants were familiar with the full range of perceptual features that characterized the test sounds. In the training phase, the original sounds were used. First, participants were exposed to all sounds at 12 pitch levels from each one of the ten instruments at an inter-onset interval of 750 ms. The order of the presentation of individual sounds and instruments was randomized. In order to further provide visual anchors, pictures of the instruments were presented concurrently. Pictures had been obtained from a web search and depicted standard tokens of the instruments in front of a white background.

In the second part of the training, participants were trained to identify sounds presented in isolation, as in the main experiment. The test contained each of the ten instruments at six randomly drawn pitch levels. In every trial,

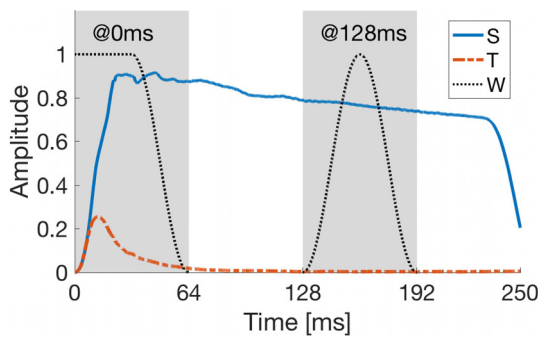

FIG. 3. (Color online) Illustration of the windows used to create the experimental signal conditions for the example of a piano tone. The figure shows the amplitude envelopes of stationary components (S), transient components (T), and the gating windows (W) with start positions at 0 ms or at 128 ms.

participants listened to a randomly drawn sound and were required to select the corresponding instrument label from a list of alternatives presented on a computer screen. Feedback about the correct response was provided with instrument labels and pictures. Overall, this amounted to 60 trials of training with response feedback and took around 12 min.

All participants continued with the main experiment, where sounds from the same ten instruments were presented at 12 pitch levels for the three signal conditions, S + T@0 ms, S@0 ms, S + T@128 ms, described above (Sec. III B 2). The signal conditions were blocked and blocks were presented in random order. There were 120 sounds per block; each block took around 25 min to complete and there were obligatory pauses of at least 5 min between blocks. Before the start of each experimental block, participants went through a passive exposure phase with the original sounds, as in the first part of the training. This exposure phase was implemented to ensure that potential differences across blocks were due to the signal conditions, and not due to memory loss of the reference that was established or consolidated during the initial training.

To avoid response bias through a fixed order of the instrument labels on the screen, the list order was randomized for each experimental block. Otherwise, the procedure was identical to the second part of the training although no feedback was provided. The experiment was self-paced.

C. Results

Figure 4 shows the average scores for the training and all experimental conditions, together with individual profiles from all participants. In the training, identification performance was high (proportion of correct identifications: $M = 0.84$). In the main experiment, average performance in the S + T@0 ms signal condition was around seven percentage points below the training score ($M = 0.77$) and slightly higher compared to the S@0 ms condition ($M = 0.71$). In the S + T@128 ms signal condition, there was a strong inflation of confusions ($M = 0.52$).

Figure 5 depicts average confusion matrices for the training phase and all experimental conditions. In the training, it is visible that, surprisingly, the cello and trumpet were frequently confused (although this only occurred in the training). In the main experiment, frequent within-family confusions occurred for the S + T@0 ms signal condition, in particular for the violin and cello (strings), and the clarinet and flute (winds). The qualitative confusion patterns were very similar for the S@0 ms signal condition. In the S + T@128 ms condition, the three impulsive instruments piano, guitar, and harp were frequently confused and even attributed to wind instruments such as the clarinet. Among the sustained (i.e., continuously excited) instruments, the flute was particularly poorly identified, and often confused with the trumpet. Four instruments were robustly identified for this condition and achieved accuracies above 0.75: the vibraphone, marimba, trumpet, and clarinet.

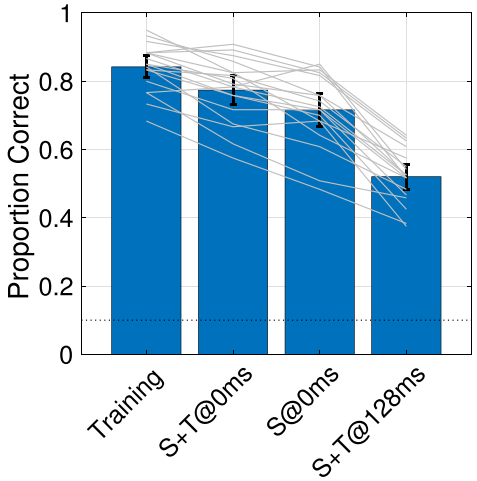

FIG. 4. (Color online) Mean identification scores from experiment 1. Individual results are plotted as gray lines and the dotted line indicates chance level. Error bars: 95% confidence interval (CI).

A repeated-measures analysis of variance (ANOVA) was conducted with the factors signal condition (S + T@0 ms, S@0 ms, S + T@128 ms) and pitch level (the statistical dependency of instrument-wise accuracies does not allow for an ANOVA on an instrument-wise level). The analysis indicated that there were significant differences between signal conditions, $F(2, 34) = 155.9, p < 0.001, \eta_p^2 = 0.90$, and of pitch, $F(11, 187) = 4.52, p < 0.001, \eta_p^2 = 0.21$, but no significant interaction between the two, $F(22, 374) = 1.52, p = 0.064, \eta_p^2 = 0.08$. *Post hoc* tests demonstrated that scores from the three signal conditions were significantly different from each other: paired $t(17) = 4.3$, $p = 0.0013$ for S + T@0 ms vs S@0 ms, $t(17) = 19.7$, $p < 0.001$ for S + T@0 ms vs S + T@128 ms, and $t(17) = 10.7$, $p < 0.001$ for S@0 ms vs S + T@128 ms (Bonferroni-corrected for multiple comparisons, $n = 3$). A comparison to the training indicated that training scores were significantly higher compared to all experimental signal conditions, paired $t(17) > 5.4$, $p < 0.001$. Visual inspection of the data did not reveal any systematic relation of identification accuracy and pitch, and scores in none of the three signal conditions significantly correlated with pitch height, $p > 0.187$ (Bonferroni-corrected, $n = 3$). This suggests that idiosyncratic stimulus features distributed across different pitch levels most likely caused the observed differences of identification scores across pitch levels.

D. Discussion

This experiment compared harmonic musical instrument identification for 64 ms-long sound segments with and without transient components taken from the onset or the middle portion of the original sound. The data indicated that removing the transient at the sound onset impaired identification scores by around 6 percentage points, whereas moving the gate from the onset to the middle portion of the sound impaired identification accuracy by 25 percentage points. Surprisingly, this effect did not appear to strictly depend on whether impulsive or sustained instruments were considered. In the signal condition that presented 64 ms segments from the middle portion of the tone (S + T@128 ms), the vibraphone and marimba were accurately identified (both impulsive) with accuracy scores above 75%, and the same held for the trumpet and clarinet (both sustained).

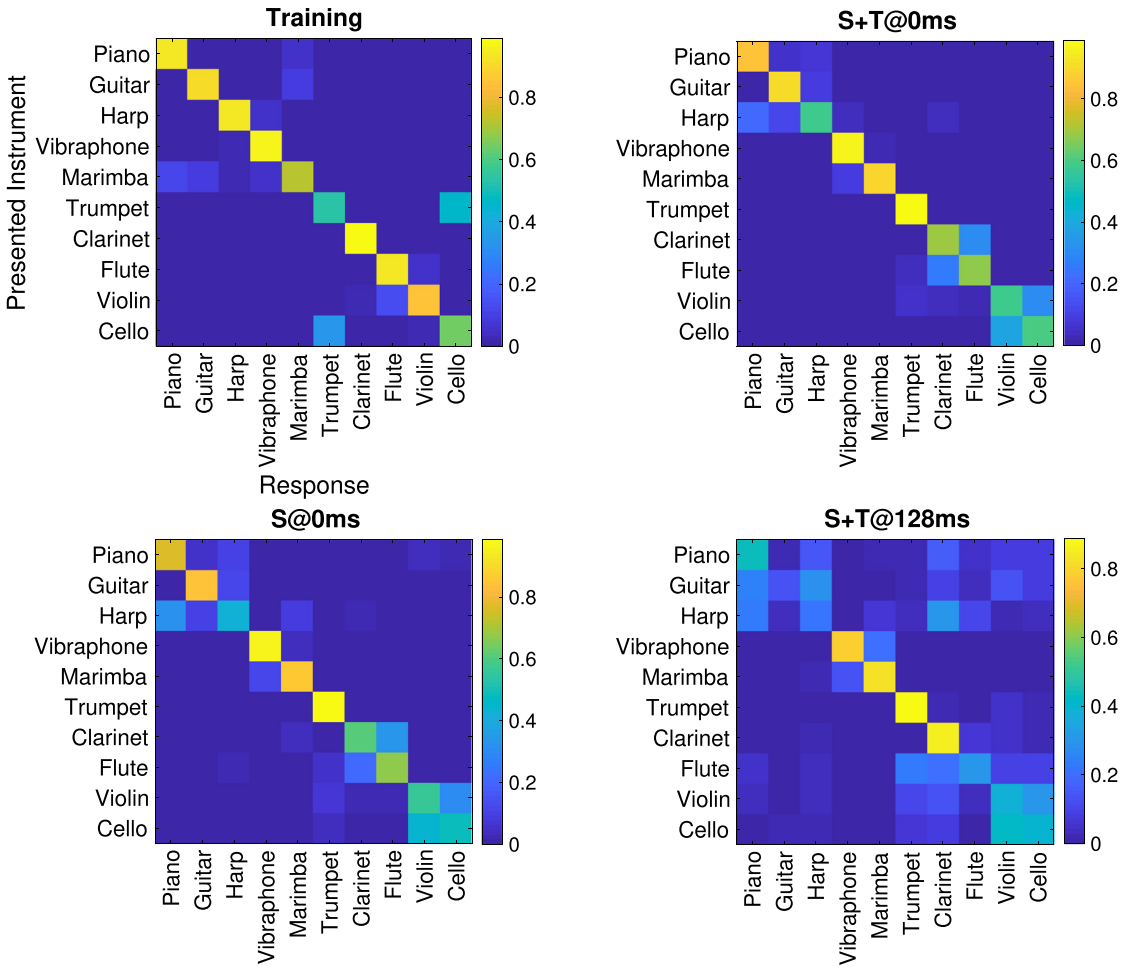

FIG. 5. (Color online) Average confusion matrices from experiment 1 including the training, normalized by the number of presentations of every instrument.

As it can be observed in Fig. 2, the energy levels of the transient components were almost negligible for the tested middle portions of sounds. Furthermore, there was a drastic drop in performance from S@0 ms to S + T@128 ms (≈S@128 ms) paired with small differences from S + T@0 ms to S@0 ms. Therefore, this pattern of results likely reflects the greater diagnosticity of the cues present in quasi-stationary sinusoidal components at the sound onset, and the lack of the transient component appears to be of smaller importance. Notably, even for sustained sounds, the steady state portion probed by the S + T@128 ms condition turned out to be less informative than the onset portion.

A potential explanation of the small effect observed for the removal of transients could be that the combined stationary and transient components (S + T) were not clearly discriminable from the stationary parts (S) alone. This question was addressed in a second experiment, which would further help to more comprehensively characterize the perceptual status of transients in musical instrument sounds.

IV. EXPERIMENT 2: TRANSIENT DISCRIMINATION

A. Rationale

The second experiment acted as a control experiment in order to test whether listeners would be sensitive to the presence of transients. Specifically, the aim was to test listeners' discrimination abilities of S from S + T, but also to measure discrimination of S + T from the original sound. This would assess the perceptual relevance of the separation algorithm's residual component. To direct listeners attention to transient information, additional foil conditions were included in the experiment, presenting amplified transients together with the stationary part.

B. Methods

1. Participants

Ten listeners (4 female, 5 male, 1 other) with self-reported normal hearing and a mean age of $M = 27.8$ yr (SD = 4.2, range: 23–37) participated. Participants had played their primary musical instrument for an average of $M = 14.8$ yr (SD = 6.8, range: 4–30) and were dedicating $M = 13.8$ h per week to musical activities (SD = 12.5, range: 2–35). Participant recruiting and compensation was identical to experiment 1.

2. Stimuli and apparatus

To keep the overall duration of the experiment within limits, only four of the ten instruments from experiment 1 were tested, two of which were impulsive (vibraphone and guitar) and two sustained (cello and trumpet). The corresponding recordings were presented at the full duration of

250 ms. There were five signal conditions, each testing the discrimination of S + T against (i) the original signal, (ii) S, (iii) 5 T, (iv) S + 10 T, and (v) S + 15 T, where S + xT indicates that the level of T was raised by x dB. The apparatus was identical to the main experiment.

3. Procedure

A 3-interval/2-alternative forced-choice task ("odd one out") was used. On every trial, there were three intervals with inter-stimulus intervals of 250 ms and participants were required to detect the odd interval. It was randomly determined whether S + T or the comparison stimulus from signal conditions (i)–(v) served as the odd stimulus. After providing their response by selecting the interval on a computer screen, participants received feedback about the correct response.

In order to maximize participant's sensitivity to potentially idiosyncratic timbral features, the presentation of instruments was blocked with a random order of the presentation of the signal conditions. The order of blocks was randomized. Every block contained 180 trials (3 intervals × 12 pitch levels × 5 signal conditions). The completion of any one block took around 25 min and there were obligatory pauses of at least 5 min between blocks.

C. Results

Performance was above chance level for all five different signal conditions, as confirmed by tailed t-tests against 0.33, $t(9) > 5.5$, $p < 0.001$. Participants robustly discriminated S + T from S, as reflected by 69% of correct identifications in this condition. Participants had greater difficulties to discriminate S + T from the original signal, yielding an average of only 42% correct responses. This result indicates that the omission of the residual from the original signal, leaving S + T, is barely detectable, which validates the general approach to use S + T as a starting point for studying timbre perception. Participants were further sensitive to an amplification of transients, as indicated by the strong effect across foil conditions. Average percentage of correct responses was 54%, 86%, and 96% for discriminating S + T from S + 5 T, S + 10 T, and S + 15 T, respectively.

A repeated-measures ANOVA was conducted to analyse differences for individual instruments. The analysis confirmed strong effects of signal condition, $F(4,36) = 171.8$, $p < 0.001, \eta_p^2 = 0.95$, instrument, $F(3,27) = 21.4, p < 0.001$, $\eta_p^2 = 0.70$, and an interaction of signal condition and instrument, $F(12,108) = 14.1, p < 0.001, \eta_p^2 = 0.61$.

The mean scores of all five signal conditions were highly different from each other, $t(9) > 4.3$, $p < 0.002$, as visible in Fig. 6. Performance for the two impulsive instruments guitar (76%) and vibraphone (74%) was generally better compared to the sustained instruments trumpet (62%) and cello (65%). Pairwise t-tests confirmed no significant differences between instruments of the same excitation type, $t(9) < 1.5$, $p > 0.16$, but all differences across excitation types were highly significant, $t(9) > 5.0$, $p < 0.001$. This means the task was generally easier for the two impulsive instruments, guitar and vibraphone.

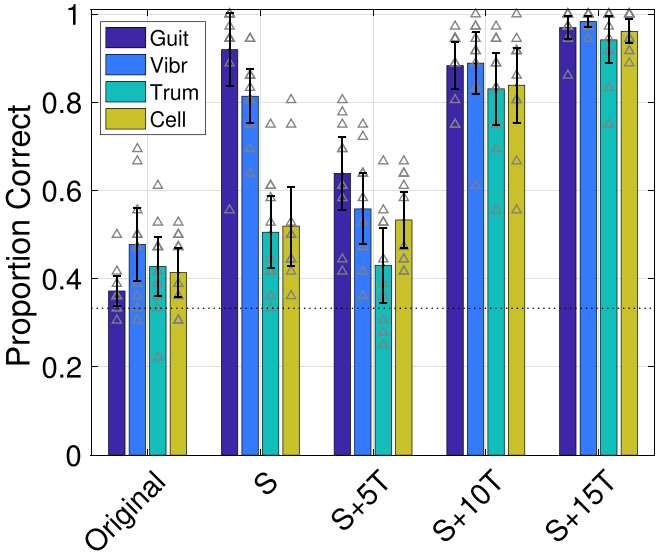

FIG. 6. (Color online) Discrimination accuracy from experiment 2. S + T was discriminated from the signal type given on the x axis, where S + xT indicates that the level of T was raised by x dB. Bar color corresponds to instruments as listed in the legend (guitar, vibraphone, trumpet, cello). Triangles correspond to performance of individual participants, the dotted line indicates chance performance. Error bars: 95% CI.

The signal conditions elicited differential effects on impulsive instruments compared to continuously excited instruments such that the interaction was due to the high scores for impulsive instruments in the S condition. Specifically, scores for S did not differ significantly from the original signal for the trumpet and the cello, $t(9) < 3.0$, $p > 0.057$ (Bonferroni-corrected for multiple comparisons, $n = 4$). But there were strong differences between the original signal and S signals for the guitar and the vibraphone, $t(9) > 7.2$, $p < 0.001$.

D. Discussion

This second experiment tested listeners' sensitivity to discriminate signals with manipulated transient components. Independent of instrument, the original sounds were only poorly discriminated from the signals that were resynthesized without residual (S + T); discrimination performance was barely above chance for this signal condition. This result implies that the residual does not appear to be very important in the current separation, which suggests that using the stationary and transients components, S + T, seems to be a good starting point for the current pursuits. More specifically, the above chance performance in both the S + T vs S and the S + T vs S + 5 T (and S + 10 T, S + 15 T) conditions indicates that listeners were sensitive to the amplification as well as to omission of transients. Note that this effect was pronounced for impulsive instruments, but, although not as strong (as indicated by the significant interaction of signal condition and instrument), it remained present for sustained instruments. In comparison to the higher performance for the signal condition that omitted the transient (S + T vs S), this indicates that listeners were much more sensitive to the presence of the transient than to the presence of the residual noise.

Only four instruments could be tested in this experiment and hence the generality of the findings is limited. It is possible that instruments with low energy transients such as the clarinet and flute would have yielded lower scores, in particular for the S vs S + T signal condition. Nonetheless, the obtained results show that it is generally not the lack of discriminability that is the underlying reason for the small effect between the S + T@0 ms and S@0 ms signal condition observed throughout instruments in experiment 1.

V. CONCLUSION

This study revisited the perceptual relevance of onsets in identification and discrimination tasks. Previous studies suggested that the onset plays a privileged role for identification, but the underlying acoustic factors had not been thoroughly tested. Here, a relatively small set of harmonic orchestral instrument sounds was used to test the importance of transient signal components. Using an algorithm to dissect transient from stationary components (Siedenburg and Doclo, 2017), acoustical analysis indicated that rapidly varying transients and quasi-stationary components are generally overlapping in time and that transient components are of relatively low energy. Importantly, these analyses indicate that the transient, defined via its short-livedness and stochastic nature, should not be confused with the onset portion of sounds as a whole—there is no point in time where transients could be neatly separated from sinusoidal components. Instead, the separation of acoustic components must take place in the time-frequency domain.

Two experiments tested the perceptual relevance of transients and quasi-stationary sinusoidal components. In experiment 1, it was shown that the omission of transient components at the onset portion of tones had a relatively small detrimental effect on instrument identification, even though experiment 2 suggested that a lack of discriminability of signals presented with and without transient components was not the underlying reason for this. Therefore, these results indicate that quasi-stationary components yield the most informative cues for instrument identification. Furthermore, shifting the position of the gate from the onset to the middle portion of the tone had a large detrimental effect on identification performance. The latter result confirms that even without the presence of transient components, onsets seem to be much more informative compared to sounds' middle portions, irrespective of the specific instrument or instrument class (impulsive vs sustained). Taken together, these findings confirm the prominent status of onsets in musical instrument identification suggested by the literature, but specify that rapidly varying transients (which often but not exclusively occur at sound onsets) have a relatively limited diagnostic value for the identification of harmonic musical instruments. In conclusion, fairly slowly varying signal components during onsets, likely the characteristic buildup of sinusoidal components in particular, provide the most valuable bundle of acoustic features for perceptual instrument identification.

A critical reader may object that the great care that musicians, sound designers, and music producers invest in the shaping of transient aspects of sound refutes this argument. This objection may be countered by noting that identification tasks require listeners to rely on informative acoustic cues for sound source identity, but not on every sound feature that may be integrated into assessments of sound quality (e.g., Pressnitzer *et al.*, 2013; Siedenburg and McAdams, 2017). Coherent with this notion, experiments 1 and 2 collectively suggested that not every class of discriminable sound feature is essential for sound source identification. In effect, sound production may deal in great length with the sculpting of timbral nuances such as high-frequency transients, even if these are only of minor importance for the inference of sound sources. Generally, this view acknowledges the multiplicity of cues available for sound source identification (Giordano *et al.*, 2010; Handel, 1995), all of which may be used opportunistically depending on the perceptual task and context at hand. Furthermore, one should not forget that this study only considered harmonic musical instruments presented in isolation. The situation may be different for non-harmonic percussion instruments and other sound-producing objects, not to speak of sound source identification in polyphonic mixtures.

A topic that should be addressed by future acoustical analyses concerns the question whether the utility of the onset (with or without transients) for instrument identification rests on perceptual or acoustical grounds. In other words, are listeners making use of informative features for identification that are only available in the onset, or do there exist equally informative features throughout the sound but listeners prefer to focus on the onset?

From a more general perspective, the current approach is in line with an upsurge of interest in signal analysis/resynthesis approaches to the study of auditory perception (McDermott and Simoncelli, 2011; Overath *et al.*, 2015; Ponsot *et al.*, 2018; Thoret *et al.*, 2017). In order to unravel the intricate workings of auditory perception these types of studies develop specific signal processing tools, which allows working with naturalistic but precisely controlled stimuli. Although this approach is principally related to the early explorations of cutting and splicing tapes (Schaeffer, 2017), today's digital tools offer an unprecedented degree of precision and versatility.

ACKNOWLEDGMENTS

The author wishes to thank the two reviewers for productive remarks on this manuscript. The author further thanks Henning Schepker, Etienne Thoret, and Trevor Agus for valuable comments on earlier versions of this manuscript, Daniel Pressnitzer and Christoph Reuter for insightful discussions, Saskia Röttges for data collection, and Simon Doclo for general support of this study. This project has received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014–2020) under the Marie Skłodowska-Curie Grant Agreement No. 747124. This project was also funded by a Carl von Ossietzky Young Researchers' Fellowship from the University of Oldenburg.

[1]See supplementary material at https://doi.org/10.1121/1.5091778 for sound examples.

Clark, M., Jr., Luce, D., Abrams, R., Schlossberg, H., and Rome, J. (**1963**). "Preliminary experiments on the aural significance of parts of tones of orchestral instruments and on choral tones," J. Audio Eng. Soc. **11**(1), 45–54.

Daudet, L. (**2005**). "A review on techniques for the extraction of transients in musical signals," in *International Symposium on Computer Music Modeling and Retrieval* (Springer, New York), pp. 219–232.

Elliott, C. A. (**1975**). "Attacks and releases as factors in instrument identification," J. Res. Music Ed. **23**(1), 35–40.

Giordano, B. L., Rocchesso, D., and McAdams, S. (**2010**). "Integration of acoustical information in the perception of impacted sound sources: The role of information accuracy and exploitability," J. Exp. Psychol. **36**(2), 462–476.

Grey, J. M. (**1977**). "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**(5), 1270–1277.

Handel, S. (**1995**). "Timbre perception and auditory object identification," in *Hearing*, edited by B. C. Moore (Academic Press, San Diego, CA), Chap. 12, pp. 425–461.

Iverson, P., and Krumhansl, C. L. (**1993**). "Isolating the dynamic attributes of musical timbre," J. Acoust. Soc. Am. **94**(5), 2595–2603.

Levine, S. N., and Smith, J. O. (**2007**). "A compact and malleable sines+transients+noise model for sound," in *Analysis, Synthesis, and Perception of Musical Sounds*, edited by J. W. Beauchamp (Springer, New York, NY), pp. 145–174.

McDermott, J., and Simoncelli, E. P. (**2011**). "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," Neuron **71**, 926–940.

Müller, M. (**2015**). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications* (Springer, Heidelberg, Germany).

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

Overath, T., McDermott, J. H., Zarate, J. M., and Poeppel, D. (**2015**). "The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts," Nature Neurosci. **18**(6), 903–911.

Ponsot, E., Arias, P., and Aucouturier, J.-J. (**2018**). "Uncovering mental representations of smiled speech using reverse correlation," J. Acoust. Soc. Am. **143**(1), EL19–EL24.

Pressnitzer, D., Agus, T. R., and Suied, C. (**2013**). "Acoustic timbre recognition," in *Encyclopedia of Computational Neuroscience: Springer Reference*, edited by D. Jaeger and R. Jung (Springer, Heidelberg, Germany), pp. 1–6.

Reuter, C. (**1995**). *Der Einschwingvorgang nichtperkussiver Musikinstrumente* (*The Onset Process of Non-Percussive Musical Instruments*) (Peter Lang Frankfurt/M).

Robinson, K., and Patterson, R. D. (**1995**). "The duration required to identify the instrument, the octave, or the pitch chroma of a musical note," Music Percept. **13**(1), 1–15.

Saldanha, E., and Corso, J. F. (**1964**). "Timbre cues and the identification of musical instruments," J. Acoust. Soc. Am. **36**(11), 2021–2026.

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

Serra, X., and Smith, J. O. (**1990**). "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," Computer Music J. **14**(4), 12–24.

Siedenburg, K., and Doclo, S. (**2017**). "Iterative structured shrinkage algorithms for stationary/transient audio separation," in *Proceedings of the 20th International Conference on Digital Audio Effects* (*DAFX-20*), Edinburgh, United Kingdom (September 5–8).

Siedenburg, K., and Dörfler, M. (**2011**). "Structured sparsity for audio signals," in *Proceedings of the 14th International Conference on Digital Audio Effects* (*DAFx-11*), Paris, France.

Siedenburg, K., Jones-Mollerup, K., and McAdams, S. (**2016**). "Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds," Front. Psychol. **6**(1977), 1–17.

Siedenburg, K., and McAdams, S. (**2017**). "Four distinctions for the auditory 'wastebasket' of timbre," Front. Psychol. **8**(1747), 1–4.

Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., and Pressnitzer, D. (**2014**). "Auditory gist: Recognition of very short sounds from timbre cues," J. Acoust. Soc. Am. **135**(3), 1380–1391.

Thoret, E., Depalle, P., and McAdams, S. (**2016**). "Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments," J. Acoust. Soc. Am. **140**(6), EL478–EL483.

Thoret, E., Depalle, P., and McAdams, S. (**2017**). "Perceptually salient regions of the modulation power spectrum for musical instrument identification," Front. Psychol. **8**(587), 1–10.

Zaunschirm, M., Reiss, J. D., and Klapuri, A. (**2012**). "A sub-band approach to modification of musical transients," Comp. Music J. **36**(2), 23–36.