JUNE 02 2021

Spectral envelope position and shape in sustained musical instrument sounds[a)]

Special Collection: Modeling of Musical Instruments

Kai Siedenburg ; Simon Jacobsen; Christoph Reuter

Check for updates

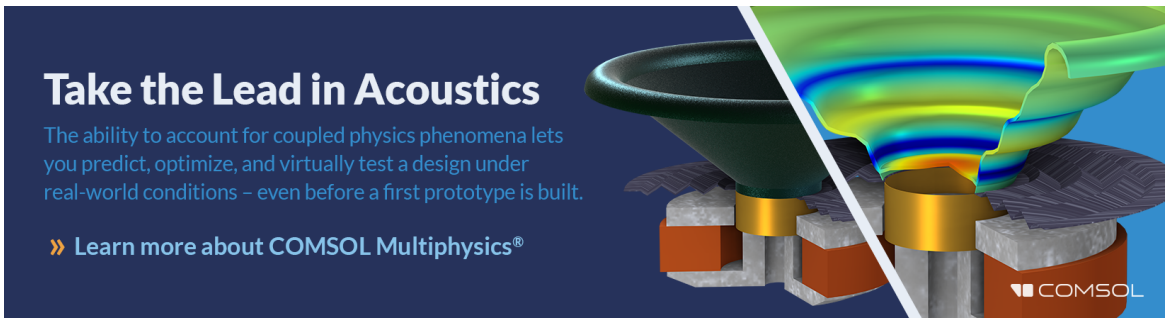
J Acoust Soc Am 149, 3715–3726 (2021)
https://doi.org/10.1121/10.0005088

CrossMark

View Online

Export Citation

17 August 2023 18:45:45



JASA ARTICLE

Spectral envelope position and shape in sustained musical instrument sounds[a)]

Kai Siedenburg,[1,b)] Simon Jacobsen,[1] and Christoph Reuter[2]
[1]*Department of Medical Physics and Acoustics, Carl von Ossietzky University of Oldenburg, 26129 Oldenburg, Germany*
[2]*Department of Musicology, University of Vienna, 1090 Vienna, Austria*

ABSTRACT:
It has been argued that the relative position of spectral envelopes along the frequency axis serves as a cue for musical instrument size (e.g., violin vs viola) and that the shape of the spectral envelope encodes family identity (violin vs flute). It is further known that fundamental frequency (F0), F0-register for specific instruments, and dynamic level strongly affect spectral properties of acoustical instrument sounds. However, the associations between these factors have not been rigorously quantified for a representative set of musical instruments. Here, we analyzed 5640 sounds from 50 sustained orchestral instruments sampled across their entire range of F0s at three dynamic levels. Regression of spectral centroid (SC) values that index envelope position indicated that smaller instruments possessed higher SC values for a majority of instrument classes (families), but SC also correlated with F0 and was strongly and consistently affected by the dynamic level. Instrument classification using relatively low-dimensional cepstral audio descriptors allowed for discrimination between instrument classes with accuracies beyond 80%. Envelope shape became much less indicative of instrument class whenever the classification problem involved generalization to different dynamic levels or F0-registers. These analyses confirm that spectral envelopes encode information about instrument size and family identity and highlight their dependence on F0(-register) and dynamic level.
© 2021 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1121/10.0005088

(Received 8 February 2021; revised 5 May 2021; accepted 9 May 2021; published online 2 June 2021)
[Editor: Nicholas J. Giordano] Pages: 3715–3726

17 August 2023 18:45:45

I. INTRODUCTION

In many acoustical scenarios, it is reasonable to assume that the spectral composition of sound signals is determined by a source signal that is linearly filtered by a resonator. Formalized in the source-filter model (Fant, 1960), a prime example is the human voice with the vocal folds generating a quasi-harmonic source signal that is filtered by the resonances of the vocal apparatus. Assuming the validity of the source-filter model implies that the spectral fine structure of sounds is decoupled from the spectral envelope, and the shape and position along the frequency axis of the latter are determined by the properties of the resonator.

Elaborating on the source-filter model, Patterson *et al.* (2010) provided an account of instrument size perception for musical sounds. Resonating systems of larger spatial dimensions not only tend to have source signals with lower F0s but also exhibit resonances at lower frequencies (Fletcher and Rossing, 2012). Accordingly, Patterson *et al.* (2010) made two central hypotheses. First, the overall size of an instrument relative to other instruments from the same instrument family (which they called *register*) affects the positioning of the filter along the frequency axis, which listeners use as a primary cue to infer size information. This would imply that smaller instruments from the same family (e.g., alto vs tenor saxophone) would generate sounds with envelopes shifted toward higher frequencies that listeners would interpret as having a brighter timbre (e.g., Saitis and Siedenburg, 2020). Second, F0-invariant shapes of the filters that are effectively implemented by instruments determine the perception of instrument family, hence allowing listeners to identify, say, a violin or a French horn. Both hypotheses were based on the premise that spectral envelopes are largely invariant to variation of F0 (Patterson *et al.*, 2010); see Fig. 1 for visualization.

Let us note that our terminology deviates from Patterson *et al.* (2010). Instead of talking about properties of the filter and the source, we here use the more agnostic terms *spectral envelope* and *F0*, such that we do not presuppose an underlying source-filter system. To further avoid confusion with the notion of *modulation scale* as in spectro-temporal modulation representations (Elhilali, 2019), we use the more pedestrian term *envelope position* to refer to what Patterson denotes as filter scale. Also, we prefer to use the term (F0-)*register* to refer to a partitioning of F0s in an instrument's playing range (and not as an index to different overall instrument sizes, which we here simply refer to as size). Finally, the notion of instrument *family* is here replaced by a more specific term instrument *class* that incorporates instruments that are physically similar but may vary in size.

[a)]This paper is part of a special issue on Modeling of Musical Instruments.
[b)]Electronic mail: kai.siedenburg@uni-oldenburg.de, ORCID: 0000-0002-7360-4249.

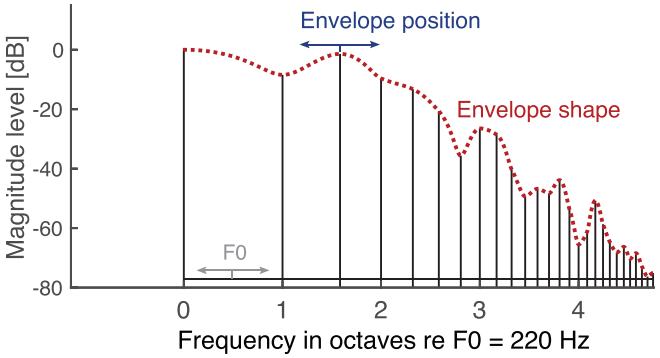


FIG. 1. (Color online) Illustration of the concepts of envelope position, envelope shape, and F0. The *x* axis indexes harmonic number. The sound stems from a bassoon with F0 of 220 Hz at *mf* dynamic level.

The account of Patterson *et al.* (2010) is appealing because it compactly associates instrument parameters (family, size) with signal parameters (envelope shape, envelope position). However, applying the source-filter model as a general framework to musical instrument sound is not without limitations. Non-linear couplings between the source generator and the resonator are critical for the generation of sustained musical instrument sounds (Fletcher, 1999). And in several instrument classes, such as wind instruments, the properties of the source signal and the resonator tend to covary with fingering positions or the player's active modification of the resonator size by valves and slides. Moreover, the F0-range of most instruments can be divided into F0-registers that are arguably characterized by distinct timbral qualities. For instance, low pitches played on a clarinet in the *chalumeau register* sound completely different from the high pitches played on the same instrument in the *clarine register* (e.g., Campbell *et al.*, 2004). Register boundaries and capabilities are closely described in orchestration treatises (e.g., Berlioz, 1844; Forsyth, 1982; Gevaert, 1885; Koechlin, 1954; Miller, 2014; Widor, 1904); see Reuter (2002) for a summary. It would be surprising if separate F0-registers were not accompanied by distinct envelope properties. Finally, it is well known that the spectral compositions of instrument sounds vary as a function of musical playing strength or effort that gives rise to differences in dynamic level such as *pp* or *ff* (Luce and Clark, 1967; Meyer, 2009; Myers *et al.*, 2012; Weinzierl *et al.*, 2018). Even despite these principled constraints, it could be that the account by Patterson *et al.* (2010) is an accurate effective model and could act as a coarse statistical approximation of spectral properties of sustained musical instrument sounds. The present study aims to test this question using regression and classification analyses on spectral envelope descriptors derived from instrument recordings.

Previous analyses of musical instrument sounds that considered F0-variance were descriptive in nature (e.g., Meyer, 2009; Reuter, 1996, 2002). Other studies have averaged spectral envelope descriptors across different F0s (Lembke and McAdams, 2015), hence assuming F0-invariance of spectral envelope properties from the outset (cf. Oehler and Reuter, 2009). Several recent studies have proposed models of acoustical correlates of timbre derived from stimuli presented in various perception tasks (e.g., Ogg and Slevc, 2019; Saitis and Siedenburg, 2020; Thoret *et al.*, 2020). In perceptual experiments, however, sounds' pitch and loudness tend to be normalized, which consequently does not allow for an analysis of effects of F0 and dynamic level. More generally, perceptual experiments typically only use small sets of sounds, which makes it hard to detect variant and invariant characteristics of signal parameters across production parameters such as F0 or dynamic level. Most perceptual studies also leave open the extent to which the few sounds selected for testing are representative of a chosen instrument (class) as a whole. In fact, only very few studies have analyzed acoustical properties of large sets of musical instrument sounds using statistically rigorous approaches [although see Peeters *et al.* (2011) and Weinzierl *et al.* (2018)].

It seems likely that acoustical regularities of instrument sounds have an immediate impact on the potential interaction of timbre with other auditory attributes in tasks such as instrument identification (Steele and Williams, 2006), dissimilarity perception (Marozeau *et al.*, 2003), pitch interval perception (Russo and Thompson, 2005), auditory short-term memory (Siedenburg and McAdams, 2018), or the evaluation of dynamic level (Fabiani and Friberg, 2011). To understand how listeners exploit timbre cues for these various tasks, we need to know about the distribution of candidate cues for the parameters F0, F0-register, and dynamic level.

Using 5640 sounds from 50 instruments of the Western orchestra, we seek to test three hypotheses that are central to Patterson *et al.* (2010):

(1) Spectral envelopes of sustained orchestral instrument sounds are invariant to variation in F0;
(2) Envelope position (inversely) encodes relative instrument size;
(3) Envelope shape distinguishes instrument family or classes.

In addition, we consider the variables of dynamic level and F0-register within instruments. To test hypotheses (1) and (2), our analysis relies on a descriptive analysis as well as regression modeling of a classical measure of spectral envelope position, namely the spectral centroid (SC) as defined in Sec. II C. We descriptively map SC vs F0 for all 50 instruments and use multiple linear regression to quantify effects of F0, instrument size, and dynamic level on SC trajectories. Considering hypothesis (3), we use support-vector machine (SVM) classification to test how well a relatively low-dimensional representation of spectral envelope shape via cepstral coefficients can serve as a basis for instrument classification. By matching and mismatching training and test conditions, the invariance of envelope shape to variation in dynamic level and F0-register is tested.

II. MATERIALS AND METHODS

A. Stimuli

The stimuli consisted of musical instrument digital interface (MIDI)-controlled recordings of acoustical

17 August 2023 18:45:45

https://doi.org/10.1121/10.0005088

instruments. To analyze several instruments with a high degree of physical similarity but different sizes, all available quasi-harmonic sustained orchestral instruments from the Vienna Symphonic Library [VSL *Super Package*; Vienna Symphonic Library (2021)] were included in the present analysis, resulting in a total of 50 instruments. Notes of 250 ms duration plus decay times of individual duration (below 1 s) were generated in the available F0-range of all instruments and were generated with the *VSL Ensemble* plug-in at an audio sampling rate of 44.1 kHz. The natural attack epoch of sounds was preserved. All sounds were generated with sustained articulation and dynamic levels *pp*, *mf*, and *ff* with MIDI velocity values 21, 85, and 127, respectively. The VSL provided the same recorder samples for all dynamic levels, because the dynamic range of recorders is generally very limited. Instruments were sorted into 12 different classes (see Table I): vocals, strings, flutes, recorders, clarinets, saxophones, oboes, bassoons, trumpets, trombones, horns, and tubas. Instrument classes contained a median number of four distinct instruments (range: 2–7 instruments). Here, the voice is considered an instrument, and hence different singers such as baritone and tenor are considered different instances of the same instrument class (vocalists sang the vowel A as in "father"). This resulted in a total number of 5640 isolated sounds to be analyzed. The median F0 of these sounds was 293 Hz (D4). Spectra from exemplary stimuli are shown in Fig. 2. A companion webpage (Siedenburg *et al.*, 2021a) features sound examples and interactive visualizations of SC vs F0 trajectories.

B. Instrument classes, sizes, and F0-register boundaries

Instruments have classically been categorized according to the means of sound generation and the material properties of the resonators (von Hornbostel and Sachs, 1914). Here, we grouped instruments into 12 rather small classes wherein instruments differ in size and compass but are physically similar otherwise [even though the distinction may not be as clear cut for the brass instruments, cf. Campbell *et al.* (2020)]. For instance, the fluegelhorn may be considered as most similar to the trumpet, but from a physical perspective, it features a conical bore and was hence grouped with the tuba instruments.

To use quantifiable measures of instrument size in the following regression analysis, we extracted estimates of instrument size from the literature. Instrument size would here act as a *proxy* for what Patterson *et al.* (2010) called *register*, which, in simplified terms, corresponds to the overall size effect of an instrument that affects both source scale and envelope position. Here, an increase in size would correspond to a downward shift of the range of feasible F0s as well as a comparable shift of the spectral envelope toward lower frequency regions. For vocals, we used estimates of vocal tract length as predictors (Mürbe *et al.*, 2011). Size estimates for string instruments were extracted from Fletcher and Rossing (2012) (p. 325). Size estimates of all

TABLE I. Instrument classes, class members, playing range in the current database, lower boundaries of instruments' middle and upper registers, and estimates of size/length of instrument (plus vocal tract length in brackets in the case of vocals).

Class	Instrument	F0-range	Register bounds	Size (cm)
Vocals	Coloratura soprano	F4–F6	F#4, F#5	167.5 [14.5]
	Soprano	C4–D6	F#4, F#5	167.5 [14.5]
	Mezzo-soprano	A3–C6	E4, E5	168 [15.5]
	Alto	F3–A5	D4, D5	171 [15.8]
	Tenor	C3–D5	C4, F#4	175 [16.5]
	Baritone	G2–A4	B3, E4	181 [16.6]
	Bass	C2–E4	G#3, D4	183 [18]
Strings	Violin	G3–Bb7	D4, E5	36
	Viola	C3–G#6	G3, A4	43
	Cello	C2–G6	G2, A3	76
	Bass	B0–G4	A1, G2	110
Flutes	Piccolo	C5–C#8	C6, C7	26
	C-flute	Bb3–D7	G4, C6	60
	Alto-flute	E3–A6	D4, A5	86
	Bass-flute	Bb2–C6	G3, C5	119
Recorders	Soprano	C5–A7	C6, C7	29
	Alto	F4–Bb6	F5, F6	42
	Tenor	C4–F6	C5, C6	59
	Bass	C3–D5	F4, D5	88
Clarinets	Eb clarinet	F3–C7	A4, C6	49
	Bb clarinet	C3–G6	D4, A4	66
	Bassethorn	F2–C6	B3, E4	106
	Bass-clarinet	Bb1–B5	D3, A3	132
	Contrb-clarinet	Bb0–F4	E2, A2	264
Saxophones	Soprano (Bb)	G3–E6	F4, G5	65
	Alto (Eb)	C3–C#6	Bb3, F5	101
	Tenor (Bb)	G2–E5	E3, A4	131
	Baritone (Eb)	C2–A4	Bb2, F4	214
	Bass (Bb)	G1–C4	E2, A3	293
Oboes	French oboe	A3–F#6	G4, G5	65
	English horn	D#3–C6	C4, D5	96
	Heckelphone	G#2–F5	E3, E5	139
Bassoons	Bassoon	Bb1–F5	Bb2, Bb3	259
	Contra-bassoon	C1–Bb3	F1, F2	593
Trumpets	Piccolo	C4–F6	A4, D#5	118
	C-trumpet	E3–C#6	G#4, G#5	132
	Cornet (Bb)	E3–D6	Bb3, F5	148
	Bass-trumpet	D2–C#5	G#3, G#4	240
Trombones	Alto (Eb)	G2–E5	B3, B4	217
	Tenor	C2–D5	E3, Bb4	269
	Bass (C)	C1–E4	C3, C4	353
	Cimbasso	D#1–G4	G#2, G#3	368
	Contrabass (Bb)	A0–C4	C#3, C4	370
Horns	Viennese horn	Bb1–F5	F3, A4	370
	French horn	A1–G5	F3, A4	396
Tubas	Fluegelhorn	E3–F6	F4, F5	148
	Wagnertuba	Bb1–G5	Bb2, Bb3	290
	Euphonium	C1–C5	Bb2, Bb3	297
	Tuba	D1–G4	F2, F3	396
	Contrb.-tuba	A0–C4	C#2, C#3	529

17 August 2023 18:45:45

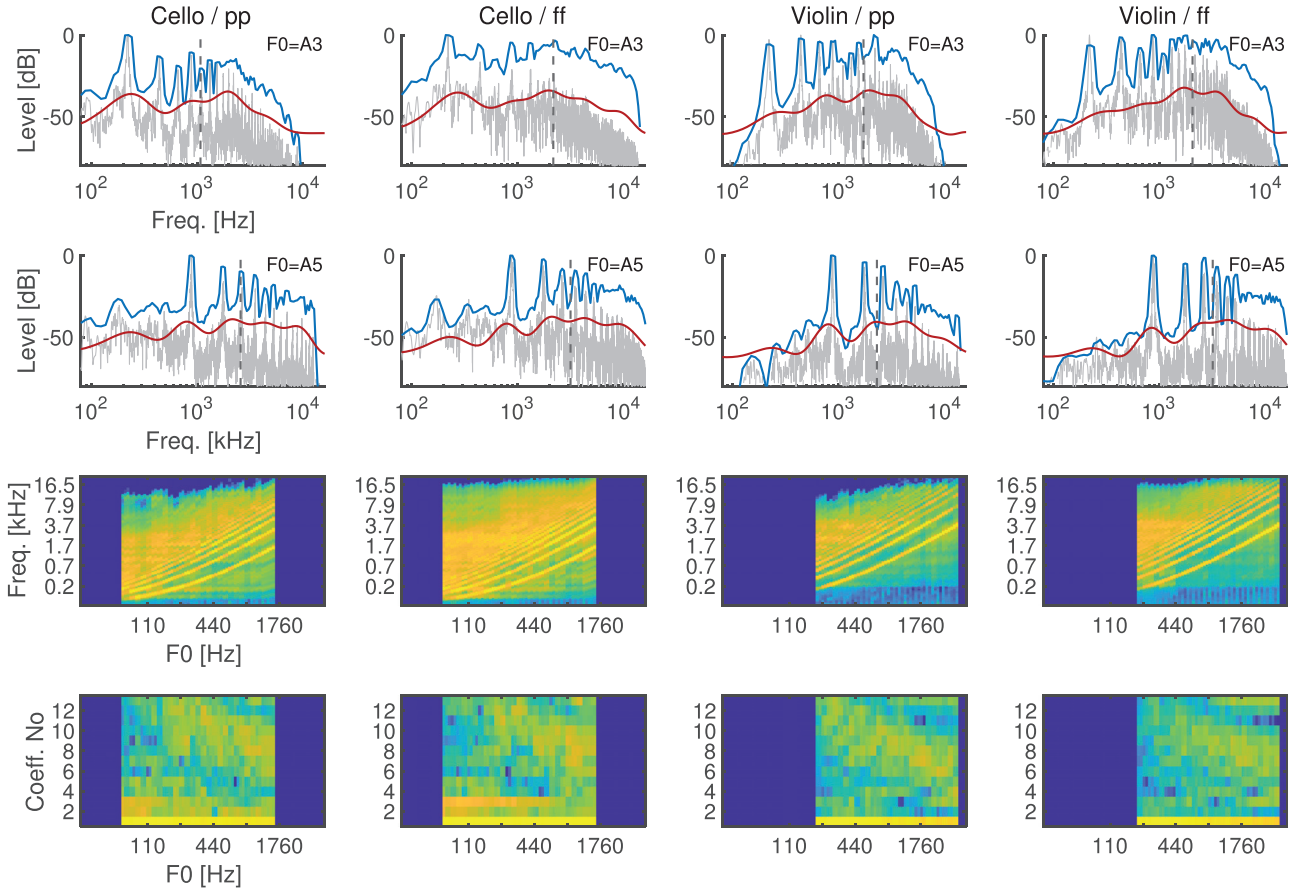


FIG. 2. (Color online) Illustration of the dependence of spectral envelope on F0, dynamic level, and instrument size for two string instruments. The two uppermost rows show spectra for cello and violin sounds at *pp* and *ff* dynamic levels with F0 = 220 Hz (row 1) and F0 = 880 Hz (row 2) Plots comprise the spectral fine structure (gray line), magnitudes of ERB-filterbands (dark blue line), spectral envelope shape as encoded by DCT-based cepstral coefficients (red line), and the SC (vertical dashed line). Row 3 shows color-coded ERB-magnitudes as a function of F0 for the entire playing range of the respective instrument. Row 4 shows the FFT-magnitude-based cepstral coefficients.

other instruments were in most cases extracted from Pape (1976).[1]

The distinct playing ranges of an instrument and its corresponding F0-registers are not always clear cut (Campbell *et al.*, 2020; Rimskij-Korsakov, 1922). Depending on a player's ability and the construction of a musical instrument, the boundaries of F0-registers found in the literature can vary by a few semitones. Nevertheless, the different timbral ranges of singing voices and musical instruments can usually be divided into a low, middle, and high register, which have proven to be stable according to orchestration treatises of the last few centuries (Reuter, 2002). Table I contains instrument classes, instruments, and instruments' playing range as provided in the VSL and, as boundary indices, the first pitch of the middle and upper registers. To account for potential differences across F0-registers, regression and classification analyses were conducted separately per F0-register.

C. Descriptors of spectral envelope position and shape

Spectral envelope position and shape were characterized by using the SC and cepstral descriptors, respectively, both of which are commonly used in audio processing (Caetano *et al.*, 2019). Sounds were analyzed by extracting magnitudes of a fast Fourier transform (FFT) from the left channels of the stereo recordings with a frequency resolution of 1 Hz. All magnitudes below the threshold of −90 dB full scale were set to zero. The resulting magnitudes were grouped by using an equivalent-rectangular-bandwidth (ERB) filterbank (Moore and Glasberg, 1983) with 128 bands. As a measure of envelope position along the frequency axis, the SC was computed as $SC = \Sigma_j f_j \cdot E_j / \Sigma_j E_j$, where f_j denotes the center frequency and E_j the magnitude of the jth ERB-band. Figure 2 illustrates these descriptors for selected sounds.[1] All computations and analyses were conducted in the MATLAB software environment. Data and scripts are available online (Siedenburg *et al.*, 2021b).

As a measure of envelope shape, the cepstrum based on the ERB-magnitude spectrum was used. ERB-magnitude spectra were rms-normalized, log10-transformed, and processed by a FFT. From the resulting vector, the magnitudes of the first 13 coefficients were used. The resulting cepstral coefficients encode spectral modulations such that the higher the coefficient number, the more fine-grained the spectral modulation that is encoded. Because only the first

17 August 2023 18:45:45

https://doi.org/10.1121/10.0005088

few coefficients are used (cf. Caetano *et al.*, 2019), this approach blurs spectral fine structure information that may be present in the ERB-spectrum and that could encode F0 (see Fig. 2). Further, the FFT-magnitude discards phase information, that is, the representation is mathematically translation-invariant—an envelope shape translated to different positions along the frequency axis is represented by identical cepstral coefficients. Conceptually, this approach is similar to the scale dimension from spectrotemporal modulation representations (Elhilali, 2019).

As a reference method, we used the discrete cosine transform (DCT, type II) to compute the cepstrum. The DCT-based cepstrum closely resembles the construction of mel-frequency cepstral coefficients (MFCCs), the only difference being that an ERB-spectrum is used instead of a mel-spectrum. In contrast to the FFT-magnitude-based cepstrum (FFTmag-CC), the DCT mathematically encodes information about envelope position. This aspect is visible in Fig. 2, where the red lines show the reconstruction of the spectral envelopes from 13 DCT-cepstral coefficients (DCT-CC) that align with the overall position of the envelope shape along the frequency axis.

D. Regression and classification

To quantify effects of F0, instrument size, and dynamic level on SC values, a multiple linear regression analysis was conducted for all 12 instrument classes and three F0-registers (lower, middle, upper) with the regressors F0, instrument size (see Table I for raw values), and dynamic level (*pp*: 21, *mf:* 85, *ff:* 127). All predictors were *z*-normalized for every instrument class (see supplementary material[1] for regression with non-normalized predictors). Estimated regression coefficients (β) served as estimates of effect size for individual instrument classes.

To clarify the interconnection between SC and FFTmag-CC/DCT-CC descriptors, we computed the R^2 fit between the SC and a linear combination of cepstral coefficients as derived from multiple regression for all sounds in the database. As a baseline, we computed the same measure for randomly generated ERB-spectra and ERB-spectra with Gaussian shape but random position and variance.

For the classification of instrument sounds into 12 classes, cepstral coefficients were used in conjunction with an SVM classifier with a standard Gaussian kernel and a one-vs-all multi-class approach (Friedman *et al.*, 2009). To discriminate between classes, SVMs learn hyperplanes that maximally separate the descriptor values from different classes. The Gaussian kernel maps values that are not necessarily linearly separable onto a new linearly separable representation.

To investigate dependencies on (1) dynamic level (*pp*, *mf*, *ff*) and (2) F0-register (lower, middle, upper), two types of training/test conditions were compared. In the "within" condition, the classifier was trained and tested on sounds from all dynamic levels or F0-registers within the test set using threefold cross-validation. In the "across" condition, the classifier was trained on two subcategories of the relevant factor and tested on the remaining level (e.g., trained on dynamic level *pp* and *mf* and tested on *ff*). By doing so for all three combinations, effectively a threefold cross-validation with mismatched training and test sets was implemented. As performance measures, we used classification accuracy (number of correct trials/number of total trials). In addition, we used F1-scores that are computed via $F1 = 2(P \cdot R)/(P + R)$, where P and R denote class-wise precision and recall, respectively. Precision is computed as the ratio of true positives over true positives plus false positives; recall is computed as the ratio of true positives over true positives plus false negatives.

III. RESULTS

A. Envelope position

Figure 3 shows all SC values as a function of F0 over the whole playing range of all instruments for the three different dynamic levels. The figure shows individual data points together with a smoothing spline and indications of F0-register boundaries. The figure suggests that most instruments from the string, flute, and recorder classes exhibit fairly strong positive associations between F0 and SC. Surprisingly, in the lower register of the bass recorder, one can observe a reversed relation such that increases in F0s are associated with a lowering of SC. For the *pp* dynamic level, effects of instrument size on envelope position are observable for the bass-trumpet and a particularly high SC trajectory for the fluegelhorn. The double bass exhibits lower SC values compared to the cello for F0s below 200 Hz. For *ff* dynamics, SC values from the clarinets partially segregate based on instrument size, the saxophones show a well-ordered increase in SC as a function of instrument size, and a similar trend exists for the bassoons. The brass instruments exhibit even more intricate patterns with SCs from C-trumpet and cornet exceeding the piccolo and bass-trumpet, and the contrabass-trombone having the highest SC trajectory between 2 and 4 kHz for the *ff* dynamic level.

To quantify effects of F0, size, and dynamic level on SC, multiple regression models were computed for every instrument class and F0-register. Figure 4 shows the corresponding regression coefficients. The model fit was rather weak for the vocals (mean $R^2 = 0.21$ averaged across the three registers) but substantially better for all other classes: trombones ($R^2 = 0.46$), clarinets ($R^2 = 0.49$), oboes ($R^2 = 0.62$), tubas ($R^2 = 0.69$), saxophones ($R^2 = 0.71$), recorders ($R^2 = 0.71$), trumpets ($R^2 = 0.74$), horns ($R^2 = 0.78$), bassoons ($R^2 = 0.80$), flutes ($R^2 = 0.83$), and strings ($R^2 = 0.88$).

Concerning the effects of F0, vocals and saxophones generally exhibited the smallest association between F0 and SC. Flutes and recorders showed the largest effects that were most pronounced in the higher register ($\beta > 0.5$). If for every instrument class a majority vote over F0-registers was used, 9 of 12 instrument classes showed positive associations between F0 and SC. Approximate F0-invariance of

17 August 2023 18:45:45

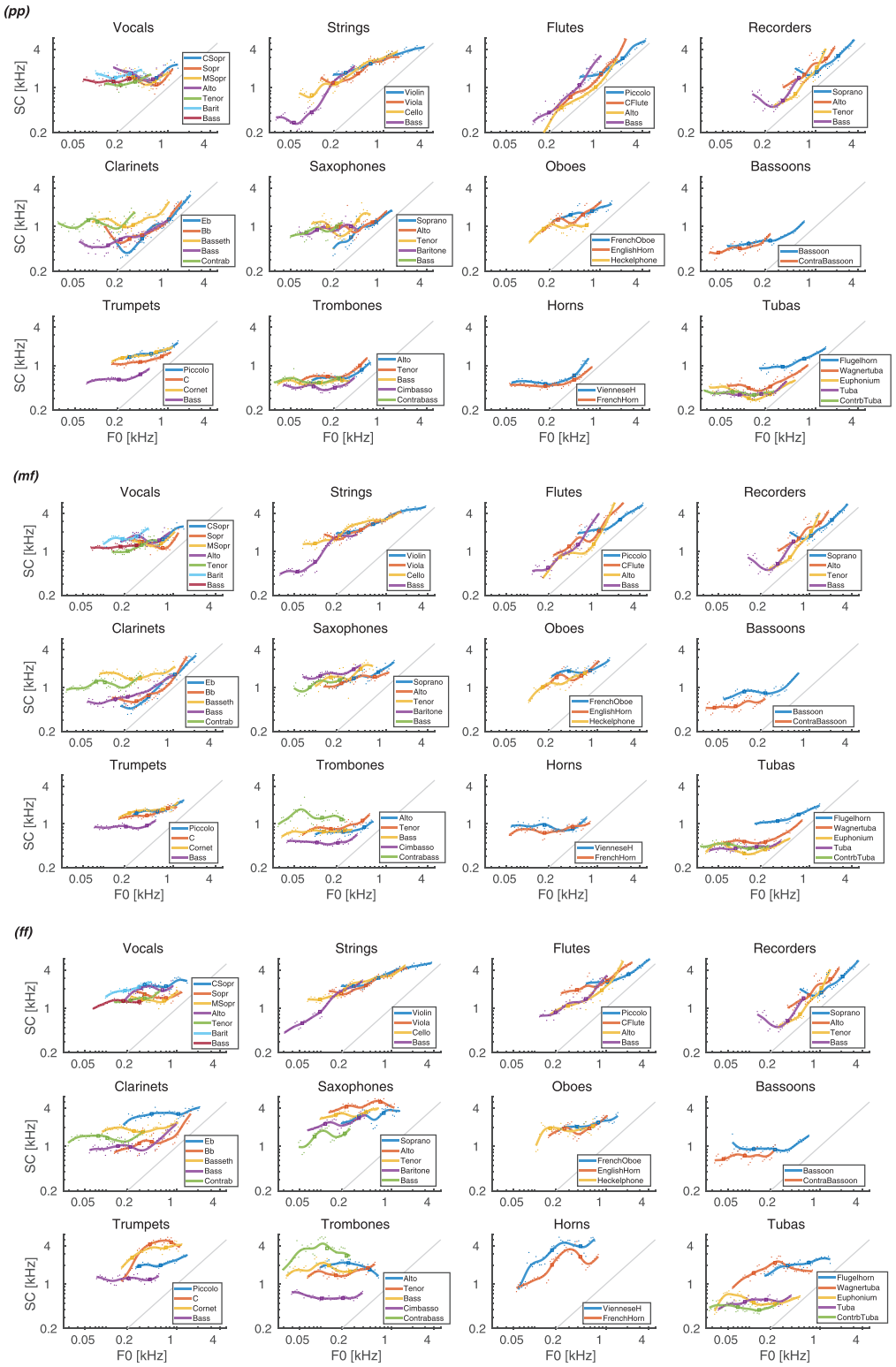


FIG. 3. (Color online) SC vs F0 for all sounds at different dynamic levels. Dots correspond to individual data points, squares to F0-register boundaries, the thin gray line corresponds to identical F0 and SC values. SC trajectories were smoothed using a cubic spline fit.

envelope position could be observed for specific F0-registers of half of the instrument classes (based on confidence intervals of regression coefficients overlapping with zero). Apart from the saxophones, F0-invariance was observed in the lower and middle register, consistent with the observation that the highest F0-registers tended to yield a strong positive association between F0 and SC.

Effects of instrument size were diverse and depended on F0-registers and instrument classes. Note that we here only measured a general size effect within instrument

17 August 2023 18:45:45

https://doi.org/10.1121/10.0005088

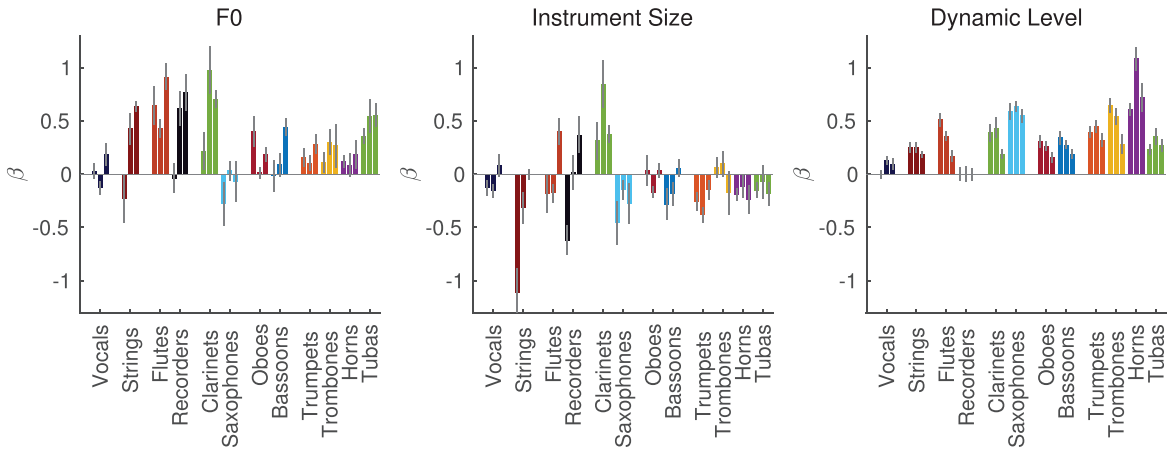


FIG. 4. (Color online) Linear effects of fundamental frequency, instrument size, and dynamic level derived from multiple linear regression with z-normalized regressors for every instrumental register. Individual bars from left to right correspond to lower, middle, and upper register, respectively. Error bars correspond to 95% confidence intervals of regression coefficients.

classes, that is, the linear association between average SC values per F0-register and the size estimates of instruments (when F0 and dynamic level were controlled). Thus, there may be sizeable differences between instruments within a class, but not necessarily a general size effect for the whole class. Specifically, there was a size effect for the lower and middle register of the vocals (recall that vocal tract length was used as a predictor of size) such that greater vocal tract length was associated with lower SCs, consistent with the hypothesis by Patterson *et al.* (2010). For the strings, the lower and middle registers showed strong size effects in the expected direction. The middle register of the flutes and the lowest register of the recorders also showed size effects as expected, but an inverse size effect for the highest register. Curiously, clarinets showed inverse size effects throughout all F0-registers, which were strongest in the middle register ($\beta = 0.8$). Oboes showed rather small and inconsistent size effects, whereas bassoons, saxophones, trumpets, horns, and tubas showed size effects for almost all registers. The trombones did not show size effects in any F0-register, which was potentially due to the extraordinarily high SCs of the contrabass-trombones for *mf* and *ff* dynamics.

Dynamic level generally exerted strong effects on SC values, although there was no effect for the lowest register of the vocals. Because the VSL provided the same recorder samples for all dynamic levels, the lack of an effect of dynamic level for the recorders was expected. From the woodwinds, the saxophones showed the largest effects of dynamic level, which were fairly consistent for all F0-registers ($\beta > 0.5$). These effects were only exceeded in strength by the middle register of the horns, where the effect was very large ($\beta = 1.0$)

Visual inspection of SC values in Fig. 3 may indicate non-negligible interactions between the factors F0, instrument size, and dynamic level, as well as potential non-linearities as a function of F0 (although one may argue that the latter are linearized by the partitioning into F0-registers). However, inclusion of several additional interaction terms would have considerably reduced the confidence of the effect estimates in the present analysis. We thus computed an additional regression with interaction terms but without the separation of F0-registers to ensure interpretability. The results of this analysis are largely congruent with the ones outlined here and are presented in the supplementary materials.[1]

To summarize the observed effects on envelope position, F0 and envelope position were positively associated for 9 of 12 instrument classes. Depending on the instrument class, effects of instrument size were diverse and varied across registers. Specifically, 7 of 12 instrument classes exhibited negative size effects for a majority of F0-registers. Effects of dynamic level appeared to be strong and consistent with positive associations for 11 of 12 instrument classes and particularly pronounced for the saxophones, trombones, and horns (Table II).

B. Envelope shape

To assess the association of spectral envelope position and shape, multiple linear regression of SC values of all

TABLE II. Summary of the presence (✓) or absence (x) of positive effects of F0 on SC, negative effects of size on SC [i.e., consistent with Patterson *et al.* (2010)], and positive effects of dynamic level on SC, all based on a majority vote among the three F0-registers.

Instrument	F0	Size	Dynamic level
Vocals	x	✓	✓
Strings	✓	✓	✓
Flutes	✓	x	✓
Recorders	✓	x	x
Clarinets	✓	x	✓
Saxophones	x	✓	✓
Oboes	✓	x	✓
Bassoons	x	✓	✓
Trumpets	✓	✓	✓
Trombones	✓	x	✓
Horns	✓	✓	✓
Tubas	✓	✓	✓

17 August 2023 18:45:45

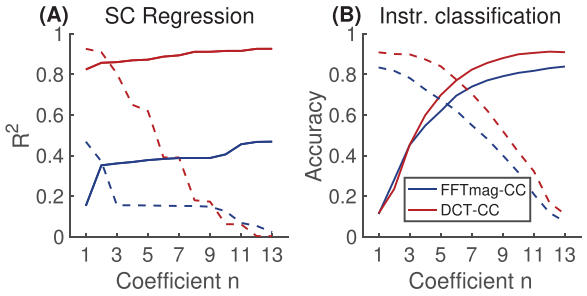


FIG. 5. (Color online) (A) R^2 values for multiple linear regression of SC values for all instrument sounds with FFTmag-CC (blue) and DCT-CC descriptors (red). Solid lines correspond to the selection of all coefficients up to the nth [1:n] for regression, and dashed lines correspond to the reverse selection [n:13]. (B) Degradation of classification accuracy as a function of type and number of included cepstral coefficients. Again, solid and dashed lines correspond to [1:n] and [n:13], respectively.

5640 sounds was computed by using cepstral coefficients as regressors. The analysis indicated that a linear combination of FFTmag-CC shared around 56% of the variance with the SC (as measured with R^2), implying that despite their mathematical translation invariance, FFTmag-CC appeared to be associated with envelope position. As shown in Fig. 5(A), lower modulation scales indeed contribute most significantly to this relationship. The same analysis further showed that a linear combination of DCT-CC in fact shared around 94% of the variance with the SC, most of which was due to the contribution of the very first coefficient. Note that the corresponding measures of association computed on uniformly random ERB-spectra yielded a shared variance of 0.14% and 77% for FFTmag-CC and DCT-CC, respectively; for Gaussian distribution functions with uniformly random positions and variances, regression yielded 5% and 86% of shared variance between SC and FFTmag-CC and DCT-CC, respectively. These results empirically illustrate that DCT-CC are not translation-invariant and hence encode information about envelope position. On the contrary, the FFTmag-CC descriptors do not represent envelope position, so that their strong association with SC reflects the empirical dependencies between envelope position and shape for acoustical instrument sounds.

Classification was conducted using a tenfold cross-validation on the full set of 5640 instrument sounds with mixed registers and dynamic levels. As shown in Table III, classification accuracy was fairly high, given the low-dimensional input representation, yielding an overall accuracy of 0.86 for FFTmag-CC and 0.95 for DCT-CC (confusion matrices are provided in the supplementary materials).[1] Note that most confusions took place within instrument families, such as the brass instruments, where tubas were often identified as trombones, for instance. Across-family confusions were frequent for the clarinets, which happened to be confused with saxophones, flutes, vocals, and oboes. The DCT-CC descriptors showed less frequent, albeit similar, patterns of confusion.

TABLE III. Classification accuracy for the mixed training/test set and the two different generalization scenarios (dynamics, F0-register) and sets of cepstral coefficients (translation-invariant FFTmag-CC vs DCT-CC). Training and test samples of the classifier were matched ("within") or mismatched ("across") according to dynamic level or F0-register (see Sec. II D). Standard deviations are given in parentheses.

		Dynamic level		F0-register	
Descriptors	Mixed	Within	Across	Within	Across
FFTmag-CC	0.86 (0.02)	0.75 (0.02)	0.55 (0.09)	0.75 (0.07)	0.34 (0.06)
DCT-CC	0.95 (0.01)	0.90 (0.02)	0.66 (0.10)	0.91 (0.01)	0.50 (0.16)

To specify the way in which instrument classification degrades with a reduction of the number of descriptors, we computed accuracy for subsets of cepstral coefficients (indexed by [1:n] or [n:13], where n was a running variable, using MATLAB notation). That is, either higher coefficients were added with increasing n ([1:n]), or lower coefficients were discarded ([n:13]). As shown in Fig. 5(B), instrument classification accuracy degraded linearly to chance level when lower-numbered coefficients were omitted for FFTmag-CC descriptors. On the other hand, when higher-numbered coefficients were added, accuracy increased more rapidly. For FFTmag-CC descriptors, the intersection between both lines was at $n = 5$ coefficients, implying that a classifier with five lower-order coefficients was as accurate as a classifier with eight higher-order coefficients. That means lower (coarser) spectral modulations were more valuable than higher (fine-grained) ones for discriminating between instrument classes.

Next, we considered the question of whether shape information was robust to changes in F0-register and dynamic level. Hence, generalization of the performance of the classifier was assessed by comparing a condition wherein training and test sets were matched in terms of dynamic level or register on the one hand with a condition wherein training and test sets contained mutually exclusive data. Table III shows the average accuracies for the two different sets of cepstral coefficients. For matched dynamic levels, the FFTmag-CC yielded an accuracy that is 15% below the DCT-CC approach. Both approaches worsened considerably when the test condition was mismatched in dynamic level. For F0-register, the effect was even more drastic: whereas classification was good for the same F0-registers with an accuracy of around 0.91 for the DCT and 0.75 for the FFTmag-CC, the accuracy dropped by around 40% for mismatched F0-registers in the test conditions.

Figure 6 shows more detailed results in form of F1-scores for every instrument class. It is notable that with the exception of the recorders and the clarinets, classification performance for all instrument classes suffered from generalization across dynamic level, suggesting that changes of dynamic level altered the spectral shape. Strong effects were observed for saxophones, oboes, bassoons, and horns, for which there were reductions of around 0.5 between F1-scores from the "within" and "across" conditions. Note that for the recorders, there is a reverse effect, likely because for the across condition, the lack of differences between recorder sounds from different dynamic levels may have led

17 August 2023 18:45:45

https://doi.org/10.1121/10.0005088

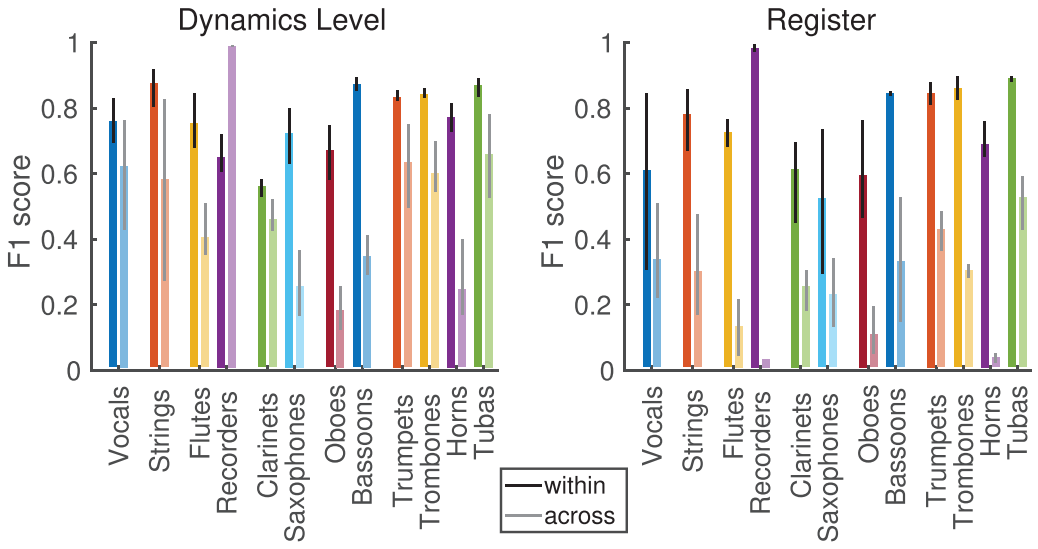


FIG. 6. (Color online) Classification performance of translation-invariant FFTmag-CC as assessed by F1-scores per instrument class. The within condition is based on threefold cross-validation specific to one dynamic level or register. Scores for the across condition correspond to training on two levels of dynamics or register and testing on the mismatched third level. Error bars display the F1-range across the three folds.

to very high classification performance. For generalization across registers, the classification of recorders declined drastically, such that F1 was close to 1 for the within condition and at around 0.1 for the across condition. Serious difficulties in generalizing across F0-register were also observed for the flutes, oboes, and horns with F1-scores below 0.2 in the across condition.

To summarize results concerning spectral shape, instrument classification using low-dimensional cepstral information was more than feasible with classification accuracies that arguably fell in the range of human listeners (Saldanha and Corso, 1964; Siedenburg, 2019). Audio representations that explicitly encoded information about spectral envelope position (DCT-CC) achieved better accuracies compared to representations that were mathematically translation-invariant (FFTmag-CC). Generalization over dynamic level and F0-register impaired classification accuracy, but the magnitude of these detriments differed strongly for specific instrument classes.

IV. DISCUSSION

We tested elementary hypotheses regarding the availability of cues of instrument size and instrument class (Patterson *et al.*, 2010) and aimed at providing a quantitative foundation regarding the interdependence of spectral envelope characteristics, F0, and dynamic level. For that purpose, we analyzed 50 sustained orchestral instruments from 12 different instrument classes using regression and classification. The present findings suggest that SC positions were strongly affected by dynamic level, an effect that was consistent across F0-registers. Increasing instrument size negatively affected SC positions [as hypothesized by Patterson *et al.* (2010)] for 7 of 12 instrument classes, although effects varied depending on F0-registers and were diminished for the highest F0-register of vocals, strings, flutes, recorders, oboes, and bassoons. Furthermore, F0 affected SC positions for 9 of 12 instrument classes, and invariance to F0 was least frequently observed for the higher F0-registers. Analogous results were obtained for the instrument classification task based on cepstral descriptors that capture spectral envelope shape. Using SVMs with Gaussian kernels and relatively low-dimensional cepstral descriptors sufficed for robust classification of 12 instrument classes beyond 80% accuracy, but accuracy degraded with generalization across dynamic levels and F0-registers. These results highlight the notion that the acoustic regularities inherent to musical instrument sounds depend on F0-register and dynamic level and that cues related to instrument size and instrument family are perturbed by these parameters.

The interaction between auditory attributes has long been studied (Melara and Marks, 1990; Russo and Thompson, 2005). In experiments with artificial stimuli, it has been observed that listeners showed symmetric interference between F0 and SC in pitch and brightness discrimination tasks (Allen and Oxenham, 2014) and that a consistent variation of pitch height and timbral brightness elicits associations with spatial dimensions (low-high), but not the variation of single dimensions alone (Pitteri *et al.*, 2017). In experiments with recorded instrument sounds, dependencies of affective qualities of sounds have been observed (McAdams *et al.*, 2017). Other studies have noted striking similarities between aspects of timbral brightness and pitch perception (Cousineau *et al.*, 2014; McDermott *et al.*, 2008; Siedenburg, 2018). Furthermore, it has been observed that listeners take into account timbre cues in the evaluation of dynamic level (Fabiani and Friberg, 2011). The present study suggests that there are complex statistical dependencies between key correlates of timbre perception, namely spectral envelope position and shape. Thus, our results provide an acoustical basis for an understanding of the interaction of auditory attributes and the association of pitch and

17 August 2023 18:45:45

timbral brightness perception in particular. Put in other terms, listeners' patterns of interference between perceptual attributes appear to be grounded in the statistics of natural sounds (cf. Młynarski and McDermott, 2019; Stilp and Kluender, 2016).

In psychophysics, it has long been known that listeners rely on size information (see, e.g., Giordano and McAdams, 2006). With regard to musical instrument sounds, Plazak and McAdams (2017) showed that for some but not all sustained instrument sounds, listeners reliably extract size differences as produced by an algorithm originally developed for speech manipulation (cf. Kawahara and Morise, 2011). Chiasson *et al.* (2017) suggested that listeners share a common perception of sound *extensity* that is supposedly inherent to the sound and independent of the size of the sound-producing object itself but is in accordance with earlier proposals of the notion of sound *volume* in the study of orchestration (Koechlin, 1954). Our results indicate that an increase in instrument size of acoustical instruments is associated with a decrease in the corresponding SC positions in 7 of 12 instrument classes, and this effect was most consistent and less affected by F0-register for the brass instruments. However, we note that the effect reversed for the clarinets.

Previous work on instrument classification has shown several precedents of successful classification of sustained musical instruments based on spectral shape alone. Brown *et al.* (2001) showed that sustained orchestral instrument sounds could be classified with accuracies above 80% by using spectral descriptors. Krishna and Sreenivas (2004) similarly achieved accuracies between 80% and 90% by using linear predictive coding and MFCC descriptors. In a study with sustained and impulsive instrument sounds from 11 instrument classes at different F0s and dynamic levels, Patil *et al.* (2012) used an SVM classifier with Gaussian kernels. With auditory spectrum descriptors, the authors achieved 79% accuracy, which increased for a more elaborate set of spectrotemporal modulation descriptors to almost 99% accuracy. Note that classification results are notoriously difficult to compare because of the many variables that affect the difficulty of the task (number of classes, size of data set, types and variability of sounds, etc.). For human listeners, Siedenburg (2019) observed 84% identification accuracy on a test set with ten instruments (constrained to a one-octave range and no variation of dynamic level). However, to approximate human levels of identification performance, a machine classifier appeared to require a detailed spectrotemporal signal representation (Siedenburg *et al.*, 2019). Given the diagnostic value of spectral features for sustained sounds that was observed in the present study, a reconsideration of the context-dependence of timbre cues used by human listeners may seem warranted.

It should be reiterated that the goal of the present study was not to obtain the best possible classification performance. Rather, the goal was to understand whether spectral shape and its diagnostic value for inferring instrument classes were much affected by the parameters of F0-register and dynamic level. We found that translation-invariant cepstral coefficients (FFTmag-CC) that primarily encode spectral shape provided good classification results, corroborating the general idea that spectral shape suffices for the perceptual characterization of instrument classes or types. The comparison between translation-invariant (FFTmag-CC) and translation-dependent (DCT-CC) descriptors further suggests that envelope position information (that is more explicitly encoded by the DCT-CC) further improves overall classification accuracy.

There has been debate regarding the use of MFCCs and the SC for describing the spectral correlates of timbre dissimilarity perception (cf. Aucouturier and Bigand, 2012; Horner *et al.*, 2011; Siedenburg *et al.*, 2016; Terasawa *et al.*, 2012). More recently, an analogous dichotomy may be located in questions around low-dimensional scalar audio descriptors and high-dimensional spectrotemporal modulation representations (Elliott *et al.*, 2013; Thoret *et al.*, 2020). Here, we illustrated that cepstral coefficients can be translation-invariant (i.e., discard envelope position information such as the FFTmag-CC) or translation-dependent (such as the DCT-CC). Due to the statistics of acoustical instrument sounds, however, both methods share a substantial fraction of variance with envelope position as measured by SC. Specifically, these statistical dependencies allowed us to almost completely reconstruct SC values by a linear combination of DCT-cepstra, which are structurally analogous to MFCCs. This insight suggests that the traditional dichotomy between MFCCs and SC descriptors could be obsolete, given that the former audio representation indeed comprises most of the essential information about the latter.

In summary, we provided empirical evidence that the position of the spectral envelope can provide cues about instrument size and that the shape of the envelope indexes instrument class or family. Importantly, envelope position and shape tend to be driven by the parameters of F0, F0-register, and dynamic level, although not all instrument classes share all of these effects. We also observed that position and shape parameters exhibit a strong association for a set of more than 5000 sounds. These analyses provide a quantitative baseline of stimulus characteristics that may inform further studies on musical instrument sound, timbre perception, and the interaction of auditory parameters. This work comprises several potential avenues for further development. Specifically, our notion of F0-register was derived from treatises on instrument acoustics and orchestration, but it is not *a priori* clear whether these accounts coincide with acoustical reality. The development of data-driven approaches to estimate register boundaries could be a next step to test whether traditional ideas in instrument acoustics and orchestration can be grounded in signal analysis.

ACKNOWLEDGMENTS

The authors thank the editors of the special issue on Modeling of Musical Instruments and the two anonymous reviewers for their helpful comments. This research was

17 August 2023 18:45:45

supported by a Freigeist Fellowship of the Volkswagen Foundation to K.S.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0005088 for a table with registers, size estimates, and detailed references and for SC values computed on a set of artificial sounds and various complementary analyses and results.

Allen, E. J., and Oxenham, A. J. (**2014**). "Symmetric interactions and interference between pitch and timbre," J. Acoust. Soc. Am. **135**(3), 1371–1379.
Aucouturier, J.-J., Bigand, E. (**2012**). "Mel Cepstrum & Ann Ova: The difficult dialog between MIR and music cognition," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, October 8–12, Porto, Portugal, pp. 397–402.
Berlioz, H. (**1844**). *Grand Traité d'Instrumentation et d'Orchestration Modernes* (Schonenberger, Paris).
Brown, J. C., Houix, O., and McAdams, S. (**2001**). "Feature dependence in the automatic identification of musical woodwind instruments," J. Acoust. Soc. Am. **109**(3), 1064–1072.
Caetano, M., Saitis, C., and Siedenburg, K. (**2019**). "Audio content descriptors of timbre," in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Springer, Cham, Switzerland), pp. 297–333.
Campbell, M., Gilbert, J., and Myers, A. (**2020**). *The Science of Brass Instruments* (Springer, Heidelberg, Germany).
Campbell, M., Greated, C. A., and Myers, A. (**2004**). *Musical Instruments: History, Technology, and Performance of Instruments of Western Music* (Oxford University, Oxford, UK).
Chiasson, F., Traube, C., Lagarrigue, C., and McAdams, S. (**2017**). "Koechlin's volume: Perception of sound extensity among instrument timbres from different families," Music. Sci. **21**(1), 113–131.
Cousineau, M., Carcagno, S., Demany, L., and Pressnitzer, D. (**2014**). "What is a melody? On the relationship between pitch and brightness of timbre," Front. Syst. Neurosci. **7**, 1–7.
Elhilali, M. (**2019**). "Modulation representations for speech and music," in *Timbre: Acoustics, Perception, and Cognition*, edited by K. Siedenburg, C. Saitis, S. McAdams, R. Fay, and A. Popper (Springer, Cham, Switzerland), pp. 335–359.
Elliott, T., Hamilton, L., and Theunissen, F. (**2013**). "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones," J. Acoustical Soc. Am. **133**(1), 389–404.
Fabiani, M., and Friberg, A. (**2011**). "Influence of pitch, loudness, and timbre on the perception of instrument dynamics," J. Acoust. Soc. Am. **130**(4), EL193–EL199.
Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague, Netherlands).
Fletcher, N. H. (**1999**). "The nonlinear physics of musical instruments," Rep. Prog. Phys. **62**(5), 723–764.
Fletcher, N. H., and Rossing, T. D. (**2012**). *The Physics of Musical Instruments* (Springer-Verlag, New York).
Forsyth, C. (**1982**). *Orchestration* (Dover, New York).
Friedman, J., Hastie, T., and Tibshirani, R. (**2009**). *The Elements of Statistical Learning*, 2nd ed. (Springer, Heidelberg, Germany).
Gevaert, F. A. (**1885**). *Nouveau Traité d'Instrumentation* (Lemoine & fils, Paris).
Giordano, B. L., and McAdams, S. (**2006**). "Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates," J. Acoust. Soc. Am. **119**(2), 1171–1181.
Horner, A. B., Beauchamp, J. W., and So, R. H. (**2011**). "Evaluation of mel-band and MFCC-based error metrics for correspondence to discrimination of spectrally altered musical instrument sounds," J. Audio Eng. Soc. **59**(5), 290–303.
Kawahara, H., and Morise, M. (**2011**). "Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework," Sadhana **36**(5), 713–727.
Koechlin, C. (**1954**). *Traité de L'Orchestration*, Vols. 1–4 (Editions Max Eschig, Paris).
Krishna, A., and Sreenivas, T. V. (**2004**). "Music instrument recognition: From isolated notes to solo phrases," in *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 17–21, Montreal, Canada, Vol. 4, pp. iv–iv.
Lembke, S.-A., and McAdams, S. (**2015**). "The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds," Acta Acust. united Acust. **101**(5), 1039–1051.
Luce, D., and Clark, M., Jr. (**1967**). "Physical correlates of brass-instrument tones," J. Acoust. Soc. Am. **42**(6), 1232–1243.
Marozeau, J., de Cheveigné, A., McAdams, S., and Winsberg, S. (**2003**). "The dependency of timbre on fundamental frequency," J. Acoust. Soc. Am. **114**(5), 2946–2957.
McAdams, S., Douglas, C., and Vempala, N. N. (**2017**). "Perception and modeling of affective qualities of musical instrument sounds across pitch registers," Front. Psychol. **8**, 1–19.
McDermott, J. H., Lehr, A. J., and Oxenham, A. J. (**2008**). "Is relative pitch specific to pitch?" Psychol. Sci. **19**(12), 1263–1271.
Melara, R. D., and Marks, L. E. (**1990**). "Interaction among auditory dimensions: Timbre, pitch, and loudness," Percept. Psychophys. **48**(2), 169–178.
Meyer, J. (**2009**). *Acoustics and the Performance of Music* (Springer, New York, NY).
Miller, R. J. (**2014**). *Contemporary Orchestration: A Practical Guide to Instruments, Ensembles, and Musicians* (Routledge, New York).
Młynarski, W., and McDermott, J. H. (**2019**). "Ecological origins of perceptual grouping principles in the auditory system," Proc. Natl. Acad. Sci. U.S.A. **116**(50), 25355–25364.
Moore, B. C., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**(3), 750–753.
Mürbe, D., Roers, F., and Sundberg, J. (**2011**). "Stimmgattung professioneller Sänger ("Voice classification in professional singers"), HNO **59**(6), 556–562.
Myers, A., Pyle, R. W., Jr., Gilbert, J., Campbell, D. M., Chick, J. P., and Logie, S. (**2012**). "Effects of nonlinear sound propagation on the characteristic timbres of brass instruments," J. Acoust. Soc. Am. **131**(1), 678–688.
Oehler, M., and Reuter, C. (**2009**). "Dynamic excitation impulse modification as a foundation of a synthesis and analysis system for wind instrument sounds," in *Mathematics and Computation in Music: First International Conference, MCM 2007* (Springer, New York), p. 189.
Ogg, M., and Slevc, L. R. (**2019**). "Acoustic correlates of auditory object and event perception: Speakers, musical timbres, and environmental sounds," Front. Psychol. **10**, 1–21.
Pape, W. (**1976**). *Instrumentenhandbuch: Streich-, Zupf-, Blas und Schlaginstrumente in Tabellenform* (Hans Gerig, Köln, Germany).
Patil, K., Pressnitzer, D., Shamma, S. A., and Elhilali, M. (**2012**). "Music in our ears: The biological bases of musical timbre perception," PLoS Comput. Biol. **8**(11), e1002759.
Patterson, R. D., Gaudrain, E., and Walters, T. C. (**2010**). "The perception of family and register in musical tones," in *Music Perception*, edited by M. Riess Jones, R. R. Fay, and A. Popper (Springer, New York), pp. 13–50.
Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (**2011**). "The timbre toolbox: Extracting audio descriptors from musical signals," J. Acoust. Soc. Am. **130**(5), 2902–2916.
Pitteri, M., Marchetti, M., Priftis, K., and Grassi, M. (**2017**). "Naturally together: Pitch-height and brightness as coupled factors for eliciting the SMARC effect in non-musicians," Psychol. Res. **81**(1), 243–254.
Plazak, J., and McAdams, S. (**2017**). "Perceiving changes of sound-source size within musical tone pairs," Psychomusicol. Music Mind Brain **27**(1), 1–13.
Reuter, C. (**1996**). *Die Auditive Diskrimination von Orchesterinstrumenten* (Peter Lang, Bern, Switzerland).
Reuter, C. (**2002**). *Klangfarbe Und Instrumentation* (Peter Lang, Bern, Switzerland).
Rimskij-Korsakov, N. (**1922**). *Principles of Orchestration* (Kalmus, New York).
Russo, F., and Thompson, W. (**2005**). "An interval size illusion: The influence of timbre on the perceived size of melodic intervals," Atten. Percept. Psychophys. **67**(4), 559–568.
Saitis, C., and Siedenburg, K. (**2020**). "Brightness perception for musical instrument sounds: Relation to timbre dissimilarity and source-cause categories," J. Acoust. Soc. Am. **148**(4), 2256–2266.

17 August 2023 18:45:45

Saldanha, E., and Corso, J. F. (**1964**). "Timbre cues and the identification of musical instruments," J. Acoust. Soc. Am. **36**(11), 2021–2026.
Siedenburg, K. (**2018**). "Timbral Shepard-illusion reveals perceptual ambiguity and context sensitivity of brightness perception," J. Acoust. Soc. Am. **143**(2), EL93–EL98.
Siedenburg, K. (**2019**). "Specifying the perceptual relevance of onset transients for musical instrument identification," J. Acoust. Soc. Am. **145**(2), 1078–1087.
Siedenburg, K., Fujinaga, I., and McAdams, S. (**2016**). "A comparison of approaches to timbre descriptors in music information retrieval and music psychology," J. New Music Res. **45**(1), 27–41.
Siedenburg, K., Jacobsen, S., and Reuter, C. (**2021a**). "Register boundaries and spectral centroids of musical instruments," http://www.univie.ac.at/muwidb/pitch_sc_register (Last viewed 5/20/2021).
Siedenburg, K., Jacobsen, S., and Reuter, C. (**2021b**). https://github.com/Music-Perception-and-Processing/spectral-envelope-study (Last viewed 5/20/2021).
Siedenburg, K., and McAdams, S. (**2018**). "Short-term recognition of timbre sequences: Music training, pitch variability, and timbral similarity," Music Percept. **36**(1), 24–39.
Siedenburg, K., Schädler, M. R., and Hülsmeier, D. (**2019**). "Modeling the onset advantage in musical instrument recognition," J. Acoust. Soc. Am. **146**(6), EL523–EL529.
Steele, K. M., and Williams, A. K. (**2006**). "Is the bandwidth for timbre invariance only one octave?" Music Percept. **23**(3), 215–220.
Stilp, C. E., and Kluender, K. R. (**2016**). "Stimulus statistics change sounds from near-indiscriminable to hyperdiscriminable," PLoS One **11**(8), e0161001.
Terasawa, H., Berger, J., and Makino, S. (**2012**). "In search of a perceptual metric for timbre: Dissimilarity judgments among synthetic sounds with MFCC-derived spectral envelopes," J. Audio Eng. Soc. **60**(9), 674–685.
Thoret, E., Caramiaux, B., Depalle, P., and McAdams, S. (**2020**). "Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre," Nat. Hum. Behav. **5**, 369–369.
Vienna Symphonic Library (**2021**). www.vsl.co.at (Last viewed 5/20/2021).
von Hornbostel, E. M., and Sachs, C. (**1914**). "Systematik der musikinstrumente. Ein versuch," Z. Ethnologie **46**(H. 4/5), 553–590.
Weinzierl, S., Lepa, S., Schultz, F., Detzner, E., von Coler, H., and Behler, G. (**2018**). "Sound power and timbre as cues for the dynamic strength of orchestral instruments," J. Acoust. Soc. Am. **144**(3), 1347–1355.
Widor, C. M. (**1904**). *Technique de L'orchestre Moderne: Faisant Suite au Traité d'Instrumentation et d'Orchestration de H. Berlioz* (H. Lemoine, Paris).

17 August 2023 18:45:45