Acoustical correlates of perceptual blend in timbre dyads and triads

Sven-Amin Lembke

Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Schulich School of Music, McGill University, Montréal, Québec, Canada

Kyra Parker

Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Schulich School of Music, McGill University, Montréal, Québec, Canada

Eugene Narmour

Department of Music, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Stephen McAdams

Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Schulich School of Music, McGill University, Montréal, Québec, Canada

Abstract

Achieving a blended timbre for particular combinations of instruments, pitches, and articulations is a common aim of orchestration. This involves a set of factors that this study jointly assesses by correlating the perceptual degree of blend with the underlying acoustical characteristics. Perceptual blend ratings from two experiments were considered, with the stimuli consisting of: 1) dyads of wind instruments at unison and minor-third intervals and at two pitch levels, and 2) triads of wind and string instruments, including bowed and plucked string excitation. The correlational analysis relied on partial least-squares regression, as this technique is not restricted by the number and collinearity of regressors. The regressors encompassed acoustical descriptors of timbre (spectral, temporal, and spectrotemporal) as well as ones accounting for pitch and articulation. From regressor loadings in principal-components space, the major regressors leading to substantial and orthogonal contributions were identified. The regression models explained around 90% of the variance in the datasets, which was achievable with less than a third of all regressors considered initially. Blend seemed to be influenced by differences across intervals, pitch, and articulation. Unison intervals yielded more blend than did non-unison intervals, and the presence of plucked strings resulted in clearly lower blend ratings than for sustained instrument combinations. Furthermore, prominent spectral features of instrument combinations influenced perceived blend.

Keywords: timbre, blend, orchestration, instrument dyads/triads, acoustical descriptors, multivariate regression

Acoustical correlates of perceptual blend in timbre dyads and triads

In orchestration, composers may consider several factors when they intend to achieve a *blended* timbre between two or more instruments playing synchronously. There is the choice of suitable instruments that can yield a blended combination, which depends on the acoustical traits of these instruments. The remaining factors involve more musical considerations: whether instruments will be playing in unison or non-unison, which instrument is assigned to the top voice in non-unison passages, in what registral range the instruments will be playing, and what kind of articulation they will employ (e.g., bowed or plucked string). When it comes to establishing general associations between the perception of timbre blend and its underlying acoustical characteristics, the joint assessment of these factors will assist in predicting the perceived degree of blend for combinations of instruments, pitches, and articulations.

Previous research has defined perceived timbre blend as the auditory fusion of concurrent instrumental sounds, where individual sounds become less distinct. The most common method to measure perceived blend employs rating scales (Kendall & Carterette, 1993; Lembke, Levine, & McAdams, in press; Lembke & McAdams, 2015; Sandell, 1995; Tardieu & McAdams, 2012). All studies found that spectral features influence blend, but employed different approaches to spectral description. One approach used the global descriptor *spectral centroid*, i.e., the amplitude-weighted frequency average of a spectrum. The composite (or sum) of the individual sounds' centroids was found to predict blend in unison dyads best (Sandell, 1995; Tardieu & McAdams, 2012), whereas for non-unison dyads, the absolute difference in individual spectral centroids served as the more reliable predictor (Sandell, 1995).

Another approach to spectral description has considered the influence of prominent spectral features, such as maxima or *formants*. Similar to the relevance of formants in describing the acoustics of the human voice (Fant, 1960), wind instruments in particular exhibit formant structures that remain largely invariant across pitch (Lembke & McAdams, 2015; D. Luce & Clark, 1967; D. A. Luce, 1975; Schumann, 1929). Their identification and description can be achieved through spectral estimations that are aggregated across an instrument's complete pitch range (Lembke & McAdams, 2015), and therefore can be considered *pitch-generalized*. Reuter (1996) has argued that similarity between instruments' formant structures can explain blend. Hardly distinguishable instrument pairings can exhibit very similar formant locations (e.g., horn and bassoon), whereas the strongly pronounced, unique formant structure of the oboe may hinder it from blending with most other instruments.

Frequency relationships between the most prominent *main formants* appear to influence blend critically (Lembke & McAdams, 2015). In dyads comprising a recorded wind-instrument sound and a synthesized analogue to that instrument, whose main-formant frequency could be shifted relative to that of the recorded sound, blend decreased drastically as the frequency of the synthesized formant exceeded that of the recorded sound. This relative dependency relates to musical performance, where accompanying musicians adjust their main formants to be lower than when playing as the leading instrument (Lembke et al., in press).

Apart from spectral properties, differences between temporal features, such as note attacks or onsets, have been found to explain blend as secondary factors for unison dyads (Sandell, 1995). However, their influence becomes more dominant as attacks turn impulsive: shorter durations and steeper attack slopes lead to reduced blend (Tardieu & McAdams, 2012).

With respect to those musical factors unrelated to timbre, blend for unison dyads is perceived as stronger than for non-unison combinations (Kendall & Carterette, 1993; Lembke et al., in press). Furthermore, the assignment of instruments to the upper and lower pitches in non-unison intervals resulted in differences in perceived blend between instrument inversions in one study (Kendall & Carterette, 1993), but lacked a comparable effect in another (Sandell, 1995). All of these studies on blend are limited to dyadic contexts, leaving open how the obtained results and proposed hypotheses would fare in combinations of three or more instruments. Little work has been published on timbre combinations in triadic contexts (Kendall, 2004; Kendall & Vassilakis, 2006, 2010), and none of these papers address issues directly related to blend.

With the aim of predicting perceived blend between arbitrary instrument combinations, linear correlation or regression can be employed to associate blend measures with single acoustical features (Sandell, 1995; Tardieu & McAdams, 2012), without, however, making it possible to assess how several acoustical descriptors could jointly contribute to the explanation of blend measures. This limitation can be overcome by *multiple linear regression* (MLR). Past attempts have succeeded in explaining up to 63% of the variance in blend ratings for mixed-instrument dyads (Sandell, 1995). Similarly, MLR models also explained up to 87% of the variance in blend ratings across dyads in which the role of local, parametric variations of the main-formant frequency was studied (Lembke & McAdams, 2015).

Yet, the MLR approach also has clear limitations. High collinearity among independent variables (regressors) or a low number of cases compared to the number of regressors may both lead to less reliable and less valid results as well as mathematically ill-defined solutions. This becomes problematic given the aim of the current paper, because many spectral descriptors are known to exhibit a high inter-correlation (Peeters, Giordano, Susini, Misdariis, & McAdams, 2011). For conventional MLR, this leaves two options: 1) disregarding the collinearity, at the risk of obtaining less reliable or invalid results or 2) eliminating regressors that are collinear to a reference regressor, i.e., one found to predict blend most strongly in simple linear regression. However, the latter approach risks excluding variables that might perform even better than the selected one once they interact with other regressors.

A viable solution to deal with collinearity is to employ a dimension-reduction technique like *principal component analysis* (PCA) that reduces a high quantity of regressors to a small number of substitute or latent variables, i.e., *principal components* (PCs), which are orthogonal to one another. These PCs can thereafter serve as regressors that represent the common aspects for groups of collinear descriptors (e.g., Giordano, Rocchesso, & McAdams, 2010).

A promising regression method that uses PCA as an integral part is *partial least-squares regression* (PLSR), which originates from the discipline of chemometrics, but has more recently been applied within the field of auditory perception (Eerola, Lartillot, & Toiviainen, 2009; Kumar, Forster, Bailey, & Griffiths, 2008; Rumsey, Zieliński, Kassier, & Bech, 2005). PLSR allows analysis of complex correlational relationships among perceptual measures and arrays of acoustical or psychoacoustical variables.

The current investigation uses PLSR in an attempt to predict blend ratings from perceptual experiments. The perceptual data are collected on a diverse set of variables that affect timbral blend and orchestration, including different instruments, pitches, and unison and non-unison intervals, as well as dyadic and triadic contexts. The set of potential regressors consists of a wide range of acoustical measures that, through several stages of PLSR models, are continually refined to retain only the most relevant regressors and, importantly, ones that are independent of each other.

Method

Partial least-squares regression (PLSR)

Predicting a single measure of blend through a set of regressors relating to acoustical descriptors can be expressed mathematically by associating the column vector of blend ratings y with an $n \times m$ matrix X, which encompasses n cases (e.g., stimulus conditions) across m regressors. Conventional MLR employs the relationship $y = X \cdot b$, with b being a vector of regression coefficients of length m. PLSR represents algorithms that employ an inherent coupling between MLR and PCA (Geladi & Kowalski, 1986), allowing large m relative to n and even collinearity among the m regressors.

PLSR decomposes X into k principal components (PCs), yielding the relationship $X = T \cdot P'$, with T representing an $n \times k$ matrix of *scores* and P an $m \times k$ matrix of *loadings*. Unlike computing a PCA on X independently and inputting the obtained scores T into MLR, PLSR achieves the component decomposition by maximization of the inherent covariance between y and X, leading to a better predictive relationship. The loadings P can be understood as vectors for the m regressors in k-dimensional space, describing the degree to which regressors contribute to individual PCs and also showing the collinearity or independence among regressors. The PLSR technique used here is SIMPLS (de Jong, 1993), in its implementation for MATLAB.[1]

Performance, predictive power, and reliability. Regression performance evaluates the variation in y that is explained by the model. The measure R^2 describes both the global and component-wise performance, with the latter quantifying the relative contribution of PCs. With increasing k, however, models are prone to overfitting the data, at the cost of predictive power when applied to other data sets. In order to assess the predictive power of models, sixfold cross validation (CV) is employed, partitioning the n cases into six subsets of similar size, building models based on five subsets, assessing the error in predicting the remaining, excluded subset, and repeating the last two steps for all permutations of subsets. CV allows the computation of an alternative measure of explained variance, Q^2 (Wold, Sjöström, & Eriksson, 2001). Just as R^2 evaluates the sum of squared deviations between the *fitted* and actual y, Q^2 evaluates the sum of squared deviations between the *predicted* and actual y, with these predictions made for the excluded subsets across all CV permutations. Together, R^2 and Q^2 can be taken as the upper and lower benchmarks of the model, respectively, in terms of explaining the data and assessing the degree of predictive power. The selection of the optimal number of components k considers two independent criteria: 1) the component-wise gain in R^2, and 2) the component-wise decrease in CV prediction error, with k being chosen when both measures cease to exhibit substantial improvements for additional PCs.

Identifying relevant and independent regressors. The current PLSR analysis aims to reduce the number of investigated regressors in X to those of greatest utility in explaining y as well as further reducing it to a selection of regressors that are relatively independent of each other. The chosen approach consists of three stages of sequentially evaluating and refining PLSR models: 1) An initial model is obtained for the original matrix X_{orig} of all regressors considered (see Table 3). 2) Based on the loadings P_{orig} from the first PLSR model, one half of the variables are identified that act as the strongest predictors. More specifically, only those regressors are retained for which the Euclidean distances across k dimensions exceed the distribution median ($Q50$), leading to the computation of another model based on the reduced matrix X_{Q50}. 3) The following stage distills the regressors down to those that explain y through ideally independent contributions. Such independence or orthogonality is achieved for loadings P that point in perpendicular directions in k-dimensional space. To this aim, imagine the resulting loadings P_{Q50} rotated so as to align the most dominant variable loading along one axis of a Cartesian coordinate system. Relative to this variable, loadings aligned along the remaining $k-1$ axes exhibit maximal independence. To obtain an ideally independent set of variables, the selection is constrained to variables such that the angles ϕ_i between variable loadings and the ith axis are less than 22.5°. This constraint yields an approximately orthogonal set of regressors X_{ortho}, on which the final PLSR model is computed.

Perceptual data sets

The regression analysis considers two data sets that originate from listening experiments in which participants provided blend ratings for dyads or triads. The two experiments were unrelated with respect to their original motivation and experimental design, yet they employed similar blend ratings, with the medians across participants taken as the dependent variable y to be modeled through PLSR.

The stimuli were presented over a standard two-channel stereophonic loudspeaker setup inside an Industrial Acoustics Company double-walled sound booth. They were drawn from recorded instrument samples from the Vienna Symphonic Library[2] (VSL), supplied as stereo WAV files (44.1 kHz sampling rate, 16-bit amplitude resolution). In separate pilot experiments, all stimuli had been both adjusted for perceptual synchrony between sounds constituting the dyads and triads and equalized for loudness within the dyad and triad sets independently. Adjustments for synchrony were based on consensus by three people for dyads and two for triads. The loudness equalization was conducted subjectively, anchored to a global reference for all dyad or triad conditions. The equalization was conducted by five people for dyads and six for triads. Gain levels were determined that equalized stimulus loudness to the global reference. These gain levels were based on median values across participants; all corresponding interquartile ranges were less than 4 dB.

For the main experiments, participants with varying degrees of musical experience were recruited from the McGill University community. All participants passed a standardized pure-tone audiogram (ISO 389–8, 2004; Martin & Champlin, 2000) ensuring that thresholds at all audiometric frequencies were less than or equal to 20 dB HL. Informed consent was obtained, and both studies were certified for ethical compliance by the McGill University Research Ethics Board II.

Dyads.

Participants. Nineteen people took part in the experiment (12 female and seven male) with a median age of 21 years (range: 18–46). Among the participants, nine considered themselves amateur musicians, two as professional musicians, and eight as non-musicians. All were compensated financially for their participation in the hour-long experiment.

Stimuli. The stimulus set comprised a total of 180 dyads that resulted from the combination of several factors. Six wind instruments, namely, (French) horn, bassoon, oboe, C trumpet, B♭ clarinet, and flute, formed the 15 possible non-identical-instrument pairs listed in Table 1. These instrument pairs occurred at two pitch levels: C4 ($f_0 = 261.6$ Hz) and G4. Furthermore, dyads comprised both unison and minor-third intervals, including the inverse voicing of instruments for the latter, resulting in a total of three interval conditions. Based on the two pitch levels, minor thirds occurred at the pitches C4-E♭4 and G4-B♭4.

All VSL samples were sustained, non-vibrato recordings, performed at *mezzoforte* dynamics, and were limited to the signal in the left channel. Both instruments were simulated as being captured by a stereo main microphone at spatially distinct locations inside a mid-sized, moderately reverberant room. Encompassing a volume of 600 m^3, the relatively absorbent room yielded a reverberation time $T_{30} = 0.4$ s; due to the configuration inside the room being fully symmetric, identical frequency responses applied to both instruments.[3] The spatial locations of instruments included both possible orientations (e.g., horn left of bassoon and vice versa). Overall, this resulted in the full-factorial combination of 15 pairs × 2 pitches × 3 intervals × 2 orientations = 180 dyads. All stimuli had a duration of 1200 ms, with artificial offsets imposed by a 100-ms linear amplitude ramp. A set of 12 representative dyad stimuli can be found in the Supplemental Material Online section.

[— Insert Table 1 about here. —]

Procedure. Participants heard individual dyads in randomized order and were asked to rate their degree of blend, employing a continuous slider scale with the verbal anchors *most blended* and *least blended* visualized on a computer screen. Ahead of the main experiment, participants had been familiarized with the degree of possible variation in blend among all dyads and had completed 15 practice trials on a separate but comparable stimulus set.

Triads.

Participants. Twenty people (15 female and five male) with a median age of 21 years (range: 19–64) completed the experiment. Thirteen participants classified themselves as amateur musicians, with the remaining seven being non-musicians. All were remunerated for the hour-long experiment.

Stimuli. The stimuli comprised 20 triads, representing only a selection of the vast multiplicity of possible instrument and pitch combinations. In order to focus on timbral characteristics, all triads formed the same chord with pitches C4, F4, and B♭4, thus controlling for contextual effects with pitch register, chroma, and height. This chord choice of stacked perfect-fourth intervals avoided standard major and minor triads and generated the same consonances (perfect fourths) and a single dissonance (minor seventh). Such quartel chords were neutral enough not to draw attention to any one melodic voice, while allowing ‘inside’, middle voices to be easily heard.

In terms of instrumentation, the triads were composed of flute, oboe, B♭ clarinet, tenor trombone and cello sounds, corresponding to the instrument families woodwinds, brass, and strings. The instrument selection for triads (see Table 2) comprised mixtures between two or three instrument families. Furthermore, the selection included all woodwind reed types (air jet, single and double reed) and two different excitation types for strings (bowed and plucked excitation; *arco* and *pizzicato*, respectively), with each distinction represented by a single instrument (e.g., oboe for double reed, cello for string instrument). Instruments would only take on pitches based on conventional voice assignments given a particular mixture. For instance, the cello only occurred at the two lower pitches, whereas the flute was always highest in pitch relative to other woodwinds. Each instrument appeared in from six to 10 triads (counting different excitation types as separate instances).

All samples were taken as stereo files from VSL, with woodwind samples comprising sustained sounds at *mezzoforte* dynamics and without vibrato. The trombone samples were similar, but at *mezzopiano* dynamics. The *arco* cello samples were recorded at *mezzoforte* dynamics. Unlike the wind instruments, they decayed after just a brief bow stroke, in order to be more similar to the *pizzicato* versions, which occurred at *forte* to allow for a longer sound decay. All cello sounds contained vibrato. The total duration for all triads was limited to 850 ms by applying an artificial 100-ms linear amplitude-decay ramp. A set of 10 representative triad stimuli can be found in the Supplemental Material Online section.

[— Insert Table 2 about here. —]

Procedure. Participants were asked to sort all triads based on their relative degree of blend along a scale continuum with the verbal anchors *most blended* and *least blended*. At the beginning, visual icons for all triads were randomly arranged on a computer screen and could be dragged around or clicked on to trigger sound playback. Participants were first asked to identify two triads perceived as exhibiting the highest or lowest blend, to assign them to the extremes of the visualized continuum and then to position all remaining triads along the continuum. The sorting was conducted twice, the first counting as a practice round meant to familiarize participants with both the experimental task and the triads, the second serving as the main experiment.

Acoustical descriptors

For each data set, a collection of acoustical measures constitute the regressors in matrix X. The measures include spectral, temporal, and spectrotemporal acoustical descriptors of timbre, as well as other potentially relevant features such as differences in fundamental frequency. Table 3 lists all the investigated descriptors, specifying how individual descriptor values were associated with dyads and triads.

Descriptor relationships within dyadic and triadic contexts. As dyads and triads consist of several constituent sounds, their individual descriptor values need to be summarized to a single regressor value per stimulus by an association of some kind. For dyads with the constituent sounds a and b and the acoustical descriptor x, the association considers the *difference* measure $\Delta x = |x_a - x_b|$ and the *composite* measure $\Sigma x = x_a + x_b$. Three associations are computed for triads with sounds a, b, and c, whose relationship along descriptor x is $x_a \leq x_b \leq x_c$. The triad *difference* considers the range between maximum and minimum, i.e., $\Delta x = x_c - x_a$. The *composite* sums all three values, i.e., $\Sigma x = x_a + x_b + x_c$. In addition, a third measure relates the *distribution* of the intermediate value x_b relative to the extremes, i.e., $\Xi\, x = 2 \cdot (x_b - x_a)/\Delta x - 1$. $\Xi\, x$ varies from -1 to 1. It is -1 when $x_a = x_b$, 0 when x_b is halfway between x_a and x_c, and 1 when $x_b = x_c$. These three regressor types apply to most of the investigated acoustical descriptors but not all, based on whether the association is appropriate or not, as indicated in Table 3.

[— Insert Table 3 about here. —]

Timbre descriptors.

Spectral descriptors. These descriptors assess properties associated with a time-averaged spectral representation. The investigated descriptors are computed on the output of one of two spectral-analysis methods: 1) analyses of the audio signals ($\tilde{}$) for individual instrument samples from VSL (e.g., oboe at G4) by use of the *Timbre Toolbox* (Peeters et al., 2011) employing harmonic analysis, and 2) pitch-generalized (°) spectral envelopes (Lembke & McAdams, 2015), which are estimated by fitting a curve to partial tones aggregated across all available pitches from VSL (e.g., oboe from B♭3 to G6) and therefore allow for the characterization of an instrument's pitch-generalized formant structure. Furthermore, the spectral descriptors can be distinguished as quantifying *global* and *local* spectral properties, as listed in Table 3. The global descriptors (S) include measures of spectral centroid (amplitude-weighted mean frequency), spectral slope (linear regression on the spectral envelope), spectral skewness (asymmetry in the spectral distribution), spectral kurtosis (peakier or flatter deviation from a Gaussian distribution), spectral spread (standard deviation of the spectral distribution), spectral roll-off (95th percentile of spectral-energy distribution), spectral decrease (spectral slope with low-frequency emphasis) and noisiness of the signal. They are described in detail in Peeters et al. (2011). The local, formant-related descriptors (F) require some elaboration.

[— Insert Figure 1 about here. —]

The formant structure derived from the pitch-generalized spectral envelope for a horn is shown in Figure 1. A set of frequencies indicate formant maxima (solid red lines; two formants are identified for this horn) and delineate their extent through lower and upper bounds (dotted lines) at which the magnitude has decreased by 3 dB. Note in the case of the horn in Figure 1 that there is no lower bound for the second formant, because the minimum point in the envelope between the two formants is not at least 3 dB below the maximum of the second formant. In this study, the focus lies on the main formant F_1, with F_{max} and F_{3dB} characterizing the frequency at its maximum magnitude and at the 3 dB upper bound, respectively (the latter appears to be more perceptually relevant; Lembke & McAdams, 2015). Two related measures, F_{slope} and $F_{\Delta mag}$, assess the relative importance of the main formant compared to the spectral-envelope regions lying above it. F_{slope} evaluates the (linear) spectral slope (grey diagonal) above the main formant in Figure 1, whereas $F_{\Delta mag}$ quantifies the level difference between the main-formant peak and the averaged magnitude of the spectral envelope above it (black arrow).

[— Insert Figure 2 about here. —]

Furthermore, the degree to which wind instruments are characterized by formant structure varies, being strongest for oboe but much weaker for clarinet and flute. The measure F_{prom} quantifies the prominence of up to two formants based on a cumulative score that increases with the number and width of formant features. As illustrated in Figure 2, the larger the total area covered by existing formant bounds (shaded rectangles), the higher the prominence of an instrument's formant structure, reflected in F_{prom} being considerably higher for oboe (blue) than for flute (red). Two additional difference measures, F_{freq} and F_{mag}, relate magnitude and frequency differences between the formants of constituent instruments. More specifically, F_{mag} quantifies the cumulative magnitude deviation between the constituent instruments' spectral envelopes at all formant frequencies they exhibit (vertical lines projected on the far right). F_{freq} evaluates the cumulative frequency difference (horizontal line projected at the top) between formants of the same order (e.g., main formant with main formant), if they exist for both sounds.

Temporal descriptors. Three descriptors characterize the time course of the amplitude envelope with respect to the attack (A) or onset portions of sounds, considering attack time and attack slope descriptors (Peeters et al., 2011).

Spectrotemporal descriptors. A pair of descriptors account for spectral variation across time, which the (static) spectral descriptors leave unaddressed. Previous research has not reported specific spectrotemporal (ST) descriptors as being relevant to blend, although temporal modulation of spectral components has been discussed in the context of blend (Reuter, 2009). In the interest of using a comprehensive set of timbre-related descriptors, two descriptors are included that involve the commonly reported *spectral flux* (ST_{flux}, Peeters et al., 2011) and the alternative measure *spectral incoherence* ($ST_{incoher}$), which quantifies the aggregate deviations of spectral magnitude between successive time frames (Horner, Beauchamp, & So, 2009).

Other descriptors and variables. The experimental designs involved factors that were likely to explain variance in median blend ratings but were not related to or not reliably measured through timbre features. Their relevance as potential regressors is assessed by several categorical variables (C), in addition to acoustical descriptors that could serve as equivalent predictors in application scenarios lacking a priori knowledge of categorical distinctions, e.g., by quantifying pitch relationships or the loudness balance between combined sounds. The categorical variables make binary or ternary distinctions and for the use with PLSR are expressed as *dummy variables* (Martens, Høy, Westad, Folkenberg, & Martens, 2001). A categorical variable is represented by as many dummy variables as there are categories, with each category's dummy variable set to 1 for cases matching the category and 0 if not. As a result, these regressors yield multiple loadings. For example, a binary categorical variable yields two loadings in opposing orientations, with "-D1" or "-D2" being appended to the variable name to symbolize the first and second categories of the variable, respectively (or -D0, -D1, and -D2 for a ternary variable).

For triads, a strong distinction was expected beforehand for the presence (D1) versus absence (D2) of *pizzicato* string sounds (C_{pizz}), as they are highly impulsive. Similarly, the distinction between unison (D1) and non-unison (D2) dyads was also expected to yield higher ratings for the former (C_{unison} and Δf_0). Additional regressors account for the lower (D1) or higher (D2) pitch level (C_{pitch}) and difference between pitches expressed in ERB units ($f_{0|ERB}$; Moore & Glasberg, 1983), interval type ($C_{interval}$; D0: unison, D1: instrument A low, instrument B high, D2: instrument B low, instrument A high), and instrument position at left (D1) or right (D2) in stereo space ($C_{position}$). In addition, the production of dyads and triads also involved determining relative mix or scaling ratios between the amplitudes of the constituent sounds forming dyads or triads, which are also quantified to assess their possible influence on the blend ratings (x_{mix}). For dyads, x_{mix} concerned a fractional gain value between 0 and 1 that applied to one instrument, while $x_{mix} -1$ scaled the other instrument. For triads, x_{mix} concerned the sound-level balance between the constituent sounds (e.g., negative slope for monotonously decreasing sound levels with ascending pitch).

Results

As mentioned under Method, PLSR analysis of a particular data set involves three stages, beginning with the original set of regressors X_{orig}, then restricting the selection to X_{Q50}, and finally attaining an approximately orthogonal selection of regressors X_{ortho}. Although statistics for all three stages are reported in Tables 4 and 5, only the results for the final stage X_{ortho} are presented in detail. In some of the following visualizations, data points representing dyads or triads use the labels in Tables I and II, respectively. A further distinction between dyads or triads containing instruments that blend more strongly and those that blend weakly is made with color to help assess how a given acoustical descriptor separates these instruments. For dyads, the instruments clarinet, bassoon, and horn lead to the highest blend ratings of comparable magnitude, whereas the trombone leads to the highest blend ratings for triads. Therefore, the horn and trombone were chosen to represent instruments that blend well with others (colored green) in the dyad and triad sets, respectively, as both brass instruments' spectral descriptions also resemble each other. Furthermore, the oboe was chosen as the exemplary instrument leading to poor blend (colored grey) in both sets.

Dyads

PLSR models predicting median blend ratings for dyads initially involved 46 regressors (X_{orig}). Elimination of loadings in P_{orig} that fall below the median threshold yielded 23 regressors in X_{Q50}. As listed in Table 4, a three-PC model explains 93% of the variance for X_{Q50}. Refining the regressors to an approximately orthogonal set, the resulting X_{ortho} consists of 14 regressors, again, leading to a three-PC model explaining 93% of the variance. The model fit in y for X_{ortho}, displayed in Figure 3, shows the variation in median blend ratings to be represented well. However, the blend ratings (x-axis) exhibit two distinct groups of data points, corresponding to unison dyads (circles) leading to substantially greater blend than non-unison dyads (diamonds). Furthermore, non-unison dyads involving horn (green) yielded slightly greater blend overall than those with oboe (grey), whereas no such distinction is observable for unison dyads.

[— Insert Table 4 about here. —]

[— Insert Figure 3 about here. —]

Figure 4 visualizes the loadings P_{ortho} (vectors) and the scores T_{ortho} (symbols) across the first two PCs. Larger symbols for scores correspond to higher blend ratings. Likewise, longer vectors represent loadings that contribute more strongly, while the vector orientation illustrates which PCs the contribution primarily affects.

[— Insert Figure 4 about here. —]

Reflecting the main distinctions in median blend ratings, the scores T_{ortho} also form two distinct groups for unison and non-unison dyads, with the corresponding categorical variable C_{unison} describing this distinction most accurately along PC 1. The acoustical descriptor Δf_0 predicts the same distinction comparably well. PC 2 appears to be influenced by two factors: 1) an additional grouping of dyads based on low and high pitch levels, described by the categorical variable C_{pitch} and the acoustical descriptor $f_{0|ERB}$, and 2) a collinear set of spectral descriptors, falling slightly oblique to the PC axis. The distinction across interval types (horizontal) and pitch levels (vertical) yields four subgroups. Along each of these obliquely aligned groups the influence of spectral features appears to lead to similar dyad constellations.

Figure 5 suggests that the spectral and pitch influence is independent (orthogonal) on the plane spanning PCs 2 and 3. The spectral regressors involve several composite (Σ) as well as difference (Δ) measures for S°_{centr} and formant-related descriptors. With regard to the resulting scores, T_{ortho} yields a grouping of dyads into those containing either horn or oboe (green/low-left vs. grey/top-right), for both unison and non-unison dyads.

[— Insert Figure 5 about here. —]

Overall, the dyad data exhibit a complex structure of underlying factors, involving interval type, pitch level, and spectral features. Across all investigated models, their performance (R^2) is remarkably well matched by their predictive power (Q^2). Given the relatively large number of cases, $N = 180$, further PLSR analyses on subsets separated by interval type are conducted, yielding $N = 60$ for unison and $N = 120$ for non-unison dyads. Separate analyses allow an assessment of whether certain spectral and pitch trends are specific to only one of the interval types.

Unison. A three-PC model on X_{Q50} involving 22 regressors leads to 46% explained variance in median blend ratings for unison dyads, exhibiting a substantially lower predictive power of only 17% explained variance. Due to a fairly wide variation in P_{Q50} orientations, the angular threshold ϕ_i determining X_{ortho} had to be increased to $|\phi_i| < 30°$ to ensure that the reduction to an approximately orthogonal set would lead to a meaningful number of contributing regressors. The resulting model with nine regressors yields a two-PC model explaining 27% of the variance, which appears a more realistic estimate of the true predictive relationship between median blend ratings and X_{ortho}, as the discrepancy between model performance and the predictive power is substantially reduced.

As shown in Figure 6, the y_{unison} fit appears a closer fit to the diagonal than for the complete dyad data (Figure 3), but the blend ratings only span a relatively narrow scale range. This may result from a reduction in the perceptual resolution among the unison dyads due to the dominant distinction between unison and non-unison dyads. The reduced resolution also makes it more likely for the variation in median blend ratings to contain increased noise levels, supported by the large discrepancy between R^2 and Q^2 in the initial models.

[— Insert Figure 6 about here. —]

PC 1 explains 22% of the variance and, as shown in Figure 7, appears to be linked to spectral composite (Σ) descriptors for main formant location (e.g., F_{max}, F_{3dB}) as well as centroid (e.g., S°_{centr}), which also distinguishes low register and high register instrument dyads (e.g., HB vs. OF). PC 2 accounts for another 5% of the variance, involving a distinction between instrument dyads with similar formant structure and those with divergent structures (e.g., HB vs. BF and HF), explained by the formant-related descriptors ΔF_{slope} and ΔF_{freq}.

[— Insert Figure 7 about here. —]

Non-unison. Twenty-three regressors in X_{Q50} yield a three-PC model explaining 55% of the variance in median blend ratings for non-unison dyads, with the predictive power corresponding to 47% of the variance explained. The reduction to X_{ortho} yields 11 regressors and a three-PC model explaining 48% of the variance, with a lower predictive power accounting for 35% of the variance. The model fit in $y_{non-unison}$ for X_{ortho}, shown in Figure 8, improved compared to the one for the complete dyad set (Figure 3), showing a better approximation to the ideal fit (dashed line).

[— Insert Figure 8 about here. —]

As shown in Figure 9, PC 1 clearly reflects a grouping of dyads based on pitch level (C_{pitch} and $f_{0|ERB}$), accounting for 33% of the explained variance. At the same time, the composite of the spectral slope $S^{\sim}_{slope}$ appears to covary with pitch change. All remaining spectral regressors appear relatively independent (orthogonal) to the pitch influence. Figure 10 illustrates that across the plane spanning PCs 2 and 3, two seemingly independent contributions of spectral regressors occur: 1) an implied triangle between the composite (Σ) regressors F_{3dB}, S°_{centr}, and the difference (Δ) descriptor F_{3dB} distinguishes dyads into those containing horn (bottom-left) and those involving oboe (top-right); 2) perpendicular to this orientation, difference in spectral slope $S^{\sim}_{slope}$ and composite in noisiness S_{noise} contribute somewhat more weakly. Together, PCs 2 and 3 account for 8% and 7% of the variance, respectively.

[— Insert Figure 9 about here. —]

[— Insert Figure 10 about here. —]

Triads

The PLSR analysis of triads first involved 61 regressors, which reduced to 30 regressors in X_{Q50}, leading to a two-PC model explaining 89% of the variance in median blend ratings and with a predictive power explaining 73% of the variance. The subsequent reduction to X_{ortho} yields another two-PC model with 15 regressors that again explains 89% of the variance, notably, gaining in predictive power compared to the previous models. As shown in Figure 11, the model fit for y appears satisfactory, given the smaller number of cases for triads ($n = 20$). Still, a compact cluster involving *pizzicato* cello (squares, bottom-left) stands in contrast to more spread out ratings for sounds lacking them (circles, right half). A trend for triads involving trombone (green) to be the most blended is apparent in each subgroup.

[— Insert Table 5 about here. —]

[— Insert Figure 11 about here. —]

The main distinction found in Figure 12 along PC 1, which accounts for 85% of the variance, concerns the presence or absence of *pizzicato* cello sounds (the categorical variable C_{pizz}), with the acoustical difference in attack slopes A_{slope} performing similarly well. Apart from A_{slope}, the composite and difference descriptors for spectrotemporal incoherence $ST_{incoher}$ and noisiness S_{noise} are somewhat correlated with C_{pizz}. This could result from both the transient attack and rapid decay of the temporal envelope of *pizzicato* sounds contributing to more noise and more spectral change over time, respectively. In addition, the inclusion of two other spectral descriptors, difference in $F_{\Delta mag}$ and distribution in S_{skew}, could be explained by differences in spectral-envelope shape for the two articulations of the cello.

[— Insert Figure 12 about here. —]

PC 2 explains the remaining 3% of the variance, appearing to relate to the distribution (Ξ) of a number of formant descriptors. These comprise frequency measures of the main formant, F_{3dB} and F_{max}, measures of balance between main formants and the remaining spectral envelope, F_{slope} and $F_{\Delta mag}$, and the overall prominence of formant structure, F_{prom}. Furthermore, the relevance of two difference measures related to formants, F_{mag} and F_{prom}, suggests that the most pronounced differences among three descriptor values could also be of importance.

Discussion

Previous research has associated blend with acoustical measures describing spectral features, as well as temporal features like the attacks or onsets of sounds under certain circumstances. The current investigation pursued a correlational analysis using

PLSR, modeling two perceptual data sets involving dyads and triads. PLSR loadings allowed us to evaluate the extent to which regressors were collinear or independent of each other. This approach helped select the most effective regressors. Applied to the complete data sets for both dyads and triads, the final models based on optimized regressor sets explain around 90% of the variance in median blend ratings. Notably, these levels of explained variance were still achieved after the elimination of non-essential regressors, i.e., more than two thirds from the original set. The variation in both data sets is best explained by a dominant factor that is unrelated to spectral features.

For dyads, the distinction between unison and non-unison intervals explains 91% of the variance, with the fundamental-frequency difference Δf_0 representing a reliable acoustical predictor. That unison dyads would lead to higher blend than for non-unison had been anticipated, given that similar effects have been found in other studies (Kendall & Carterette, 1993; Lembke et al., in press). The pronounced difference obtained in the current results, however, seems to exceed those previously reported, which could be related to the current study being the only one in which unison and non-unison were presented in a common stimulus set, whereas in other studies both interval types had been grouped into separate experimental blocks (Kendall & Carterette, 1993; Lembke et al., in press) or had even been tested in separate experiments (Sandell, 1995).

In addition, even the second-most important factor in explaining the variation among dyads, $f_{0|ERB}$, is unrelated to spectral features, as it reflects differences in pitch height, accounting for 2% in all dyads and 33% when considering only the non-unison dyads. The fact that the contribution of the pitch level is limited to the non-unison dyads implies that it may not affect blend of unison dyads. For non-unison dyads, it is also worth noting that inverting the assignment of instruments to the two pitches had no effect on blend ratings. This negative finding goes counter to many claims in orchestration treatises that the order of pitch assignment affects blend. It thus supports the conclusion by Sandell (1995) that timbral inversion does not appear to influence blend; only a single finding argues in its favor (Kendall & Carterette, 1993).

With regard to triads, the presence of a *pizzicato* cello evoked a strong decrease in blend ratings, whereas even triads including cello sounds excited by a single, brisk bow stroke led to substantially more blend. Again, this distinction had been anticipated, given that increasingly impulsive sounds have been associated with comparable decreases in blend (Tardieu & McAdams, 2012). Regarding the description of onset articulations, the difference in attack slopes A_{slope} is strongly collinear with the categorical distinction C_{pizz}, explaining about 85% of the variance; additional collinearity with spectrotemporal or noise features can be assumed to co-occur as byproducts of the abrupt changes in temporal envelopes.

With both data sets being dominantly influenced by pitch or temporal features (e.g., attack), spectral descriptors only occur as secondary or even tertiary sources of variation in the modeled blend ratings. In perceptual tasks comparable to those employed in these experiments, participants may focus their attention on the dominant distinctions across stimuli at the cost of perceptual resolution for the less pronounced differences.

As the spectral factors likely only affected blend ratings in these regions of reduced perceptual resolution, the possible role of behavioral noise needs to be considered. Indeed, clear discrepancies between model performance R^2 and predictive power Q^2 indicate that the initial PLSR models could have been overfitting to noise artifacts instead of systematic factors of variation. For example, stripped of the dominant factor, the unison and non-unison subsets of data account for no more than 50% of the variance (R^2). The unison-dyad data suggest that the true performance is substantially lower as the predictive power is generally quite low and likely reflects random variation or factors not captured by the tested regressors. In summary, the identified tendencies for spectral regressors can be assumed to be valid for the obtained proportions of explained variance, but they should be considered preliminary until confirmed in additional datasets yielding greater resolution in the perceptual ratings.

Three spectral descriptors stand out in explaining the PLSR models for both data sets, namely, the centroid of the pitch-generalized spectral envelope S°_{centr} and the two main-formant descriptors F_{max} and F_{3dB}, notably representing spectral features that have previously been found to be relevant (Lembke et al., in press; Lembke & McAdams, 2015; Reuter, 1996; Sandell, 1995; Tardieu & McAdams, 2012). Differences exist concerning the types of association between descriptor values of the instruments constituting dyads or triads. For unison dyads, the composite (Σ) measures for all three descriptors became relevant in explaining 22% of the variance, which is in agreement with the same association explaining other perceptual results for unison dyads (Sandell, 1995; Tardieu & McAdams, 2012).

Non-unison dyads yield a more complex relationship and involved the composite for S°_{centr} and F_{3dB} complemented by the difference in F_{3dB}, overall contributing 15% of the variance. The relevance of the difference measure (Δ) is in agreement with the absolute spectral-centroid difference having previously been reported as the strongest predictor for non-unison dyads (Sandell, 1995). The particular combination of composite and difference measures suggests that as S°_{centr} and F_{3dB} increased, so did the divergence of F_{3dB} between the individual instruments, with both possibly contributing to decreased blend. For instance, oboe paired with horn yields a higher composite centroid due to the oboe's higher main formant, which at the same time increases the frequency distance to the horn's low main formant, whereas for horn and bassoon, both main formants are relatively low and, moreover, practically coincide in frequency.

The results for triads expand previous knowledge beyond dyadic contexts. Even if spectral features only account for 3% of the variance, some new insight is gained from the *distribution* (Ξ) of three descriptor values for several formant measures serving as the strongest predictor, suggesting that relative position of the sound having an intermediate descriptor value among all three sounds may indeed be useful in describing instrument combinations with more than two instruments.

Overall, the global descriptor S°_{centr} and the main-formant location F_{3dB} indicate that prominent spectral-envelope properties represent reliable correlates to blend across various instruments, pitches, and polyphonic combinations. Being the first investigation to test for the relevance of global and local spectral descriptors jointly, both domains seem helpful as regressors in a predictive application. Across all datasets, the descriptor loadings P confirmed that most spectral descriptors were partially correlated, at the same time, allowing the identification of descriptors that appeared independent of S°_{centr} and F_{3dB}, namely, the spectral slope, $S^{\sim}_{slope}$, and noisiness, S_{noise} (Figure 10), as well as the formant-based spectral slope, F_{slope}, and formant frequency difference between constituent sounds, F_{freq} (Figure 7). These additional descriptors could be of special interest in achieving more complete prediction models, although their relevance seems to depend on the stimulus context. A similar analysis approach on a wider data set is needed to confirm these trends, and possibly even to give further insight into the role of associations (Σ, Δ, Ξ) relevant for different musical scenarios. Furthermore, the apparent utility of pitch-generalized descriptors, i.e., all F descriptors and S°_{centr} as opposed to $S^{\sim}_{centr}$, implies that a case-by-case signal analysis on individual pitches may not be necessary, but instead, a prediction application could rely on a comprehensive, offline database of pitch-generalized instrument descriptions alone, which would significantly facilitate computation.

When considering the relative locations of instrument combinations along the PCs that correlate with spectral features, a recurring pattern of dyads or triads including oboe (grey), on the one side, opposed to combinations involving horn or trombone (green), on the other, becomes apparent. Dyads or triads containing oboe are often less blended, whereas combinations with horn or trombone (e.g., bassoon and horn, clarinet and horn, trombone and trombone) are among the most blended ones. If we consider the notion of *blendability* of a particular instrument, the oboe should be considered a poor 'blender', which can be explained spectrally by its prominent and unique formant structure. Similar observations linking oboe to poor blend have been made in previous perceptual investigations (Kendall & Carterette, 1993; Reuter, 1996; Sandell, 1995; Tardieu & McAdams, 2012) as well as 'prescriptions' found in orchestration treatises (Koechlin, 1954; Reuter, 2002). On the other hand, the horn is generally considered an easily blendable instrument, again reflected in perceptual results

(Reuter, 1996; Sandell, 1995). The relatively 'dark' timbre of the horn could support a general hypothesis of lower centroids leading to more blend (Sandell, 1995), at the same time supporting the argument that similar main-formant locations explain the good blend obtained between horn and bassoon (Lembke et al., in press; Reuter, 1996).

In addition, Figure 12 illustrates that the distribution (Ξ) along formant descriptors (F) distinguishes triads with two identical instruments (e.g., two trombones plus clarinet, two *pizzicati* or two *arco* celli plus clarinet) from more diverse combinations, without, however, directly reflecting how these combinations vary in blend (visualized size of the symbols for scores predicted by the models). Nevertheless, it does imply that timbral similarity, if not identity, aids blending. In summary, once factors related to pitch intervals or onset articulation are taken into account, spectral features do seem to represent the main underlying factor governing whether instrument combinations blend or not, with pitch-generalized spectral descriptions possibly conveying the timbral signature traits of instruments.

Conclusion

The present investigation shows that the perception of blended timbres in dyadic and triadic contexts correlates with a number of acoustical factors. Analyses using PLSR converged on an apparently reliable selection of independent predictors. The importance of factors such as pitch interval type, pitch, and articulation (e.g., impulsive vs. gradual note attack) became apparent. In addition, a group of spectral descriptors that exhibit the strongest predictive abilities could be identified from a wide range of descriptors, namely, the global spectral centroid and the upper frequency bound of main formants, which may represent the relevant features informing instrumentation choices. This wide variety of predictors suggests that in blend-prediction applications aimed at realistic musical scenarios, all factors should be taken into account. Given an appropriate acoustical characterization of instruments and details of how they are combined and employed musically (e.g., in unison or non-unison, the articulation and dynamic markings), these properties could suffice to predict the associated degree of blend.

One main challenge for future research is determining the effective weighting between these different factors of influence. Whether the clear dominance of interval type or impulsiveness of attacks over spectral features, which became apparent in the current investigation, would extend to more complex musical contexts remains to be explored. It can be assumed that the growing complexity that a listening scenario involving musical contexts would present, given the simultaneous presence of other musical parameters, could significantly alter the relative importance of factors found in listening experiments employing isolated dyadic or triadic stimuli.

For instance, a composer may assign a unison blend between two instruments to a melodic voice while juxtaposing this against a chordal, non-unison accompaniment layer whose instruments are chosen to blend amongst themselves into a homogeneous timbre. On another level, the melody may become more distinct from the accompaniment due to the distinction between unison and non-unison, which may also be desired. This case scenario illustrates that blend-related factors need not stand in competition with each other like they do in the investigated perceptual data, but instead could operate on independent levels, fulfilling separate functions within the musical context.

For the composer, working with blend is not a matter of favoring unison intervals over non-unison intervals, but being able to employ it at individual levels of the musical scene (e.g., melody, accompaniment, or contrasting the two). Within each level, blend is achieved by relying on the same principles, i.e., similarity in spectral description as well as articulatory features (e.g., note attacks). This hypothetical scenario encourages future work on blend-prediction models to rely on perceptual data obtained from stimuli involving musical contexts (Kendall & Carterette, 1993; Lembke et al., in press; Reuter, 1996), as it provides a more realistic setting from which weights between blend-related factors could be estimated. We thus propose the need for a meta-analytical investigation into a diverse range of perceptual blend data, in an attempt to move toward generally applicable blend-prediction techniques.

Footnotes

[1]MATLAB, Mathworks. The *plsregress* function from the Statistics and Machine Learning Toolbox was used. URL: https://uk.mathworks.com/products/matlab.html. Last accessed: August 21, 2017.

[2]URL: http://vsl.co.at/. Last accessed: August 21, 2017.

[3]See Appendix C in Lembke (2015) for details.

Supplemental Material

Representative examples for the dyad and triad stimuli are available online as supplemental material, which can be found as part of the online version of this article at http://msx.sagepub.com. Access the sound files through the hyperlink “Supplemental material”. Their filenames follow the naming convention found in Tables 1 and 2.

Author Note

The authors would like to thank Bennett K. Smith for programming the control software for both experiments and Emma Kast for the recruitment and running of participants for the experiment involving triads. We also thank two anonymous reviewers and the editor for their valuable feedback on an earlier version of this article. The findings reported in this article were conducted as part of the first author’s doctoral research and are also featured in his thesis (Lembke, 2015).

Correspondence concerning this article should be directed to Sven-Amin Lembke, currently at De Montfort University, Clephan Building, Room 00.07b, Leicester LE1 9BH, United Kingdom; email: sven-amin.lembke@dmu.ac.uk.

Funding

This research was partly funded by an ACN CREATE undergraduate research award to Kyra Parker as well as a Canadian National Sciences and Engineering

Research Council grant (RGPIN 312774-2010) and a Canada Research Chair to Stephen McAdams.

References

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251–263.

Eerola, T., Lartillot, O., & Toiviainen, P. (2009, October). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proc. 10th International Society of Music Information Retrieval (ISMIR) Conference* (pp. 621–626). Kobe, Japan.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, *185*, 1–17.

Giordano, B. L., Rocchesso, D., & McAdams, S. (2010). Integration of acoustical information in the perception of impacted sound sources: The role of information accuracy and exploitability. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(2), 462–476.

Horner, A. B., Beauchamp, J. W., & So, R. H. Y. (2009). Detection of time-varying harmonic amplitude alterations due to spectral interpolations between musical instrument tones. *Journal of the Acoustical Society of America*, *125*(1), 492–502.

ISO 389–8. (2004). *Acoustics: Reference zero for the calibration of audiometric equipment—Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones* (Tech. Rep.). Geneva, Switzerland: International Organization for Standardization.

Kendall, R. A. (2004, August). Musical timbre in triadic contexts. In *Proc. 8th International Conference on Music Perception and Cognition (ICMPC)* (pp. 600–602). Evanston, Illinois.

Kendall, R. A., & Carterette, E. C. (1993). Identification and blend of timbres as a basis for orchestration. *Contemporary Music Review*, *9*(1), 51–67.

Kendall, R. A., & Vassilakis, P. (2006). Perceptual acoustics of consonance and disssonance in multitimbral triads. *Journal of the Acoustical Society of America*, *120*(5), 3276–3276.

Kendall, R. A., & Vassilakis, P. N. (2010). Perception and acoustical analyzes of traditionally orchestrated musical structures versus non-traditional counterparts. *Journal of the Acoustical Society of America*, *128*(4), 2344–2344.

Koechlin, C. (1954). *Traité de l'orchestration: En quatre volumes (Treatise of orchestration: In four volumes)*. Paris: M. Eschig.

Kumar, S., Forster, H. M., Bailey, P., & Griffiths, T. D. (2008). Mapping unpleasantness of sounds to their auditory representation. *Journal of the Acoustical Society of America*, *124*(6), 3810–3817.

Lembke, S.-A. (2015). *When timbre blends musically: perception and acoustics underlying orchestration and performance* (PhD thesis). McGill University.

Lembke, S.-A., Levine, S., & McAdams, S. (in press). Blending between bassoon and horn players: An analysis of timbral adjustments during musical performance. *Music Perception*, *35*(2).

Lembke, S.-A., & McAdams, S. (2015). The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds. *Acta Acustica united with Acustica*, *101*(5), 1039–1051.

Luce, D., & Clark, J. (1967). Physical correlates of brass-instrument tones. *Journal of the Acoustical Society of America*, *42*(6), 1232–1243.

Luce, D. A. (1975). Dynamic spectrum changes of orchestral instruments. *Journal of the Audio Engineering Society*, *23*(7), 565–568.

Martens, H., Høy, M., Westad, F., Folkenberg, D., & Martens, M. (2001). Analysis of designed experiments by stabilised PLS Regression and jack-knifing. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 151–170.

Martin, F., & Champlin, C. (2000). Reconsidering the limits of normal hearing. *Journal of the American Academy of Audiology*, *11*(2), 64–66.

Moore, B. C., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, *74*(3), 750–753.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The

Timbre Toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, *130*(5), 2902–2916.

Reuter, C. (1996). *Die auditive Diskrimination von Orchesterinstrumenten - Verschmelzung und Heraushörbarkeit von Instrumentalklangfarben im Ensemblespiel (The auditory discrimination of orchestral instruments: Fusion and distinguishability of instrumental timbres in ensemble playing).* Frankfurt am Main, Germany: P. Lang.

Reuter, C. (2002). *Klangfarbe und Instrumentation: Geschichte–Ursachen–Wirkung (Timbre and instrumentation: History–causes–effect).* Frankfurt am Main, Germany: P. Lang.

Reuter, C. (2009). The role of formant positions and micro-modulations in blending and partial masking of musical instruments. *Journal of the Acoustical Society of America*, *126*(4), 2237–2237.

Rumsey, F., Zieliński, S., Kassier, R., & Bech, S. (2005). On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *Journal of the Acoustical Society of America*, *118*(2), 968–976.

Sandell, G. J. (1995). Roles for spectral centroid and other factors in determining “blended" instrument pairings in orchestration. *Music Perception*, *13*(2), 209–246.

Schumann, K. E. (1929). *Physik der Klangfarben (Physics of timbres)* (professorial dissertation). Universität Berlin.

Tardieu, D., & McAdams, S. (2012). Perception of dyads of impulsive and sustained instrument sounds. *Music Perception*, *30*(2), 117–128.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130.

Table 1
Fifteen dyads across pairs of the six investigated instruments.

Dyad	Instrument pair	
HB	horn	bassoon
HO	horn	oboe
HT	horn	trumpet
HC	horn	clarinet
HF	horn	flute
BO	bassoon	oboe
BT	bassoon	trumpet
BC	bassoon	clarinet
BF	bassoon	flute
OT	oboe	trumpet
OC	oboe	clarinet
OF	oboe	flute
TC	trumpet	clarinet
TF	trumpet	flute
CF	clarinet	flute

Table 2
Twenty triads and their constituent instruments and assigned pitches.

	Instruments & pitches		
Triad	C4	F4	B♭4
AAF	cello (*arco*)	cello (*arco*)	flute
AAC	cello (*arco*)	cello (*arco*)	clarinet
PPC	cello (*pizz.*)	cello (*pizz.*)	clarinet
PPO	cello (*pizz.*)	cello (*pizz.*)	oboe
PAF	cello (*pizz.*)	cello (*arco*)	flute
PAO	cello (*pizz.*)	cello (*arco*)	oboe
ACF	cello (*arco*)	clarinet	flute
AOF	cello (*arco*)	oboe	flute
ACO	cello (*arco*)	clarinet	oboe
PCO	cello (*pizz.*)	clarinet	oboe
TTF	trombone	trombone	flute
TTC	trombone	trombone	clarinet
TTO	trombone	trombone	oboe
TCO	trombone	clarinet	oboe
PTT	cello (*pizz.*)	trombone	trombone
PAT	cello (*pizz.*)	cello (*arco*)	trombone
ATF	cello (*arco*)	trombone	flute
ATC	cello (*arco*)	trombone	clarinet
PTC	cello (*pizz.*)	trombone	clarinet
PTO	cello (*pizz.*)	trombone	oboe

Table 3

Acoustical descriptors investigated for dyads and/or triads (marked by 'x' in the rightmost columns), related to the global spectrum (S), formants (F), the attack portion of the temporal envelope (A), spectrotemporal variation (ST), as well as categorical variables (C). Descriptor values for individual sounds forming dyads or triads were associated with a single regressor value by difference Δ, composite Σ, distribution Ξ (triads only) or as specified otherwise.

Abbrev.	Description	Unit	Association	Dyad	Triad
$S^{\sim}_{centr}$	spectral centroid[a]	Hz	Δ, Σ, Ξ	x	x
S°_{centr}	spectral centroid[b]	Hz	Δ, Σ, Ξ	x	x
$S^{\sim}_{slope}$	spectral slope[a]	Hz^{-1}	Δ, Σ, Ξ	x	x
S°_{slope}	spectral slope[b]	Hz^{-1}	Δ, Σ, Ξ	x	x
S_{skew}	spectral skew[a]	-	Δ, Σ, Ξ	x	x
S_{kurtos}	spectral kurtosis[a]	-	Δ, Σ, Ξ	x	x
S_{spread}	spectral spread[a]	Hz	Δ, Σ, Ξ	x	x
S_{roll}	spectral roll-off[a]	Hz	Δ, Σ, Ξ	x	x
$S_{decrease}$	spectral decrease[a]	-	Δ, Σ, Ξ	x	x
S_{noise}	noisiness[a]	-	Δ, Σ, Ξ	x	x
F_{max}	main-formant maximum[b]	Hz	Δ, Σ, Ξ	x	x
F_{3dB}	main-formant upper bound[b]	Hz	Δ, Σ, Ξ	x	x
F_{slope}	spectral slope above main formant[b]	Hz^{-1}	Δ, Σ, Ξ	x	x
$F_{\Delta mag}$	level difference F_1 vs. above[b]	dB	Δ, Σ, Ξ	x	x
F_{prom}	formant prominence[b]	-	Δ, Σ, Ξ	x	x
F_{freq}	formant frequency deviations[b]	Hz	Δ	x	x
F_{mag}	formant magnitude deviations[b]	dB	Δ	x	x
A_{time}	attack time	s	Δ, Ξ	x	x
$A_{log(time)}$	log. attack time	s	Δ, Ξ	x	x
A_{slope}	attack slope	s^{-1}	Δ, Ξ	x	x
ST_{flux}	spectral flux[a]	-	Δ, Σ, Ξ	x	x
$ST_{incoher}$	spectral incoherence[a]	-	Δ, Σ, Ξ	x	x
C_{unison}	unison or non-unison	-	binary	x	
Δf_0	f_0 difference	Hz	Δ	x	
C_{pitch}	pitch level	-	binary	x	
$f_{0\|ERB}$	f_0, auditory scaling	ERB[c] rate	C4 or G4	x	
$C_{interval}$	interval type	-	ternary	x	
$C_{position}$	instrument positions	-	binary	x	
C_{pizz}	including *pizzicato* or not	-	binary		x
x_{mix}	amplitude mix or balance	-	scaled value	x	x

Note.

[a] $S^{\sim}$ based on signal analysis for individual pitches.
[b] S° based on pitch-generalized spectral-envelope estimate.
[c] ERB: equivalent rectangular bandwidth (Moore & Glasberg, 1983).

Table 4
Performance (R^2) and predictive power (Q^2) of the PLSR model for dyads as well as component-wise contribution along the first three PCs for the three stages X_{orig}, X_{Q50}, X_{ortho} involving a sequential reduction of the number of regressors m.

y dyads	X regressors	m	R^2	Q^2	PC 1	PC 2	PC 3
all	X_{orig}	46	.94	.91	.88	.04	.01
	X_{Q50}	23	.93	.92	.90	.03	$<.01$
	X_{ortho}	14	.93	.93	.91	.02	$<.01$
unison	X_{orig}	44	.56	.18	.33	.14	.10
	X_{Q50}	22	.46	.17	.26	.12	.09
	X_{ortho}	9	.27	.16	.22	.05	-
non-unison	X_{orig}	45	.60	.40	.42	.10	.08
	X_{Q50}	23	.55	.47	.39	.14	.03
	X_{ortho}	11	.48	.35	.33	.08	.07

Table 5

Performance (R^2) and predictive power (Q^2) of the PLSR model for triads as well as component-wise contribution along up to three PCs. Three stages X_{orig}, X_{Q50}, X_{ortho} involve a sequential reduction of the number of regressors m.

y triads	X regressors	m	R^2	Q^2	PC 1	PC 2	PC 3
all	X_{orig}	61	.90	.64	.86	.04	-
	X_{Q50}	30	.89	.73	.84	.05	-
	X_{ortho}	15	.89	.76	.85	.03	-

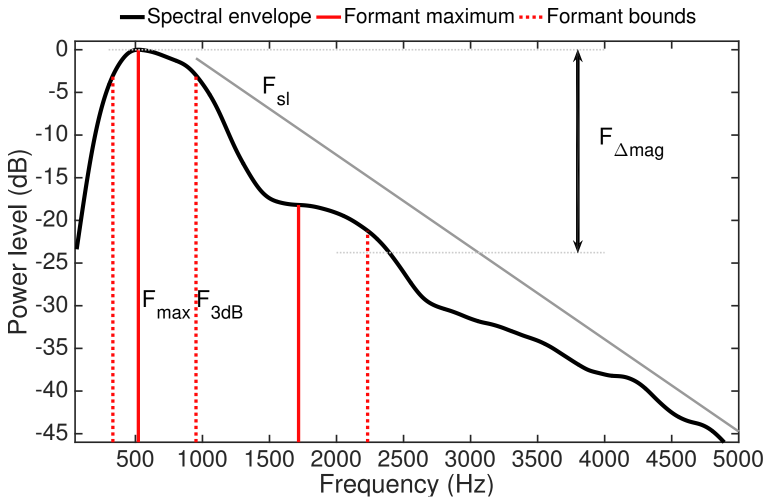

Figure 1. Pitch-generalized spectral envelope of a horn with identified frequencies for formant maxima and 3 dB bounds (see Lembke & McAdams, 2015). F_{max} and F_{3dB} characterize the maximum and upper bound, respectively, for the dominant, lower *main* formant. F_{slope} represents the spectral slope above the main formant. $F_{\Delta mag}$ quantifies the magnitude difference between F_{max} and and the average magnitude above F_{3dB}.

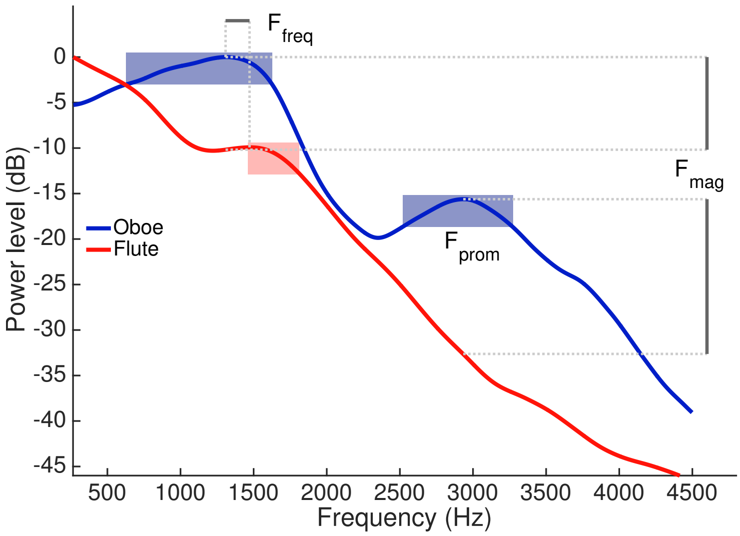

Figure 2. Pitch-generalized spectral envelopes of oboe (blue) and flute (red). F_{prom} characterizes the existence and clarity of formant features (e.g., 3 dB formant bounds); the larger the total shaded area, the more prominent the instrument's formant structure. F_{mag} evaluates the total magnitude difference between spectral envelopes at formant frequencies. F_{freq} quantifies the deviation between formant frequencies.

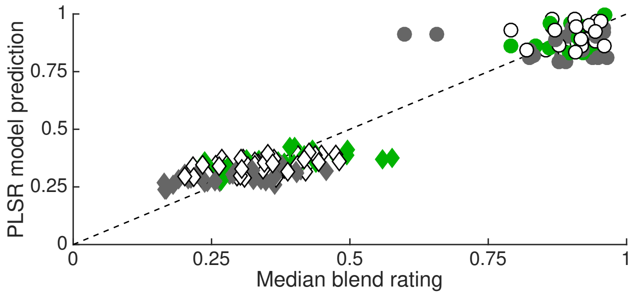

Figure 3. Dyad model fit of y variables for X_{ortho}. Legend: circles, unison; diamonds, non-unison; grey involves oboe; green involves horn (excl. HO).

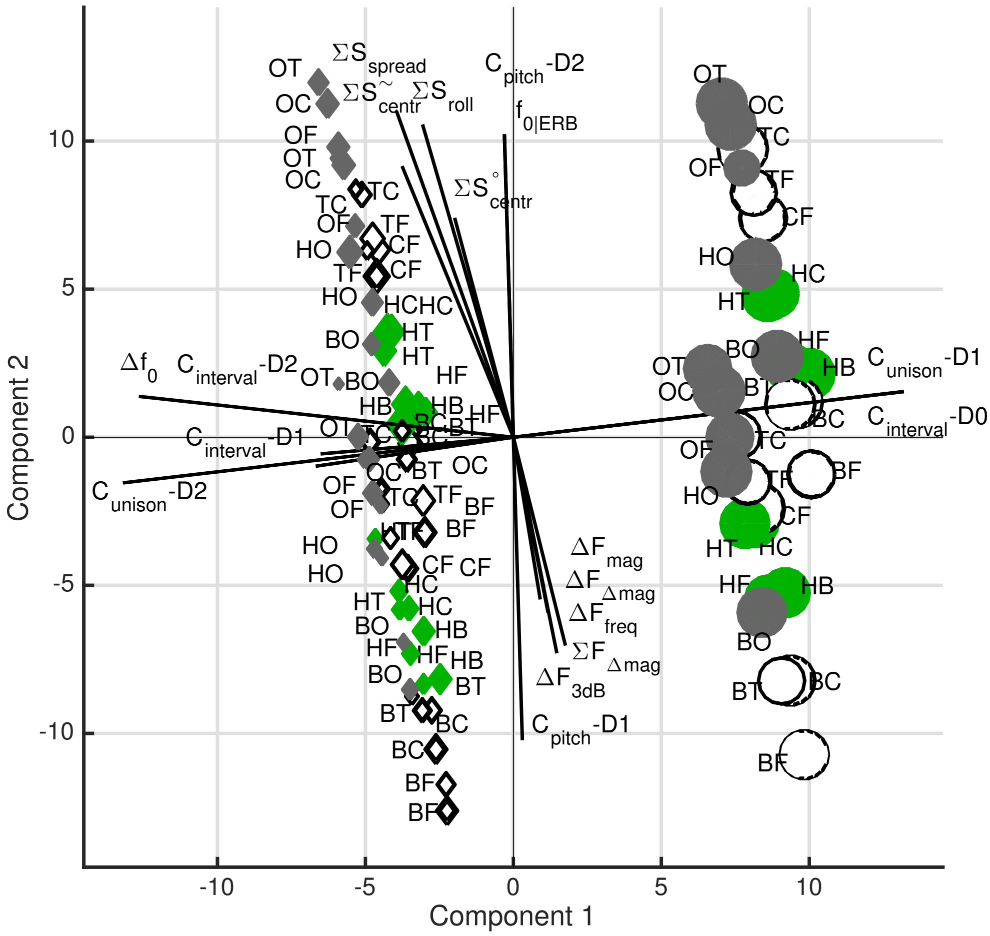

Figure 4. PLSR loadings P_{ortho} (vectors) and scores T_{ortho} (symbols) for PCs 1 and 2 with dyads. Legend: circles, unison; diamonds, non-unison; their size represents the relative degree of blend; dark grey involves oboe; green involves horn (excl. HO).

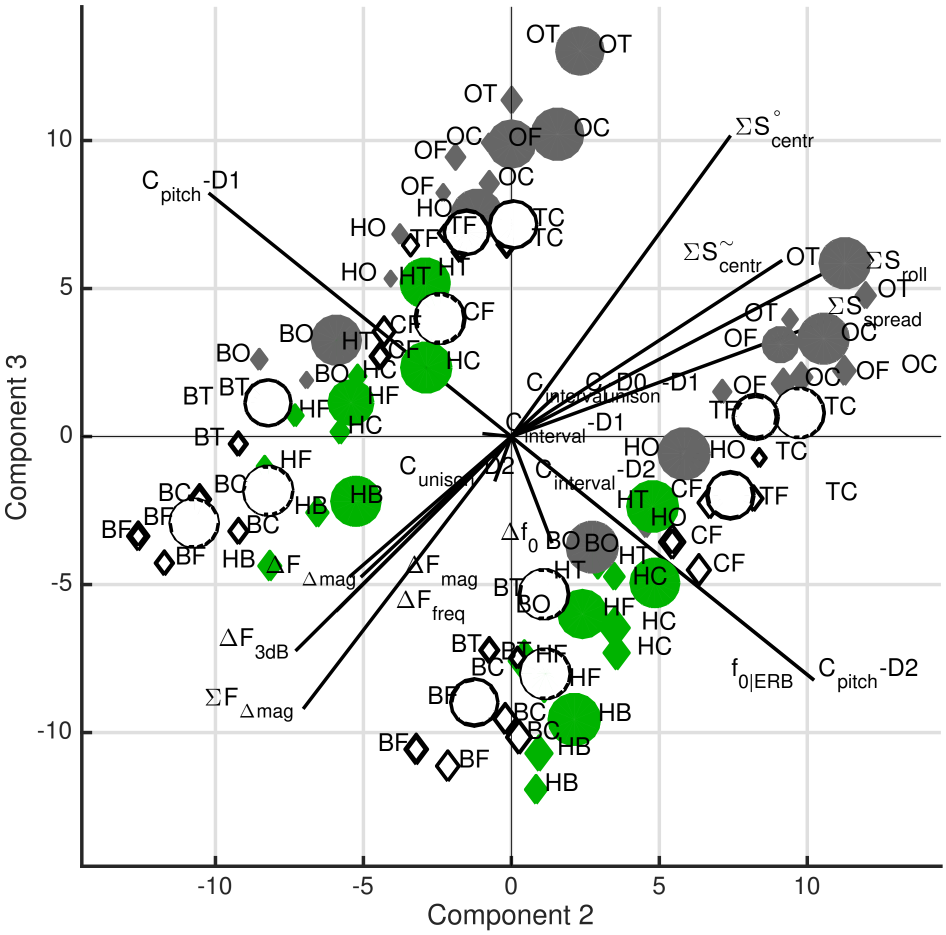

Figure 5. Dyad P_{ortho} and T_{ortho} for PCs 2 and 3. See Figure 4 for legend.

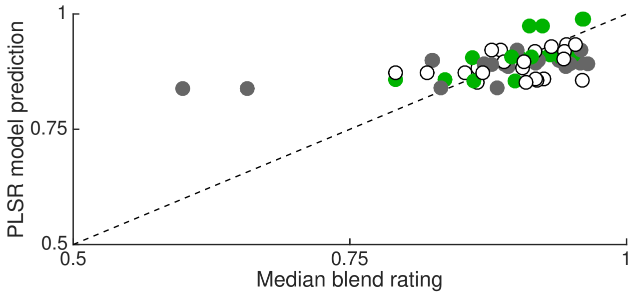

Figure 6. Unison-dyad model fit of y variables for X_{ortho}. See Figure 3 for legend.

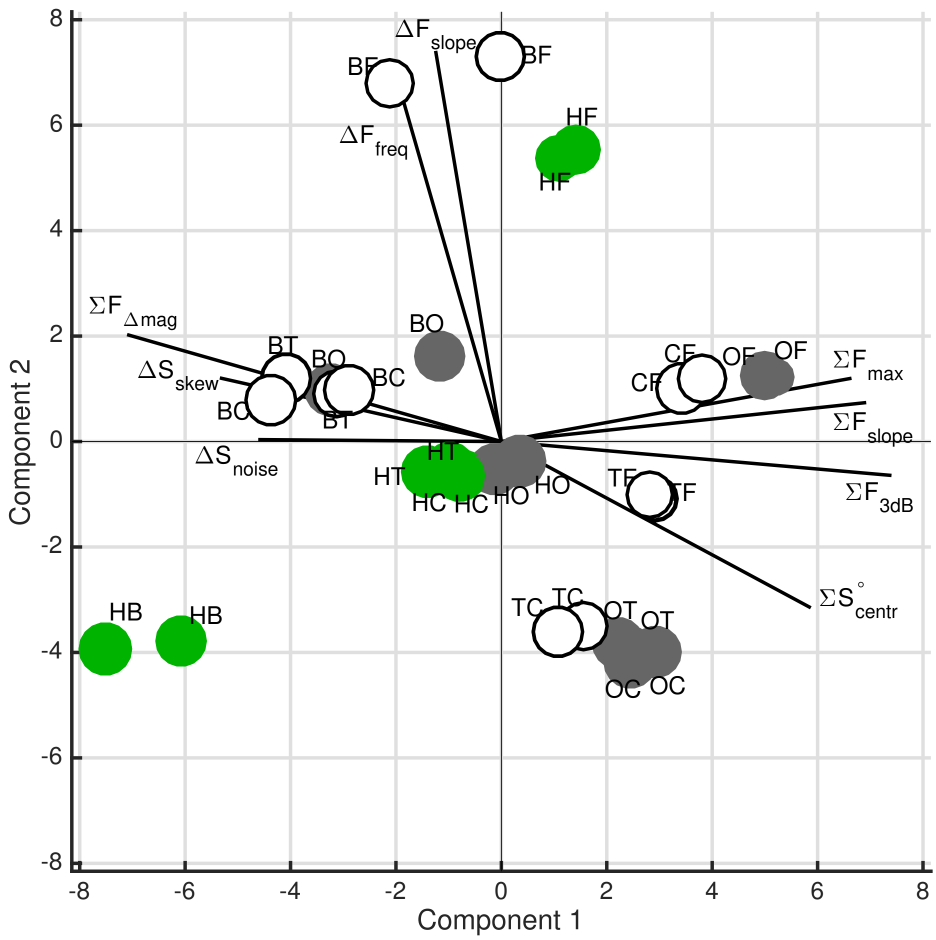

Figure 7. Unison-dyad P_{ortho} and T_{ortho} for PCs 1 and 2. See Figure 4 for legend.

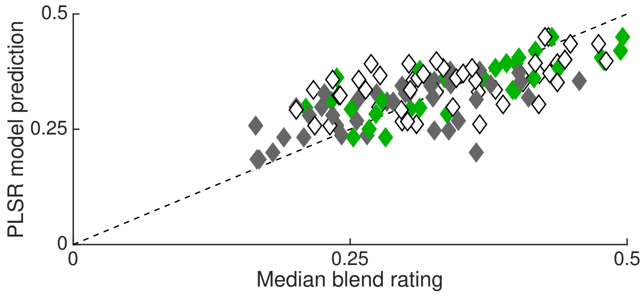

Figure 8. Non-unison-dyad model fit of y variables for X_{ortho}. See Figure 3 for legend.

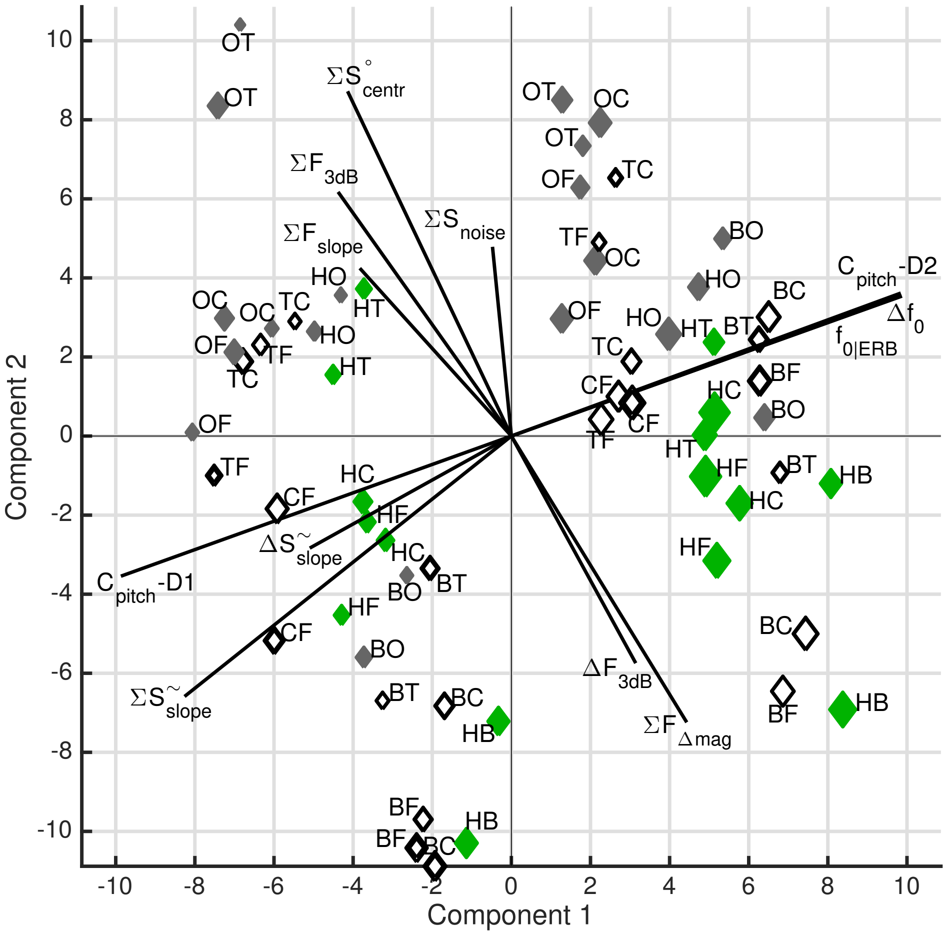

Figure 9. Non-unison-dyad P_{ortho} and T_{ortho} for PCs 1 and 2. See Figure 4 for legend.

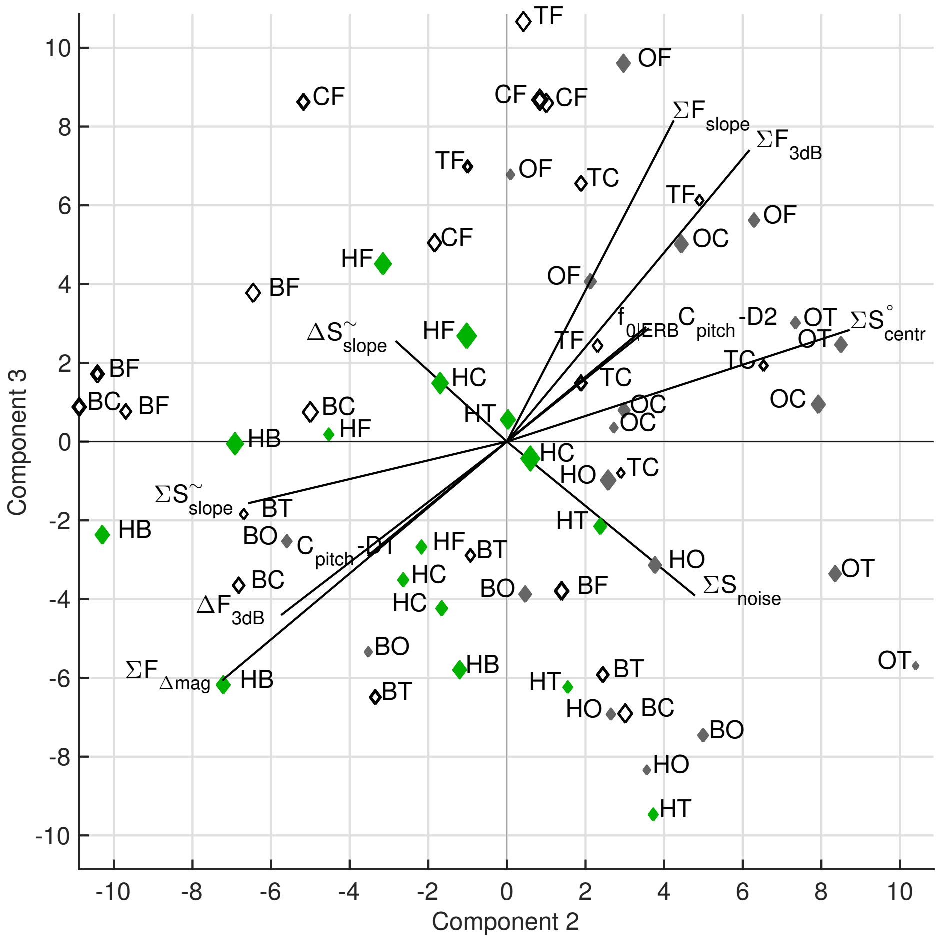

Figure 10. Non-unison-dyad P_{ortho} and T_{ortho} for PCs 2 and 3. See Figure 4 for legend.

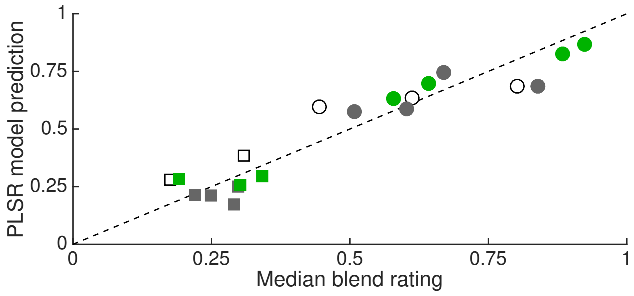

Figure 11. Triad model fit of y variables for X_{ortho}. Legend: squares, including *pizzicati*; circles, excluding *pizzicati*; grey involves oboe; green involves trombone (excl. PTO, TTO, TCO).

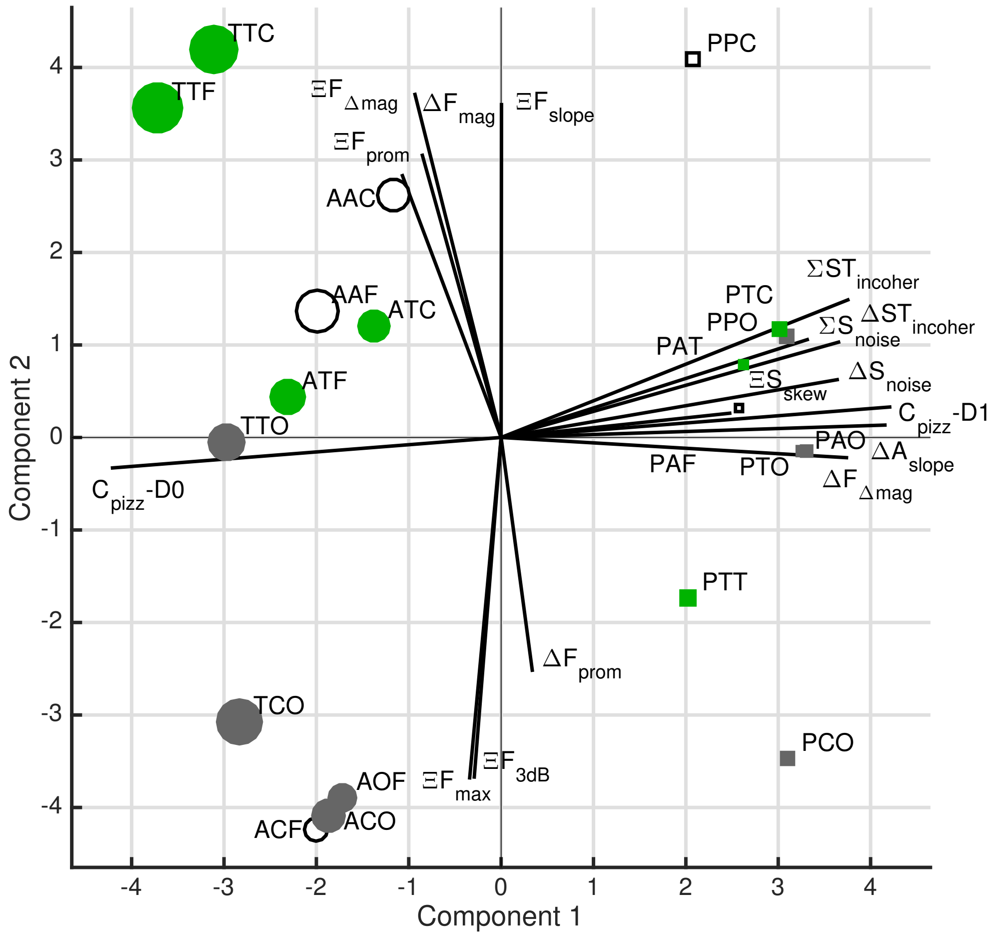

Figure 12. PLSR loadings P_{ortho} (vectors) and scores T_{ortho} (symbols) for PCs 1 and 2 with triads. Legend: squares, including *pizzicati*; circles, excluding *pizzicati*; symbol size represents relative degree of blend; grey involves oboe; green involves trombone (excluding PTO, TTO, TCO).