Perception of Dyads of Impulsive and Sustained Instrument Sounds

Damien Tardieu
IRCAM-STMS-CNRS, Paris, France

Stephen McAdams
McGill University, Montréal, Canada

Perception of instrumental blend is important for understanding aspects of orchestration, but no work has studied blends of impulsive and sustained instruments. The first experiment identified the factors that influence the rating of blendedness of dyads formed of one sustained sound and one impulsive sound. Longer attack times and lower spectral centroids increased blend. The contribution of the impulsive sound's properties to the degree of blend was greater than that of the sustained sound. The second experiment determined the factors that influence similarity ratings among dyads. The mean spectral envelope and the attack time of the dyad best explained the dissimilarity ratings. However, contrary to the first experiment, the spectral envelope of the sustained sound was more important than that of the impulsive sound. Multidimensional scaling of dissimilarity ratings on blended dyads yielded one dimension correlated with the attack time of the dyad and another dimension whose spectral correlate was different for two different clusters within the space, spectral spread for one and spectral flatness for the other, suggesting a combined categorical-analogical organization of the second dimension.

Received March 30, 2011, accepted January 29, 2012.

Key words: orchestration, timbre space, sound descriptors, blend, instrument dyads

Orchestration holds a special place in music composition. This art of timbre manipulation, of going from the musical notation to its acoustic realization, relies more than any other compositional activity on the composer's experience, knowledge of instrument timbres, and ability to predict the timbre of instrument sound mixtures. This particularity makes its formalization very difficult, if not impossible, as pointed out by authors of orchestration treatises such as Berlioz's: "This art can no more be taught than the writing of beautiful melodies... What suits the various instruments best, what is or is not practicable, easy or difficult, muffled or resonant, this can be taught... When it comes to combining them in groups ... and the art of mixing them in order to modify the sound of one with that of another and produce from the whole a particular timbre unobtainable on any instrument on its own... one can only point to the example of great composers and draw attention to the way they did it." (Berlioz, 1855/2002, p. 6) More than a century later, and despite a tremendous evolution in its practice, orchestration remains mainly an intuitively empirical discipline.

Historically, the main way to teach orchestration was the treatise form (e.g., Berlioz, 1855/2002; Casella & Mortari, 1958; Koechlin, 1954; Rimski-Korsakov, 1913). Generally these treatises extensively describe the different possibilities of the various instruments of the orchestra. The description of orchestration itself, as a way to combine instruments, is usually less important and mainly consists of descriptions of the sound of various instrument combinations, illustrated by examples from the repertoire. The description is very empirical, and there is no attempt to formalize orchestration. From this perspective, Charles Koechlin's (1954) treatise is an exception. It is different from the others as it clearly exposes goals for orchestration tasks and proposes methods to attain those goals. Koechlin proposes the concept of "balance" to describe a perceptual attribute resulting from an orchestration. Balance is not a goal per se, but for Koechlin, it is important to know how to attain balanced orchestration in order to be able to write unbalanced orchestrations intentionally. Koechlin then suggests two scales to describe instruments that control orchestral balance: volume and intensity. Whereas intensity is often studied in the other treatises, volume is specific to Koechlin. He does not give a precise definition, but his ranking of instruments along this scale gives an intuitive idea. Koechlin uses this scale to obtain various instrumental layers. If instruments playing simultaneously have the same volume and intensity, then they will be perceived as belonging to the same layer. Thus, the

Music Perception volume 30, issue 2, pp. 117–128. issn 0730-7829, electronic issn 1533-8312. © 2012 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's rights and permissions website, http://www.ucpressjournals.com/reprintinfo.asp.DOI: 10.1525/mp.2012.30.2.117

originality of Koechlin's treatise is to propose dimensions that sort instruments and predict the effect of their combinations. He proposes the first true formalization of aspects of orchestration.

More recently, the development of knowledge concerning timbre perception and auditory scene analysis has brought new insights into perceptual aspects of orchestration. Experiments on mixtures of instrument tones have been used to explore the perception of instrumental tone blend and the emergent timbre of instrumental mixtures. Before describing these experiments we review some important results on timbre perception and auditory scene analysis.

TIMBRE PERCEPTION

Timbre similarity is usually studied using dissimilarity rating tasks (cf. McAdams, 1993). The dissimilarity ratings are then fed into a multidimensional scaling (MDS) algorithm to find a distance model to represent them (Grey, 1977; McAdams et al., 1995; Wessel, 1979). The dimensions in the model are correlated with audio descriptors to derive a psychophysical model of timbre. The dimensions have most often been related to temporal envelope, spectral envelope and spectral variations. The most common audio descriptors are spectral centroid, logarithm of the attack time, and either spectral flux, spectral spread, or spectral irregularity (also called spectral deviation) (Krimphoff, McAdams, & Winsberg, 1994 ; Peeters et al., 2000). Whereas the spectral centroid is widely accepted, the role of attack time is more questionable. Iverson and Krumhansl (1993) ran three pairwise similarity rating experiments on whole instrumental sounds, on the attack part, and on the remaining part of the sounds after removal of the attack. Dissimilarity matrices for the three stimulus sets were fairly strongly correlated ($r \geq .74$), showing that, according to the authors, the salient attributes for timbral similarity judgments are present throughout the tones and that the attack time per se does not have the special role for similarity it has for instrument identification. This result may also be explained by the strong correlation between some descriptors of the attack part of the temporal envelope and other descriptors of the remaining part of the envelope. For instance, attack time and the shape of the decay are usually strongly correlated in instrument sounds when both sustained and impulsive sounds are presented in a stimulus set (sharp attacks are usually associated with impulsive sounds that have a resonant decay).

AUDITORY SCENE ANALYSIS

Following Bregman's (1994) work, many researchers have focused on trying to discover the mechanisms by which the auditory system segregates sounds coming from different sources and integrates the acoustic information coming from the same sound source. The grouping of auditory events, either sequentially or concurrently, seems to rely on one main guiding principle; events that are similar in their spectrotemporal properties should be grouped together because they probably come from the same source (McAdams, 1999). For pure tones, the factors influencing sequential streaming are frequency and spatial location, the factors influencing concurrent grouping are harmonicity, onset synchrony, and spatial location (Bregman, 1994). For complex sounds, the situation is somewhat less clear. There is general consensus that pitch and timbre have an influence on fusion.

Iverson (1995) extensively studied the effect of static and dynamic attributes of timbre on sequential grouping. He found that tones with highly similar spectra were less segregated perceptually (i.e., were grouped into a single auditory stream) than were tones with dissimilar spectra, and that tones with shorter attacks were more segregated than were tones with gradual attacks. He also found that tones with low spectral centroid streamed less than high centroid tones. In line with this result, Bey and McAdams (2003) found that the distance separating sounds in timbre space was monotonically related to the ability of listeners to hear out a target sequence interleaved with a distractor sequence. Interestingly, the sound descriptors influencing sequencial streaming highlighted by Iverson (1995) have also been found to influence simultaneous grouping for instruments and singers (Goodwin, 1980; Sandell, 1995).

EXPERIMENTS ON INSTRUMENT MIXTURES

Following the development of auditory scene analysis, researchers began to study orchestration from a scientific point of view. Bregman (1991) tried to apply auditory scene analysis rules to an explanation of some orchestration techniques on the basis of concurrent and sequential grouping factors. For instance, by creating two distinct streams, the composer can write two dissonant lines without creating perceptual dissonance; in that case, sequential grouping prevails over simultaneous grouping, the latter resulting in dissonance. This study is the first to show that rules of orchestration can be explained by cognitive science and conversely that it might be possible to use science to define orchestration rules. With this aim, Kendall and Carterette (1991) tried to identify how dyads of instrument sounds are perceived. Participants were asked to rate the dissimilarity between all the pairwise combinations of five wind instrument tones: oboe, clarinet, flute, alto saxophone, and trumpet in six different contexts. A multidimensional scaling analysis was performed on the results. The axes were then identified by a musicologist as being nasal/not nasal, bright/rich, and simple/complex. This work demonstrates that tone dyads can be placed in a multidimensional timbre space in the same way as isolated tones. They also found that the positions of the dyads in the space could be partially predicted by the vector sum of the positions of the sounds composing the mixture in the same space, raising the possibility that the timbre of a combination of sounds can be predicted from its constituent timbres. In a second paper (Kendall & Carterette, 1993), the authors attempted to link these dimensions to the perceived blend of the dyads and to the identification of the instruments composing the dyad. Blend ratings were well predicted by the distance between the instruments in a timbre space, showing that sounds with similar timbres are more likely to blend. In a similar experiment, Sandell (1995) attempted to identify audio descriptors that explain blend between instruments. The instrumental sounds used were drawn from those of Grey (1977). Participants were asked to rate the blend of dyads of instrument sounds. Results were then correlated with nine audio descriptors. When the sounds were presented in unison, the centroid of the mixture was the descriptor that best explained blend ratings followed by attack contrast and loudness correlation. When they were presented with an interval of a minor third, the variance was explained (in decreasing order) by the difference in centroid, attack contrast, composite centroid, and release synchrony. The conclusion is thus quite similar to that of Kendall and Carterette (1993): similar timbres blend better. This experiment also showed that dark sounds blend better. A similar effect was also noted by Goodwin (1980) when analyzing the blending strategy of choir singers.

The link between spectral centroid and blend perception could be a confirmation of the analysis by Chiasson (2007) of Koechlin's (1954) notion of volume. In an analysis of Koechlin's treatise, Chiasson (2007) compared the volume scale with the axes of a timbre space. He suggested that volume could be explained by the spectral centroid or the spectral spread of sounds and that these audio descriptors could be important for orchestration analysis. One might then hypothesize that the layers described by Koechlin are groups of instruments that blend well. Therefore, a balanced orchestration may be an orchestration in which all the instruments blend into a single timbre. The observation that balanced orchestrations are produced by instruments with the same volume could thus be explained by the fact that sounds with similar spectral centroid blend best.

All of the previously cited experiments deal with sustained (or continuant) sounds. But a very common technique in orchestration consists of using dyads of impulsive and sustained sounds to create a sound with particular temporal and spectral descriptors. In this case, the previously cited results are not really useful. In this article, we propose two experiments to study the perception of this kind of instrumental combination.

COMPUTER-AIDED ORCHESTRATION CONTEXT

The two experiments presented in this article were conceived with the problem of designing perceptually relevant computer-aided orchestration (CAO) systems in mind. To help the composer explore the timbral possibilities of the orchestra, we propose the following problem: Given a target sound, how do we find a combination of instrument notes that sounds close to this target (for more details concerning this problem see Carpentier, Tardieu, Assayag, Rodet, & Saint-James, 2007; Tardieu & Rodet, 2007). This problem raises many questions, some of which we address here: How is a combination of instrument tones perceived in terms of blend and emergent timbre? Can we predict this perception from acoustic attributes of the dyad or of the tones themselves?

We propose two experiments on dyads composed of an impulsive and a sustained sound. The first experiment attempts to identify the factors that influence the blending of such mixtures. From the previously described literature, it can be expected that the centroid and the attack time of the sounds would have an influence on blending. The second experiment seeks to determine the factors that influence the perception of dissimilarity between blended dyads. The expected factors are descriptors of the spectral envelope, such as spectral centroid, and of the temporal envelope, such as attack time.

Experiment 1: Blend Rating

PARTICIPANTS

Participants ($N = 23$) were recruited from the Schulich School of Music of McGill University. Before the experiment, they received an audiogram to test their hearing (ISO 389–8, 2004; Martin & Champlin, 2000). Three participants were rejected because their hearing thresholds were at least 20 dB above standard at a given test frequency. They were paid $5 CAD. Those who completed the main experiment were paid $10 CAD.

These included 10 males and 10 females with a median age of 21 years (range 18–42) and a median number of years of musical practice of 12 years (range 5–30).

STIMULI

Eleven sustained sounds and 11 impulsive sounds were used. Sustained sounds were extracted from the Studio Online database at IRCAM (Ballet & Borghesi, 1999). Impulsive sounds were extracted from the Vienna Symphonic Library database (*www.vsl.co.at*). These sounds were chosen to cover a wide range of timbres. Instruments with sustained sounds included bassoon, B♭ clarinet, oboe, flute, C trumpet, French horn, trombone, violin, cello, and double bass. Instruments with impulsive sounds included flute "pizzicato" (produced by fingering a specific pitch and producing a hard, brief "T" gesture with the mouth), woodblock (played with a hard string-wound mallet and a wooden mallet), marimba (played forte and piano), vibraphone (played forte and piano), two different violin pizzicati, and two harp sounds, one playing normally and the other producing a harmonic at the octave. The list of sounds, their families, and abbreviations are given in Table 1. The pitch was fixed at C#4 (a fundamental frequency of approximately 277 Hz). Sounds were cut to have a maximum duration of 2.5 s and a 200 ms decreasing linear ramp was applied. The decay times of the marimba and vibraphone sounds were thus shortened, but the slopes of the natural decay were not altered up to the ramp. Before starting the loudness equalization task, all of the sounds were equalized using a loudness estimation method based on Moore, Glasberg, and Baer (1997) for the instantaneous loudness and on N_6 for the estimation of the global loudness (Fastl, 1993); that is, the loudness value that is reached or exceeded 6% of the time over the duration of the tone. Then, six listeners equalized the loudnesses of the sounds by moving a slider on a computer screen to make the loudness of a sound equal to that of a reference sound. For the impulsive sounds, the reference was the forte vibraphone, and for the sustained sounds it was the trombone. These references were chosen because they sounded louder than the others to the authors. The impulsive and sustained sounds were equalized independently. All possible combinations of an impulsive sound and a sustained sound were then created by adding the sounds. A total of 121 stimuli were thus produced. The sounds were encoded in a 16-bit AIFF format with a sampling rate of 44,100 Hz. They were played at a mean level of 67 dB SPL as measured with a Bruel & Kjaer Type 2205 sound level meter coupled with a Bruel & Kjaer Type 4153 artificial ear.

TABLE 1. **Name, Family and Abbreviation of the Instruments Used in the Experiments.**

Instrument	Family	Abbreviation
B♭ Clarinet	Woodwinds	Cla
Bassoon	Woodwinds	Bsn
Flute	Woodwinds	Flt
Oboe	Woodwinds	Obo
Double bass	Strings	Cba
Double bass, harmonic	Strings	Cbh
Cello	Strings	Vcl
Violin	Strings	Vln
C Trumpet	Brass	Tpt
French horn	Brass	Fhn
Trombone	Brass	Tbn
Woodblock, hard string-wound mallet	Block	Wbs
Woodblock, wooden mallet	Block	Wbw
Flute pizzicato	Woodwinds	Flp
Violin, Bartok pizzicato	Strings	Vlb
Violin, normal pizzicato	Strings	Vlp
Marimba forte	Bar	Maf
Marimba mezzo forte	Bar	Mam
Vibraphone forte	Bar	Vbf
Vibraphone piano	Bar	Vbp
Harp, pizzicato, harmonic	Strings	Hph
Harp, pizzicato	Strings	Hpp

PROCEDURE

The experimental session consisted of a familiarization phase and an experimental phase. The participant read the experimental instructions and asked any questions necessary for clarification. Then the 121 sounds were presented in random order to familiarize the participant with the range of variation of blend among the sounds that were to be rated. For each experimental trial, participants heard a dyad, which they rated for blend on a continuous scale: the left end of the scale was noted "not blended" to indicate the absence of blend, and the right was noted "very blended" to indicate a perfect blend. Answers were given by adjusting a slider with the mouse. Each participant rated the 121 stimuli four times in four sessions. For each session, dyads were presented in random order. There was no break between sessions. The total duration of the experiment was about 45 min. The participant was seated in a booth in front of the computer. The experiment was controled by a PsiExp program (Smith, 1995) running on a Macintosh G5 computer.

The stimuli were presented via Sennheiser HD280 earphones connected to a Grace Design m904 digital amplifier, which converted and amplified the signal received from the computer.

SOUND ANALYSIS

The audio descriptors are taken from the Timbre Toolbox (Peeters, Giordano, Susini, Misdariis, & McAdams, 2011). Each of the seven descriptors used (see below) is calculated on the impulsive sound, the sustained sound, and the mixture. The difference between the values of a given descriptor for the sustained and impulsive sounds is also used, for a total of 28 descriptors for each dyad. These descriptors will be classified as impulsive descriptors, sustained descriptors, mixture descriptors, and difference descriptors, respectively. All of the time-varying spectral descriptors are extracted using a 60-ms Blackman window and a hop size of 20 ms. Each time-varying function is then reduced to a single value by taking the mean over time weighted by loudness. Loudness is computed using the simplified version of the Moore et al. (1997) model described in (Peeters, 2004). Note that because each individual sound appears in several dyads, the sustained and impulsive sound descriptors are the same for several dyads.

Spectral centroid. Let $X(k)$ be the amplitude spectrum computed on a logarithmic frequency scale of a time frame, where k is the frequency bin. The spectral centroid s_c of the frame is:

$$s_c = \frac{\sum_{k=0}^{K-1} f(k)\,|X(k)|}{\sum_{k=0}^{K-1} |X(k)|} \tag{1}$$

where $f(k)$ is the frequency corresponding to bin k and K is the index of the bin corresponding to the Nyquist frequency (22,050 Hz).

Spectral spread is defined as the standard deviation of the spectrum about the spectral centroid:

$$s_s = \sqrt{\frac{\sum_{k=0}^{K-1} (f(k)-s_c)^2\,|X(k)|}{\sum_{k=0}^{K-1} |X(k)|}} \tag{2}$$

Spectral flatness is defined as the ratio between the geometric mean and the arithmetic mean of the amplitude spectrum. It is a measure of the sinusoidality of the sound. Let $Y(k)$ be the amplitude spectrum of a time frame computed on a linear frequency scale, where k is the frequency bin and K is the index of the bin corresponding to the Nyquist frequency (22,050 Hz).

$$s_f = 10 \cdot \log_{10} \frac{\left(\prod_{k=0}^{K-1} |Y(k)^2|\right)^{\frac{1}{K}}}{\frac{1}{K}\sum_{k=0}^{K-1} |Y(k)^2|} \tag{3}$$

Mel spectrum is a multidimensional descriptor that is computed in the following way:

$$S(m) = \sum\nolimits_{k=0}^{K-1} |Y(k)|\,H(k,m) \tag{4}$$

where $H(k,m)$ is a filter bank of overlapping triangular filters. We used 70 filters with centers equally spaced on an approximated Mel scale:

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{5}$$

It should be noted that this multidimensional descriptor is only used to compute the Euclidean distance between the Mel spectra of two sounds.

Log-attack time. Attack time estimation is achieved with the weakest-effort method of Peeters et al. (2011). Indeed, these authors found that the usual method, a fixed threshold for the start and maximum for the end of the attack, both derived from the energy function, was not robust for real sounds. For instance, the estimation of the start of the attack can be disturbed by additional background noise and the estimation of the end of the attack can be difficult for some instruments, like trumpet, which often have a continuously increasing envelope.

Temporal increase is defined as the mean slope over the attack portion of the energy envelope (Peeters et al., 2011).

Temporal decrease is a measure of the signal energy decrease. Signal energy is modeled by:

$$\hat{e}(t) = A \cdot \exp(-\alpha(t - t_{\max}))\; t > t_{\max} \tag{6}$$

where $t_{\max}$ is the time at which the maximum energy is attained. The temporal decrease α is estimated by fitting a line to the logarithm of the energy envelope. This descriptor discriminates between sustained and impulsive sounds. For impulsive sounds, it also describes the shape of the release of the sound. We hypothesize that this shape has an influence on perceptual blend.

Results

Each participant's data consisted of 484 blend ratings. The analysis proceeded in three stages. Interparticipant correlations on the ratings were computed using the Pearson correlation coefficient, and a cluster analysis of the correlations was used to detect participants who performed very differently from the others. Data sets that were systematically uncorrelated with all other sets might indicate participants who had not adopted a systematic rating strategy or those who misunderstood the instructions. These participants were eliminated from further analysis. Subsequently, analyses of variance (ANOVA) were performed to assess the effect of the kind of instrument contained in the dyad on the perceived blend. The Geisser-Greenhouse correction (Greenhouse & Geisser, 1959) was applied to compensate for inhomogeneity of covariances due to repeated measures. *F* statistics are cited with uncorrected degrees of freedom. If ε is less than one, its value is cited, and the probability is determined with the corrected degrees of freedom. Finally, ratings were correlated with audio descriptors.

Cluster analysis. The correlations between the rating vectors of all pairs of participants were computed (df = 482). The correlation matrix was submitted to a hierarchical cluster analysis using the nearest neighbor (single link) algorithm. Two participants were clearly isolated from the rest of the participants (see Figure 1). The data for these two outliers were eliminated from the subsequent analysis. The average correlations between these participants and the others were .08 and −.09. The average interparticipant correlation was .51 (*CI* [.30, .72]) for the remaining participants.

Blend ratings. The mean blend ratings for each dyad are listed in Table 2. The dyad that blended the best was French horn combined with vibraphone played piano (.85). The dyad that blended the worst was violin with Bartók pizzicato combined with trumpet (.20). Overall, vibraphone piano, with a mean blend rating of .70, was the impulsive instrument that blended the best whatever the sustained instrument, whereas Bartók pizzicato on the violin blended the worst (.25). Bassoon was the sustained instrument that blended the best with a mean blend rating of .59 and cello blended the worst (.38).

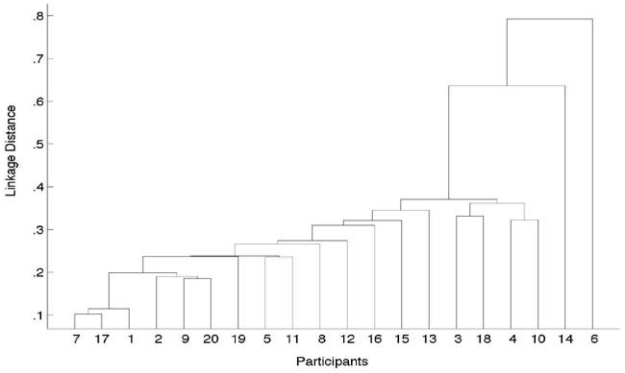

FIGURE 1. Dendrogram for the cluster analysis in Experiment 1.

ANOVA. To assess the effect of the instrument on the blend rating, a three-way ANOVA with repeated measures on Impulsive Instrument, Sustained Instrument, and Session was performed. The effect of the factor Impulsive Instrument, $F(10, 170) = 25.44, \varepsilon = .18$, $p < .001$, $\eta_p^2 = .60$, was stronger than the effect of the factor Sustained Instrument, $F(10, 170) = 9.91, \varepsilon = .25$, $p < .001, \eta_p^2 = .37$. This would suggest that the choice of the impulsive instrument is more important than the choice of the sustained instrument in the control of the perceived blend for such dyads. We also observed a significant effect of the interaction of these two factors, $F(100, 1700) = 4.53, \varepsilon = .12, p < .001$, showing that the choice of the instruments cannot be made independently. An interesting point here is that the mean ratings for the factor Impulsive Instrument (see Figure 2) were strongly correlated with the spectral centroid of the corresponding sound, $r(9) = -.93$, $p < .001$. This indicates that a bright impulsive sound (high spectral centroid) will hardly blend, whatever the sustained sound is. This finding is in agreement with the results of Sandell (1995) on sustained sound dyads.

Correlation with audio descriptors. We tested the correlation of the blend ratings with the set of audio descriptors. We used only unidimensional descriptors. Note that because the individual sounds are the same in many dyads, the sustained and impulsive sound descriptors are equal for many dyads. Also, due to the strong correlation between temporal and spectral descriptors for impulsive sounds, it is impossible to separate temporal and spectral effects in the blend judgments.

The highest correlation was obtained for the spectral centroid of the impulsive sound, $r(119) = -.79, p < .001$. This is consistent with observations made in the previous section showing that a bright sound blends less. If we consider only mixture and difference descriptors, temporal descriptors correlate better. The temporal increase of the mixture, $r(119) = -.75$, $p < .001$, and difference between temporal decreases, $r(119) = .73$, $p < .001$, show moderate correlations with blend ratings, indicating that dyads with a slow attack or dyads composed of sounds with the same kind of envelope decay blend more. However, because the temporal descriptors of the sustained sounds do not vary much, these descriptors are strongly correlated with the corresponding descriptors computed on the impulsive sound. Therefore, these

TABLE 2. Mean Blend Rating for Each Dyad.

	Bsn	Cba	Cbh	Cla	Fhn	Flt	Obo	Tbn	Tpt	Vcl	Vln
Flp	.40	.25	.36	.33	.36	.28	.24	.33	.24	.26	.25
Hph	.75	.50	.44	.69	.76	.53	.53	.61	.50	.42	.46
Hpp	.49	.34	.35	.49	.46	.36	.31	.40	.30	.28	.29
Maf	.58	.45	.37	.53	.54	.49	.45	.49	.41	.40	.47
Mam	.70	.51	.39	.68	.75	.58	.57	.62	.49	.47	.54
Vbf	.78	.55	.44	.73	.75	.62	.62	.74	.55	.53	.54
Vbp	.83	.65	.47	.79	.85	.70	.74	.80	.67	.56	.69
Vlb	.30	.24	.34	.27	.28	.23	.21	.25	.20	.24	.22
Vlp	.71	.40	.44	.68	.67	.56	.47	.63	.39	.43	.48
Wbs	.52	.28	.35	.42	.50	.35	.30	.47	.28	.25	.28
Wbw	.40	.32	.31	.30	.40	.31	.27	.34	.23	.28	.22

Note: Columns are sustained instruments, rows are impulsive instruments. White corresponds to blend ratings lower than .5, light grey to ratings between .5 and .7, dark grey to ratings higher than .7.

results could also be interpreted by saying that dyads containing an impulsive sound with a long attack or a slow decay blend better. Actually, the small ranges of the sustained sound descriptors compared to impulsive sound descriptors make the effect of the latter more salient and may hide the effect of the former. Finally, and practically, we suggest that this kind of mixture blends well either when the spectral centroid of the impulsive sound is low (the sound is not bright) or when the impulsive sound has a soft attack; these two descriptors being strongly correlated for this sound set.

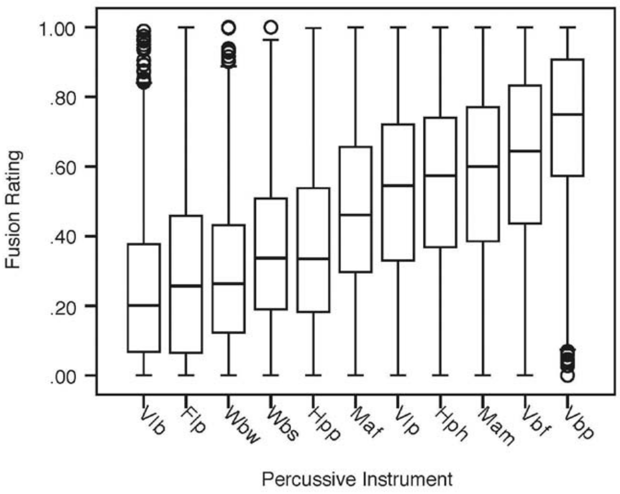

FIGURE 2. Mean blend rating for the factor Impulsive Instruments. The center horizontal line represents the median. The middle two horizontal lines represent the upper and lower limits of the interquartile range. The outer whiskers represent the highest and lowest values that are not outliers. Outliers, represented by 'o' signs, are values that are more than 1.5 times the interquartile range.

Experiment 2

In Experiment 2, we investigated the perceptual dissimilarities among a subset of blended dyads in order to determine the underlying acoustic features.

PARTICIPANTS

The 25 participants came from the Schulich School of Music of McGill University. Fourteen of the participants had also taken part in Experiment 1. The participants who had not participated in the previous experiment all had normal hearing as measured with an audiometer. The participants were paid $10 CAD for their participation. Participants included 10 males and 15 females with a median age of 21.5 years (range 18–42) and a median number of years of musical practice of 14 years (range 7–30).

STIMULI

Sixteen sounds were selected based on the results of the previous experiment. We selected sounds that blended well, but with various timbres. Because the perception of blend is highly dependent on the instrument timbre, we were not able to select only the dyads that blended the best, and we made a tradeoff between degree of blend and timbral diversity. The selected dyads are presented in Table 3. Because vibraphone blends better than any other impulsive instrument, the selection method we used led to a high number of dyads containing vibraphone. The sounds were encoded in 16-bit AIFF format with a sampling rate of 44,100 Hz. They were played at a mean level of 63 dB SPL as measured with a Bruel & Kjaer Type 2205 sound-level meter. The difference in sound level between Experiments 1 and 2 is due to the fact that Experiment 1

TABLE 3. Selected Dyads for Experiment 2 and Their Mean Blend Ratings from Experiment 1.

Impulsive sound	Sustained sound	Mean blend rating
Wbs	Bsn	.55
Hph	Cla	.67
Hph	Bsn	.71
Maf	Bsn	.58
Mam	Flt	.58
Mam	Fhn	.72
Mam	Tbn	.63
Vbf	Cba	.53
Vbf	Vcl	.54
Vbp	Tpt	.64
Vbp	Fhn	.82
Vbp	Obo	.71
Vbp	Vln	.66
Vlp	Cla	.67
Vlp	Bsn	.68
Vlp	Tbn	.64

was conducted using headphones, whereas Experiment 2 was conducted using speakers in a different testing space.

PROCEDURE

The experimental session consisted of two phases: a familiarization phase and an experimental phase. Participants read the experimental instructions and asked any questions necessary for clarification. Then the 16 dyad sounds were presented in random order to familiarize the participants with the range of variation among the timbres to be rated. On each experimental trial, the participant's task was to compare two dyads and rate directly their degree of dissimilarity on a continuous scale ranging from "very close" to "very distant." Ratings were made with a computer mouse controlling a cursor on a slider on the computer screen. The pair could be played as many times as desired before entering the rating. All 120 pairs of the 16 sounds (excluding identical pairs) were presented for dissimilarity ratings in a different random order for each participant. The order of presentation of the two dyads in each pair was also randomized for each trial. Participants were allowed to take a break at any time during the experimental session, which lasted about 45 min.

The participant was seated in an isolated, sound-treated room in front of the computer. The experiment was controlled by a PsiExp program running on a Macintosh G5 computer. The stimuli were presented via a pair of Dynaudio BM 15A speakers driven by a Grace Design m904 digital amplifier, which converted and amplified the signal received from the computer through an M-Audio Audiophile 192 sound card. The same signal was sent to both speakers, which were located about 2 m from the listener at an angle of about ±45°.

RESULTS

Each participant's data consisted of a vector of 120 paired comparisons among 16 sounds. The analysis proceeded in four stages. Interparticipant correlations derived from the dissimilarity matrices were computed, and a cluster analysis of the correlations was used to detect participants who performed very differently from the others. These participants were eliminated from further analysis. Subsequently, analyses of variance were performed to assess the effect of the kind of instrument on the perceived dissimilarity. Then a multidimensional scaling analysis was performed using the CLASCAL algorithm (McAdams et al., 1995; Winsberg & De Soete, 1993). Finally similarity ratings were correlated with the audio descriptors described above.

Cluster analysis. The correlations between the dissimilarity vectors of all pairs of participants were computed. The correlation matrix was submitted to a hierarchical cluster analysis using the nearest neighbor (single linkage) algorithm. One participant was clearly isolated from the rest of the participants (see Figure 3). The mean correlation between this outlier and the other participants was .34. The average interparticipant correlation for the remaining participants was .52 (*CI* = .34, .70). The data of this outlier were eliminated from subsequent analyses.

ANOVA. A two-way ANOVA on the independent variables Both Impulsive Instruments Belong to the Same Family and Both Sustained Instruments Belong to the Same Family with mean dissimilarity ratings as dependent

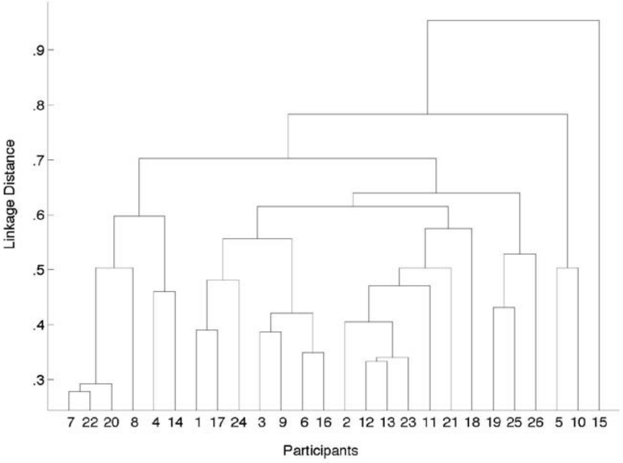

FIGURE 3. Dendrogram for the cluster analysis in Experiment 2.

variable was performed. Both factors had an influence, $F(1, 118) = 17.63, p < .001$, and $F(1, 118) = 25.13, p < .001$, , respectively, but no interaction between them was observed. In both cases, dissimilarity was lower if the instruments belonged to the same family (median ± interquartile range = .43 ± .16 for impulsive instruments, .39 ± .43 for sustained instruments) than when they belonged to different families (.63 ± .23 for impulsive instruments, .61 ± .24 for sustained instruments). However, the effect on the overall similarity was a bit stronger for the sustained sound. This may indicate that the sustained instrument makes a greater contribution to the dissimilarity ratings and thus that the sustained part of the sound affects dissimilarity perception more than the attack part does in this context. It is interesting to note that this result contradicts that of the previous experiment: although the impulsive sound contributes more strongly to perceptual blend, the sustained sound contributes more strongly to the emergent timbre of the mixture.

Multidimensional scaling. The data from the 24 selected subjects were analyzed with the CLASCAL algorithm (McAdams et al., 1995; Winsberg & De Soete, 1993), which models the dissimilarity ratings with a distance model that contains dimensions shared among all sounds, specificities for specific sounds and perceptual weights on each dimension and the set of specificities for an estimated number of latent classes of participants. The analysis converges to a two-dimensional space with specificities and three latent classes of participants. Figure 4 shows the two-dimensional space obtained by CLASCAL, and Table 4 shows the exact coordinates and the specificities. The first thing we noticed was the presence of two clusters along the first dimension. All the sounds in the right cluster contain a vibraphone. Thus, the first axis may reflect a categorization of the sounds on the basis of the presence or absence of the vibraphone. Interestingly, inside the *no vibraphone* cluster, the sustained instrument seems to be more important than the impulsive instrument in the dissimilarity ratings. Indeed, dyads containing the same sustained instrument but different impulsive instruments are very close together. On the one hand, vibraphone, when it is present, has a very strong influence on dissimilarity ratings, more important than the sustained instrument of the dyad. On the other hand, for dyads not containing vibraphone, the sustained instrument is more important. It is thus possible that two different strategies were used in the ratings depending on the presence/absence of the vibraphone. In the following, we explore this hypothesis by correlating audio descriptors and differences between descriptors with the coordinates of the MDS space and the dissimilarity ratings, respectively.

Correlation with audio descriptors. The attack time of the mixture is very strongly correlated with the first dimension of the MDS space, $r(14) = .94, p < .001$ (see Figure 5). Thus, the vibraphone/no-vibraphone categorization can be simply based on the attack time of the mixture: dyads with a low attack time are on the left and dyads with a high attack time (those containing vibraphone) are on the right. Inside the no-vibraphone cluster, the proximities can be explained by another acoustic attribute. Indeed, they rely mostly

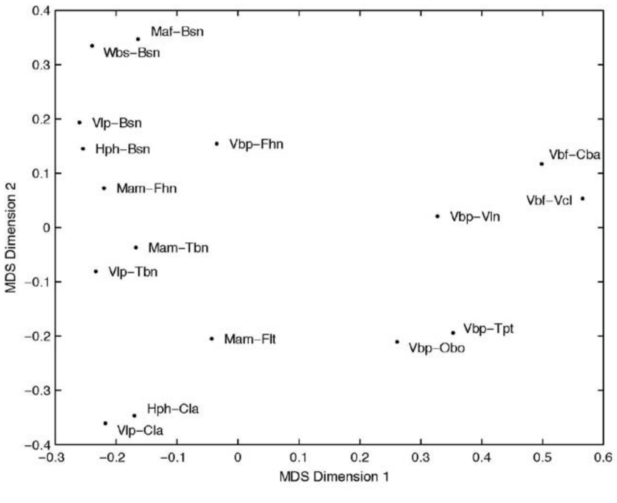

FIGURE 4. Timbre space in two dimensions: A spatial model with specificities and three latent classes derived from dissimilarity ratings on 16 timbres by 25 participants.

TABLE 4. Timbre Coordinates Along Common Dimensions and Corresponding Specificities.

Stimuli		Dim 1	Dim 2	Specif
Wbs	Bsn	−.24	.33	.041
Hph	Cla	−.17	−.35	.010
Hph	Bsn	−.25	.15	.000
Maf	Bsn	−.16	.35	.000
Mam	Flt	−.04	−.21	.094
Mam	Fhn	−.22	.07	.000
Mam	Tbn	−.17	−.04	.068
Vbf	Cba	.50	.12	.016
Vbf	Vcl	.57	.05	.000
Vbp	Tpt	.35	−.19	.076
Vbp	Fhn	−.03	.15	.092
Vbp	Obo	.26	−.21	.068
Vbp	Vln	.33	.02	.070
Vlp	Cla	−.22	−.36	.002
Vlp	Bsn	−.26	.19	.000
Vlp	Tbn	−.23	−.08	.047

Note: The values of the specificities are the square root of the value estimated in Eq. 10 (McAdams et al., 1995) in order for them to be of comparable magnitude to the coordinates along the common dimensions.

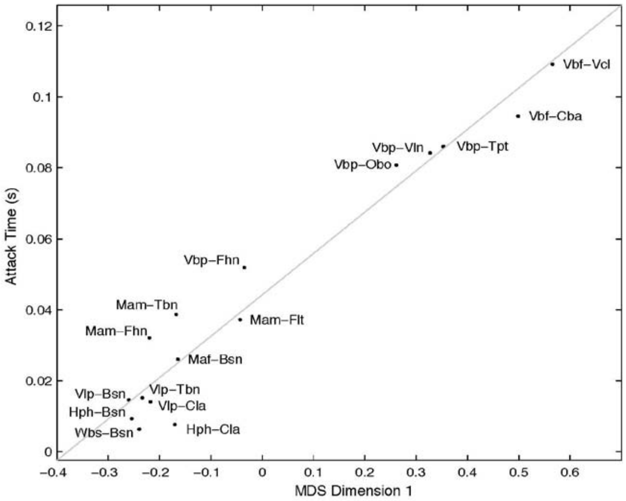

FIGURE 5. First axis of the timbre space versus attack time of the dyad.

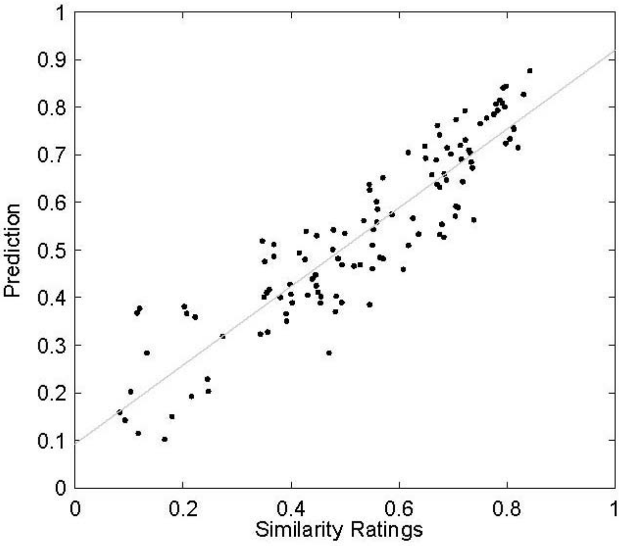

FIGURE 6. Prediction of similarity ratings using a linear combination of Mel log spectra and attack time.

on the second MDS dimension, which correlates strongly with the spectral spread of the mixture $r(9) = -.96$, $p < .001$. Concerning the right cluster, although interpreting a correlation coefficient derived from five samples should be done with caution, the second dimension correlates very strongly with spectral flatness of the mixture, $r(3) = -.95$, $p < .001$.

Another way to find the acoustic attributes underlying the dissimilarity ratings is to compute the Euclidean distance between dyads for each audio descriptor described in Experiment 1, and then to compute the correlation between the obtained distances and the mean dissimilarity ratings (Iverson & Krumhansl, 1993). Since we compute the correlation between distances, we can use both unidimensional and multidimensional descriptors. Note that in the case of unidimensonal descriptors, the Euclidean distance is just the absolute difference. The highest correlation is obtained with the log-amplitude Mel spectra of the mixture, $r(118) = .81$, $p < .001$. The same descriptor also gives a strong correlation, $r(118) = .74$, $p < .001$, when computed between the two sustained sounds. The correlation is very weak for the impulsive sounds, $r(118) = .32$, $p < .001$. Whereas the influence of spectral attributes of the sounds was somehow hidden in the MDS space, it becomes obvious when we perform the correlation directly on the dissimilarity ratings. Concerning temporal attributes, the attack time gives a strong correlation, $r(118) = .76$, $p < .001$, confirming the previous analysis of the MDS results.

Finally, as shown in Figure 6, a very good prediction of the ratings can be obtained by combining both descriptors in a linear regression, $r(118) = .91$, $p < .001$. To obtain this prediction, we computed the distance d_m between dyads using log-amplitude Mel spectra on the one hand and the distance d_a using attack time on the other hand and computed the linear regression between these distances and the dissimilarity ratings (s): $s = ad_m + bd_a + c$.

General Discussion

In the two experiments, we highlight the audio descriptors underlying the perception of blend and the perception of emergent timbre for dyads composed of one impulsive and one sustained sound. In both cases, the descriptors are typical of instrumental timbre. The attack time is very important in both experiments, being one of the two most important factors for predicting both blend and emergent timbre perception. This confirms again the central role of attack, and more generally time-varying properties of sounds in timbre perception (McAdams et al., 1995), auditory scene analysis (Iverson, 1995), or instrument fusion (Sandell, 1995). The importance of this descriptor for blend can be related to the importance of onset differences for simultaneous grouping (Darwin, 1981). Slight onset differences between two sounds favor the segregation of the two sounds. In our case, a slow attack makes it difficult to identify the starting point of a sound and thus favors blend. We might therefore hypothesize that the influence of onset asynchrony on simultaneous grouping depends on the onset characteristics and more specifically on onset duration. The second feature found in both experiments relates to the spectral envelope of the sound. The spectral centroid of the impulsive sound explains the degree of fusion, whereas the spectral spread and spectral envelope of the mixture explain similarity ratings.

Overall the results confirm and extend previous results from the literature. It is interesting to note that results on blend for sustained instrument sounds found by Kendall and Carterette (1993) and Sandell (1995) can be partially extended to impulsive sounds. The main difference is the much greater importance of attack time for impulsive sounds. The second experiment also confirms the fact that a perceptual space can be derived for concurrent sound dyads, as shown by Kendall and Carterette (1993). The perceptual correlates of the dimensions of this space were attack time for the first dimension and a spectrum-based descriptor for the second. These dimensions are the same as those found in previous experiments with single tones by McAdams et al. (1995), except that attack time is measured linearly instead of logarithmically. It is worth noting that the attack time computed on the whole dyad better explains perceived similarity than does this descriptor when computed on only one of the sounds composing the dyad. This is an indication of the high degree of blend of the dyads and also of the effectiveness of the attack time measurement.

The space obtained with MDS suggests that ratings could have been made on the basis of a categorization in which the first descriptor was the presence/absence of the vibraphone and the second was a spectral property of the sustained instruments, meaning that two dyads containing vibraphone and a similar sustained instrument would be perceived as similar and two dyads not containing vibraphone, but containing a similar sustained instrument would also be similar. This interpretation suggests that longer impulsive tones or tones with slower decays, such as the vibraphone, have more influence on the overall similarity than shorter sounds. The correlations with acoustic attributes also indicate that different acoustic attributes may have been used in each category. However, when correlating directly the dissimilarity ratings with distances in the descriptor space, we found a unique linear regression, based on the attack time and the spectral envelope, that explains very well the ratings for all pairs. This apparent contradiction can be explained by the fact that the information contained in the two acoustic attributes – spectral spread and spectral flatness – is also contained in the spectral envelope. So when we compute correlations using distances in the descriptor space, we only need the spectral envelope as a spectral feature.

Finally, concerning orchestration, we can summarize all the results by saying that because blend is more influenced by the impulsive instrument, whereas the overall timbre is more influenced by the sustained sound, the composer could have two nearly independent parameters to control these dyads: perceived blend can be controlled by choosing the impulsive instrument, and the overall timbre can be controlled by choosing the sustained sound.

Author Note

Portions of this research were described in a dissertation submitted to the Université Pierre et Marie Curie, Paris, France, in partial fulfillment of Damien Tardieu's Ph.D. degree. This work was supported by a travel grant from the Centre de Coopération Interuniversitaire Franco-québécoise (CCIFQ), French RIAM project Sample Orchestrator and French Oseo Project QUAERO grants to D. Tardieu and by the Canada Research Chair program and a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN 312774-05) to S. McAdams. The authors wish to thank Yan Maresz for his help choosing the sounds and for sharing his knowledge on orchestration. They also want to thank Geoffroy Peeters for his help with computing the sound descriptors.

Correspondence concerning this article should be addressed to Damien Tardieu, STMS-IRCAM-CNRS, 1 place Igor Stravinsky 75004 Paris, France (E-MAIL: Damien.Tardieu@ircam.fr) or Stephen McAdams, McGill University, 555 Sherbrooke St. W., Montréal, Québec, Canada H3A 1E3 (E-MAIL: smc@music.mcgill.ca).

References

BALLET, G., & BORGHESI, R. (1999). Studio Online 3.0: An Internet "killer application" for remote access to Ircam sounds and processing tools. *Journées d'informatique musicale* [Computer music days] (pp. 123–131). Issy-les-Moulineaux, France.

BERLIOZ, H. (2002). *Berlioz's orchestration treatise: A translation and commentary (Cambridge musical texts and monographs)* (Hugh Macdonald, Ed.). Cambridge, UK: Cambridge University Press. (Original work published 1855)

BEY, C., & MCADAMS, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 267–279.

BREGMAN, A. S. (1991). Timbre, orchestration, dissonance et organisation auditive [Timbre, orchestration, dissonance and auditory organization]. In J. B. Barrière (Ed.), *Le timbre, métaphore pour la composition* [Timbre: A metaphor for composition] (pp. 204–215). Paris, France: Christian Bourgeois.

BREGMAN, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

CARPENTIER, G., TARDIEU, D., ASSAYAG, G., RODET, X., & SAINT-JAMES, E. (2007). An evolutionary approach to computer-aided orchestration. In M. Giacobini (Ed.), *Applications of evolutionary computing* (Vol. 4448, pp. 488–497). Berlin, Germany: Springer/Heidelberg.

CASELLA, A., & MORTARI, V. (1958). *La technique de l'orchestre contemporain* [The technique of the contemporary orchestra] (Pierre Petit, Trans.). Paris, France: Ricordi.

CHIASSON, F. (2007). L'universalité de la méthode de Koechlin [*The universality of Koechlin's method*]. In M-H Benoit-Otis (Ed.), *Charles Koechlin, compositeur et humaniste* [Charles Koechlin, composer and humanist] (pp. 397–414). Paris, France: Vrin.

DARWIN, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology Section A*, *33*, 185–207.

FASTL, H. (1993). Loudness evaluation by subjects and by a loudness meter. In R. T. Verrillo (Ed.), *Sensory research: Multimodal perspectives* (pp. 199–210). Hillsdale, NJ: Lawrence Erlbaum.

GOODWIN, A. W. (1980). An acoustical study of individual voices in choral blend. *Journal of Research in Music Education*, *28*, 119–128.

GREENHOUSE, S. W., & GEISSER, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112.

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

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.

IVERSON, P. (1995). Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology-Human Perception and Performance*, *21*, 751–763.

IVERSON, P., & KRUMHANSL, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, *94*, 2595–2603.

KENDALL, R. A., & CARTERETTE, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, *8*, 369–404.

KENDALL, R. A., & CARTERETTE, E. C. (1993). Identification and blend of timbre as a basis for orchestration. *Contemporary Music Review*, *9*, 51–67.

KOECHLIN, C. (1954). *Traité de l'orchestration* [Treatise of orchestration]. Paris, France: Max Eschig.

KRIMPHOFF, J., MCADAMS, S., & WINSBERG, S. (1994). Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique [Characterization of the timbre of complex sounds. II. Acoustic analyses and psychophysical quantification]. *Journal de Physique*, *4*, 625–628.

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

MCADAMS, S. (1993). Recognition of sound sources and events. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychologie of human audition* (pp. 146–198). Oxford, UK: Oxford University Press.

MCADAMS, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, *23*(3), 85–102.

MCADAMS, S., WINSBERG, S., DONNADIEU, S., SOETE, G. D., & KRIMPHOFF, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*, 177–192.

MOORE, B. C. J., GLASBERG, B. R., & BAER, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, *45*, 224–240.

PEETERS, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project* (Tech. Rep.). Paris, France: IRCAM.

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*, 2902–2916.

PEETERS, G., MCADAMS, S., & HERRERA-BOYER, P. (2000). Instrument description in the context of MPEG-7. In I. Zannos (Ed.), *International Computer Music Conference 2000* (pp.166–169). Berlin, Germany: International Computer Music Association.

RIMSKI-KORSAKOV, N. A. (1913). *Principles of orchestration* (M. Steinberg, Ed.). Mineola, NY: Dover Publications Inc.

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

SMITH, B. (1995). PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation. *Proceedings of the Society for Music Perception and Cognition Conference* (pp. 83–84). Berkeley, CA: University of California.

TARDIEU, D., & RODET, X. (2007). An instrument timbre model for computer aided orchestration. *Workshop on applications of signal processing to audio and acoustics* (pp. 347–350). New Paltz, NY: IEEE.

WESSEL, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal*, *3*(2), 45–52.

WINSBERG, S., & DE SOETE, G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika*, *58*, 315–330.