The archived file is not the final published version of the article.

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

© (2015) S. Hirzel Verlag/European Acoustics Association
The definitive publisher-authenticated version is available online at:

http://www.ingentaconnect.com/content/dav/aaua

Digital Object Identifier http://dx.doi.org/10.3813/AAA.918898
Readers must contact the publisher for reprint or permission to use the material in any form.

Manuscript
Click here to download Manuscript: acustica_manuscript_rev1_final.tex

The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds

Sven-Amin Lembke[1)], Stephen McAdams[1)]

[1)] Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montréal, Québec, Canada H3A 1E3
sven-amin.lembke@mail.mcgill.ca

Abstract

Certain combinations of musical instruments lead to perceptually more blended timbres than others. Orchestration commonly seeks these combinations and can benefit from generalized acoustical descriptions of perceptually relevant features that allow the prediction of blend. Previous research on correlating such instrument-specific features with the perception of blend shows an important role of spectral-envelope characteristics, leaving unanswered, however, whether global or local characteristics are more important (e.g., spectral centroid or formant structure). This paper reports how wind instruments can be characterized through pitch-generalized spectral-envelope descriptions that exhibit their formant structure and how this is represented in an auditory model. Two experiments employing blend-production and blend-rating tasks study the perceptual relevance of formants to blend, involving dyads of a recorded instrument sound and a parametrically varied synthesized sound. Frequency relationships between formants influence blend critically, as does the degree of formant prominence. In addition, multiple linear regression relying primarily on local spectral-envelope characteristics explains 87% of the variance in blend ratings. A perceptual model for the contribution of spectral characteristics to perceived blend is proposed.

1 Introduction

Knowledge of instrument timbre leads composers to select certain instruments over others to fulfill a desired purpose in orchestrating a musical work. One such purpose is achieving a *blended* combination of instruments. The blending of instrumental timbres is thought to depend mainly on factors such as the synchrony between note onsets, partial tones aligned along the harmonic series, and specific combinations of instruments [1]. Whereas the first two factors depend on compositional decisions and their precise execution during musical performance, the third factor strongly relies on instrument-specific acoustical characteristics. A representative characterization of these features would thus facilitate explaining and theorizing perceptual effects related to blend. In agreement with past research [1–3], blend is defined as the perceptual fusion of concurrent sounds, with a corresponding decrease in the distinctness of individual sounds. It can involve different practical applications, such as *augmenting* a dominant timbre by adding other subordinate timbres or creating an entirely novel, *emergent* timbre [4]. This paper addresses only the first scenario, as the latter likely involves more than two instruments.

Along a perceptual continuum, maximum blend is most likely only achieved for concurrent sounds in pitch unison or octaves. Even though other intervals may be rightly assumed to make two instruments more distinct, certain instrument combinations would still exhibit higher degrees of blend than others. On the opposite extreme of this continuum, a strong distinctness of individual instruments leads to the perception of a heterogeneous, non-blended sound. Assuming auditory fusion to rely on low-level, bottom-up processes (related to auditory scene analysis, see [5]), increasingly strong and congruent perceptual cues for blend should counteract even deliberate attempts to identify individual sounds.

Previous research on timbre perception has shown a dominant importance of spec-

tral properties. Timbre similarity has been linked to spectral-envelope characteristics [6]. Similarity-based behavioral groupings of stimuli reflect a categorization into distinct spectral-envelope types [7] or the exchange of spectral envelopes between synthesized instruments results in an analogous inversion of positions in multidimensional timbre space [8]. Furthermore, Strong & Clark [9] reported increasing confusion in instrument identification (e.g., oboe with trumpet) whenever prominent spectral-envelope traits are disfigured, making instruments resemble each other more. With regard to blending, Kendall & Carterette [2] established a link between timbre similarity and blend, by relating closer timbre-space proximity between pairs of single-instrument sounds to higher blend ratings for the same sounds forming dyads. 'Darker' timbres have been hypothesized to be favorable to blend [4, 10], quantified through the global spectral-envelope descriptor *spectral centroid*, with 'dark' referring to lower centroids. Strong blend was found to be best explained by a low centroid *composite*, i.e., the centroid sum of the sounds forming a dyad.

By contrast with global descriptors, attempts to explain blending through local spectral-envelope characteristics focus on prominent spectral maxima, also termed *formants* [11, 12] in this context. Reuter [3] reported behavioral findings in favor of timbre blend occurring whenever formant regions between two instruments coincide. His explanation argues that this coincidence avoids *incomplete masking* [13], which inversely hypothesizes that the non-coincidence of formant locations prevents auditory fusion due to incomplete mutual masking of the presumedly salient formants between instruments, facilitating the detection of their distinct identities.

As prominent signifiers of spectral envelopes, formants have been employed widely to describe wind instruments [3, 7, 8, 12, 14–19]. Like the formant structure found in the human voice [11, 12, 20], formants in wind instruments are located at absolute frequency regions, which remain largely unaffected by pitch change [14, 17, 18]. This invariance may

in fact allow for the generalized acoustical description for these instruments and together with assessing its potential constraints (e.g., instrument register, dynamic marking), it will be of value to musical applications (e.g., [21]). Furthermore, it is meaningful to assess how such prominent spectral features are represented at an intermediary stage between acoustics and perception, i.e., at a sensorineural level, simulated by computational models of the human auditory system. Auditory models can account for effects related to spectral masking, i.e., to what neural excitation pattern a spectrum of a single or compound sound leads. For instance, excitation patterns typically involve an asymmetric upward spread in frequency, but the shape of excitation still varies both as a function of frequency and excitation level. The Auditory Image Model (AIM) simulates different stages of the peripheral auditory system, covering the transduction of acoustical signals into neural responses and the subsequent temporal integration across auditory filters yielding the *stabilized auditory image* (SAI), which provides the closest representation relating to acoustical spectral-envelope traits for human-voice and musical-instrument sounds [22]. AIM's most recent development employs *dynamic, compressive gammachirp* (DCGC) filterbanks, which account for both frequency and level dependency of basilar excitation by adapting filter shape accordingly [23]. AIM may therefore aid in assessing the relevance of hypotheses concerning blend, as previous theories had also employed representations or explanations which took spectral-masking effects into account [3, 4].

This paper addresses whether pitch-invariant spectral-envelope characterization is relevant to blending. Section 2 introduces the chosen approach to spectral-envelope description, its corresponding representation through auditory models, and how in the perceptual investigation the spectral description is operationalized in terms of parametric variations of formant frequency location. Section 3 outlines the design of two behavioral experiments that investigate the relevance of local variations of formant structure to blend perception, with their specific methods and findings presented in Sections 4

and 5, respectively. Finally, the combined results from acoustical and perceptual investigations are discussed in Section 6, leading to the establishment of a spectral model for blend in Section 7.

2 Spectral-envelope characteristics

A corpus of wind instrument recordings was used to establish a generalized acoustical description for each instrument. The orchestral instrument samples were drawn from the Vienna Symphonic Library[1] (VSL), supplied as stereo WAV files (44.1 kHz sampling rate, 16-bit dynamic resolution), with only left-channel data considered. The investigated instruments comprised (French) horn, bassoon, C trumpet, B♭ clarinet, oboe, and flute, with the available audio samples spanning their respective pitch ranges in semitone increments. Because the primary focus concerned spectral aspects, all selected samples consisted of long, sustained notes without vibrato. As spectral envelopes commonly exhibit significant variation across dynamic markings, all samples included only *mezzoforte* markings, representing an intermediate level of instrument dynamics.

2.1 Spectral-envelope description

Past investigations of pitch-invariant spectral-envelope characteristics pursued comprehensive assessments of spectral analyses encompassing extended pitch ranges of instruments [14, 17, 18]. The spectral-envelope description employed in this paper was based on an empirical estimation technique relying on the initial computation of power-density spectra for the sustained portions of sounds (excluding onset and offset), followed by the detection of partial tones, i.e., their frequencies and power levels. A curve-fitting procedure employing a *cubic smoothing spline* (piecewise polynomial of order 3) applied to the composite distribution of partial tones over all pitches yielded the spectral-envelope esti-

[1]URL: http://vsl.co.at/. Last accessed: April 12, 2014.

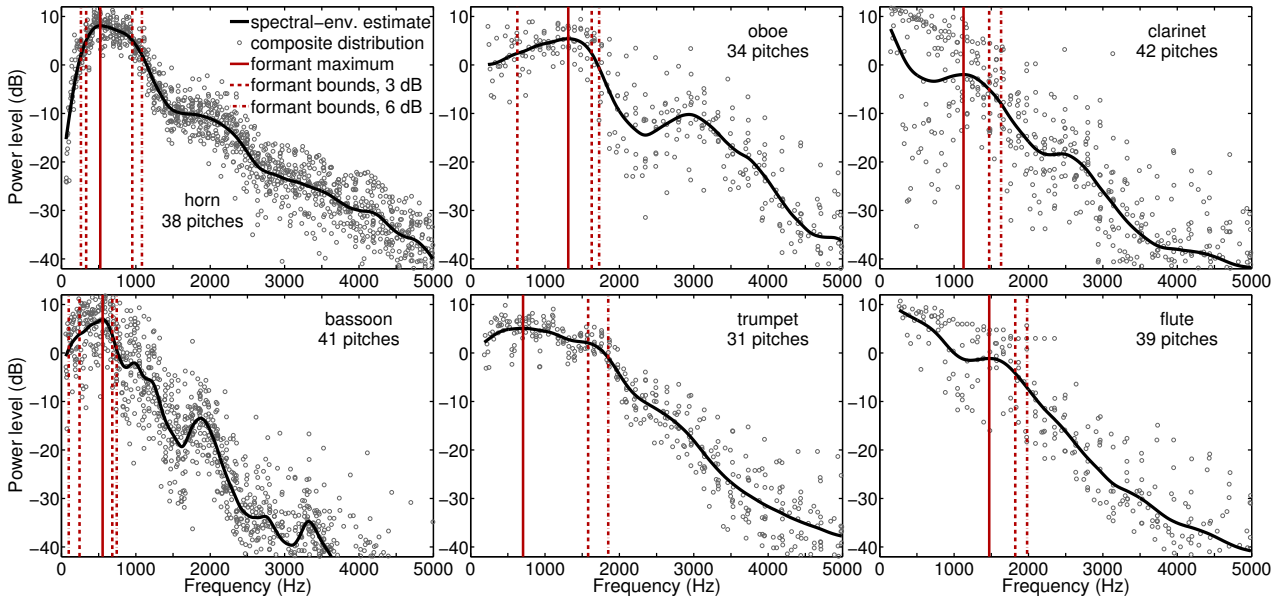

Figure 1: Estimated spectral-envelope descriptions for all six instruments (labelled in individual panels). Estimates are based on the composite distribution of partial tones compiled from the specified number of pitches across the range of each instrument.

mates. The procedure balanced the contrary aims of achieving a detailed spline fit and a linear regression, involving iterative minimization of deviations between the estimate and the composite distribution until an optimal criterion was met. These *pitch-generalized* spectral-envelope estimates then served as the basis for the identification and categorization of formants. The *main formant* represented the most prominent spectral maximum with decreasing magnitude towards both lower and higher frequencies or if not available, the most prominent spectral plateau, i.e., the point exhibiting the flattest slope along a region of decreasing magnitude towards higher frequencies. Furthermore, descriptors for the main formant F were derived from the estimated spectral envelope. They comprised the frequencies of the formant maximum F_{max} as well as *upper* and *lower* bounds (e.g., $F_{3dB}^{\rightarrow}$ and $F_{3dB}^{\leftarrow}$) at which the power magnitude had decreased by either 3 dB or 6 dB relative to F_{max}.

The spectral-envelope estimates for all investigated instruments generally suggested pitch-invariant trends, as shown in Fig. 1. A narrower spread of the partial tones (circles) around the estimate (curve) argues for a stronger pitch-invariant trend. The lower-pitched instruments horn and bassoon (left panels) exhibited strong tendencies for prominent spectral-envelope traits, i.e., formants. Higher-pitched instruments yielded two different kinds of description. Oboe and trumpet (middle panels) displayed moderately weaker pitch-invariant trends, nonetheless exhibiting main formants, with that of the trumpet being of considerable frequency extent compared to more locally constrained ones reported for the other instruments. Although still following an apparent pitch-invariant trend, the remaining instruments, clarinet and flute (right panels), displayed only weakly pronounced formant structure, with the identified formants more resembling local spectral plateaus. Furthermore, the unique acoustical trait of the clarinet concerning its low, *chalumeau* register prevented any valid assumption of pitch invariance to be made for the lower frequency range. This register is characterized by a marked attenuation of the lower even-order partials whose locations accordingly varied as a function of pitch. Fig. 1 also displays the associated formant descriptors (vertical lines), from which it can be shown that the identified main formant for the clarinet (top-right panel) was located above the pitch-variant low frequencies.

2.2 Auditory-model representation

If pitch-invariant spectral-envelope characteristics are perceptually relevant, they should also become apparent in a representation closer to perception, like the output of a computational auditory model. Using the AIM, SAIs were derived from the DCGC basilar-membrane model, comprising 50 filter channels, equidistantly spaced along equivalent-rectangular-bandwidth (ERB) rate [24] and covering the audible range up to 5 kHz[2].

[2]As band-limited analysis economized computational cost and no prominent formants above 5 kHz were found, the audio samples were sub-sampled by a factor of 4 to a sampling rate of 11025 Hz only

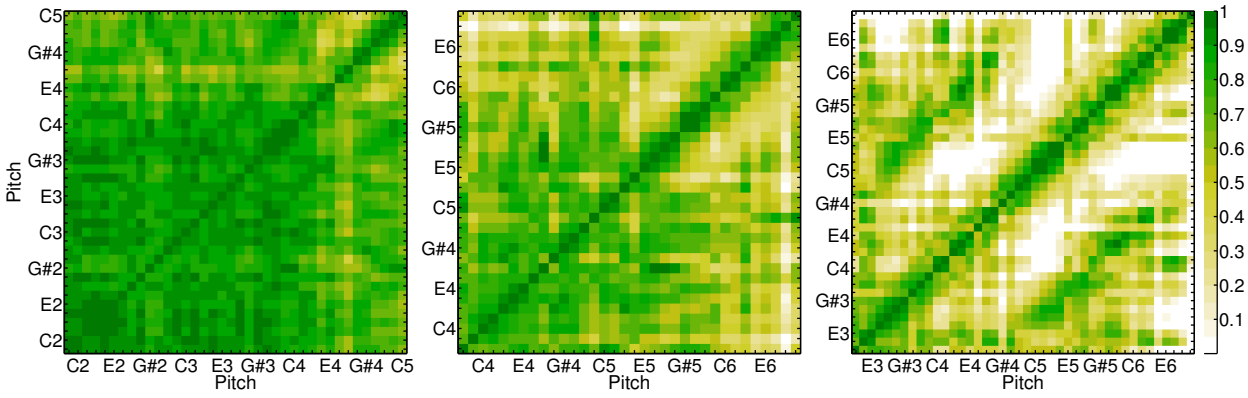

Figure 2: SAI correlation matrices for horn, oboe, and clarinet (left-to-right). Correlations (Pearson r) between SAI magnitudes across 50 filter channels consider all possible pitch combinations, with obtained r falling within $[0, 1]$ (see legend, far-right).

A time-averaged SAI magnitude profile was obtained by computing the medians across time frames per filter channel, which resembled the auditory excitation pattern [22].

A strong similarity among SAIs across an extended range of pitches was taken as an indicator for pitch-invariant tendencies. Pearson correlation matrices for all possible pitch combinations were computed, comparing the profiles of SAI magnitudes over filter channels. In addition, this approach also aided in identifying the limits of pitch invariance, as adjacent regions exhibiting weaker correlations delimited instrument registers where SAIs varied as a function of pitch. Three representative cases are illustrated in Fig. 2. For horn (left panel) and bassoon (not shown), broad regions of pitch-invariant SAI profiles became apparent (dark square), spanning large parts of their ranges up to pitches of about D4. Oboe (middle panel) and trumpet (not shown) exhibited more constrained and fragmented regions of high SAI similarity, contrasted by increasingly pitch-variant SAI profiles above A4. For these four instruments, pitch-invariant characterization appeared to be more prevalent and stable in lower pitch regions, from which low-pitched instruments in particular would benefit. All of these instruments lost pitch-invariant

for the purposes of analysis with AIM.

tendencies in their high registers. The remaining instruments, clarinet (right panel) and flute (not shown), lacked widespread pitch-invariant SAI characteristics, as strong patterns of correlation were only obtained between directly adjacent pitches (diagonal) and not across wider pitch regions.

2.3 Parametric variation of main-formant frequency

In order to study the contribution of local variations of spectral characteristics, a synthesis model was employed that provided parametric control over separate spectral-envelope components. The synthesis infrastructure relied on a source-filter model and was realized for real-time modification of the control parameters [25], based on which the spectral envelope remained invariant to pitch changes. During synthesis, the filter structure was fed a harmonic source signal of variable fundamental frequency, containing harmonics up to 5 kHz. The filter structure consisted of two independent filters, modeling the main formant on the one hand and the remaining spectral-envelope regions on the other. A parameter allowing the main formant to be shifted in frequency relative to the remaining regions was implemented as an absolute deviation ΔF in Hz from a predefined origin, i.e., $\Delta F = 0$. Analogue models for each instrument were designed for $\Delta F = 0$ by matching the frequency response of the composite filter structure to the spectral-envelope estimates, as illustrated in Fig. 3 for the horn (dashed black line), superimposed over its corresponding estimate (solid grey line). The analogues were not meant to deliver realistic emulations of the instruments per se, but rather to achieve a good fit between the main formants of the analogue and spectral-envelope estimate. Limiting differences in shape between main formants helped to deduce the measured perceptual differences that resulted from frequency relationships between them. It should be noted that the synthesis filter structure for the clarinet excluded its pitch-variant lower frequency region (see Section 2.1). It only modeled the formant above that region as well as the remaining

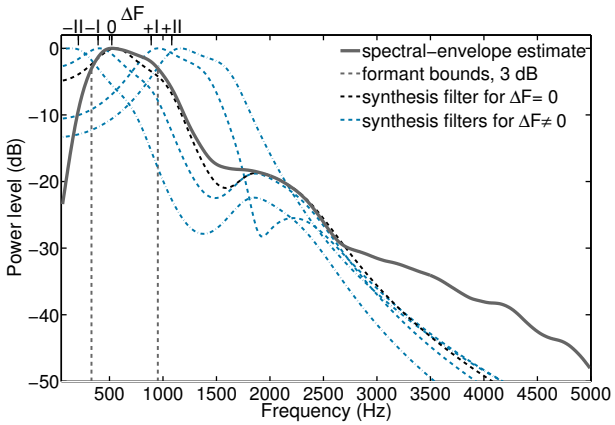

Figure 3: Spectral-envelope estimate of horn and filter magnitude responses of its synthesis analogue. The original analogue is modeled for $\Delta F = 0$; the other responses are variations of ΔF. The top axis displays the equivalent scale for the five ΔF levels investigated in Experiment B.

spectral envelope towards higher frequencies in order to orient the investigation toward specifically testing the relevance of the identified, albeit less pronounced, formant.

3 General methods

The perceptual relevance of main-formant frequency to blending was tested for sound dyads. All dyads comprised a *sampled* instrument and its *synthesized* analogue model. In a given dyad, the instrument sample was constant, and its synthesized analogue was variable with respect to the parameter ΔF. Variations with $\Delta F > 0$ shifted the main formant of the synthesized sound higher in frequency relative to the sampled instrument's main formant and, accordingly, $\Delta F < 0$ corresponded to shifts toward lower frequencies. Two perceptual experiments were conducted to investigate how ΔF variations relate to blend. In Experiment A, participants controlled ΔF directly and were asked to find the ΔF that gave optimal blend, whereas in Experiment B, listeners provided direct blend ratings for predefined ΔF variations. Using the instruments presented in Section 2, the robustness of perceptual effects was assessed over the two experimental tasks for different

pitches, unison and non-unison intervals, and stimulus contexts. Given six instruments, and various pitches and intervals, it was impractical to test all possible combinations. An exploratory approach was chosen instead, with not all instruments being tested across all factors. Whereas this could reveal blend-related dependencies concerning ΔF across the factors of interest, it did not allow generalizing across all instruments to the same degree as well as determining perceptual thresholds. Still, each factor was studied with at least three instruments for greater generalizability. Pitches were chosen to represent common registers of the individual instruments. Non-unison intervals included both smaller and larger intervals.

The methods both experiments share in common are presented in this section, before addressing their specifics and results in the following sections.

3.1 Participants

Due to the demanding experimental tasks, participants of both experiments were musically experienced listeners. They were recruited primarily from the Schulich School of Music, McGill University. Their backgrounds were assessed through self-reported degree of formal musical training, accumulated across several disciplines, e.g., instrumental performance, composition, music theory, and/or sound recording. All participants passed a standardized hearing test [26, 27]. No participant took part in both experiments.

3.2 Stimuli

All stimuli involved dyads, comprising one *sampled* (drawn from VSL) and one *synthesized* sound. For a sample at any given pitch, the spectral envelope was approximated by the pitch-generalized description from Section 2.1, which resulted in the main formants of sampled and synthesized sounds resembling each other for $\Delta F = 0$. With regard to the temporal envelope, both instruments were synchronized in their note onsets, fol-

lowed by the sustain portion and ending with an artificial 100-ms linear amplitude decay ramp applied to both instrument sounds. The sampled sound retained its original onset characteristics, whereas across all modeled analogues, the synthesized onsets were characterized by a constant 100-ms linear amplitude ramp. Stimuli were presented over a standard two-channel stereophonic loudspeaker setup inside an Industrial Acoustics Company double-walled sound booth, with the instruments simulated as being captured by a stereo main microphone at spatially distinct locations inside a mid-sized, moderately reverberant room (see [25] for details).

3.3 Procedure

Experimental conditions were presented in randomized order within blocks of repetitions. A specific condition could not occur twice in succession between blocks. The main experiments were in each case preceded by 10 practice trials under the guidance of the experimenter, to familiarize participants with the task and with representative examples of stimulus variations. Dyads were played repeatedly throughout experimental trials, allowing participants to pause playback at any time.

3.4 Data analysis

With respect to investigated factors, Experiment A evaluated the influence of the factors instrument register and interval type. Experiment B assessed pitch-invariant perceptual performance across a number of factors and furthermore correlated the perceptual data with spectral-envelope traits. Separate analyses of variance (ANOVAs) were conducted for each instrument, testing for statistically significant main effects within factors and interaction effects between them. A criterion significance level $\alpha = .05$ was chosen and, if multiple analyses on split factor levels or individual post-hoc analyses were conducted, Bonferroni corrections were applied. In repeated-measures ANOVAs, the Greenhouse-

Table 1: Seventeen dyad conditions from Experiment A across instruments, pitches, and intervals (top-to-bottom). Intervals in semitones relative to the specified reference pitch.

horn			bassoon				oboe	trumpet				clarinet				flute
C3			A2			D5	C4	C4			B5	E3			D5	C4
0	6	7	0	-2	-3	0	0	0	6	7	0	0	-2	-3	0	0

Geisser correction (ε) was applied whenever the assumption of sphericity was violated. In addition, Experiment A also considered one-sample t-tests against a mean of zero for testing differences to $\Delta F = 0$. Statistical effect sizes η_p^2 and r are reported for ANOVAs and t-tests, respectively. The analyses considered participant-based averages for trial repetitions of identical conditions.

4 Experiment A

4.1 Method

4.1.1 Participants

The experiment was conducted with 17 participants, 6 female and 11 male, with a median age of 27 years (range 20-57). Fifteen participants reported more than 10 years of formal musical training, with 10 indicating experience with wind instruments. Participants were remunerated with 15 Canadian dollars.

4.1.2 Stimuli

Table 1 lists the 17 investigated dyad conditions (column entries of bottom row). All instruments included unison intervals (0 semitones, ST). With regard to additional factors, three levels of the Interval factor compared unison intervals to consonant (7 or -3 ST) and dissonant (6 or -2 ST), non-unison intervals. Two levels of the Register factor

contrasted low (A2, C4 or E3) to high (D5 or B5) instrument registers for unison dyads, with the high-register pitches being derived from the pitch-variant regions identified in Section 2.2. The sampled sound remained at the indicated reference pitch, whereas the synthesized sound varied relative to it to form the non-unison intervals. All dyads had constant durations of 4900 ms. The level balance between instruments was variable and determined by the participant.

4.1.3 Procedure

A production task required participants to adjust ΔF directly, in order to achieve the maximum attainable blend, with the produced value serving as the dependent variable. User control was provided via a two-dimensional graphical interface, including controls for ΔF and the level balance between instruments. The slider controls for $\Delta F = f_{slider} + \Gamma$ provided a constant range of 700 Hz, with $f_{slider} \in [-350, +350]$, and including a randomized roving offset $\Gamma \in [-100, +100]$ between trials. As visualized in Fig. 4 (top), minimal or maximal Γ limited the range covered by all trials to 500 Hz (solid thick grey line), with all possible ΔF deviations spanning a range of 900 Hz (dashed thick line). Participants completed a total of 88 experimental trials (22 conditions[3] $\times$ 4 repetitions) taking about 50 minutes and including a 5-minute break after about 44 trials.

4.2 Results

4.2.1 General trends

For all six instruments, participants associated *optimal* blend with deviations $\Delta F \leq 0$. Fig. 6 (diamonds in lower part) illustrates the means for optimal ΔF, from which two different patterns become apparent among instruments. ΔF are displayed relative to a

[3]Only 17 conditions investigated ΔF; the remainder studied other formant properties that lie outside the focus of this paper and will be published separately.

scale of equivalent variations tested in Experiment B, with the scale value *0* corresponding to $\Delta F = 0$.[4] The grey lines indicate each instrument's respective slider range. For the instruments horn, bassoon, oboe, and trumpet (left panel), optimal blend was found in direct proximity to $\Delta F = 0$. For the unison intervals of horn and bassoon, ΔF did not differ significantly from zero; ΔF for the other two instruments were located slightly lower [$t(16) \leq -5.6$, $p < .0001, r \geq .82$]. By contrast, optimal ΔF for the clarinet and flute (right panel) were relatively distant from $\Delta F = 0$, in line with significant underestimations [$t(16) \leq -3.8$, $p \leq .0015, r \geq .69$]. In summary, ΔF values leading to optimal blend were limited to cases in which the formant of the synthesized instrument was at or below that of the sampled instrument.

4.2.2 Instrument register and interval type

The influence of instrument register on the optimal ΔF was investigated for trumpet, bassoon, and clarinet at pitches corresponding to instrument-specific low and high registers. One-way repeated-measures ANOVAs for Register yielded moderately strong main effects for trumpet and bassoon [$F(1, 16) \geq 19.2, p \leq .0005, \eta_p^2 \geq .55$], due to an increase of the optimal ΔF in the high register. A less pronounced effect was obtained for the clarinet [$F(1, 16) = 5.3, p = .0358, \eta_p^2 = .25$].

The investigation of possible differences in optimal ΔF between interval types involved comparisons between unison and non-unison intervals as well as a distinction between consonant and dissonant for the latter. One-way repeated-measures ANOVAs on Interval conducted for bassoon, clarinet, and trumpet only led to a weak main effect for the trumpet [$F(2, 32) = 3.7$, $p = .0347$, $\eta_p^2 = .19$]. Post-hoc tests for the three possible comparisons yielded a single significant difference between the interval sizes 0 and 6 ST [$t(16) = -3.5, p = .0033, r = .65$].

[4] ΔF were linearly interpolated to a scale of equi-distant levels, e.g., $-I$, *0*, and $+I$ corresponding to the numerical scale values -1, 0, and 1, respectively.

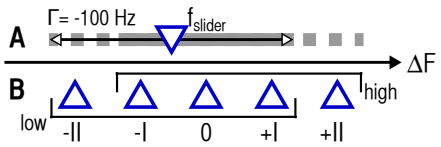

Figure 4: ΔF variations investigated in Experiments A and B. A: Participants controlled f_{slider}, which provided a constant range of 700 Hz (white arrows). Γ (e.g., -100 Hz) represented a randomized roving parameter, preventing the range from always being centered on $\Delta F = 0$. B: Participants rated four dyads varying in ΔF, drawn from the *low* or *high* context. The contexts represented subsets of four of the total of five predefined ΔF levels.

Table 2: Twenty-two dyad conditions from Experiment B across instruments, pitches, and intervals (top-to-bottom). Intervals in semitones relative to the specified reference pitch.

horn				bassoon				oboe			trumpet				clarinet				flute		
C3		B♭3		A2		D4		C4	G♯4	E5	C4		B♭4		E3		A4		C4	G♯4	E5
0	6	0	6	0	-2	0	-2	0	0	0	0	6	0	6	0	-2	0	-2	0	0	0

5 Experiment B

5.1 Method

5.1.1 Participants

The experiment was completed by 20 participants, 9 female and 11 male, with a median age of 22 years (range 18-35). Fifteen participants reported more than 10 years of formal musical training, with 11 indicating experience with wind instruments. Participants were remunerated with 20 Canadian dollars.

5.1.2 Stimuli

Table 2 lists the 22 investigated dyad conditions. The Interval factor investigated two levels, comparing unison to non-unison (6 or -2 ST, dissonant) intervals. Depending on the instrument, the Pitch factor involved two (horn, bassoon, trumpet, clarinet) or three (oboe, flute) levels. In the case of horn, bassoon, trumpet, and clarinet, there were

two levels of Interval for each level of Pitch. In addition, this experiment included two factors that were related to ΔF variations alone, which applied to all conditions listed in Table 2. The first was synonymous with ΔF, as it explored a total of five ΔF levels, including $\Delta F = 0$ and two sets of predefined moderate and extreme deviations above and below it, i.e., the ΔF levels hereafter labeled *0*, $\pm I$, and $\pm II$. The second factor grouped the five levels contextually into two subsets of four, which are denoted as *low* and *high* contexts and defined in Fig. 4 (bottom).

Employing the formant descriptors from Section 2.1, the investigated ΔF levels were expressed on a common scale of spectral-envelope description, which provided a better basis of comparison than taking equal frequency differences in Hz, as the frequency extent of formants across instruments varied considerably. Fig. 5 provides examples for all resulting ΔF levels of the horn. The four levels $\Delta F \neq 0$ were defined as frequency distances between the formant maximum F_{max} and measures related to the location and width of its bounds (e.g., $F_{6dB}^{\rightarrow}$ and ΔF_{3dB}, respectively). For example, the positive deviation $\Delta F_1(+I)$ was the distance between the formant maximum and its upper bound minus 10% of the width between the 3 dB bounds. If spectral-envelope descriptions lacked lower bounds (e.g., trumpet, clarinet, flute), the frequency located below F_{max} that exhibited the lowest magnitude was taken as a substitute value.

Unlike the dyads in Experiment A, the synthesized sound always remained at the reference pitch, whereas the sampled sound varied its pitch for non-unison intervals, because this tested the assumption of pitch-invariant description for the recorded sounds more thoroughly. The dyads had a constant duration of 4700 ms. In addition, the conditions listed in Table 2, including the associated five ΔF levels per condition, had predetermined values for the level balance between sounds and had also been equalized for loudness. The first author determined the level balance, aiming for good balance between both sounds while maintaining discriminability between ΔF levels, which was

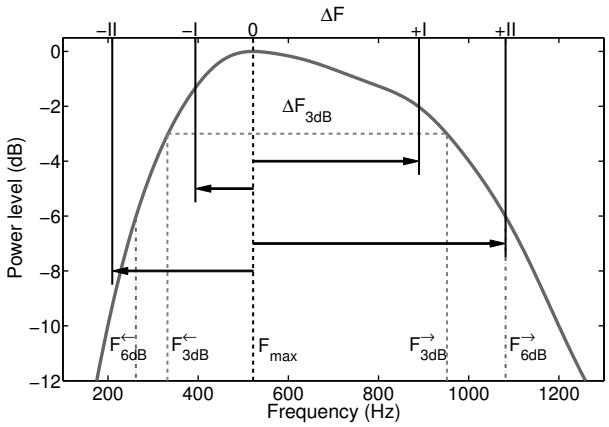

Figure 5: ΔF levels from Experiment B, defined relative to a spectral-envelope estimate's formant maximum and bounds. $\Delta F(\pm I)$ fall 10% inside of ΔF_{3dB}'s extent. $\Delta F(+II)$ aligns with $F_{6dB}^{\rightarrow}$, whereas $\Delta F(-II)$ aligns with either $80\% \cdot F_{6dB}^{\leftarrow}$ or 150 Hz, whichever is closer to F_{max}.

subsequently verified by the second author. Loudness equalization was conducted subjectively in a separate pilot experiment, anchored to a global reference dyad for all conditions and ΔF levels. For all stimuli, gain levels were determined that equalized stimulus loudness to the global reference. These gain levels were based on median values, which were determined either after their corresponding interquartile ranges fell below 4 dB or after running a maximum of 10 participants.

5.1.3 Procedure

A relative-rating task required participants to compare ΔF levels for a given condition from Table 2. In each experimental trial, participants were presented four dyads and asked to provide four corresponding ratings. The four dyads represented one of the two ΔF contexts labeled *high* and *low* in Fig. 4. A continuous rating scale was employed, which spanned from *most blended* to *least blended* (values 1 to 0, respectively) and served as the dependent variable. Participants needed to assign two dyads to the scale extremes (e.g., *most* and *least*); the remaining two dyads were positioned along the scale continuum relative to the chosen extremes. Playback could be switched freely between the

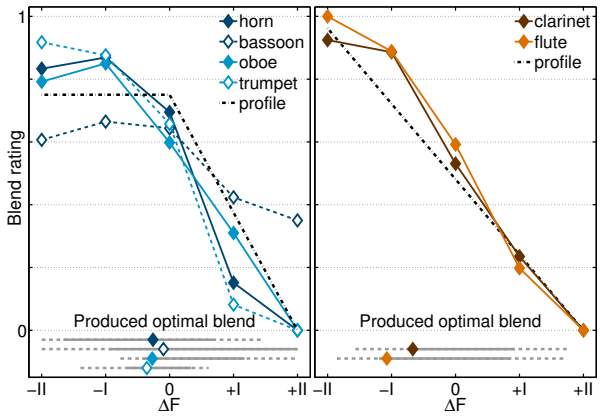

Figure 6: Perceptual results for the different instruments, grouped according to two typical response patterns (left and right panels). Experiment A (diamonds, bottom): mean ΔF for produced optimal blend, transformed to a continuous scale of ΔF levels. The grey lines indicate slider ranges (compare to Fig. 4, top). Experiment B (curves): median blend ratings across ΔF levels and idealized profile.

four dyads, with the visual order of the selection buttons and rating scales for individual dyads randomized between trials. Participants completed 120 trials (30 conditions[5] $\times$ 2 contexts $\times$ 2 repetitions) taking about 75 minutes, including two 5-minute breaks after about 40 and 80 trials.

5.2 Results

5.2.1 General trends

Group medians of ratings aggregated across the factors Pitch, Interval, and Context illustrate how blend varies as a function of ΔF. Fig. 6 suggests that participants mainly associated higher degrees of blend with the levels $\Delta F \leq 0$, whereas much lower ratings were obtained for $\Delta F > 0$. In terms of higher degrees of blend, two typical rating profiles as a function of ΔF emerged (shown as the idealized dashed-and-dotted curves in Fig. 6): 1) For the instruments horn, bassoon, oboe, and trumpet (left panel), medium

[5]Only the 22 conditions investigating ΔF are reported here.

to high blend ratings were obtained at and below $\Delta F = 0$, above which ratings decreased markedly, resembling the profile of a *plateau*. 2) The instruments clarinet and flute (right panel) exhibited a monotonically decreasing and approximately *linear* rating profile as ΔF increases, in which $\Delta F = 0$ did not appear to assume a notable role. These differences in *plateau* vs. *linear* profiles for the two instrument subsets are analyzed more closely in the following sections, also taking into account potential effects due to the other factors.

5.2.2 Blend and pitch invariance

Spectral characteristics that remain stable with pitch variation, such as formants, may have a pitch-invariant perceptual relevance. To test this, whenever the profiles of blend ratings over ΔF remained largely unaffected by different pitches, intervals, and ΔF contexts, the perceptual results were assumed to be pitch-invariant. For instance, Fig. 7 suggests this tendency for the horn, in which the *plateau* profile was maintained over all factorial manipulations. As a first step, the main effects across ΔF were tested to confirm that ratings served as reliable indicators of perceptual differences. Given these main effects, perceptual robustness to pitch variation was fulfilled if no $\Delta F \times$ Pitch or $\Delta F \times$ Interval interaction effects were found across both ΔF contexts. An absence of main effects between ΔF contexts would indicate further perceptual robustness.

As the Context factor only involved ΔF levels common to both the *high* and *low* contexts, namely *0* and $\pm I$ (see Fig. 4, bottom), the ratings for these levels required range normalization and separate analyses from the remaining factors. For the instruments involving the Interval factor, these were conducted on split levels of that factor. Furthermore, the experimental task imposed the usage of the rating-scale extremes, which resulted in several violations of normality due to skewed distributions for the dyads selected as extremes. As a result, all main and interaction effects were tested with a battery of five independent repeated-measures ANOVAs on the raw and transformed ratings. The data transformations included non-parametric approaches of rank

Table 3: Range of ANOVA main effects along ΔF across all six instruments.

		low context		high context	
Effect	Stat.	conserv.	liberal	conserv.	liberal
	F	86.0	82.6	165.1	165.1
Clarinet	df	1.6,30.7	2.1,39.1	3,57	3,57
(strong)	ε	.54	.69	-	-
	p	<.0001	<.0001	<.0001	<.0001
	η_p^2	.82	.81	.90	.90
	F	16.4	16.8	12.6	15.2
Bassoon	df	3,57	3,57	1.4,27.1	3,57
(weak)	ε	-	-	.48	-
	p	<.0001	<.0001	.0005	.0001
	η_p^2	.46	.47	.40	.44

transformation [28] and prior alignment of 'nuisance' factors [29].[6] The statistics for the most liberal and conservative p-values are reported (e.g., *conserv.*|*liberal*), with the conservative finding being assumed valid if statistical significance is in doubt.

Strong main effects were found for all instruments, which indicated clear differences in perceived blend among the investigated ΔF levels. Table 3 lists ANOVA statistics for the range between strongest (clarinet) and weakest (bassoon) main effects among the instruments, which reflects analogous differences in the utilized rating-scale ranges in Fig. 6. Furthermore, the rating profiles of the instruments horn, bassoon, oboe, and trumpet fulfilled the criteria for pitch-invariant robustness, as they appeared unaffected by pitch-related variation. There was only one exception from a complete absence of

[6]Given the unavailability of non-parametric alternatives for repeated-measures, three-way ANOVAs that include tests for interaction effects, an approach was chosen that assesses tests over multiple variants of dependent-variable transformations, presuming that the most conservative test in the ANOVA battery minimizes accepting false positives. Rank transformation is a common approach in non-parametric tests, such as the one-way Friedman test [28]. Issues with tests for interaction effects losing power in the presence of strong main effects were addressed through 'alignment' of the raw data prior to rank transformation [29]. For instance, a test for the interaction A × B would align its 'nuisance' factors by removing the main effects for A and B. The four data transformations processed the raw data with or without alignment and for two ranking methods. The first method computed *global* ranks across the entire data set, i.e., across participants and conditions, whereas the second method evaluated *within-participant* ranks across conditions per participant.

Table 4: ANOVA effects for clarinet and flute suggesting an absence of pitch invariance.

Clarinet		low context		high context	
Effect	Stat.	conserv.	liberal	conserv.	liberal
	F	2.9	3.4	3.8	4.4
ΔF	df	3,57	3,57	2.0,38.7	3,57
×					
Interval	ε	-	-	.68	-
	p	.044	.024	.030	.008
	η_p^2	.13	.15	.17	.19
Flute		low context		high context	
Effect	Stat.	conserv.	liberal	conserv.	liberal
	F	2.8	4.3	4.4	7.2
ΔF	df	3.9,75.0	6,114	6,114	6,114
×					
Pitch	ε	.66	-	-	-
	p	.031	.0006	.0005	$<.0001$
	η_p^2	.13	.18	.19	.28
	F	4.9	15.7	-	-
	df	1,19	1,19	-	-
Context[a]	ε	-	-	-	-
	p	.039	.0008	-	-
	η_p^2	.21	.45	-	-

[a]The column header *low context* does not apply in this case.

effects interacting with ΔF: a moderate main effect for Context for trumpet was found only at non-unison intervals [$F(1, 19) = 10.48|25.04$, $p = .0043|.0001$, $\eta_p^2 = .355|.569$]. By contrast, the rating profiles for clarinet and flute did not exhibit pitch invariance, as they clearly violated the criteria across both ΔF contexts. The interaction effects with ΔF and a main effect for Context leading to their disqualification are described in Table 4.

The instruments displaying pitch invariance were the same ones with plateau rating profiles, possibly attributing a special role to $\Delta F = 0$ as defining a boundary governing blend. To further support this assumption by joint analysis of the four instruments,

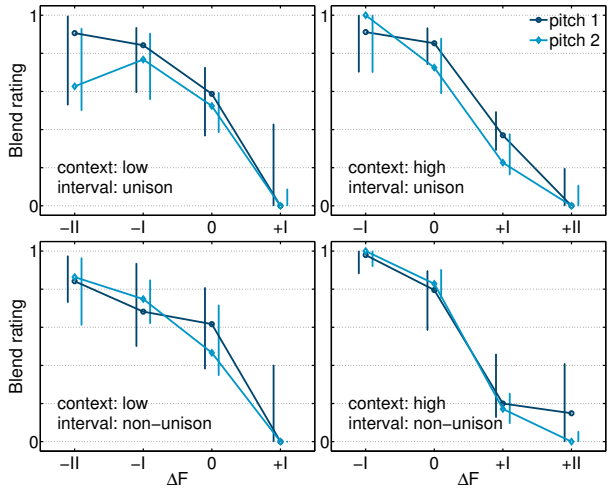

Figure 7: Medians and interquartile ranges of blend ratings for horn across ΔF levels and the factorial manipulations Pitch × Context × Interval.

two hierarchical cluster analyses were employed that grouped ΔF levels based on their similarity in perceptual ratings or auditory-model representations. The first cluster analysis reinterpreted rating differences between ΔF as a dissimilarity measure. This measure considered effect sizes of statistically significant non-parametric post-hoc analyses (Wilcoxon signed-rank test) for pairwise comparisons between ΔF levels, i.e., greater statistical effects between two ΔF levels were expressed as being more dissimilar in the perceived degree of blend. For non-significant differences, dissimilarity was assumed to be zero. The second analysis relied on correlation coefficients (Pearson r) between dyad SAI profiles across ΔF levels (see Fig. 9 for examples). The dissimilarity measure considered the complement value $1-r$, and as all correlations fall within the range $[0, 1]$, no special treatment for negative correlations was required. Both cluster analyses employed complete-linkage algorithms. The dissimilarity input matrices were obtained by averaging 30 independent data sets, aggregated across the four instruments, and the factors Context, Pitch, and Interval. As shown in Fig. 8, both analyses led to analogous solutions in which the two levels $\Delta F > 0$ are maximally dissimilar to a compact cluster associating the three levels $\Delta F \leq 0$. In other words, ΔF associated with low and high

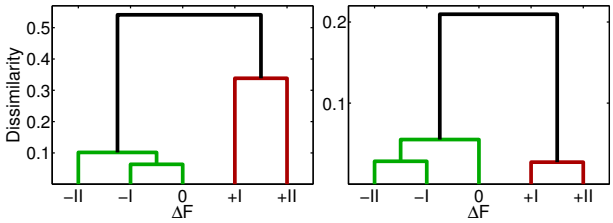

Figure 8: Dendrograms of ΔF-level groupings for the pitch-invariant instruments. Dissimilarity measures are derived from perceptual ratings (left) and auditory-modelled dyad SAI profiles (right).

degrees of blend group into two distinct clusters, clearly relating to the *plateau* profile, where $\Delta F = 0$ defines the boundary to higher degrees of blend.

5.2.3 Blend and its spectral correlates

Explaining blending between instruments with the help of spectral-envelope characteristics could eventually allow the prediction of blend through these instrument-specific traits. In addition, it could help understand the way in which spectral characteristics contribute perceptually. Given this aim, multiple linear regression was employed to model the median blend ratings through a number of variables or *regressors*. The regression models assessed the relative contributions of regressors describing both global and local spectral-envelope traits. Global descriptors involved the commonly reported *spectral centroid* and *spectral slope* [30], whereas local descriptors concerned the formant characterization discussed in Section 2.1. Because a dyad yielded two descriptor values across its constituent sounds, the regressor measure had to associate the two in some way. For the spectral centroid, two measures were considered, namely, *composite* (sum) and *absolute difference* [4]. Although the centroid composite relates to the 'darker'-timbre hypothesis mentioned in the Introduction, the centroid difference had still been found to best explain blend in non-unison intervals [4], which left some uncertainty as to which of these two measures was more appropriate in explaining blend in general. The remaining spectral regressors were implemented as polarity-preserving differences

between descriptors, with the *sampled* instrument serving as the reference (ref) and the *synthesized* instrument being variable (var) across ΔF. For example, the difference of descriptor values d_x for instrument x would correspond to $\Delta d = \ d_{\mathrm{ref}} - d_{\mathrm{var}}$.

Regression models for two separate subsets of the perceptual data were explored: pitch-invariant (horn, bassoon, oboe, trumpet) and pitch-variant instruments (clarinet, flute). The datasets comprised all conditions tested across factors and instruments, with $N = 118$ and $N = 54$ for the pitch-invariant and -variant subsets, respectively. Regressor variables were pooled from the spectral-envelope descriptors and additional variables that were included to account for potential confounding factors, e.g., pitch, interval. If these factors did not contribute as regressors, this would further support a perceptual role for pitch invariance. The initial pool of regressors comprised 32 variables, subsequently reduced to a pre-selected set that exhibited inter-variable correlations $|r| < .7$, in order to avoid pronounced collinearity among regressors. The pre-selection was determined by first identifying the variable that in simple linear regression exhibited the highest R^2 and subsequently adding all remaining variables that yielded permissible inter-variable correlations. Table 5 lists the pre-selected variables entered into the regression, which comprised spectral-envelope descriptors (nos. 1-6) and variables representing other potential factors of influence (nos. 7-12). Stepwise multiple-regression algorithms with both *forward-selection* and *backward-elimination* schemes were considered, which converged on optimum models by iteratively adding or eliminating regressors, respectively. Models with similar combinations of regressors to the optimum models were explored as well. In anticipation of reporting the results, the inclusion of a binary regressor for ΔF context $C_{lo/hi}$ benefited all investigated regression models, as it corrected the systematic offset of scaled ratings between the low and high contexts (see Fig. 7).

In simple linear regression, the strongest spectral-envelope descriptors all concerned local formant characterization and did not involve the global descriptors. Among the

Table 5: Variables entering stepwise-regression algorithm to obtain models reported in Table 6.

No.	Variable	Description
1	$\Delta L_{3dB}^{\rightarrow}$	derivate of $F_{3dB}^{\rightarrow}$, Equation 1
2	$\Delta L_{F_1 vs F_2}$	ΔL betw. formants F_1 & F_2
3	$\Delta S_{slope_{F1}}{}^{ab}$	spectral slope above $F_{3dB}^{\rightarrow}$
4	$\Delta S_{slope}{}^{b}$	global spectral slope
5	$\lvert\Delta S_{centroid}\rvert^{\ b}$	absolute centroid difference
6	$\sum S_{centroid}{}^{b}$	centroid composite
7	$ERB_{rate}{}^{c}$	reference pitch in Table 2
8	$I_{(non)unison}$	interval category (binary)
9	$I_{ST}{}^{a}$	interval size in semitones
10	$C_{lo/hi}$	ΔF context (binary)
11	mix_{ratio}	balance betw. instruments
12	$AM_{depth}{}^{bd}$	amplitude modulation depth

[a] Not for pitch-variant subset, inter-variable correlation $|r| > .7$
[b] Computed as described in [30]
[c] Accounting for pitch
[d] Accounting for perceivable beating between partial tones

formant descriptors, the highest correlations were obtained for the main-formant upper bound $F_{3dB}^{\rightarrow}$, applied to both pitch-invariant [$R^2(118) = .656$, $p < .0001$] and pitch-variant subsets [$R^2(54) = .713$, $p < .0001$]. Notably, the formant maximum F_{max} did not perform better than $F_{3dB}^{\rightarrow}$, likely due to differing skewness properties between instrument formants (see Fig. 1). At the same time, the utility of $F_{3dB}^{\rightarrow}$ implies that it could assume an important role in explaining blend, as perhaps the perceptually most salient feature of formants. It performed slightly better for the pitch-variant than for the pitch-invariant subset, because $\Delta F_{3dB}^{\rightarrow}$ essentially follows a strictly monotonic function across the investigated ΔF levels, which apparently models the *linear* blend profile better. To improve the modeling for the *plateau* blend profile, derivate descriptors of $F_{3dB}^{\rightarrow}$ were explored. The most effective derivate $\Delta L_{3dB}^{\rightarrow}$ related the upper-bound frequencies $F_{3dB}^{\rightarrow}$ of the two instruments to a difference in the reference instrument's spectral-envelope magnitude $L_{\mathrm{ref}}(F_{3dB}^{\rightarrow})$, as formalized in Equation 1. In other words, this measure evaluated magnitude differences relative to the reference instrument, at frequencies appearing to be of particular perceptual relevance, which therefore may relate to spectral-masking effects (e.g., *incomplete masking* [13]).

$$\Delta L_{3dB}^{\rightarrow} = L_{\mathrm{ref}}\left(F_{3dB|\mathrm{ref}}^{\rightarrow}\right) - L_{\mathrm{ref}}\left(F_{3dB|\mathrm{var}}^{\rightarrow}\right) \tag{1}$$

The obtained solutions from stepwise multiple regression yielded identical models for both instrument subsets, involving the regressors $\Delta L_{3dB}^{\rightarrow}$, absolute spectral-centroid difference $|\Delta S_{centroid}|$, and context $C_{lo/hi}$. A slight gain in performance was achieved by substituting the $|\Delta S_{centroid}|$ based on audio signals from individual sounds with a variant computed on the pitch-generalized spectral-envelope estimates. Table 6 displays these optimized regression models for pitch-invariant and pitch-variant subsets, both leading to about 87% explained variance (adjusted R^2). The standardized regression-slope coefficients β_{std} indicate the relative contribution of regressors, with the relative weights

Table 6: Multiple-regression models best predicting timbre-blend ratings for two instrument subsets.

Pitch-invariant subset		R^2_{adj}=.87	
			$F(3, 116)$=272.4, p <.0001
Regressors	β_{std}	t	p
$\Delta L^{\rightarrow}_{3dB}$	1.00	28.6	<.0001
$\lvert\Delta S_{centroid}\rvert$	.26	7.7	<.0001
$C_{lo/hi}$	.27	7.9	<.0001
Pitch-variant subset		R^2_{adj}=.88	
			$F(3, 52)$=134.2, p <.0001
Regressors	β_{std}	t	p
$\Delta L^{\rightarrow}_{3dB}$	1.03	20.0	<.0001
$\lvert\Delta S_{centroid}\rvert$	.16	3.3	.0018
$C_{lo/hi}$	.34	6.8	<.0001

being very similar for both subsets. In these models, $\Delta L^{\rightarrow}_{3dB}$ acted as the strongest predictor for the blend ratings, contributing about five times more than $|\Delta S_{centroid}|$, which furthermore did not perform better than $C_{lo/hi}$. These findings clearly argue for local spectral-envelope descriptors to be more meaningful than global ones in explaining blending in the investigated dyads. Moreover, the remaining global descriptor *spectral slope* appeared to play no role. Furthermore, finding both instrument subsets to be modeled equally well through the same spectral-envelope descriptors points to a general utility of pitch-generalized descriptions for all instruments. Despite the findings in Section 5.2.2 arguing against pitch-invariant perceptual robustness for clarinet and flute, the obtained regression models excluded the Pitch and Interval variables, thus appearing to be less relevant to explaining the blend ratings.

6 General discussion

Orchestrators would benefit from acoustical descriptions of instruments that correspond to the perceptual processes involved in achieving blended timbres. Section 2 suggests that common orchestral wind instruments are reasonably well described through pitch-generalized spectral-envelope estimates, which furthermore show the instruments horn, bassoon, oboe, and trumpet to be characterized by prominent formant structure. Auditory models employing stabilized auditory images (SAI) confirm that for strong formant characterization and for lower to middle pitch ranges, the pitch-invariant characterization is stable. In higher instrument registers, however, SAI profiles indicate limitations to pitch-invariant characterization. Other instruments, like clarinet and flute, yield SAI profiles clearly varying as a function of pitch, implying that this pitch dependency may also extend to perception.

The perceptual investigation in Sections 3 to 5 confirms the acoustical implications, showing that strong formant characterization results in main formants becoming perceptually relevant to blending. Given a dyad in which a putative main formant is variable in frequency relative to a fixed reference formant, the investigated instruments display two archetypical profiles based on their formant prominence. For the pitch-variant clarinet and flute, blend increases as a monotonic, quasi-*linear* function if the variable formant moves from above to below the reference. For pitch-invariant instruments, the frequency alignment between the variable formant and the reference ($\Delta F = 0$) functions as a boundary, delimiting a region of higher degrees of blend at and below the reference and contrasted by a marked decrease in blend when the variable formant exceeds it, which overall resembles a *plateau* profile. The pitch invariance even extends to different interval types, as the *plateau* profile remains unaffected in non-unison intervals, regardless of their degree of consonance. However, the findings suggest that the perceptual relevance of spectral-envelope estimates diminishes in higher instrument registers. The

limited sampling of conditions across the investigated factors prevented a more comprehensive evaluation of all instruments. Whereas the findings allow general relationships for formant frequency to be deduced, more comprehensive investigations are needed to attain more generalizable quantification of absolute frequency ranges and thresholds.

In correlating acoustical and perceptual factors, spectral-envelope characteristics alone explain up to 87% of the variance in blend ratings. In addition, local spectral traits seem to be more powerful acoustical predictors of blend than global traits. The formant descriptor for the upper formant bound $F_{3dB}^{\rightarrow}$, when expressed as a derivate descriptor $\Delta L_{3dB}^{\rightarrow}$, acts as the strongest predictor for the blend ratings, regardless of whether instruments belong to the pitch-invariant group or not. With regard to clarinet and flute, the departures from pitch invariance found in Section 5.2.2 contradict the utility of pitch-generalized spectral-envelope description in predicting blend ratings, as reported in Section 5.2.3. Taking both findings into account, this for one argues that the descriptor $F_{3dB}^{\rightarrow}$ still succeeds in explaining blend well even for clarinet and flute. On the other hand, the same instruments display a greater perceptual sensitivity to the Pitch and Interval factors, likely associated with their less pronounced formant structure. Overall, strong formant prominence leads to more drastic changes in blend.

The prediction of blend using $\Delta L_{3dB}^{\rightarrow}$ still presumes that one of the instruments serves as a reference formant, as the employed difference descriptors are anchored to the *sampled* instrument. The dependence on a reference leaves some ambiguity, because an arbitrary combination of two instruments would lead to contradictory predictions of blend if both instruments were given equal importance in serving as the reference. Given the context of both experiments, it can be assumed that the sampled instrument, acting as a constant anchor, had been biased into serving as the reference by combining it with a variable synthesized instrument. In addition, a possible perceptual explanation could concern audio samples of instruments playing non-vibrato generally still exhibit-

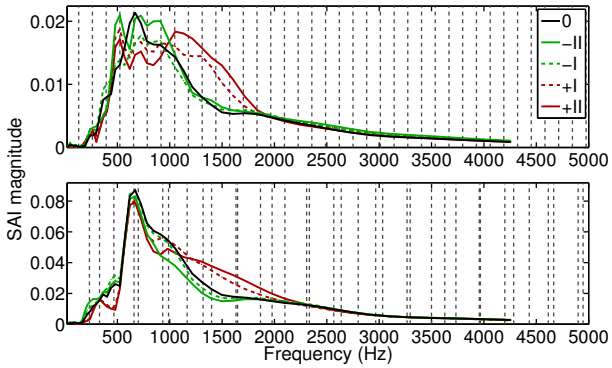

Figure 9: SAI profiles of dyads for all ΔF levels (Experiment B), depicting two experimental conditions for horn. Top: pitch 1, unison; bottom: pitch 2, non-unison; the grid lines correspond to partial-tone locations.

ing coherent micro-modulations of partial tones. These modulations have been shown to contribute to a stronger and more stable auditory image [31] and may thus bias the more stable image toward acting as the reference, especially as the synthesized partials remain static over time. Even in the context of blending in musical performance, one instrument assumes the role of the leading voice, in which it possibly serves as the reference, whereas an accompanying instrument avoids exceeding the lead instrument's main-formant frequency. Likewise, returning to the notion of blend leading to *augmented* timbres [4], the dominant timbre to be embellished by another would seem predestined to function as the reference, based on which the less dominant timbre should not exhibit formant frequencies exceeding those of the reference.

Finally, the results allow a reassessment of previous explanations for blend. The 'darker'-timbre hypothesis [4] is directly reflected in the obtained *linear* blend profile, in which lower ΔF increases blend and at the same time causes a decrease in the spectral-centroid composite. By contrast, this hypothesis is not well explained by the *plateau* profile, as blend ratings remain similarly high for $\Delta F \leq 0$. The alternative hypothesis of coincidence of formant regions [3] would have predicted stronger blend ratings for $\Delta F = 0$ than for all other levels, which in the perceptual results is only achieved for the

levels $\Delta F > 0$. While the hypothesis with respect to spectral variations ΔF only achieves partial fulfillment, it shows more agreement in the corresponding SAI representations. As shown in two example cases for horn in Fig. 9, the dyad SAI profiles for the levels $\Delta F > 0$ are distinguishable from the remaining levels through clear deviations between 1 and 2 kHz[7] and located just above the horn's estimated $F_{3dB}^{\rightarrow}$. Remarkably, the formant shifts related to $\Delta F < 0$ (Fig. 3) are not reflected in the corresponding dyad SAI profiles (Fig. 9), which instead exhibit direct alignment below 500 Hz for all three levels $\Delta F \leq 0$. Therefore, only $\Delta F > 0$ seems to lead to *incomplete masking* [13], revealing the identity of the synthesized instrument, whereas the spectral-envelope variations $\Delta F < 0$ evoke little change compared to the dyad excitation pattern for $\Delta F = 0$.

Of still greater importance, the auditory system, as modeled by AIM using the DCGC, seemingly involves a high-pass characteristic that attenuates spectral-envelope regions below 500 Hz, affecting the perceived magnitudes of the respective partial tones (grid lines). This implies that in the region below 500 Hz, frequency deviations between main formants no longer affect the achieved degree of blend, as reflected both in Fig. 9 and the perceptual findings. Horn and bassoon would especially benefit from this, as their main formants are centered around 500 Hz. Oboe and trumpet, both exhibiting higher $F_{3dB}^{\rightarrow}$, can be assumed to benefit to a lesser degree. In summary, main formants located in proximity to 500 Hz will benefit more, on top of which lower pitches would also increase the number of partial tones falling into this favorable frequency region. This reflects tendencies for pitch-invariant traits in SAI correlations (see Section 2.2) to be more pronounced at lower pitch ranges and for instruments of lower register, which would lend support to the 'darker-timbre' hypothesis in terms of limiting the spectral centroid

[7] Concerning the output from the AIM, a misalignment between actual sinusoidal frequencies and the corresponding SAI peaks was observed. Through personal communication with the developer of the utilized AIM implementation, this was explained as being an inherent property of the dynamic-compression filters. A correction function was derived by computing SAIs for various sinusoidal frequencies and fitting the two frequency scales, yielding the linear function $f_{\mathrm{SAI}} = 1.17 \cdot f + 28.2$ Hz $[r^2(50) = .999, p < .0001]$. In Fig. 9, the correction manifests itself in the compressed frequency extent for the SAIs.

in frequency.

7 Conclusion

Evidence from acoustical and psychoacoustical descriptions of wind instruments and from perceptual validation shows that relative location and prominence of main formants affect the perception of timbre blend critically. Furthermore, these pitch-invariant spectral characteristics explain and predict the perception of blend to a promisingly high degree. Remaining discrepancies between the acoustic and perceptual domains can be explained through apparent constraints of the simulated auditory system. In conclusion, a perceptual model for the contribution of local spectral-envelope characteristics to blending is proposed, keeping in mind that it would serve as an instrument-specific component in a more complex, general perceptual model involving compositional and performance factors, as initially discussed in Section 1. The main factors influencing the perception of spectral blend are summarized:

1. Frequency relationships between upper bounds of main formants are critical to blend. Among several instruments, one is expected to serve as the reference (e.g., lead instrument, dominant timbre), above which the presence of other instruments' formants would strongly result in decreased blend.

2. Prominence of the main formants governs whether these relationships lead to *plateau* or *linear* blend profiles, and in the first case pitch-invariant perceptual robustness extends to non-unison intervals.

3. Spectral-envelope relationships below 500 Hz may be negligible, due to constraints of the auditory system. At the same time, blend decreases at higher pitches due to a degraded perceptual robustness of formants.

This hypothetical model still requires further investigation concerning a more systematic study on 1) the apparent constraints of the auditory system as modeled by AIM, 2) how in musical practice one instrument may function as the reference, 3) establishing a more specific description of formant prominence, and 4) addressing the contribution of loudness balance between instruments to blend. These future investigations will further validate and refine the proposed perceptual model and will improve computational prediction tools for the instrument-specific, spectral component of blend. Orchestration practice will benefit from these research efforts even beyond aiming for blend, as knowledge of favorable instrument relationships also informs orchestrators as to how to avoid it.

References

[1] G. J. Sandell: Concurrent timbres in orchestration: a perceptual study of factors determining blend. PhD dissertation. Northwestern University, 1991.

[2] R. A. Kendall, E. C. Carterette: Identification and blend of timbres as a basis for orchestration. Contemp. Music Rev. **9** (1993) 51–67.

[3] C. Reuter: 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). P. Lang, Frankfurt am Main, 1996, 339.

[4] G. J. Sandell: Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration. Music Percept. **13** (1995) 209–246.

[5] A. S. Bregman: Auditory scene analysis: the perceptual organization of sound. MIT Press, Cambridge, MA, 1990, 519–521.

[6] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, J. Krimphoff: Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol. Res. **58** (1995) 177–192.

[7] L. Wedin, G. Goude: Dimension analysis of the perception of instrumental timbre. Scand. J. Psychol. **13** (1972) 228–240.

[8] J. M. Grey, J. W. Gordon: Perceptual effects of spectral modifications on musical timbres. J. Acoust. Soc. Am. **63** (1978) 1493–1500.

[9] W. Strong, M. Clark: Synthesis of wind-instrument tones. J. Acoust. Soc. Am. **41** (1967) 39–52.

[10] D. Tardieu, S. McAdams: Perception of dyads of impulsive and sustained instrument sounds. Music Percept. **30** (2012) 117–128.

[11] L. Hermann: Phonophotographische Untersuchungen (Phonophotographic investigations). Pflüger, Archiv für die Gesammte Physiologie des Menschen und der Thiere **58** (1894) 264–279.

[12] C. Stumpf: Die Sprachlaute (The speech sounds). Springer, Berlin, 1926.

[13] J. P. Fricke: Zur Anwendung digitaler Klangfarbenfilter bei Aufnahme und Wiedergabe (On the application of digital timbre-filters during recording and playback). Proc. 14th Tonmeistertagung, Munich, Germany, 1986, 135–148.

[14] K. E. Schumann: Physik der Klangfarben (Physics of timbres). Professorial dissertation. Universität Berlin, Berlin, 1929.

[15] E. L. Saldanha, J. F. Corso: Timbre cues and the identification of musical instruments. J. Acoust. Soc. Am. **36** (1964) 2021–2026.

[16] W. Strong, M. Clark: Perturbations of synthetic orchestral wind-instrument tones. J. Acoust. Soc. Am. **41** (1967) 277–285.

[17] D. Luce, J. Clark: Physical correlates of brass-instrument tones. J. Acoust. Soc. Am. **42** (1967) 1232–1243.

[18] D. A. Luce: Dynamic spectrum changes of orchestral instruments. J. Audio Eng. Soc. **23** (1975) 565–568.

[19] J. Meyer: Acoustics and the performance of music. 5th ed. Springer, New York, 2009, 438.

[20] G. Fant: Acoustic theory of speech production. Mouton, The Hague, 1960.

[21] C. Reuter: Klangfarbe und Instrumentation: Geschichte – Ursachen – Wirkung (Timbre and instrumentation: history – causes – effect). P. Lang, Frankfurt am Main, 2002, 584.

[22] R. van Dinther, R. D. Patterson: Perception of acoustic scale and size in musical instrument sounds. J. Acoust. Soc. Am. **120** (2006) 2158–2176.

[23] T. Irino, R. D. Patterson: A dynamic compressive gammachirp auditory filterbank. IEEE Trans. Audio Speech Lang. Process. **14** (2006) 2222–2232.

[24] B. C. Moore, B. R. Glasberg: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. **74** (1983) 750–753.

[25] S.-A. Lembke, S. McAdams: A spectral-envelope synthesis model to study perceptual blend between wind instruments. Proc. 11th Congrès Français d'Acoustique / IOA annual meeting / Acoustics 2012, Nantes, France, 2012, 1025–1030.

[26] F. Martin, C. Champlin: Reconsidering the limits of normal hearing. J. Am. Acad. Audiol. **11** (2000) 64–66.

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

[28] W. J. Conover, R. L. Iman: Rank transformations as a bridge between parametric and nonparametric statistics. Am. Stat. **35** (1981) 124–129.

[29] J. J. Higgins, S. Tashtoush: An aligned rank transform test for interaction. Nonlinear World **1** (1994) 201–221.

[30] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, S. McAdams: The Timbre Toolbox: extracting audio descriptors from musical signals. J. Acoust. Soc. Am. **130** (2011) 2902–2916.

[31] S. McAdams: Spectral fusion, spectral parsing and the formation of auditory images. PhD dissertation. Stanford University, 1984.