When timbre blends musically: perception and acoustics underlying orchestration and performance

Sven-Amin Lembke

Music Technology Area, Department of Music Research
Schulich School of Music, McGill University
Montreal, Canada

December 2014

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

2014/12/11

Abstract

Blending or contrasting instrumental timbres are common techniques employed in orchestration. Both bear a direct relevance to the perceptual phenomenon of auditory fusion, which in turn depends on a series of acoustical cues. Whereas some cues relate to musical aspects, such as timing and pitch relationships, instrumentation choices more likely concern the acoustical traits of instrument timbre. Apart from choices made by composers and orchestrators, the success of timbre blending still depends on precise execution by musical performers, which argues for its relevance to musical practice as a whole.

This thesis undertakes a comprehensive investigation aiming to situate timbre blend in musical practice, more specifically addressing the perceptual effects and acoustical factors underlying both orchestration and performance practice. Three independent studies investigated the perception of blend as a function of factors related to musical practice, i.e., those derived from musical context and realistic scenarios (e.g., pitch relationships, leadership in performance, room acoustics).

The first study establishes generalized spectral descriptions for wind instruments, which allow the identification of prominent features assumed to function as their timbral signatures. Two listening experiments investigate how these features affect blend by varying them in frequency, showing a critical perceptual relevance. The second study considers two other listening experiments, which evaluate perceived blend for instrument combinations in dyads and triads, respectively. Correlational analyses associate the obtained blend measures with a wide set of acoustic measures, showing that blend depends on pitch and temporal relationships as well as the previously identified spectral features. The third study extends the previous ones, addressing factors related to musical performance by investigating the timbral adjustments performers employ in blending with one another, as well as their interactive relationship. Timbral adjustments can be shown to be made towards the musician leading the performance.

All studies contribute to a greater understanding of blend as it applies to musical and orchestration practice. Their findings expand previous research and provide possible explanations for discrepancies between hypotheses made in the past. Together, the conclusions drawn allow us to propose a general perceptual theory for timbre blend as it applies to musical practice, which considers the musical material and spectral relationships among instrument timbres as the determining factors.

Résumé

Fusionner ou différencier les timbres instrumentaux sont des techniques d'orchestration courantes. Elles présentent toutes deux un intérêt direct pour le phénomène de fusion auditive, qui dépend d'une série d'indices acoustiques. Alors que certains indices sont liés aux aspects musicaux comme la synchronisation ou les relations de hauteurs perçues, les choix d'instrumentation sont davantage liés aux traits acoustiques du timbre de l'instrument. En plus des choix faits par les compositeurs et les orchestrateurs, le succès de la fusion des timbres tient de la précision de l'exécution des instrumentistes, ce qui renforce encore sa pertinence pour la pratique musicale en général.

Cette thèse présente une étude approfondie de la place de la fusion des timbres dans le jeu musical, et s'intéresse plus particulièrement aux effets perceptifs et aux facteurs acoustiques sous-jacents à l'orchestration et à la pratique instrumentale. Trois études indépendantes ont été conduites pour étudier la perception de la fusion en fonction de facteurs liés à la pratique musicale, c'est-à-dire, découlant du contexte musical et de scénarios réalistes comme les relations entre les hauteurs perçues, le leadership pendant le jeu, l'acoustique de la salle.

La première étude propose des descriptions spectrales généralisées pour les instruments à vent, ce qui permet l'identification des descripteurs les plus importants pouvant représenter leur signature de timbre. Deux tests d'écoute étudient leur influence sur la fusion en les faisant varier en fréquence, ce qui démontre leur pertinence sur le plan perceptif. La seconde étude est fondée sur deux autres tests d'écoute ayant pour but d'évaluer la fusion perceptive lors de combinaison d'instruments, respectivement présentés en dyade et en triade. Des analyses de corrélation montrent une association entre les mesures obtenues sur la fusion et de nombreuses mesures acoustiques, et montrent que la fusion dépend de la hauteur et des relations temporelles mais également des caractéristiques spectrales identifiées précédemment. La troisième étude complète les précédentes en ce sens qu'elle s'intéresse aux facteurs liés à la performance musicale en étudiant les ajustements de timbre auxquels les musiciens ont recours lorsqu'ils cherchent à fusionner leurs jeux, et comment ces ajustements sont interdépendants. Il est possible de montrer que ces ajustements de timbre sont exécutés en fonction du musicien qui guide la performance.

Toutes ces études contribuent à une meilleure compréhension de la fusion, appliquée au jeu musical et à l'orchestration. Les résultats obtenus permettent de compléter les recherches existantes sur le sujet en ce sens qu'ils apportent des explications possibles aux divergences existant entre les différentes hypothèses formulées par le passé. Finalement, les conclusions de cette thèse permettent d'établir une théorie perceptive générale pour la fusion de timbre en contexte musical, qui pose le matériel musical et les relations spectrales entre timbres instrumentaux comme facteurs déterminants.

Acknowledgments

This thesis would not have been possible without the assistance of many helpful people. Firstly, I am very grateful to Stephen McAdams for supporting my wish to pursue a doctorate under his supervision. I have greatly benefitted from his intellectual and inspirational guidance, his methodological rigor and skepticism, as well as his encouragement of individuality in defining research projects. Next, the Music Perception and Cognition Lab has been an enormously fertile ground, owing to the diverse range of disciplines its research talents come from. The lab environment has been extremely beneficial to the development of research ideas, through at all times constructive criticism and the selfless bravery of its members piloting early stages of experiments. I would like to thank Song Hui Chon and Kai Siedenburg as my immediate 'timbre peers' as well as Bruno Giordano, Hauke Egermann, Nils Peters, Michel Vallières, Meghan Goodchild, David Sears, Cecilia Taher, Yinan Cao, Chelsea Douglas, Jason Noble, and others for always lending an open ear for such diverse issues as Max/MSP, statistics, and music theory. I am indebted to my former *co-bureau*, Indiana Wollman, for translating the abstract to French, and rekindling my violin playing in times when research on wind instruments dominated my life. I would also like to thank my former undergraduate research assistant Kyra Parker for playing a crucial role in the success of Experiment 3, and Eugene Narmour for allowing me to see alternative notions of blend through Experiment 4. I also acknowledge Jamie Webber for his compilation of blend-related citations across various orchestration treatises, which proved a helpful aid in contextualizing my findings into orchestration practice. I would like to again thank Kyra Parker and also Emma Kast for running participants for Experiment 4. And last but certainly not least, encountering Bennett Smith's office door open has always been a blessing, as that would mean technical issues would find their resolution. I also thank him for programming the software for Experiments 3 and 4 and, expressly, for his well-appreciated sense of humor.

The lively and bright environment the Music Technology Area represents has played another important role. As one of the professors, I thank Philippe Depalle for his always helpful signal-processing advice and insider knowledge into certain mysteries of AudioSculpt. Among my peers, I would like to acknowledge Bertrand Scherrer for the hundreds of times his partial-tone detection algorithm came to use, Jason Hockman for sharing his tools for automated detection of note onsets, Charalampos Saitis for graduating just a year ahead

of me and guiding me around potential pitfalls, and IDMIL for the many lusophone friends I have made over the years. Writing this thesis has been rather smooth, and I owe it to the persons introducing me to LaTeX and those improving the experience, namely, Mark Zadel, Finn Upham, and Marcello Giordano.

Across the road at the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), I would first like to thank Scott Levine, formerly with the Sound Recording Area, whose technical contribution to Experiment 5 has been invaluable to its success, which reached a scope of complexity impossible to realize alone and in such a short time. I would also like to thank Martha de Francisco for her co-supervision of Scott's and my research project, as well as serving as the second reader to this thesis. CIRMMT's technical team has always been a great help, for which Harold Kilianski, Yves Méthot, and Julien Boissinot receive my gratitude. Similarly, its administrative workers, Jacqui Bednar and Sara Gomez, have always been very helpful and dedicated. Just a floor below CIRMMT, at the music school's administration, Hélène Drouin has always proved to have the answers to any unclarity in formal matters concerning the doctoral degree. From the institutional side, I would like to acknowledge the Schulich School of Music and CIRMMT for student scholarships, fellowships, and awards over the years. Lastly, I also would like to acknowledge professors Denys Bouliane and John Rea for allowing me to audit their courses on orchestration, which served as an eye-opener to where timbre really fits in musically, and also Jacqueline Leclair for her kind assistance in finding wind-instrument players.

Before moving to Montréal I had two homes on two different continents; in the meantime, it has become three on three. In all these homes and beyond, I would like to thank my friends and family for their support and company and for always being there when needed.

Contribution of authors

This is a *manuscript*-based thesis. Its core chapters comprise research articles formatted for publication in scientific journals. They either have already been submitted or are in preparation for submission.

- Chapter 2: Lembke, S.-A. and McAdams, S. (Under review). The role of local spectral-envelope characteristics in perceptual blending of wind-instrument sounds. *Acta Acustica united with Acustica*.
- Chapter 3: Lembke, S.-A., Parker, K., Narmour, E. and McAdams, S. (In preparation). Acoustical correlates of perceptual blend in timbre dyads and triads. *Journal of the Acoustical Society of America*.
- Chapter 4: Lembke, S.-A., Levine, S. and McAdams, S. (In preparation). Blending between bassoon and horn players: an analysis of timbral adjustments during musical performance. *Music Perception*.

Among the co-authors, Stephen McAdams functioned as the thesis supervisor as well as the director of the laboratory in which all of the research was conducted. His contribution concerns all stages of research, i.e., overseeing the conception and design of experimental paradigms, the discussion of analysis approaches, and interpretation of results, as well as financing the research with regard to technical facilities and the remuneration of participants. Kyra Parker was an undergraduate research assistant, who under my supervision helped design, conduct, and analyze Experiment 3 and during a following summer internship also initiate the regression analysis reported in Chapter 3. Eugene Narmour motivated the conception of the design for Experiment 4, which allowed the experiment to be designed in such a way as to allow the study of blend in triads. Scott Levine, a former master's student in the Sound Recording program, and I were awarded student funding from CIRMMT, to work on a project related to Chapter 4. His contribution involved the conception of the virtual performance environment, its realization (e.g., recording of RIRs, real-time convolution), and supervision of the technological aspects of Experiment 5. My contribution as principal author involves the initial conception and design of all experiments reported in Chapters 2, 3, and 4, conducting all reported acoustical and statistical analyses, as well as authoring all parts of this thesis.

Contents

List of Figures

List of Tables

List of Acronyms

AIM	Auditory Image Model
ANOVA	Analysis of variance
ASA	Auditory scene analysis
CV	Coefficient of variation
dB	Decibel
DV	Dependent variable
DCGC	Dynamic, compressive gammachirp
FFT	Fast Fourier transform
HRTF	Head-related transfer function
Hz	Hertz
MDS	Multidimensional scaling
MLR	Multiple linear regression
PC	Principal component
PCA	Principal components analysis
PLSR	Partial least-squares regression
RIR	Room impulse response
RMS	Root-mean-square
SAI	Stabilized Auditory Image
STFT	Short-time Fourier transform
TE	True envelope
VSL	Vienna Symphonic Library
XC	Cross-correlation coefficient

Chapter 1

Introduction

Writing music for the orchestra seems like a most liberating venture, given its sheer unlimited possibilities of expression. It may, however, quickly turn out to be an equally challenging endeavor. This collective of fifty to a hundred musicians confronts composers with myriad variables, requiring decisions spanning all parameters of musical expression. The challenge lies not in the musical material, where ideas are expressed in rhythms and across pitches, because, if not venturing into a cacophony of seventy-part counterpoint, the musical material usually yields a manageable number of musical voices. Trying to expand this limited number onto an orchestral *tutti* is the actual feat, because in order to maintain the clarity of the musical ideas, the individual voices will have to be replicated at various levels of the musical texture, by unison or octave doubling, melodic coupling or chordal expansion. And important questions arise: Which instruments should be paired to achieve a less 'sharp' timbre for the melodic line? Would a chordal passage sound more homogeneous if a certain combination of instruments were chosen? How can the following musical idea be better distinguished from its antecedent? The experienced orchestrators will seek the answers by relying on their extensive knowledge of instrument *timbre*. Alluding to its synonym *tone color*, or its German term *Klangfarbe*, we draw a visual analogy, with the problem confronting orchestrators illustrated in Figure 1.1. We are given a texture of several instruments, overlapping in space (or time), yielding various combinations of their respective tone colors. While the yellow and red instruments 'blend' into their complementary orange, the pairing of blue and yellow renders both quite distinct, yielding a 'contrasting' mixture. *Blend* and *contrast* may therefore serve as two valuable, basic concepts in orchestration, finding constant usage, even if oftentimes only fulfilling secondary purposes.

Fig. 1.1 *Orchestra* from the portfolio *Revolving Doors* by Man Ray, 1926.[1]

1.1 Timbre blending in music

Unlike exploiting symbolic roles attributed to particular instruments (e.g., *heroic* trumpet), attaining a blended timbre (or its opposite) fulfills a more functional role by contributing to a sonic goal. As it concerns auditory properties, it relates to the perception of musical timbre and its instrument-specific acoustical correlates. Blend has been argued to be an aspect of orchestration for which a shared understanding of its utility is found across orchestration treatises, allowing methodologies for its perceptual investigation to be developed (Sandell, 1991). The notion of blend has been argued to be related to a wide range of sonic goals, such as *augmenting*, *softening*, *imitating* or *inventing* timbres (Sandell, 1991). For the most common cases of blend, Sandell (1995) distinguishes between the creation of *augmented* timbres, in which a dominant timbre is 'enriched' by the timbral quality of another, and *emergent* timbres, where two or more timbres combine to create a novel timbre (related to *inventing* timbres). Reuter (1996) considers only a single category of blend as *Schmelzklang* ('fused' or 'molten' sound). On the other hand, the contrasting, non-blended case corresponds to *heterogeneous* (Sandell, 1995) timbres or *Spaltklang* ('split' sound; Reuter, 1996). Given these varying notions for blend, it is meaningful to nonetheless establish a working definition for it, which generally concerns the case of two or more concurrent timbres achieving an integrated timbral percept, with the constituent timbres losing their individual distinctness, although the integrated percept may still bear some resemblance to its constituents.

In orchestration practice, instrumental blend is first conceived in the mind of a composer or orchestrator, then jointly executed by performers, with the final aim of being perceived as blend by the recipient, i.e. the listener. Commonly, intermediate parties may also be involved, such as a conductor or sound-recording engineer, acting as mediators towards achieving the intended blend result at the listener location. Orchestrators operate on an idealized, conceptual level, with their chosen instrument combinations intended to lead to blend in practice. Moreover, their choices depend on musical factors, encompassing pitch register, dynamic marking, and articulation, all of them linked to instrument-specific acoustical traits. Furthermore, the recommendations found across orchestration treatises are similarly subject to these instrument-specific factors (Rimsky-Korsakov, 1964; Koech-

[1]Painting. New York: Museum of Modern Art. URL: *http://www.wikiart.org/en/man-ray/orchestra-from-the-portfolio-revolving-doors-1926*. Last accessed: December 1, 2014.

lin, 1959). When a novel timbre *emerges* from the mixture of multiple instruments, the outcome may also rely heavily on compositional factors, as discussed for the example of Ravel's *Boléro* (Bregman, 1990). By contrast, the more common case of *augmenting* timbres even bears the potential of extending the notion of blend to arbitrary non-unison combinations. For instance, a chordal passage scored for brass could be expected to blend more than a combination of highly diverse timbres. During musical performance, musicians are entrusted with the actual realization of an orchestrator's idealized blend. This involves at least two performers situated in an interactive relationship, enabling each to adjust their individual instrument timbre to achieve the intended blend. Furthermore, each performer experiences an individual perception of the blend achieved during performance, based on room-acoustical and musical factors. For instance, role assignments as *leading* or *accompanying* musician may determine how timbral adjustments between performers are going to take place. In summary, the investigation of the perception of blend, as it is mediated by acoustical factors, opens an intriguing research project directly relevant to the heart of musical practice.

1.2 Timbre perception and its acoustical correlates

Timbre will here be considered a perceptual quality corresponding to the auditory experience of sounds, which is somewhat detached from other sound attributes. However, common usage associates broader definitions that take generalized descriptions of musical instruments into account. In order to address the notion of timbre as it applies to musical practice adequately, it is therefore advantageous to describe it as completely as possible. This would consider implicit generalizations orchestrators and other musicians rely on in their knowledge of instrumental timbre, which likely involves acoustical commonalities within certain pitch registers, dynamic markings, and articulations. In addition, when musicians perform with their instruments, especially in the case of orchestras with spatial extent, the role of room acoustics becomes increasingly relevant to the shaping of perceived timbre. As a result, these factors will also be briefly addressed in the discussion of the perception of timbre.

1.2.1 Defining timbre

Past research has been unable to attain a general definition for musical timbre that would qualify as describing its role in music adequately, mainly due to its complex and multidimensional nature (McAdams, 1993; Handel, 1995; Hajda et al., 1997; Hajda, 2007; Patterson et al., 2010). The widely referenced ANSI-definition (ANSI, 1973), with timbre being delimited as that sound attribute conceptually detached from pitch, loudness, and duration, is essentially a "definition [...] by exclusion" (Handel, 1995; p. 426). One has to acknowledge that while we may have a clear perceptual notion of pitch for different timbres, our perception of timbre is generally confounded by concurrent variations in pitch, rendering timbre a hard-to-grasp perceptual phenomenon. The requirement of empirical research for *constitutive* and *operational* definitions to derive methods and models can be seen as a primary motivation behind the ANSI-definition (Hajda et al., 1997). More universal but also more vague attempts at definition like "the way it sounds" (Handel, 1995; p. 426) are merely phenomenological and hard to operationalize as variables in empirical research (Hajda et al., 1997). At the same time, the ANSI-definition disqualifies itself for the description of musical sounds not exhibiting distinct pitch (Bregman, 1990), and furthermore, it manifests clear limitations to investigating musical timbre in melodies (Hajda, 2007) as well as across instrument registers or families (Patterson et al., 2010). Defining timbre for its role in music becomes even more complex due to it affecting both categorical sound source identification and qualitative evaluation along continuous perceptual dimensions.

Musical timbre may commonly even be associated with describing an entire instrument or family (Patterson et al., 2010), with some attempts already made to extend the single-pitch definition by referring to a conjunction of such timbres constituting an instrument, as the concepts of *source timbre* (Handel and Erickson, 2004) or *macrotimbre* (Sandell, 1998 reported in Hajda, 2007) illustrate. With the aim of developing a working definition for musical timbre as it applies to a wide range of musical instruments across their timbral range and as it relates to musical contexts, there is a need to broaden past definitions, with the same applying to the breadth of research methodologies.

1.2.2 Perceptual investigation of timbre

Known to be multidimensional and accounting for its *dual categorical and continuous nature* (Hajda et al., 1997), two main approaches have been followed in perceptual research on timbre: the *identification* of instruments and the rating of *similarity* between instrument pairs. These two experimental tasks have found wide application in the investigation of timbre perception (McAdams, 1993; Handel, 1995; Hajda et al., 1997; Hajda, 2007).[2] There is in fact a clear correspondence between high degrees of timbre similarity and greater likelihood for false identification of instruments, but at the same time, not every discriminable difference in timbre similarity may bear an effect on instrument categorization (McAdams, 1993).

Studies investigating timbre similarity through ratings for sound pairs have employed *multidimensional scaling* (MDS) to obtain geometrical models, so-called *timbre spaces*, which are assumed to reflect the underlying perceptual dimensions for timbre that can also be correlated with potential acoustic descriptors. Numerous such studies have been conducted, which does not allow an exhaustive discussion. The most relevant findings are readily available in review articles (McAdams, 1993; Handel, 1995; Hajda et al., 1997; McAdams, 2013). Among the reliable acoustical descriptors for perceived timbre, i.e., those exhibiting correlations with the underlying dimensions in exploratory (McAdams et al., 1995) and confirmatory studies (Caclin et al., 2005), the most prominent will be briefly introduced: The *spectral centroid*, which expresses the central tendency of a spectrum through an amplitude-weighted frequency average, has been found to be the most reliable correlate explaining principal dimensions of timbre spaces. To a lesser degree, spectro-temporal variation (e.g., *spectral flux*) has in some cases been shown to correlate with perceptual dimensions. Finally, descriptors for attack or onset time serve as the principal correlates for the temporal amplitude envelope, which also appears to depend on the variability along this dimension in the stimulus set.[3] These salient features of timbre perception are also expected to have a relevance to the blending between multiple simultaneous timbres (see Section 1.3.3).

[2] Various other approaches such as *matching*, *discrimination*, and *verbal description* have also been considered, but not to the same extent (McAdams, 1993; Hajda et al., 1997).

[3] Formulaic expressions for the main descriptors can be found in Hajda et al. (1997); Hajda (2007); Peeters et al. (2011).

Potential signature traits of instrument timbre

An overwhelming number of studies accredit a high importance to spectral-envelope shape, notably always showing a relevance to timbre perception, as opposed to temporal or spectrotemporal features seeming to depend on the particular stimulus contexts investigated. In seeking an acoustical description that would encompass the signature traits of instruments, i.e., features that could be generalized as describing the instrument across an extended pitch range, spectral features seem most promising. A common limitation to most studies is that only a single pitch has been investigated, which, due to all stimuli usually being equalized in pitch, has also represented some instruments in atypical registers. Nonetheless, the sheer quantity of findings can still be taken as a strong argument in favor of the importance of the spectrum. The following examples of studies provide qualitative arguments for the importance of spectral features, as they govern instrument identity as well as the similarity relationships among instruments. Wedin and Goude (1972) applied cluster analysis to similarity ratings and obtained three clusters, which strikingly corresponded to a grouping based on spectral-envelope shape, with groups organized into flat spectra, spectra with strong fundamental frequencies and a monotonic decrease of the remaining partial amplitudes, and spectra exhibiting a maximum centered above the fundamental. The latter group corresponds to wind instruments, where the maximum represents a prominent spectral feature, also referred to as a *formant*. The perceptual relevance of the spectral envelope is further suggested in an MDS study, where an exchange of spectral envelopes between trumpet and trombone led to a corresponding change in timbre-space positions, arguing that differences between pronounced spectral-envelope features, such as formants, appear to strongly affect similarity judgments of timbre (Grey and Gordon, 1978). To address one example from identification studies, Strong and Clark (1967a) employed an identification task on synthesized woodwind and brass instruments, which provided them with the means to interchange temporal and spectral components between the stimuli.[4] In order to study the effect of an oboe's pronounced two formants on identification accuracy, the secondary formant or the spectral valley between the two formants was removed. The omission of the secondary formant increased false identifications in general, whereas the removal of the valley selectively increased confusions with the trumpet, whose spectral envelope

[4] Although the use of synthesized sounds to emulate real orchestral instruments can generally be questioned, their synthesis method was based on modeling formants, with their supplied spectral diagrams being in agreement with other results.

approximates the outline of the oboe's, except for the valley (see also Figure 2.1, middle panels). In conclusion, these examples, as well as numerous findings in other MDS studies consistently suggesting a perceptual relevance of spectral centroid, argue for a strong significance of spectral-envelope shape in the perception of musical timbre.

Musical context

Apart from most research having investigated only single pitches, it also has almost exclusively focused on isolated sounds, which presents another limitation to musical practice. As Hajda (2007; p. 257) notes, "[m]usical timbre does not operate as a series of unrelated, isolated entities" and, as a result, perceptual cues found to be relevant in isolated contexts may not generalize to musical contexts. A possible reason considers the engagement of a listener in musical scenarios as essentially being an *online* task, whereas experiments conducted on isolated contexts employ *offline* tasks, in which participants are given unlimited time to attend to minute timbral differences. Furthermore, over the course of timbral variations across pitches, dynamics, and articulations, musical contexts could provide less ambiguous and more reliable perceptual cues over the general timbral identity of instruments, as opposed to isolated notes potentially bearing timbral specificities of their own.

An early study investigating the effect of musical context on timbre perception tested the discriminability of slight modifications of resynthesized sounds (Grey, 1978). The discrimination accuracy was compared between three stimulus contexts: isolated notes, and musical contexts with single and multiple voices. Changes to attack or articulation features were increasingly disregarded with growing context complexity, whereas detectability of spectral modifications appeared to be robust across all contexts. As not all modifications were applied to all instruments, the results do not necessarily generalize across instruments. Furthermore, in a discrimination task, the investigated variations are usually kept small, leaving open how relevant this would be in musical practice. Nonetheless, the findings argue for reduced detectability of 'irrelevant' cues in musical as opposed to isolated contexts. Kendall (1986) tested identification performance[5] for instruments in isolated and musical legato phrases. At the same time, stimuli were tested for different variants of the temporal envelope (e.g., lacking attack, lacking sustain). Whereas identification performance in isolated contexts was comparable across all stimulus conditions, identification in musical

[5]The exact task was actually determining whether the instruments, playing two successive presentations of single notes or musical phrases, were the same or different.

contexts was generally better than in isolated contexts, except for the stimulus condition lacking sustained portions. For one, this suggests a greater availability of perceptual cues in musical contexts, as the signature traits of instruments may be better conveyed in musical phrases that offer variation across pitch, dynamic markings, and articulation. Furthermore, the absence of the sustained portions leading to deteriorated identification accuracy even in musical contexts argues for them to carry the essential cues to instrument identity, i.e., they are likely associated with spectral characteristics. However, it has also been noted that the removal of attack portions effectively replaces them by a short artificial attack, which could itself contribute to the obtained performance differences (McAdams, 1993). With regard to instrument identification in musical phrases, Reuter (1996) reports that identification accuracy can also be modulated by idiosyncratic musical figurations of instruments, with confusions in identification biased towards selecting the instrument typically associated with a certain figuration, whenever the figurations and the playing instrument did not agree. For instance, a noise-masked rendition of a synthesized oboe playing melodic figurations typical for a flute, misled participants into identifying it as a flute. In summary, these examples show the relevance of musical context and the fact that *online* scenarios affect timbre perception differently than do isolated, *offline* contexts, and may even point out that the notion of *musical timbre* as it concerns instrument identity might not be attributable to auditory properties alone.[6]

1.2.3 Timbre as a function of acoustical factors

Orchestras contain a vast universe of timbres, with only few orchestrators having shown to have mastered, and none having exhausted, the full potential of timbral variety, partly due to its boundless combinatorial possibilities. Instead of knowing or imagining how all possible combinations of pitch, dynamic markings, and, articulation would sound, instrumentalists as well as orchestrators likely rely on some form of generalized knowledge of instrument timbre. This knowledge may relate to implicitly internalized ideas of the acoustical systems involved and also how these systems interact with room acoustics.

[6] Musical context of course involves not only sequential, but also concurrent, occurrences of notes, which is relevant to blend and is discussed in Section 1.3.1.

Instruments as acoustical systems

Although perceived timbre should be primarily related to the resulting acoustic signals, it can still be assumed that musicians acquire some implicit knowledge of instrument acoustics through frequent interaction with their instruments. And even orchestrators' knowledge of instrumentation may have developed an implicit vocabulary for the acoustical signature traits of instruments, based on some generalizations across pitch registers, dynamic markings, and articulations. It would be valuable to correlate perceived timbre to these generalizable descriptions, likely representing the underlying *structural invariants* (McAdams, 1993) of instruments. Timbral variety reflects the differences among acoustical systems found across instruments. Sound generation is based on some form of excitation, which can vary greatly, e.g., air pulses through reeds, bowing across strings, hitting hammers or mallets on membranes, plates, or strings. Still, melodic instruments share in common that the excitation couples to resonators, which together determine pitch as well as its spectral characteristics, with the excitation energy being proportional to the dynamic intensity. These principles can be discussed in more conceptual terms (Patterson et al., 2010) or can involve the detailed description of musical acoustics (Benade, 1976; Fletcher and Rossing, 1998). The excitation and resonator components are oftentimes also expressed as *source-filter models* (see also Appendix D). These models have more recently been described as yielding *pulse-resonance* sounds, which illustrate how instrument sound evolves across different registers as well as among relatives of an instrument family (Patterson et al., 2010). These timbre relationships can be simply expressed by the physical dimensions of *source scale* and *filter scale*, with their joint magnitudes linked to an instrument's size and thus determining its pitch range, their ratio determining register within its pitch range, and a general spectral-envelope shape characterizing the instrument as a whole. These models all suggest that spectral envelopes are relatively stable with respect to pitch change, and, in the context of wind instruments, prominent spectral-envelope features resembling local maxima have been termed *formants*, by analogy with the human voice. Formant structure in wind instruments has been discussed in the German literature for about a century (Stumpf, 1926; Schumann, 1929; Mertens, 1975; Reuter, 1996; Meyer, 2009), with somewhat less widespread references made in English publications (Saldanha and Corso, 1964; Strong and Clark, 1967a; Luce and Clark, 1967; Wedin and Goude, 1972; Grey and Gordon, 1978; Brown et al., 2001). Pioneering research by Schumann (1929) established

a set of rules concerning how formant structure governs the spectral envelopes of wind instruments across their pitch and dynamic range. Whereas reported formant regions for wind instruments find large agreement across the literature (Reuter, 2002), members of the orchestral string-instrument family exhibit strong resonance structure, stemming from their body plates and enclosed air cavity, but these are highly individual among instruments. As a result, no two violins sound alike, but in the orchestral context, the choric use of string instruments could lead to an 'averaged' spectral envelope, which however, will seldom exhibit prominent spectral features comparable to that of certain wind instruments (see Appendix B). These differences may even explain why in teaching orchestration, wind-instrument timbre requires more careful consideration than orchestrating for string sections.

Timbre and room acoustics

Given the size of orchestras and the spatial distance between musicians, the influence of room acoustics on timbre becomes increasingly relevant, both from the perspective of players and the audience. Instruments function as sound sources, radiating sound waves into space, with those impinging on sufficiently large and rigid surrounding surfaces being reflected back into the room, while also being modified with respect to frequency through absorption. A model describing how the sound from the instrument is modified by the room at any listening position involves computing the mathematical convolution between the signal emitted by the sound source and the *room impulse response* (RIR). In the frequency domain, it acts like a filter, whereas the time course of RIRs consists of an initial, delayed impulse for the direct sound wave, followed by myriad impulses from surface reflections, growing in temporal density and decaying in amplitude, in essence, shaping the reverberation pattern. RIRs are variable for different spatial configurations between source and receiver and furthermore depend on the sound-directivity patterns of both the source and receiver. Instruments vary in their radiation directivity, which, moreover, are frequency-dependent (Meyer, 2009). Therefore, an identical instrument at different locations or two different instruments at identical locations yield distinct RIRs. Returning to the notion of source-filter models, the RIR essentially corresponds to cascading the instrument with an additional room filter. For large ensembles and performance spaces, timbre is clearly a function of room-acoustical factors, and its effect should be taken into account, especially

as this timbral *coloration* through room acoustics has been shown to be perceptible for orchestral instruments in a discrimination task (Goad and Keefe, 1992).

1.3 Previous research related to blend

After *timbre* has been established as it relates to musical practice, the following sections place it within the context of orchestration practice aiming for *blend* between timbres. Its perceptual and acoustical implications to musical practice are discussed.

1.3.1 Blend as a part of the auditory scene

As blend presumedly involves the auditory fusion of concurrent sounds, its perceptual processes operate within the larger framework of *auditory scene analysis* (ASA, Bregman, 1990). At the outset, ASA deals with the perceptual challenge of decoding an indiscriminate summation of acoustical signals entering both human ears into distinct informational units. For concurrent events, separate informational entities are formed by *fusion* of their constituent elements, whereas the association of sequential events involves the perceptual grouping into temporal *streams*. These two perceptual processes stand in competition to one another, with sequential grouping into streams capable of affecting the spectral fusion of simultaneous components and vice versa. The relevance of ASA to instrumental blend is twofold: First, the establishment of a timbre identity (e.g., that of an instrument) already depends on how partial tones fuse into a single timbral entity. At a higher level, the same principles apply to the blending between individual instrument timbres. Second, if one aims to situate the perception of blend in a musical context, both simultaneous and sequential processes need to be taken into account, and they will also be valuable to the attempt to establish theories of blend that generalize to musical practice.

Perceptual fusion of simultaneous tones

General principles Fundamental research on ASA employed elementary acoustic signals such as pure tones[7] and noise. Given the case of two or three pure tones being variable in frequency or temporal location, two fundamental principles can be established: For one,

[7] The usage of the term *tone* will hereafter apply to pure tones serving as elementary components that may *fuse* into timbral identities. This is meant to distinguish it from the term *partial* (tone), which already presumes *fusion*.

larger frequency separation tends to *segregate* tones into separate units and thus act against fusion. On the other hand, temporal asynchrony between tones also acts against fusion. These basic factors can still be modulated or even suppressed by several other factors, such as spatial difference, harmonicity, and temporal modulation, which, moreover, exert mutual interactions in a hierarchical system of dominance and subordination. Spatial separations between tones with frequency separations as small as 7% prevent their fusion. Similarly, a segregation of tones can also result if one tone changes its spatial position over time. However, there have also been examples of spatial cues being suppressed in cases where discordant correspondence to other cues occur, with Bregman (1990; p. 302) stating that "the human auditory system does not give an overriding importance to the spatial cues for belongingness but weighs these cues against all others. When the cues all agree, the outcome is a clear perceptual organization, but when they do not, we can have a number of outcomes." Harmonic frequency relationships among tones, i.e., all occurring at integer multiples of a common fundamental frequency, have been shown to achieve fusion even for wider frequency separations. A similar unifying influence is achieved by temporal modulation over the duration of concurrent tones employing the Gestalt-psychology principle of *common fate*, which presumes that components originate from a common source if they evolve temporally in a similar and coherent way. For example, coherent *micromodulations* of frequency as small as 0.5% on a group of tones achieves their fusion (McAdams, 1984). With coherent micromodulation even achieving fusion of inharmonic or spatially separated pure tones, it can be rated as one of the most unifying perceptual cues to fusion.

Principles applied to reality The general principles of ASA were studied using experimental paradigms that employed repetition of short sequences, which in turn increased the magnitude of the observed perceptual effects (Bregman, 1990). At the same time, perceptual tolerance towards incoherence between cues is reduced, yielding exaggerated observed effects compared to what would apply in more realistic scenarios. Strongly discordant relationships between perceptual cues (e.g., temporally alternating spatial positions of a harmonic tone complex) are also less likely to occur in reality, where different cues are generally concordant, although minor incongruences do indeed occur. Revisiting the case of sound radiating from a source into a room, it propagates outward and is reflected by surrounding surfaces of varying materials, which results in multiple instances of the

original sound impinging on a listener at various delays and in spectrally altered form (see Section 1.2.3). Although the available cues exhibit a considerable degree of incongruity, on a higher level the listener might rely on common-fate cues, such as common temporal amplitude patterns, and still establish a distinct source identity for the sound. In other words, this would correspond to translating the discordant *classical*, acoustical cues into unambiguous *world structure cues*, which emerge from ASA processes (Bregman, 1990). In summary, in increasingly realistic settings, where occurrences of strongly discordant cues are less likely, a general reluctance to accept incongruent and thus irrelevant cues could be hypothesized.

Based on the outlined principles, timbre can be understood as an emergent quality after its constituent tones get fused together. The robust perceptual fusion into timbral identities is generally ensured for instrumental sounds by reliable cues based on common fate and harmonicity (Sandell, 1991), which can be extended to the realm of room acoustics by the previous discussion of world structure cues. As a result, minor discrepancies among auditory cues, such as slight deviations from harmonicity and even asynchronous onsets of different partials, do not pose a risk to maintaining stable fusion, with the identity of isolated instrument timbre being generally unchallenged.

Sequential grouping of tones

The fundamental principles governing sequential streaming comprise frequency separation and the temporal rate of occurrence. Greater frequency separations or faster rates both increase the tendency toward stream segregation. However, there is no single boundary between perceiving one or two streams, as streaming effects also depend on attentional processes, which has led to the discovery of two task-dependent boundaries (van Noorden, 1975 reported in Bregman, 1990): 1) If the task is to attempt to uphold the perception of a single, unified stream until a forced segregation into different streams occurs, one considers the *temporal coherence boundary*. This boundary is a function of both frequency separation and tempo and is seen as a perceptual limit. 2) If a listener is asked to attend to a single stream among multiple streams until no longer possible, one obtains the *fission boundary*. This boundary is roughly independent of tempo and a function of frequency separation alone. Inside the region defined by the two boundaries, attentional focus and stimulus context determine whether separate or unified streams are perceived. Bregman

(1990) suggests that the former boundary involves a purely perceptual *primitive stream segregation*, whereas the latter concerns *schema-based* streaming, involving attentional and thus cognitive processes.

In the absence of temporal differences and distinctions along other factors, two overlapping tone complexes, conceived as being separate, may in fact not be perceptually discriminable. This is a case where sequential grouping may influence the *segregation* into separate tone complexes by relating them to prior occurrences. Bregman (1990) refers to the approach as the *old-plus-new* heuristic, comparing a current combination of tones to what preceded it and basing simultaneous grouping on this evaluation, which itself can be associated with another Gestalt principle known as *good continuity*. If in this example one of the 'conceptual' tone complexes had appeared in isolation preceding the presentation of both complexes, it could have resulted in the segregation into two timbral identities by way of identifying the repeated tone complex as a good continuation and effectively grouping it into a stream.

With regard to increasingly realistic scenarios, alternating instrumental timbres of equal pitch have been shown to segregate into independent streams (Iverson, 1995) and even for the case of incongruent cues, timbre dissimilarity can dominate over pitch proximity in segregating streams (Bregman and Levitan, 1983, reported in Bregman, 1990). Furthermore, the latter study varied timbre by changes to formant frequencies, which suggests spectral features to serve as important cues for streaming. Similar results have been reported for instrumental sounds in simple melodic sequences, with differing main-formant locations for two wind instruments contributing to their segregation, whereas agreement between main-formant locations led to a grouping into a single stream (Reuter, 2003).

Blend between timbres

As established in Section 1.3.1, the timbral identities for musical-instrument sounds are quite robust, largely due to strong unifying cues of common fate and harmonicity. Given that some of these cues are unique to each sound (e.g., coherent micromodulations), they would likely contribute to segregation if these sounds were presented concurrently. In order to achieve blend, these tendencies would need to be overcome by stronger, higher-order cues that promote the fusion between timbres, which could rely on exploiting perceptual ambiguities the timbres may exhibit along certain factors. It can be reasonably assumed that

blend among instrumental timbres never achieves the same degree of fusion as for tones into timbral identities, but it has also been argued by Bregman (1990) that in musical terms, it might be feasible to assume a mode of *chimeric* perception, where, for instance, synchronized onsets of several sound sources all contribute to the identity of a single musical note. Despite retaining some degree of timbral independence, blend operates at a higher level, conveying some form of unified informational unit or layer. Thus, sound attributes pertaining to a single sound source are no longer restricted to *exclusive-allocation* but might in fact contribute to varying levels of abstraction as in the case of *duplex perception* (Bregman, 1990). In orchestration, this could even relate to the blend in non-unison combinations, where on a larger level of the musical texture, a blended chordal accompaniment may serve as a background layer against which other musical layers are contrasted. Thus, to some extent blend between instrumental timbres may indeed rely on *perceptual illusions* (Bregman, 1990; Sandell, 1991), which exploit perceptual ambiguities along a number of blend-related factors and could be strengthened further by unifying cues emerging from the musical context.

1.3.2 Factors contributing to blend

In his investigation of blend for orchestral instruments, Sandell (1991) discussed a list of factors related to blend. This list is expanded upon in the following paragraphs, complementing it with findings from more recent empirical research. Some of these factors naturally bear a strong resemblance to some of the ones known from general ASA research, only in this case applied to already established timbral identities. In addition, these factors are related to higher-level features involving acoustical and musical aspects.

Spectral similarity Several studies have argued for the similarity in spectra between instruments to be related to higher degrees of blend (Sandell, 1995; Reuter, 1996; Tardieu and McAdams, 2012) and they are discussed in greater detail in Section 1.3.3, with the role of spectral features in blend also being the main focus of the research reported in this thesis.

Onset synchrony Synchronous note onsets or attacks are also thought to contribute to blend (Sandell, 1991), as they provide common-fate cues. Concerning its relevance within musical contexts, Reuter (1996) suggested that temporal forward-masking could

render succeeding onsets inaudible, implying that its relevance is secondary to spectral characteristics (see also Section 1.2.2). Masking of onsets could become even more relevant when one considers the effective lengthening of note decays through room reverberation. Given that attack characteristics themselves also mediate blend (Tardieu and McAdams, 2012), onset asynchrony between more impulsive attacks could be assumed to affect blend more critically than between asynchronous notes with more gradual onsets.

Dynamics With decrease in musical dynamic markings (e.g., *ff*–*mf*–*p*), instrument spectra are generally known to exhibit reduced intensities for higher partials, which is confirmed when comparing spectral-envelope slopes across dynamic markings in Appendix B. Sandell (1991) argues that softer dynamics, which lead to 'darker' timbres, may blend more. An acoustically more informative explanation for this could be given by the finding that for softer dynamic markings, secondary formant intensities (located higher in frequency than main formants) are reduced, rendering the spectral envelopes less pronounced in high frequencies (Schumann, 1929).

Pitch In musical terms, pitch proximity relates to interval size. Growing pitch separation is expected to reduce blend between timbres (Sandell, 1991), although at the same time this could also be a function of the degree of consonance or dissonance, which involves the relationships among the combined partials (Stumpf, 1890; Dewitt and Crowder, 1987 reported in Sandell, 1991) and could be effectively related to their degree of harmonicity. As a result, pitch combinations in unison or octave intervals can be assumed to lead to higher blend, due to their coincident partial-tone frequencies. At the same time, intonation could become a critical factor for these cases. Furthermore, pitch height is also mentioned as a factor that is relevant to instrument register. Instruments are known to vary in timbre across registers (see Section 1.2.3), which may affect their ability to blend; for instance, in the high registers, the wide spacing of partials increasingly obscures formant structure (Reuter, 1996).

Non-unison voicing If blend is to be achieved across several pitches in non-unison, such as in voice coupling or homophonic accompaniment, several alternatives of instrument combinations are possible. Based on examples discussed in orchestration treatises, *interlocking* voicing has been argued to be most effective (Kennan and Grantham, 1990; Sandell, 1991;

Reuter, 1996). Given the case of two instruments playing two voices each, interlocked voicing leads to one voice of each instrument always being encapsulated by two voices of the other instrument. Furthermore, the more general *harmonization* rule encourages the spacing of voice texture along the harmonic series. Within the context of achieving blend across wider pitch ranges, the possibility of *bridging* instruments contributing to greater blend by acting as cohesive elements against divergent pitch and timbral relationships has also been proposed (Sandell, 1991). As concerns the number of timbres involved, there is agreement that the number should be limited to a few because a large timbral variety might counter the desired blending effect of both interlocked voicing or bridging timbres (Sandell, 1991; Reuter, 1996), although this could also be mediated by spectral similarity.

Performance factors In musical practice, instrumental blend is achieved through musicians *performing* together. It may therefore be valuable for experimental investigations of blend to also consider performance parameters either in the stimulus production for listening experiments (Kendall and Carterette, 1991, 1993) or in the technical execution of experiments involving production tasks such as musicians performing to blend (Goodwin, 1980). Two performers aiming to achieve the maximal attainable blend would try to optimize factors contributing to greater common fate and coherence cues, such as intonation, spectral similarity, onset synchrony, and articulation. Overall, these factors all comprise those previously addressed, although they occur during actual musical performance, i.e., independent of prior conceptual considerations during composition and orchestration.

Spatial separation In musical performance, the influence of the physical separation of instruments in space is inevitable and is present both in live concert situations and in stereophonic recordings. In terms of room acoustics, spatial separation provides two principally different sets of factors: First, inter-channel time and level differences provide localization cues.[8] Although spatial separation might appear to play a significant role in hindering blend and facilitating sequential streaming, it has been reported to apply only to angular separations greater than 120° (Song and Beilharz, 2007). Given strong agreement concerning other ASA-related cues, spatial separation cues have been shown to

[8] Inter-*channel* considers the generic case of multiple audio channels being involved. More specifically, this could represent the case of differences between two stereo-microphone channels, but also inter-*aural* differences that relate to binaural hearing.

be disregarded and become subordinate in perceptual fusion (Bregman, 1990; Hall et al., 2000; see also Section 1.3.1). Second, another factor stems from distinct RIRs, which correspond to unique *coloration* defined by the spatial configurations between instrument and listener (see Section 1.2.3). Although coloration should be assumed to be perceptually relevant (Goad and Keefe, 1992), some commonalities across different spatial constellations (e.g. reverberation times across frequencies) might indeed exist, perhaps even leading to improved blend through 'common room' cues. By contrast, two theories argue that binaural cues might in fact allow the auditory system to conduct a *de-coloration* of sources (Moore, 1997; Watkins, 1998 reported in Flanagan and Moore, 2000). In the context of ASA, Bregman (1990) notes that binaural cues support stream segregation by reducing the perceived dissonance of highly cluttered sonic environments. Hence, it remains unclear whether coloration through room acoustics aids or hinders perceptual blend, especially as different viewpoints for musicians (e.g., between performers, conductor) or other listeners (e.g., audience, sound-recording engineer) are all mediated by room-acoustical variation.

1.3.3 Perceptual investigations of blend

Most of timbre research has precluded the study of concurrent presentations of instrumental sounds. As a result, there has only been a handful of studies with *blend* as their main research focus (Goodwin, 1980; Sandell, 1991, 1995; Kendall and Carterette, 1993; Reuter, 1996; Tardieu and McAdams, 2012).[9] In the interest of brevity, their general commonalities and differences are presented, as more specific issues will be addressed in the main chapters.

Experimental tasks Previous studies have assessed the degree to which two instruments blend by employing two experimental tasks: 1) direct ratings of blend on a continuous scale, and 2) indirect assessment of blend from the inability to identify constituent instruments in dyads. Among the studies employing rating scales, two used the verbal anchors *oneness* and *twoness*, with highest degree of blend being attributed to the former (Sandell, 1991; Kendall and Carterette, 1993; Sandell, 1995).[10] The usage of the label *twoness* is seen as problematic, as it might be mistaken as facilitated detection of two distinct sound sources

[9] Goodwin (1980) studied blend in choral singing applied to a production task, and this work therefore does not directly compare to the other studies investigating perception of blend between orchestral instruments.

[10] Sandell (1991) and Sandell (1995) report the same experiments, although only the latter clearly confirms the usage of the mentioned verbal labels.

or pitches as opposed to judging only timbral differences, i.e. one might be able to detect two distinct pitches, but not clearly hear out the individual timbres. Similarly, the label *oneness* would not prevent the label from seeming appropriate in the case of complete masking of one timbre by the other, which arguably would not correspond to blend between timbres. However, both studies give no reason to believe that these concerns have manifested themselves in the obtained results. Avoiding these issues, Tardieu and McAdams (2012) used the verbal anchors *very blended* and *not blended*.

The identification task allows the indirect measurement of blend through inference that increasing inability or confusion in correctly identifying constituent instruments in a mixture argues for a high degree of blend. Kendall and Carterette (1993) supplied participants with 10 alternatives of instrument pairs to associate with the presented timbre dyad, with wrong identification being taken as an indicator of indistinguishability between timbres. The authors were mainly interested in complementing and comparing these data to direct blend ratings acquired on the same stimulus set. Reuter (1996) asked participants to identify the two presented instruments on two identical lists providing a list of instrument options. He operationalized blend (*Schmelzklang*, see Section 1.1) as the case in which both identification judgments were assigned to the same instrument. However, disregarding other potential instrument confusions seems like an overly limited approach, as is also the narrow understanding of *Schmelzklang*, which would not extend to *emergent* timbres (see Section 1.1). In addition, the identification of a single instrument could again correspond to the case in which one of the timbres was completely masked by the other. Sandell (1991) raised a concern regarding the usage of identification tasks for characterizing blend, in that identification performance has been shown to be variable across different instruments, which could prove to be a confounding factor in accurately characterizing blend across different instruments to equal degrees. One would need to 'correct' the relative identification rate in blends with those in isolated sounds.

Experimental stimuli The four investigations of blend between instrumental timbres exhibit differences not only in terms of experimental tasks, but also concerning some of the factors discussed in Section 1.3.2. Their individual experimental details, from which the resulting differences become apparent, are summarized in Table 1.1.[11] Some important

[11]The factors investigated in Kendall and Carterette (1993) were based on Kendall and Carterette (1991), which also includes the description of stimulus production and context.

strengths or limitations of the studies are briefly addressed. As in other research on timbre perception, the investigation of isolated sounds is less generalizable to musical scenarios (see Section 1.2.2), with two studies having included musical contexts among their stimuli. In addition, two studies also considered blend for non-unison cases, which extends to more practical orchestration scenarios of coupled voicing or homophonic accompaniment. One of the strongest limitations in some of the studies is that their results were obtained and interpreted based on a very limited pitch range, which even includes some instruments in atypical, extreme registers. Furthermore, relatively short note durations may not be representative of many cases in musical practice. Especially in non-melodic, more homophonic contexts, blended timbres would involve half- or whole-note durations. The individual studies bear limitations with regard to being less generalizable to the instruments beyond the investigated pitches, registers, dynamics, and articulation. Nonetheless, each study made a contribution towards a better understanding of blend.

Main findings Varying experimental tasks and methodologies to operationalize blend as well as other differences among the reported studies limit the extent to which direct comparisons can be drawn between them. As a result, the most valuable contributions for each study are presented separately.

Kendall and Carterette (1993) have shown that increased confusion in identifying the constituent instruments in dyads corresponded to the same dyads resulting in higher degrees of blend; there was a strong negative correlation between blend ratings and identification accuracy. Furthermore, good agreement between timbre spaces based on blend ratings with a previously acquired timbre space for similarity ratings (imagined by a musicology professor in Kendall and Carterette, 1991) argues for timbral blend and timbral similarity to be intrinsically related. With respect to the ratings, main effects for different instrument pairs as well as stimulus contexts (e.g., unison or non-unison for isolated notes and musical context) were obtained, which furthermore lead to interaction effects. In other words, blend was found to vary as a function of instrument pairing and furthermore could be mediated by musical context. Interestingly, a post-hoc comparison also suggested that unison dyads yield higher blend than do non-unison dyads, which also translated to more confusion in identification for unison than non-unison cases. Although the authors announced a separate publication dedicated to the acoustical analysis of the stimuli, which would have allowed correlation analyses with the behavioral blend measures, no such article

Study	Sandell (1991; 1995)	Kendall & Carterette (1993)	Reuter (1996)	Tardieu & McAdams (2012)
Isolated-note context	✓	✓		✓
Musical context		✓	✓	
Unison interval	✓	✓	✓	✓
Non-unison interval	✓	✓		
Register / pitch range	E♭4, unison; C♯4:E4, major third	B♭4, unison; B♭4-F5, unison melody; B♭4:D5, major third; G4-F5, 2-part harmony	diatonic C-G & G-D scales in C major; C2-G6, depending on instrument	C♯4
Instrument families	woodwinds, brass, strings (solo)	woodwinds (incl. saxophone)	woodwinds, brass, strings (section)	woodwinds, brass, strings (solo), percussion
Combination	dyadic	dyadic	dyadic	dyadic, paired as sustained & impulsive
Note durations	200-300 ms	650 ms or 2600 ms, half-notes or two whole-notes at 92 BPM, respectively	300 ms, eighth-note at 100 BPM	2500 ms
Spatial separation		✓		
Performed together		✓		
Stimulus source	resynthesized	recorded	recorded	recorded

Table 1.1 Experimental details to several perceptual investigations of blend.

has ever been published.

Sandell (1991, 1995) conducted three perceptual experiments. The first two found that the spectral centroid explains the obtained blend ratings best, correlating in two different ways: For unison dyads, the composite spectral centroid, which describes a *darker* or *brighter* overall timbre, suggested that higher blend is obtained for *darker* timbral combinations, i.e. lower centroid values. By contrast, for blend combinations at an interval of a minor third, the centroid difference between the two constituent sounds served as the strongest correlate to blend ratings. The inconsistency of two different spectral-centroid measures in either context partly motivated the third experiment, which was meant to provide clarification. However, the reported results, which argue in favor of a greater relevance of composite centroid, do not on closer examination appear that convincing, because only about half of the investigated cases display the pattern supporting that conclusion. In the absence of more compelling findings, it may be assumed that spectral centroid as a descriptor of the global spectral envelope may not capture some more differentiated spectral relationships that would explain the obtained blend ratings better. In support of the 'darker'-timbre hypothesis, Tardieu and McAdams (2012) also found centroid composite to be related to more blend. Furthermore, this study made the unique contribution of assessing the influence of different degrees of impulsiveness (e.g., plucked vs. bowed string, different mallets and idiophones) in sounds to blend. They showed that distinctions among impulsive sounds had a greater impact on blend than similar distinctions among sustained sounds, and that increasing impulsiveness rendered dyads less blended.

Assuming formant structure in wind instruments to have a perceptual relevance (see Section 1.2.2), Reuter (1996) found indications for high degrees of blend when two wind instruments exhibited similar formant regions, i.e., coincided in frequency, whereas string instruments, which lack formant structure, were found to blend well amongst themselves. By employing FFT manipulations to reduce spectra to either just the formant regions or the inverse, residual case, these principles were further supported. The inclusion of a wide pitch range for the dyad stimuli also suggested that for high registers, blend generally deteriorates, as the wide spacings of partials may render the formants less salient. Based on these findings, Reuter (1996) hypothesized a perceptual theory for blend: 1) instruments displaying coincident formant regions tend to blend well; 2) instruments displaying divergent formant regions tend to segregate; 3) instruments characterized by *spectral fluctuations* (e.g., string instruments) blend well amongst themselves, and lastly, 4) blend between the

latter and formant-dominated wind instruments is dependent on a sufficiently high sound level for the string instruments.

Although not studying blend between instruments, Goodwin (1980) delivers interesting insights into blend not evaluated through listening tests, but instead as produced by soprano singers during musical performance. The investigation compared how sopranos sang the same passage in solo or in a choral scenario, showing that in order to blend, singers modified the formant structure toward a 'darker' timbre. Whereas Sandell (1995) correctly interprets this as another example of lower composite centroids leading to blend, Goodwin employs a more differentiated explanation, related to local modifications of the formant structure. More specifically, singers selectively attenuated the second and third formants relative to the first formant, which is a common technique employed by them to ensure blend, termed *vowel modification*. In summary, this study illustrates the potential uncertainty as to which spectral description may be more appropriate in acoustically explaining blend-related effects.

1.4 Research aims

Findings from previous perceptual investigations are inconclusive in explaining timbre blend between instruments through specific spectral-envelope characteristics, arguing for the relevance of either global (Sandell, 1995) or local (Reuter, 1996) characteristics. Furthermore, half of the studies have only considered a very limited range of pitch register and dynamic markings, which prevents generalizations of the findings to extended pitch and dynamic ranges of instruments. Focusing on the *augmented*-blend scenario (see Section 1.1), my doctoral research aims to expand current knowledge, by situating the notion of timbre blend into increasingly realistic musical scenarios, addressing two central topics: 1) orchestrators' choice of instrument combinations as being closely associated to generalizable, instrument-specific acoustical traits and 2) the actual realization of blend during musical performance, i.e., what acoustical or musical factors modulate the realization of blend and the perception of individual performers.

My main research question concerns what spectral-envelope characteristics influence and explain blend between orchestral instruments and, furthermore, whether global (e.g., spectral centroid) or local (e.g., formant structure) traits are more important. These aspects will be investigated in several stages. Chapter 2 establishes an acoustical description that suc-

ceeds in assessing and quantifying spectral properties of instruments across their pitch and dynamic ranges, also considering sensorineural representations from computational models of the human peripheral auditory system (Irino and Patterson, 2006). Furthermore, two perceptual experiments (Experiments 1 and 2) investigate how parametric variations of local spectral-envelope shape, i.e., main formants, affect the perceived degree of blend. These experiments involve different behavioral tasks, registral ranges, and instruments, in order to allow a greater generalizability of the findings. In addition, perceptual results are correlated with the acoustic descriptors contributing most to blend. Chapter 3 reports a similar correlational analysis based on two other experiments (Experiments 3 and 4) for which blend ratings were obtained for dyadic and triadic pairings of arbitrary instrument sounds. Unlike the first two experiments, these two consider larger-scale differences applying to entire spectral envelopes. The obtained ratings are then used to explore a wide set of acoustic properties in regression analysis, to identify the most meaningful predictors of blend.

Chapter 4 undertakes a new exploration into the influence of musical performance on blend, as it concerns the actual realization of an orchestrator's conception. Performance of blend involves at least two musicians situated in an interactive relationship, enabling each to adjust their individual instrument timbre to achieve the intended blend. Furthermore, each performer experiences an individual perception of the blend achieved during performance, based on room-acoustical and musical factors. For instance, role assignments as *leading* or *accompanying* musician may yield asymmetric dependencies between performers (Goebl and Palmer, 2009; Keller and Appel, 2010). The investigation focuses on a performance experiment involving bassoon and horn players (Experiment 5). The main research question addresses what timbral adjustments performers employ with their individual instrument, given the aim of achieving blend. The experiment considers both musical and acoustical factors. With regard to the former, performers will be assigned to either leading or accompanying roles and, furthermore, either playing in melodic unison or in non-unison phrases. Acoustical factors will concern whether performances take place in either mid-sized or large venues and whether the acoustical feedback between performers is impaired or not.

Drawing on the results from the individual investigations, Chapter 5 concludes my doctoral research through an in-depth discussion of all investigated and known factors related to blend, providing a more complete understanding of how blend is perceived,

characterized acoustically, and, moreover, relates to musical practice in terms of both orchestration and performance. The joint investigation of timbre blend as concerns factors relevant to orchestration and performance practice will provide valuable insight into what aspects of blend assume important roles in realistic musical scenarios. Furthermore, the obtained results are thereupon compared to the actual use of blend in musical practice, addressing the motivations in orchestration and comparing the observed perceptual utility of certain instruments and instrument combinations to their discussion in orchestration treatises. Together, this widens the understanding of the perceptual phenomenon of timbre blend and allows the proposition of a general perceptual model as it applies to musical practice and orchestral music.

Chapter 2

Role of local spectral-envelope variations on blend

This chapter establishes a method of spectral-envelope estimation and description that allows the evaluation of instruments across their pitch and dynamic ranges. Using the acoustical description, two listening experiments (Experiments 1 and 2) investigate how parametric variations of local spectral-envelope shape, i.e., main formants, affect the perceived degree of blend. The content is based on the following research article:

Lembke, S.-A. and McAdams, S. (Under review). The role of local spectral-envelope characteristics in perceptual blending of wind-instrument sounds. *Acta Acustica united with Acustica*.

2.1 Introduction

Implicit 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 note-onset synchrony, partial-tone harmonicity, and specific combinations of instruments (Sandell, 1991). 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 (Kendall and Carterette, 1993; Sandell, 1991; Reuter, 1996), 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 (Sandell, 1995). 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 non-unison intervals may be rightly assumed to make two instruments more distinct, certain instrument combinations 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, 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 spectral properties. Timbre similarity has been linked to spectral-envelope characteristics (McAdams et al., 1995). Similarity-based behavioral groupings of stimuli reflect a categorization into distinct spectral-envelope types (Wedin and Goude, 1972) or show that the exchange of spectral envelopes between synthesized instruments results in an analogous inversion of positions in multidimensional timbre space (Grey and Gordon, 1978). Furthermore, Strong and Clark (1967b) 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 and Carterette (1993) 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 (Sandell, 1995; Tardieu and McAdams, 2012), quantified through the global spectral-envelope descriptor *spectral centroid*, with 'dark' referring to lower centroids. Strong blend was found to be best explained by 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* in this context. Reuter (1996) reported that blend occurs whenever the formants of two instruments coincide in frequency, hypothesizing that the non-coincidence would prevent auditory

fusion due to incomplete concealment of these presumedly salient spectral traits, thus facilitating the detection of distinct instrument identities.

As prominent signifiers of spectral envelopes, formants have been applied to the acoustical description of orchestral wind instruments and, like the formant structure found in the human voice, they exhibit frequency locations that are largely invariant to pitch change (Schumann, 1929; Saldanha and Corso, 1964; Strong and Clark, 1967a; Luce and Clark, 1967; Wedin and Goude, 1972; Luce, 1975; Grey and Gordon, 1978; Reuter, 1996; Meyer, 2009). 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. 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. The most advanced development of the Auditory Image Model (AIM) employs *dynamic, compressive gammachirp* (DCGC) filterbanks that adapt filter shape to signal level (Irino and Patterson, 2006). Its *auditory images* show a direct correspondence to acoustical spectral-envelope traits for human-voice and musical-instrument sounds (van Dinther and Patterson, 2006). AIM may aid in assessing the relevance of hypotheses concerning blend due to previous theories not taking auditory filters and spectral-masking effects into account.

This paper addresses whether pitch-invariant spectral-envelope characterization is relevant to blending. Section 2.2 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 2.3 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 2.4 and 2.5, respectively. Finally, the combined results from acoustical and perceptual investigations are discussed in Section 2.6, leading to the establishment of a spectral model for blend in Section 2.7.

2.2 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 comprise (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 concerns spectral aspects, all selected samples consist of long, sustained notes without vibrato. As spectral envelopes commonly exhibit significant variation across dynamic markings (see Appendix B), all samples include only *mezzoforte* markings, representing an intermediate level of instrument dynamics.

2.2.1 Spectral-envelope description

Past investigations of pitch-invariant spectral-envelope characteristics pursued comprehensive assessments of spectral analyses encompassing extended pitch ranges of instruments (Schumann, 1929; Luce and Clark, 1967; Luce, 1975). The spectral-envelope description employed in this paper is 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 a partial-tone detection routine. 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 yields the spectral-envelope estimates. The procedure balances the contrary aims of achieving a detailed spline fit and a linear regression, involving iterative minimization of deviations between estimate and the composite distribution until an optimal criterion is met (see Appendix A). The spectral-envelope estimates then serve as the basis for the identification and categorization of formants. The *main formant* represents 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 are derived from the estimated spectral envelope. They comprise the frequencies of the formant maximum F_{max} as well as *upper* and *lower* bounds (e.g., $F^{\rightarrow}_{3dB}$ and $F^{\leftarrow}_{3dB}$) at which the power magnitude decreases by either 3 dB or 6 dB relative to F_{max}.

The spectral-envelope estimates for all investigated instruments generally suggest pitch-invariant trends, as shown in Figure 2.1. A narrower spread of the partial tones (circles)

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

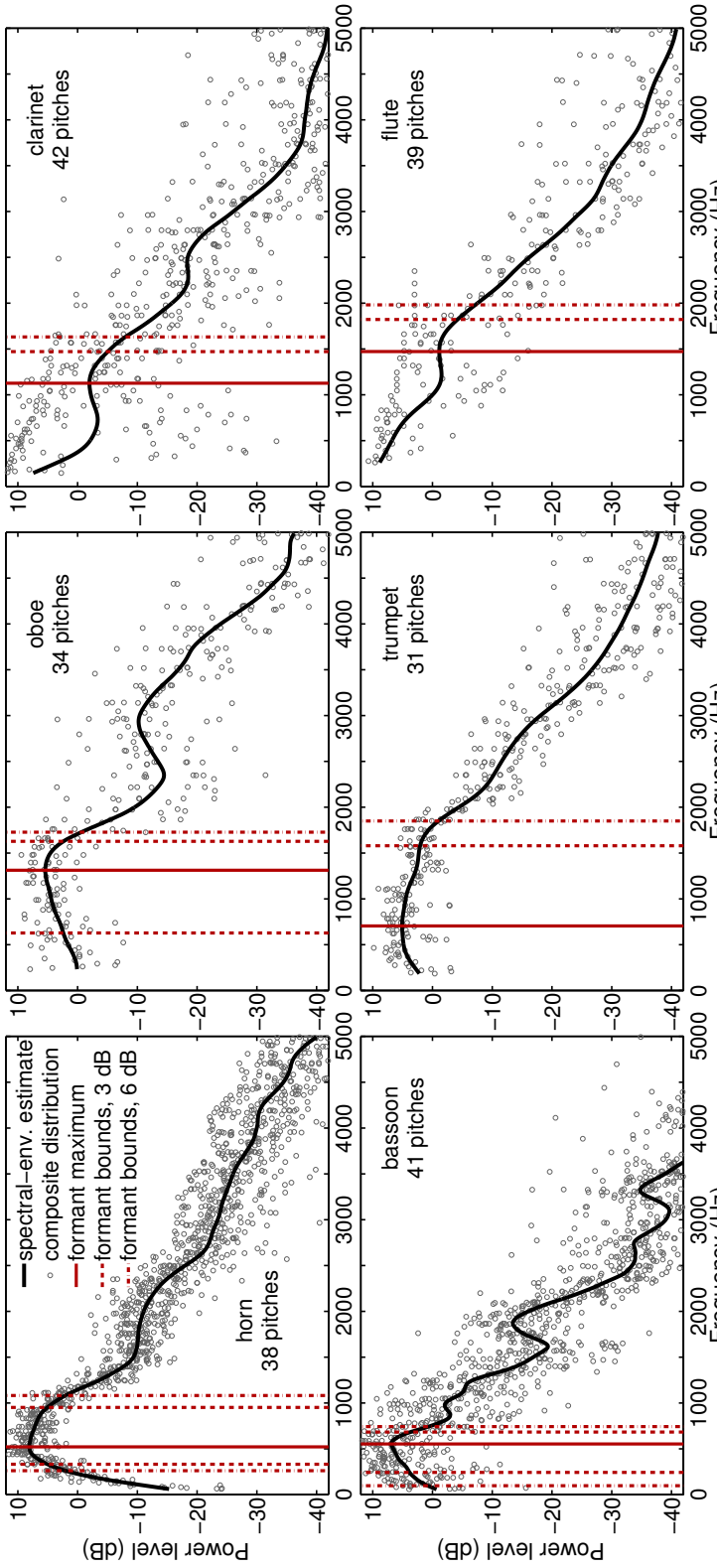

Fig. 2.1 Estimated spectral-envelope descriptions for all six instruments (labelled in individual panels). Estimates are based on the composite distribution of partial tones compiled from the specified number of pitches for each instrument.

around the estimate (curve) argues for a stronger pitch-invariant trend. The lower-pitched instruments horn and bassoon (left panels) exhibit strong tendencies for prominent spectral-envelope traits, i.e., formants. Higher-pitched instruments yield two different kinds of description. Oboe and trumpet (middle panels) display 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), display 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 prevents 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 vary as a function of pitch. Figure 2.1 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) is located above the pitch-variant low frequencies.

2.2.2 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. The 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 spectral envelopes. The SAIs are derived from the DCGC basilar-membrane model, comprising 50 filter channels, equidistantly spaced along equivalent-rectangular-bandwidth (ERB) rate (Moore and Glasberg, 1983) and covering the audible range up to 5 kHz[2]. A time-averaged SAI magnitude profile is obtained by computing the medians across time frames per filter channel, which resembles the auditory excitation pattern (van Dinther and Patterson, 2006).

A strong similarity among SAIs across an extended range of pitches is taken as an indicator for pitch-invariant tendencies. Pearson correlation matrices for all possible pitch

[2]As band-limited analysis economizes 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 for the purposes of analysis with AIM.

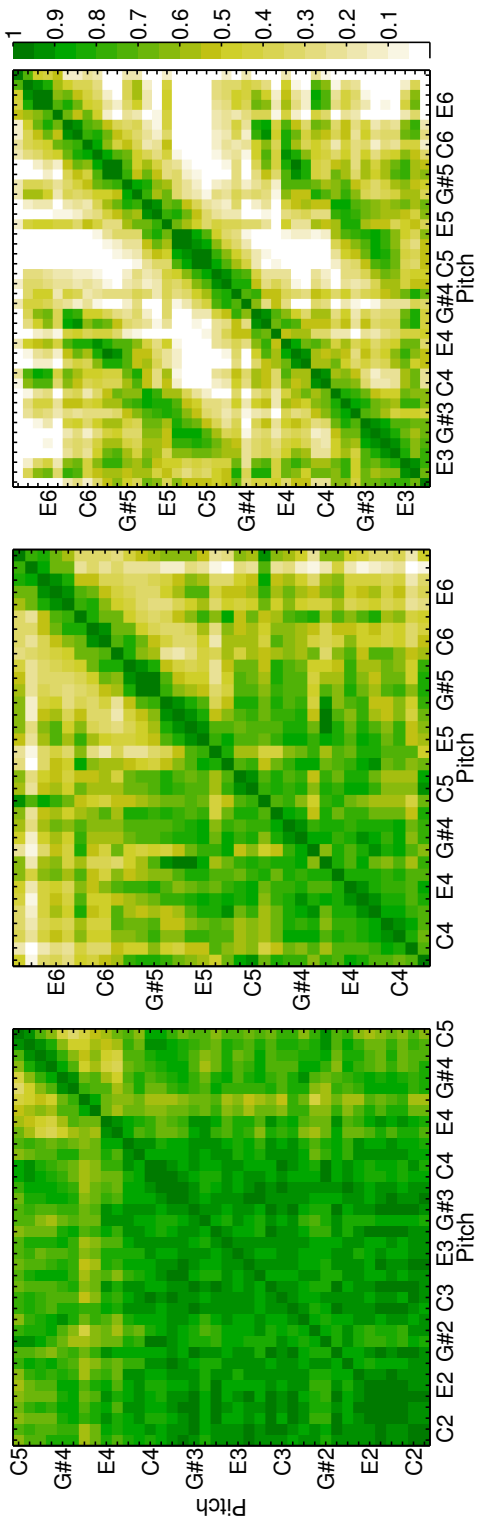

Fig. 2.2 SAI correlation matrices for horn, oboe, and clarinet (left-to-right). Correlations (Pearson r) consider all possible pitch combinations, with the obtained r falling within $[0, 1]$ (see legend, far-right).

combinations are computed, comparing the profiles of SAI magnitudes over filter channels. In addition, this approach also aids in identifying the limits of pitch invariance, as adjacent regions exhibiting weaker correlations delimit instrument registers where SAIs vary as a function of pitch. Three representative cases are illustrated in Figure 2.2. For horn (left panel) and bassoon (not shown), broad regions of pitch-invariant SAI profiles become apparent (dark square), spanning large parts of their ranges up to pitches of about D4. Oboe (middle panel) and trumpet (not shown) exhibit 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 appears to be more prevalent and stable in lower pitch regions, from which low-pitched instruments in particular would benefit. All of these instruments lose pitch-invariant tendencies in their high registers. The remaining instruments, clarinet (right panel) and flute (not shown), lack widespread pitch-invariant SAI characteristics, as strong patterns of correlation are only obtained between directly adjacent pitches (diagonal) and not across wider pitch regions.

2.2.3 Parametric variation of main-formant frequency

In order to study the contribution of local variations of spectral characteristics, a synthesis model is employed that provides parametric control over separate spectral-envelope components. The synthesis infrastructure is based on a source-filter model and realized for real-time modification of the control parameters (see Appendix D). During synthesis, the filter structure is fed a harmonic source signal of variable fundamental frequency, containing harmonics up to 5 kHz. The filter structure consists 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 is implemented as an absolute deviation ΔF in Hz from a predefined origin, i.e., $\Delta F = 0$. Analogue models for each instrument are designed for $\Delta F = 0$ by matching the frequency response of the composite filter structure to the spectral-envelope estimates, as illustrated in Figure 2.3 for the horn (dashed black line), superimposed over its corresponding estimate (solid grey line). The analogues are 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 helps to deduce the measured perceptual differences that result from frequency

relationships between them. It should be noted that the synthesis filter structure for the clarinet excludes its pitch-variant lower frequency region (see Section 2.2.1). It only models the formant above that region, as well as the remaining spectral envelope towards higher frequencies, in order to orient the investigation toward specifically testing the relevance of the identified, albeit less pronounced, formant.

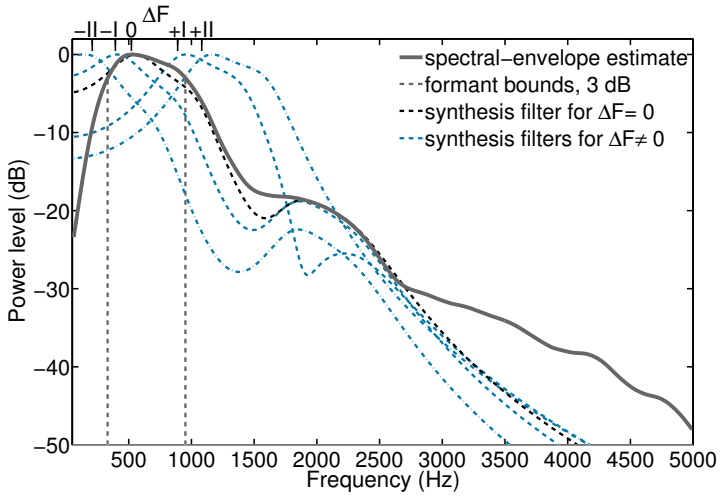

Fig. 2.3 Spectral-envelope estimate of horn and filter magnitude responses of its synthesis analogue. The original analogue is modeled for $\Delta F = 0$; the other responses are variations of ΔF. The top axis displays the equivalent scale for the five ΔF levels investigated in Experiment 2.

2.3 General methods

The perceptual relevance of the main-formant characteristics outlined in the previous section to blending is tested for sound dyads. All dyads comprise a *sampled* instrument and its *synthesized* analogue model. In a given dyad, the instrument sample is constant, and its synthesized analogue is variable with respect to the parameter ΔF. Variations with $\Delta F > 0$ shift the main formant of the synthesized sound higher in frequency relative to the sampled instrument's main formant and, accordingly, $\Delta F < 0$ corresponds to shifts toward lower frequencies. Two perceptual experiments are conducted to investigate how ΔF variations relate to blend. In Experiment 1, participants control ΔF directly and are asked to find the ΔF that gives optimal blend, whereas in Experiment 2, listeners provide direct blend ratings for predefined ΔF variations. Using the instruments presented

in Section 2.2, the robustness of perceptual effects is assessed over the two experimental tasks for different pitches, unison and non-unison intervals, and stimulus contexts. With six instruments and several investigated factors, a full-factorial experimental design would have been impractical. An exploratory sampling of scenarios was chosen instead, with not all instruments being tested across all factors. However, each factor is studied with at least three instruments. Pitches are chosen to represent common registers of the individual instruments. Non-unison intervals include 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.

2.3.1 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 (Martin and Champlin, 2000; ISO 389–8, 2004). No participant took part in both experiments.

2.3.2 Stimuli

All stimuli involve dyads, comprising one *sampled* (drawn from VSL) and one *synthesized* sound. For a sample at any given pitch, the spectral envelope is approximated by the pitch-invariant description from Section 2.2.1, which results in the main formants of sampled and synthesized sounds resembling each other for $\Delta F = 0$. With regard to the temporal envelope, both instruments are synchronized in their note onsets, followed by the sustain portion and ending with an artificial 100-ms linear amplitude decay ramp applied to both instrument sounds. The sampled sound retains its original onset characteristics, whereas across all modeled analogues, the synthesized onsets are characterized by a constant 100-ms linear amplitude ramp. Stimuli are 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

Appendix C).

2.3.3 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.

2.3.4 Data analysis

With respect to the investigated factors, Experiment 1 evaluates the influence of the factors instrument register and interval type. Experiment 2 assesses pitch-invariant perceptual performance across a number of factors and furthermore correlates 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 are conducted, Bonferroni corrections are applied. In repeated-measures ANOVAs, the Greenhouse-Geisser correction (ε) is applied whenever the assumption of sphericity is violated. In addition, Experiment A also considered one-sample t-tests against a mean of zero for testing differences from $\Delta F = 0$. Statistical effect sizes η_p^2 and r are reported for ANOVAs and t-tests, respectively. The analyses consider participant-based averages for trial repetitions of identical conditions.

2.4 Experiment 1

2.4.1 Method

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 CAD.

Stimuli

Table 2.1 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 compare unison intervals to consonant (7 or -3 ST) and dissonant (6 or -2 ST), non-unison intervals. Two levels of the Register factor contrast 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 2.2.2. 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 to maximize perceived blend.

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

Table 2.1 Seventeen dyad conditions from Experiment 1 across instruments, pitches, and intervals (top-to-bottom). Intervals in semitones relative to the specified reference pitch.

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 Figure 2.4 (top), minimal or maximal Γ limits 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.

2.4.2 Results

General trends

All six instruments exhibit the common trend that *optimal* blend is associated with deviations $\Delta F \leq 0$. In other words, optimal blend is limited to only one side of the possible frequency relationships between main formants, relative to the case of formant coincidence ($\Delta F = 0$). Two different patterns for the optimal ΔF become apparent among instruments. Figure 2.6 (diamonds in lower part) illustrates these by situating the mean optimal ΔF obtained in Experiment 1 relative to a continuous scale of equivalent ΔF levels from Experiment 2, with *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 falls in direct proximity to $\Delta F = 0$. For the unison intervals of horn and bassoon, ΔF does not differ significantly from zero; the other two instruments underestimate it [$t(16) \leq -5.6$, $p < .0001, r \geq .82$]. By contrast, optimal ΔF for the clarinet and flute (right panel) are relatively distant from the $\Delta F = 0$, in line with significant underestimations [$t(16) \leq -3.8$, $p \leq .0015, r \geq .69$].

Instrument register and interval type

The influence of instrument register on the obtained ΔF leading to optimal blend was investigated for trumpet, bassoon, and clarinet at pitches corresponding to instrument-specific low and high registers. One-way repeated-measures ANOVAs for Register yield moderately strong main effects for trumpet and bassoon [$F(1, 16) \geq 19.2, p \leq .0005, \eta_p^2 \geq .55$], showing that the chosen ΔF differs in the high register. A less pronounced effect was obtained for the clarinet [$F(1, 16) = 5.3, p = .036, \eta_p^2 = .25$].

Differences in optimal ΔF between interval types were also investigated, which involved comparisons between unison and non-unison intervals as well as a distinction between

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

[4] ΔF are 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.

consonant and dissonant for the latter. One-way repeated-measures ANOVAs on Interval conducted for horn, bassoon, clarinet, and trumpet only lead to a weak main effect for the trumpet [$F(2,32)=3.7, p=.035, \eta_p^2=.19$]. Post-hoc tests for the three possible comparisons yield a single significant difference between the interval sizes 0 and 6 ST [$t(16)=-3.5$, $p=.003, r=.65$].

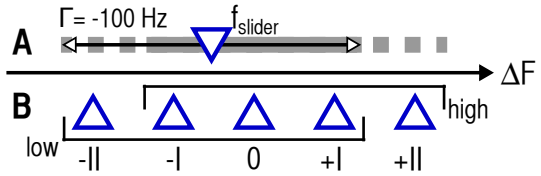

Fig. 2.4 ΔF variations investigated in Experiments 1 and 2 (labelled 'A' and 'B', respectively). A: Participants control f_{slider}, which provides a constant range of 700 Hz (white arrows). Γ (e.g., -100 Hz) represents a randomized roving parameter, preventing the range from always being centered on $\Delta F=0$. B: Participants rate four dyads varying in ΔF, drawn from the *low* or *high* context. The contexts represent subsets of four of the total of five predefined ΔF levels.

2.5 Experiment 2

2.5.1 Method

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 CAD.

Stimuli

Table 2.2 lists the 22 investigated dyad conditions. The Interval factor investigates two levels, comparing unison to non-unison (6 or -2 ST, dissonant) intervals. Depending on the instrument, the Pitch factor involves two (horn, bassoon, trumpet, clarinet) or three (oboe, flute) levels, in the former case, interwoven with the levels of Interval. In addition, this experiment includes two factors that are related to ΔF variations alone, which apply to all

conditions listed in Table 2.2. The first is synonymous with ΔF, as it explores 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 groups the five levels contextually into two subsets of four, which are denoted as *low* and *high* contexts and defined in Figure 2.4 (bottom).

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	

Table 2.2 Twenty-two dyad conditions from Experiment 2 across instruments, pitches, and intervals (top-to-bottom). Intervals in semitones relative to the specified reference pitch.

Employing the formant descriptors from Section 2.2.1, the investigated ΔF levels are expressed on a common scale of spectral-envelope description, which provides a better basis of comparison than taking equal frequency differences in Hz, as the frequency extent of formants across instruments varies considerably. Figure 2.5 provides examples for all resulting ΔF levels of the horn. The four levels $\Delta F \neq 0$ are defined as frequency distances between the formant maximum F_{max} and measures related to the location and width of its bounds (e.g., $F^{\rightarrow}_{6dB}$ or ΔF_{3dB}). For example, the positive deviation $\Delta F_1(+I)$ is the distance between the formant maximum and its upper bound minus 10% of the width between the 3 dB bounds. If spectral-envelope descriptions lack lower bounds (e.g., trumpet, clarinet, flute), the frequency located below F_{max} that exhibits the lowest magnitude is taken as a substitute value.

Unlike the dyads in Experiment 1, the synthesized sound always remains at the reference pitch, whereas the sampled sound varies its pitch for non-unison intervals, because this tests the assumption of pitch-invariant description for the recorded sounds more thoroughly. The dyads have a constant duration of 4700 ms. In addition, the conditions listed in Table 2.2, 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 subsequently verified by the second author. Loudness equalization was conducted subjectively, anchored to a global reference dyad for all conditions and ΔF levels. For all stimuli, the equalization gain levels

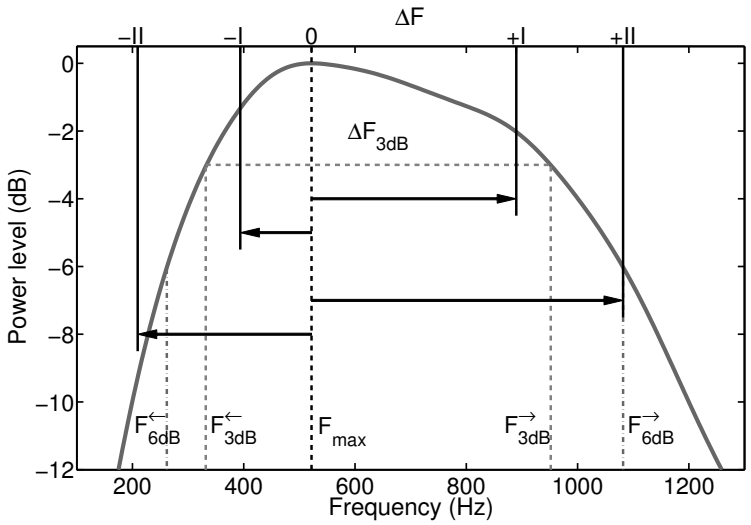

Fig. 2.5 ΔF levels from Experiment 2, defined relative to a spectral-envelope estimate's formant maximum and bounds. $\Delta F(\pm I)$ fall 10% inside of ΔF_{3dB}'s extent. $\Delta F(+II)$ aligns with $F_{6dB}^{\rightarrow}$, whereas $\Delta F(-II)$ aligns with either $80\% \cdot F_{6dB}^{\leftarrow}$ or 150 Hz, whichever is closer to F_{max}.

considered medians that were obtained either after their corresponding interquartile ranges fell below 4 dB or after running a maximum of 10 participants.

Procedure

A relative-rating task required participants to compare ΔF levels for a given condition from Table 2.2. In each experimental trial, participants were presented with four dyads and asked to provide four corresponding ratings. The four dyads represent one of the two ΔF contexts. A continuous rating scale was employed, which spanned from *most blended* to *least blended* (values 1 to 0) 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 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]Only the 22 conditions investigating ΔF are reported here.

2.5.2 Results

General trends

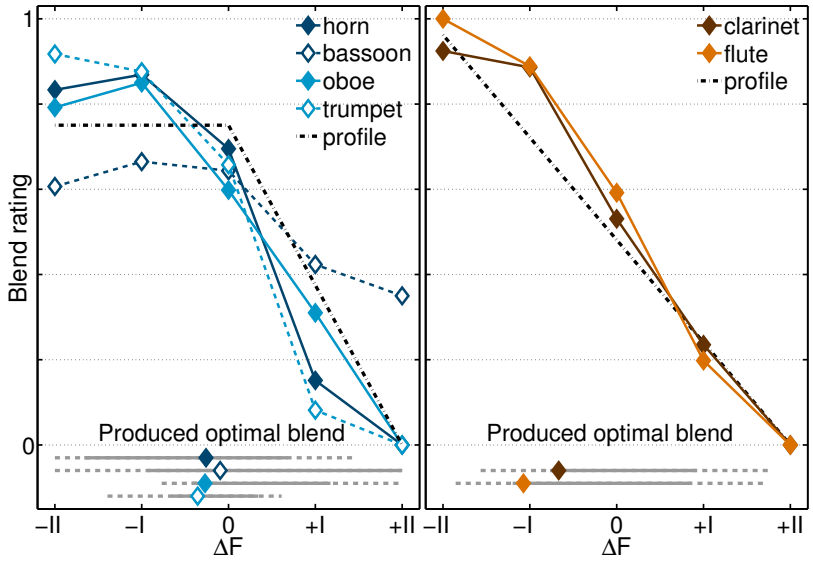

Fig. 2.6 Perceptual results for the different instruments, grouped according to two typical response patterns (left and right panels). Experiment 1 (diamonds, bottom): mean ΔF for produced optimal blend, transformed to a continuous scale of ΔF levels. The grey lines indicate slider ranges (compare to Figure 2.4, top). Experiment 2 (curves): median blend ratings across ΔF levels and typical profile.

Trends across ΔF levels are evaluated for median ratings that are collapsed across the factors Pitch, Interval, and Context. Potential effects along these additional factors are addressed in the next section. As shown in Figure 2.6, high blend is never associated with the levels $\Delta F > 0$, which always yield low ratings. In terms of higher blend, two typical rating profiles (dashed-and-dotted curves) as a function of ΔF emerge: 1) For the instruments horn, bassoon, oboe, and trumpet (left panel), the ratings resemble the profile of a *plateau*. Medium to high blend ratings are obtained at and below $\Delta F = 0$, with a marked decrease of ratings above this boundary. 2) The instruments clarinet and flute (right panel) exhibit a monotonically decreasing and approximately *linear* rating profile as ΔF increases, in which the $\Delta F = 0$ level does 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.

Blend and pitch invariance

Whenever the profiles of blend ratings over ΔF remain largely unaffected by different pitches, intervals, and ΔF contexts, we will assume that perceptual performance is pitch-invariant. For instance, Figure 2.7 suggests this tendency for the horn, in which the *plateau* profile is maintained over all factorial manipulations. First, the main effects across ΔF are tested to confirm that ratings serve as reliable indicators of perceptual differences. Given these main effects, perceptual robustness to pitch variation is fulfilled if no $\Delta F \times$ Pitch or $\Delta F \times$ Interval interaction effects are found across both ΔF contexts. An absence of main effects between ΔF contexts would indicate further perceptual robustness. As the Context factor only involves ΔF levels common to both the *high* and *low* contexts, namely *0* and $\pm I$ (see Figure 2.4, bottom), the ratings for these levels require range normalization and separate analyses from the remaining factors. For the instruments involving the Interval factor, these are conducted on split levels of that factor. The experimental task imposed the usage of the rating-scale extremes, which resulted in several violations of normality. 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 include non-parametric approaches of rank transformation (Conover and Iman, 1981) and prior alignment of 'nuisance' factors (Higgins and Tashtoush, 1994).[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 along ΔF are found for all instruments, indicating this factor's utility in measuring perceptual differences. Table 2.3 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 Figure 2.6. Fur-

[6] Given the unavailability of non-parametric alternatives for a repeated-measures, three-way ANOVA which 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 Type I errors. Rank transformation is a common approach in non-parametric tests, such as the one-way Friedman test (Conover and Iman, 1981). Issues with tests for interaction effects losing power in the presence of strong main effects are addressed through 'alignment' of the raw data prior to rank transformation (Higgins and Tashtoush, 1994). 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 global and condition-based ranking methods.

Effect	Stat.	low context conserv.	low context liberal	high context conserv.	high context 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

Table 2.3 Range of ANOVA main effects along ΔF across all six instruments.

thermore, the instruments horn, bassoon, oboe, and trumpet are found to meet the requirements that argue for pitch-invariant robustness of the rating profiles. There is only one exception from a complete absence of effects interacting with ΔF: a moderate main effect for Context is found for trumpet only at non-unison intervals [$F(1,19) = 10.48|25.04$, $p = .0043|.0001$, $\eta_p^2 = .355|.569$]. By contrast, the remaining instruments clarinet and flute depart from exhibiting pitch-invariant robustness, as they clearly violate the assumptions across both ΔF contexts. The interaction effects with ΔF and a main effect for Context leading to their disqualification are described in Table 2.4.

The pitch-invariant group in fact represents the same instruments for which the blend-rating profiles resemble a plateau, attributing a special role to $\Delta F = 0$ as a perceptually relevant boundary. To further confirm this observation by joint analysis of the four instruments, two hierarchical cluster analyses are employed that group ΔF levels based on their similarity in perceptual ratings or auditory-model representation. The first cluster analysis reinterprets rating differences between ΔF as a dissimilarity measure. This measure considers effect sizes of statistically significant non-parametric post-hoc analyses (Wilcoxon signed-rank test) for pairwise comparisons between ΔF levels. For non-significant differences, dissimilarity is assumed to be zero. The second analysis relies on correlation coefficients (Pearson r) between dyad SAI profiles across ΔF levels (see Figure 2.9 for examples). The dissimilarity measure considers the complement value $1 - r$, and as all correlations fall within the range $[0, 1]$, no special treatment for negative correlations is

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	-	-

Table 2.4 ANOVA effects for clarinet and flute leading to the departure from pitch-invariant robustness.

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

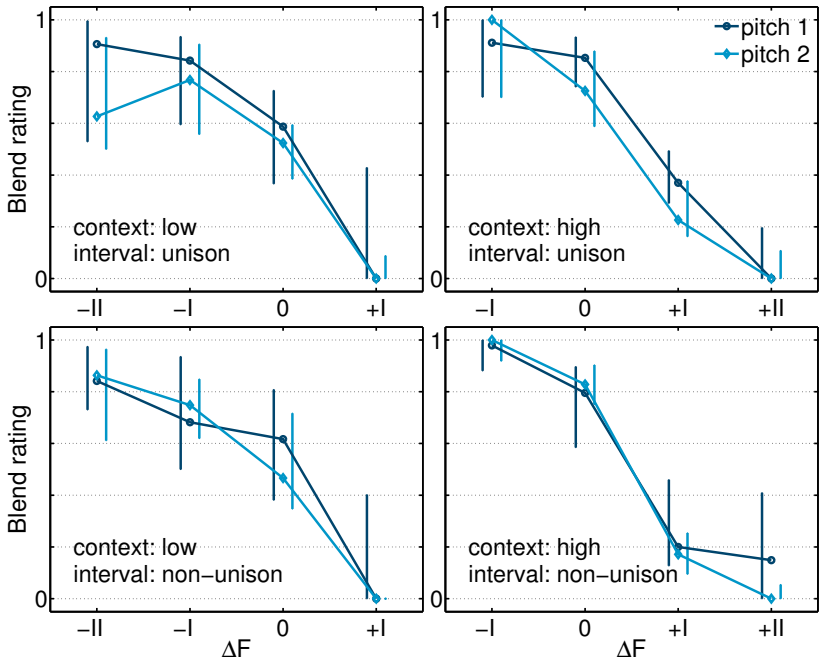

Fig. 2.7 Medians and interquartile ranges of blend ratings for horn across ΔF levels and the factorial manipulations Pitch × Context × Interval.

required. Both cluster analyses employ complete-linkage algorithms. The dissimilarity input matrices were obtained by averaging 30 independent data sets, compiled across the four instruments, and the factors Context, Pitch, and Interval. As shown in Figure 2.8, both analyses lead 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 levels associated with low and high degrees of blend group into two distinct clusters, clearly relating to the *plateau* profile, where $\Delta F = 0$ defines its high-blend boundary.

Blend and its spectral correlates

Explaining blending between instruments with the help of spectral-envelope characteristics would eventually allow these instrument-specific traits to predict blend. In addition, it would aid in establishing a perceptual model that addresses the contribution of spectral characteristics. Given this aim, multiple linear regression was employed to model the median blend ratings. The regression models allow an assessment of the relative contributions of regressors describing both global and local spectral-envelope traits. Global descriptors

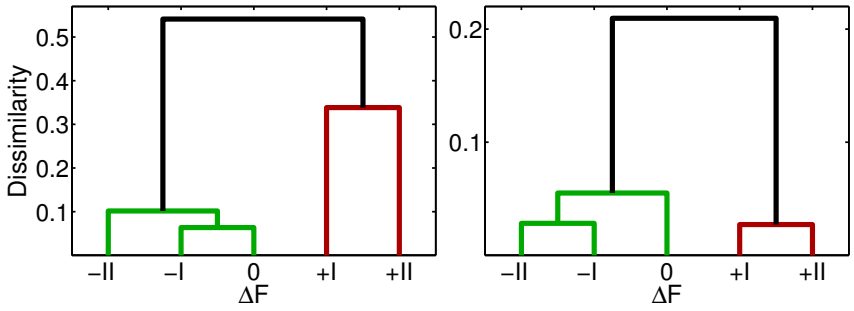

Fig. 2.8 Dendrograms of ΔF-level groupings for the pitch-invariant instruments. Dissimilarity measures are derived from perceptual ratings (left) and auditory-modelled dyad SAI profiles (right).

involve the commonly reported *spectral centroid* and *spectral slope* (Peeters et al., 2011), whereas local descriptors concern the formant characterization discussed in Section 2.2.1. Because a dyad yields two descriptor values across its constituent sounds, the regressor measure has to associate the two in some way. For the spectral centroid, two measures are considered, namely, *composite* (sum) and *absolute difference* (Sandell, 1995). Although the former relates to the 'darker'-timbre hypothesis mentioned in Section 2.1, the latter was still found to best explain blend in non-unison intervals, which leaves some uncertainty as to which of these two measures is more appropriate in explaining blend in general. All other regressors are implemented as polarity-preserving differences between descriptors, with the *sampled* instrument serving as the reference (R) and the *synthesized* instrument being variable (V) across ΔF. For example, the difference of descriptor values d_x for instrument x corresponds to $\Delta d = \ d_R - d_V$.

Regression models are investigated for two separate subsets of the behavioral data: pitch-invariant and pitch-variant instruments. Regressor variables are pooled from the spectral-envelope descriptors and additional variables that are included to account for potential confounding factors, e.g., pitch, interval. If these factors do not contribute as regressors, this would further support a pitch-invariant perceptual robustness. The initial pool of regressors comprised 32 variables, subsequently reduced to a pre-selected set that exhibits inter-variable correlations $|r| < .7$. The pre-selection was determined by first identifying the variable that in simple linear regression exhibits the highest R^2 and subsequently adding all remaining variables that yield permissible inter-variable correlations. Table 2.5 lists the pre-selected variables entered into the regression, which comprise

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 converge on optimum models by iteratively adding or eliminating regressors, respectively. In addition, customized models considering similar regressors were explored as well. In anticipation of reporting the results, the inclusion of a binary regressor for ΔF context $C_{lo/hi}$ benefits all investigated regression models, as it succeeds in correcting for the systematic offset of scaled ratings between the low and high contexts (see Figure 2.7).

No.	Variable	Description
1	$\Delta L_{3dB}^{\rightarrow}$	derivate of $F_{3dB}^{\rightarrow}$, Equation 2.1
2	$\Delta L_{F_1 vs F_2}$	ΔL between formants F_1 & F_2
3	$\Delta S_{slope_{F1}}$ [a b]	spectral slope above $F_{3dB}^{\rightarrow}$
4	ΔS_{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 2.2
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 between instruments
12	AM_{depth} [b d]	amplitude modulation depth

Table 2.5 Variables entering stepwise-regression algorithm to obtain models reported in Table 2.6.

[a] Not for pitch-variant subset, inter-variable correlation $|r| > .7$
[b] Computed as described in Peeters et al. (2011)
[c] Accounting for pitch
[d] Accounting for perceivable beating between partial tones

The strongest spectral-envelope descriptors in simple regression models all concern local formant characterization and do not involve the global descriptors. Among the formant descriptors, the highest correlations are obtained for the main-formant upper bound $F^{\rightarrow}_{3dB}$, 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} does not perform better than $F^{\rightarrow}_{3dB}$, likely due to differing skewness properties between formants of different instruments (see Section 2.2.1). At the same time, the utility of $F^{\rightarrow}_{3dB}$ implies that it could assume an important role in explaining blend, as perhaps the perceptually most salient feature of formants. It performs slightly better for the pitch-variant than for the pitch-invariant subset, which is explained by the fact that $\Delta F^{\rightarrow}_{3dB}$ essentially follows a monotonic, quasi-linear function across ΔF. This inherent property apparently models the *linear* blend profile better. In order to improve the performance for the *plateau* blend profile as well, derivate descriptors of $F^{\rightarrow}_{3dB}$ were explored. The most effective derivate $\Delta L^{\rightarrow}_{3dB}$ relates the difference between upper bounds $F^{\rightarrow}_{3dB}$ of the two instruments to a corresponding difference in the reference instrument's spectral-envelope magnitude L_R, as formalized in Equation 2.1.

$$\Delta L^{\rightarrow}_{3dB} = L_R\left(F^{\rightarrow}_{3dB|R}\right) - L_R\left(F^{\rightarrow}_{3dB|V}\right) \tag{2.1}$$

The final regression solutions yield identical models for both instrument subsets, involving the regressors $\Delta L^{\rightarrow}_{3dB}$, absolute spectral-centroid difference $|\Delta S_{centroid}|$, and context $C_{lo/hi}$. A slight gain in performance is achieved by substituting the audio-signal-based $|\Delta S_{centroid}|$ with a variant computed on the spectral-envelope estimates. Table 2.6 displays these optimized regression models for pitch-invariant and pitch-variant subsets, both leading to 87% explained variance. The patterns for the standardized regression-slope coefficients β_{std} are very similar for both subsets. In these models, $\Delta L^{\rightarrow}_{3dB}$ acts as the strongest predictor for the blend ratings, contributing about five times more than $|\Delta S_{centroid}|$, which furthermore does not perform better than the correcting influence of $C_{lo/hi}$. These findings clearly argue for local spectral-envelope descriptors to be more meaningful than global ones in explaining blending. Moreover, the remaining global descriptor *spectral slope* appears 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-invariant descriptions for all instruments. Despite the findings in Section 2.5.2 arguing against pitch-invariant perceptual robustness for clarinet and flute, the obtained regression models exclude the

Pitch and Interval factors.

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

Table 2.6 Best obtained multiple-regression models predicting timbre-blend ratings, for two instrument subsets.

2.6 General discussion

Orchestrators would benefit from acoustical descriptions of instruments that correspond to the perceptual processes involved in achieving blended timbres. Section 2.2 suggests that common orchestral wind instruments are reasonably well described through pitch-invariant 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 and continuously spans extended pitch regions. 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 2.3 to 2.5 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-invariant perceptual robustness 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 robustness ceases in higher instrument registers.

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 $F^{\rightarrow}_{3dB}$, when expressed as a derivate descriptor $\Delta L^{\rightarrow}_{3dB}$, 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 perceptual robustness to pitch found in Section 2.5.2 contradict the general utility of pitch-invariant spectral-envelope description in predicting blend ratings, as reported in Section 2.5.2. Taking both findings into account, this for one argues that the descriptor $F^{\rightarrow}_{3dB}$ 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 $\Delta L^{\rightarrow}_{3dB}$ 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 exhibiting coherent micromodulations of partial tones. These modulations have been shown to contribute to a stronger

and more stable auditory image (McAdams, 1984) 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 while an accompanying instrument avoids exceeding the lead instrument's main-formant frequency. Likewise, returning to the notion of blend leading to *augmented* timbres (Sandell, 1995), the dominant timbre in such a mixture would seem predestined to function as the reference.

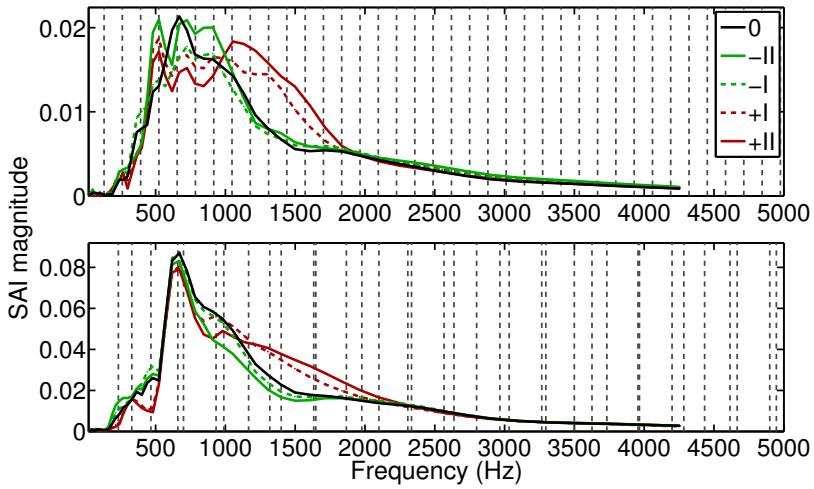

Fig. 2.9 SAI profiles of dyads for all ΔF levels (Experiment 2), depicting two experimental conditions for horn. Top: pitch 1, unison; bottom: pitch 2, non-unison; the grid lines correspond to partial-tone locations.

Finally, the results allow a reassessment of previous explanations for blend. The 'darker'-timbre hypothesis (Sandell, 1995) 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 (Reuter, 1996) 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 ΔF relationships only achieves partial fulfillment, it shows more agreement in the corresponding SAI representations. As shown in two example cases for horn in Figure 2.9, 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 $\Delta F < 0$ (Figure 2.3) are not reflected in the corresponding dyad SAI profiles (Figure 2.9), which instead exhibit direct alignment below 500 Hz for all three levels $\Delta F \leq 0$.

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 Figure 2.9 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. This reflects tendencies for pitch-invariant traits in SAI correlations (see Section 2.2.2) to be more pronounced at lower pitch ranges, and would support the 'darker-timbre' hypothesis in terms of a pitch-related containment of the spectral centroid.

2.7 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 2.1.

[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 Figure 2.9, the correction manifests itself in the compressed frequency extent for the SAIs.

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 of 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 as well as 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.

Chapter 3

Acoustical correlates for blend in mixed-timbre dyads and triads

This chapter considers a correlational analysis which associates the perceived degree of timbre blend with a broad range of acoustical descriptors. The perceptual data were collected in two listening experiments (Experiments 3 and 4), investigating blend ratings for dyadic and triadic combinations of orchestral instruments, respectively. The content is based on the following research article:

Lembke, S.-A., Parker, K., Narmour, E. and McAdams, S. (In preparation). Acoustical correlates of perceptual blend in timbre dyads and triads. *Journal of the Acoustical Society of America*.

3.1 Introduction

In orchestration practice, composers need to consider several factors when they intend to achieve a *blended* timbre between two or more instruments playing concurrently. At first, there is the choice of suitable instruments that yield a blended combination, with this *suitability* related to acoustical traits of the instruments. The remaining factors involve more musical considerations: whether instruments will be playing in unison or non-unison; whether in the latter case, one or the other instrument is assigned the top voice; in what registral range the instruments will be playing; what kind of articulation they will employ (e.g., bowed or plucked string). When it comes to developing orchestration aids, such as predicting the expected degree of perceived blend for a given combination of instruments, pitches, and articulations, which relies on their acoustical description, the investigation

and validation of such tools should be based on equally diverse sets of perceptual blend judgments.

Previous research has defined perceived timbre blend as the auditory fusion of concurrent instrumental sounds, with individual sounds becoming less distinct. Two main approaches to measuring the degree of blend perceptually have been employed: 1) an indirect measure by deducing blend from increasing confusion in the identification of instruments in a mixture (Kendall and Carterette, 1993; Reuter, 1996) or 2) directly measuring blend through rating scales (Kendall and Carterette, 1993; Sandell, 1995; Tardieu and McAdams, 2012; Chapters 2 and 4).

All studies have hypothesized spectral features to influence blend, although with differences with respect to the employed spectral description. Using a global descriptor of the amplitude-weighted frequency average, the *spectral centroid*, the composite (or sum) of the individual sounds' centroids was found to predict blend in unison dyads best (Sandell, 1995; Tardieu and McAdams, 2012), whereas for non-unison dyads, the absolute difference in individual spectral centroids served as the more reliable predictor (Sandell, 1995). Another approach to spectral description identifies and analyzes the influence of prominent spectral features, such as maxima or *formants*. Their identification considers spectral estimations that are aggregated across an instrument's complete pitch range, and therefore can be considered *pitch-generalized*, where wind instruments in particular exhibit formant structures that remain largely invariant across pitch (see Chapter 2 and Appendix A). Similarity in the formant structure between instruments has been linked to explaining blend (Reuter, 1996), with hardly distinguishable instrument pairings exhibiting very similar formant locations (e.g., horn and bassoon), whereas the strongly pronounced and quite unique formant structure of the oboe may hinder it from blending with most other instruments. Focusing on the influence of the most prominent *main formants*, their frequency relationship between instruments appears to affect blend critically. In dyads comprising a recorded wind instrument sound and a synthesized analogue to that instrument, whose main-formant frequency could be shifted relative to that of the recorded sound, blend decreased drastically as the synthesized formant exceeded that of the recorded sound in frequency (Chapter 2). This relative dependency of one instrument on another relates to musical performance, where accompanying musicians adjust their main formants to be lower than when playing as the leading instrument (Chapter 4).

Apart from spectral properties, differences between temporal features, such as note

attacks or onsets, have been found to explain blend as secondary factors for unison dyads (Sandell, 1995), but their influence becomes more dominant as attacks turn impulsive, i.e., exhibit shorter durations and steeper attack slopes, leading to reduced degrees in blend (Tardieu and McAdams, 2012).

With respect to factors of a musical nature and those unrelated to timbre, blend for unison dyads is perceived as stronger than for non-unison combinations (Kendall and Carterette, 1993; Chapter 4). Furthermore, the assignment of instruments to the upper and lower pitches in non-unison intervals has resulted in differences in perceived blend between instrument inversions in one study (Kendall and Carterette, 1993), but lacked a comparable effect in another (Sandell, 1995). All of these studies on blend were limited to dyadic contexts, leaving open how the obtained results and proposed hypotheses would fare in combinations of three or more instruments. Little work has been published on timbre combinations in triadic contexts (Kendall, 2004; Kendall and Vassilakis, 2006, 2010), and none of these papers addresses issues directly related to blend.

With the aim of predicting perceived blend between arbitrary instrument combinations, linear correlation or regression can be employed to associate blend measures with single acoustical features (Sandell, 1995; Tardieu and McAdams, 2012), without, however, assessing the possibility of a combination of descriptors to model the behavioral data. Modeling the data on multiple descriptor variables would furthermore assess the relative contributions of different acoustical features to blend. Past attempts utilizing *multiple linear regression* (MLR) have succeeded in explaining up to 63% of the variance in blend ratings for mixed-instrument dyads (Sandell, 1995). The previously mentioned study investigating the impact of local, parametric variations of main-formant frequency in dyads sought to also correlate the resulting acoustical changes with blend ratings, leading to MLR models explaining up to 87% of the variance (Chapter 2). Yet, the MLR approach has many limitations, as a high collinearity among independent variables (regressors) as well as a low number of cases compared to the number of regressors leads to less reliable and valid results, as well as mathematically ill-defined solutions. This becomes problematic to the aim of the current paper, because many spectral descriptors are known to exhibit a high inter-correlation (Peeters et al., 2011), which for conventional MLR leaves two options: 1) disregarding the collinearity, at the risk of obtaining less reliable or invalid results or 2) eliminating regressors that are collinear to a reference regressor, i.e., one found to predict blend most strongly in simple linear regression. However, this latter approach risks excluding variables that

might perform even better than the selected one once they interact with other regressors. A viable solution to deal with collinearity is to employ a dimension-reduction technique like *principal components analysis* (PCA) that reduces a high quantity of regressors to a small number of substitute or latent variables, i.e., *principal components* (PCs), which are orthogonal to one another. These PCs can thereafter serve as regressors that represent the common aspects for groups of collinear descriptors (e.g., Giordano et al., 2010). A promising regression method that uses PCA as an integral part is *partial least-squares regression* (PLSR), which originates from the discipline of chemometrics, but more recently has been applied within the field of auditory perception (Rumsey et al., 2005; Kumar et al., 2008; Eerola et al., 2009), allowing analysis of complex correlational relationships between perceptual measures and arrays of acoustical or psychoacoustical variables.

The current investigation uses PLSR in an attempt to predict blend ratings from perceptual experiments. The perceptual data are collected on a diverse set of variables that affect timbral blend and orchestration, including different instruments, pitches, unison and non-unison intervals, as well as dyadic and triadic contexts. The set of potential regressors consists of a wide range of acoustical measures that, through several stages of PLSR models, are continually refined to retain only the meaningful set of regressors and, importantly, ones that are independent of each other. Section 3.2 introduces the PLSR approach, the two investigated perceptual data sets, and the set of regressors considered. The results are presented in Section 3.3, followed by a general discussion and concluding remarks in Sections 3.4 and 3.5, respectively.

3.2 Methods

3.2.1 Partial least-squares regression (PLSR)

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

into k PCs, yielding the relationship $X = T \cdot P'$, with T representing an $m \times k$ matrix of *loadings* and P an $n \times k$ matrix of *scores*. Unlike computing a PCA on X independently and inputting the obtained P into MLR, PLSR achieves the component decomposition by inherent maximization of the covariance between y and X, leading to a better predictive relationship. The PLSR implementation used here is SIMPLS (de Jong, 1993).

Performance, predictive power, and reliability

Regression performance evaluates the variation in y that is explained by the model, commonly considering R^2. This measure describes both the global and component-wise performance, with the latter quantifying the relative contribution of PCs. However, with increasing k models are prone to over-fit the data, at the cost of predictive power when applied to other data. In order to assess the predictive power of models, sixfold cross validation (CV) is employed, partitioning the n cases into six subsets of similar size, building models based on five subsets, assessing the error in predicting the remaining subset, and repeating the last two steps for all permutations of subsets. CV also involves ten Monte-Carlo repetitions and, furthermore, allows the computation of an alternative measure of explained variance Q^2 (Wold et al., 2001). Similar to R^2 being based on the sum of squared deviations between the modeled and actual y, Q^2 relies on the summed squared CV prediction error. Together, R^2 and Q^2 can be taken as the upper and lower benchmarks of the model, in terms of explaining the data and assessing the degree of predictive power, respectively. The selection of the optimal number of components k considers two independent criteria: 1) the component-wise gain in R^2, and 2) the component-wise decrease in CV prediction error, with k being chosen when both measures cease to exhibit substantial improvements for additional PCs.

The loadings T can be seen as vector coordinates of the m regressors in k-dimensional space, describing the degree to which regressors contribute to individual PCs and also showing the collinearity or independence among regressors. A stronger degree of variation in perceptual measures y is assumed to reflect the main perceptually relevant factors, whereas weaker variation may be due to measurement noise alone. Higher-order PCs, accounting for smaller component-wise R^2, may therefore be modeling noise, which would compromise the reliability in interpreting the relationships found among the loadings T. As a measure of reliability, we adopt a resampling technique to assess the influence of noise (similar

techniques have been used in obtaining confidence intervals for PLSR statistics, Martens et al., 2001). The resampling technique fulfills two objectives: 1) to estimate the magnitude of loadings induced by artificial noise and 2) to assess the stability of T coordinates in the presence of that noise. For this purpose, two artificial noise variables are added to X, corresponding to random sequences with uniform and normal distributions, respectively. Across 50 resampling iterations, with unique random sequences per iteration, the aggregate influence of the noise variables on T is evaluated robustly with the medians and interquartile ranges along the k dimensions. The resampling PLSR is conducted separately from the one computed on the X without added noise and, hence, delivers confidence estimates of how T coordinates would vary in 50% of the cases given the presence of noise in the data.

Identifying relevant and independent regressors

The current PLSR analysis aims to reduce the number of investigated regressors in X to those contributing most strongly to explaining y as well as reducing it further to a selection of regressors that are relatively independent of each other. The chosen approach consists of three stages of sequentially evaluating and refining PLSR models: 1) an initial model is obtained for the original matrix X_{orig}; 2) based on the loadings T_{orig}, only those variables are retained for which the Euclidean distances across k dimensions exceed the distribution median ($Q50$), leading to the computation of another model based on X_{Q50}; 3) the matrix T_{Q50} is rotated to align the dominant variable loading along the nearest PC axis, which yields T_{rota}. The rotation achieves maximal independence between regressors for variable loadings aligned along the individual PC axes. The variables are constrained such that the angles ϕ_i between variable loadings and the ith PC axis are less than 22.5°. This constraint yields an approximately orthogonal set of regressors, allowing the final model to be computed on X_{ortho}.

3.2.2 Perceptual data sets (Experiments 3 and 4)

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

were tested in the same venue and involved the same target population. The stimuli were presented over a standard two-channel stereophonic loudspeaker setup inside an Industrial Acoustics Company double-walled sound booth (see Appendix C) and relied on recorded instrument samples from the Vienna Symphonic Library[1] (VSL), supplied as stereo WAV files (44.1 kHz sampling rate, 16-bit dynamic resolution). In addition, all stimuli were adjusted for perceptual synchrony between sounds constituting the dyads and triads and were equalized for loudness within the dyad and triad sets independently. Adjustments for synchrony were based on consensus across two to three persons; the subjective equalization of loudness to a reference sound was conducted by up to six persons, relying on median equalization gains after the corresponding interquartile ranges fell below 4 dB. Participants were recruited from the McGill University community, involving varying degrees of musical experience, with all participants having passed a standardized audiogram (Martin and Champlin, 2000; ISO 389–8, 2004) ensuring that thresholds at all audiometric frequencies were less than 20 dB HL.

Dyads

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

Stimuli The stimuli comprised a total of 180 dyads that resulted from the combination of several factors. Six wind instruments, namely, (French) horn, bassoon, oboe, C trumpet, B♭ clarinet, and flute, formed the fifteen possible non-identical-instrument pairs listed in Table 3.1. These instrument pairs occurred at two pitch levels, i.e., C4 and G4. Furthermore, dyads comprised both unison and minor-third intervals, including the inverse order of instruments for the latter, resulting in a total of three interval conditions. Based on the two pitch levels, minor thirds occurred at the pitches C4-E♭4 and G4-B♭4. All VSL samples were sustained, non-vibrato recordings, performed at *mezzoforte* dynamics, and were limited to the signal in the left channel. Both instruments were simulated as being captured by a stereo main microphone at spatially distinct locations inside a mid-sized,

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

moderately reverberant room (see Appendix C). The spatial position between instruments (e.g., horn left of bassoon) included both possible orientations. Overall, this resulted in the full-factorial combination of 15 pairs × 2 pitches × 3 intervals × 2 positions = 180 dyads. All stimuli had a duration of 1200 ms, with artificial offsets imposed by a 100-ms linear amplitude ramp.

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

Table 3.1 Fifteen dyads across pairs of the six investigated instruments.

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

Triads

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

Stimuli The stimuli comprised 20 triads constituted from sounds of flute, oboe, B♭ clarinet, tenor trombone and cello (*arco*, *pizzicato*). These instruments corresponded to the instrument families woodwinds (air jet, single and double reed), brass, and strings (bowed and plucked excitation). All triads formed the same chord with pitches C4, F4, and B♭4. The investigated triads represent only a selection of all possible combinations between instruments and pitches. The selection (see Table 3.2) represents various combinations of sustained (blown, bowed) and impulsive (plucked) sounds and of string, woodwind and brass families. Each instrument appears in from six to 10 triads (counting *arco* and *pizz.* as separate instances). All samples were taken as stereo files from VSL, with woodwind samples comprising sustained sounds at *mezzoforte* dynamics and without vibrato. The trombone samples were similar, but at *mezzopiano* dynamics. The *arco* cello samples were recorded at *mezzoforte* dynamics and, unlike the wind instruments, decayed after just a brief bow stroke, in order to be more similar to the *pizzicato* versions, occurring at *forte* to allow for a longer sound decay; all cello sounds contained vibrato. The total duration for all triads was limited to 850 ms by applying an artificial 100-ms linear amplitude-decay ramp.

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

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

Table 3.2 Twenty triads and their constituent instruments and assigned pitches.

3.2.3 Acoustical descriptors

For each data set, a collection of acoustical measures serves as the regressor matrix X. The measures are based on acoustical descriptors comprising all relevant domains to the acoustical description of timbre, i.e., including spectral, temporal, and spectro-temporal features, as well as other potentially relevant features unrelated to timbre (e.g., accounting for pitch differences). Table 3.3 lists all the investigated descriptors, specifying how individual descriptor values were associated in dyads and triads.

Descriptor relationships within dyadic and triadic contexts

As dyads and triads consist of several constituent sounds, their individual descriptor values need to be summarized to a single regressor value per dyad or triad by an association of some kind. For dyads with the constituent sounds a and b and the acoustical descriptor x, the association considers the *difference* measure $\Delta x = |x_a - x_b|$ and the *composite* measure $\Sigma x = x_a + x_b$. Consider triads with sounds a, b, and c, whose relationship along descriptor x is $x_a \leq x_b \leq x_c$. The triad *difference* considers the range between maximum and minimum, i.e., $\Delta x = x_c - x_a$. The *composite* sums all three values, i.e., $\Sigma x = x_a + x_b + x_c$. In addition, a third measure relates the *distribution* of the intermediate value x_b relative to the extremes, i.e., $\Xi x = 2 \cdot (x_b - x_a)/\Delta x - 1$. Ξx yields normalized values with 0 corresponding to x_b being centered between x_a and x_c, and -1 and +1 corresponding to $x_b = x_a$ and $x_b = x_c$, respectively. These three regressor types apply to most of the investigated acoustical descriptors but not all, based on whether the association is appropriate or not as indicated in Table 3.3.

Timbre descriptors

Spectral descriptors assess properties associated with a time-averaged spectral representation. The investigated descriptors are computed on the output of one of two spectral-analysis methods: 1) analyses of the audio signals for individual samples from VSL (e.g., oboe at G4) by use of the *Timbre Toolbox* (Peeters et al., 2011) employing harmonic analysis, and 2) pitch-generalized spectral-envelope estimates (see Chapter 2 and Appendix A) based on all available pitches from VSL (e.g., oboe from B♭3 to G6), which allowed the identification and description of formant structure. Furthermore, the spectral descriptors can be distinguished as quantifying *global* and *local* spectral properties, as listed in

Abbreviation	Description	Unit	Association	Dyad	Triad
$S_{ct}^{\sim}$	spectral centroid[a]	Hz	Δ, Σ, Ξ	✓	✓
S_{ct}°	spectral centroid[b]	Hz	Δ, Σ, Ξ	✓	✓
$S_{lp}^{\sim}$	spectral slope[a]	Hz^{-1}	Δ, Σ, Ξ	✓	✓
S_{lp}°	spectral slope[b]	Hz^{-1}	Δ, Σ, Ξ	✓	✓
S_{kw}	spectral skew[a]	-	Δ, Σ, Ξ	✓	✓
S_{ku}	spectral kurtosis[a]	-	Δ, Σ, Ξ	✓	✓
S_{pr}	spectral spread[a]	Hz	Δ, Σ, Ξ	✓	✓
S_{ro}	spectral roll-off[a]	Hz	Δ, Σ, Ξ	✓	✓
S_{dc}	spectral decrease[a]	-	Δ, Σ, Ξ	✓	✓
S_{ns}	noisiness[a]	-	Δ, Σ, Ξ	✓	✓
F_{max}	main-formant maximum[b]	Hz	Δ, Σ, Ξ	✓	✓
F_{3dB}	main-formant upper bound[b]	Hz	Δ, Σ, Ξ	✓	✓
F_{sl}	spectral slope above main formant[b]	Hz^{-1}	Δ, Σ, Ξ	✓	✓
$F_{\Delta mag}$	level difference F_1 vs. above[b]	dB	Δ, Σ, Ξ	✓	✓
F_{promi}	formant prominence[b]	-	Δ, Σ, Ξ	✓	✓
F_{freq}	formant-frequency deviations[b]	Hz	Δ	✓	✓
F_{mag}	formant-magnitude deviations[b]	dB	Δ	✓	✓
A_t	attack time	s	Δ, Ξ	✓	✓
$A_{lg(t)}$	log. attack time	-	Δ, Ξ	✓	✓
A_{sl}	attack slope	amplitude/s	Δ, Ξ	✓	✓
ST_{fl}	spectral flux[a]	-	Δ, Σ, Ξ	✓	✓
ST_{in}	spectral incoherence[a]	-	Δ, Σ, Ξ	✓	✓
C_{UniNon}	unison or non-unison	-	binary	✓	
Δf_0	f_0 difference	Hz	Δ	✓	
C_{PiLoHi}	pitch level	-	binary	✓	
$f_{0\|ERB}$	f_0, auditory scaling	ERB[c] rate	C4 or G4	✓	
C_{In012}	interval type	-	ternary	✓	
C_{PosLR}	instrument positions	-	binary	✓	
$C_{PizzNon}$	plucked or non-plucked	-	binary		✓
x_{mix}	amplitude mix factor	-	ratio[d]	✓	✓

Table 3.3 Acoustical descriptors investigated for dyads and/or triads (marked by ✓ in the rightmost columns), related to the global spectrum (S), formants (F), the temporal attack (A), spectro-temporal variation (ST) as well as categorical variables (C). Descriptor values for individual sounds forming dyads or triads were associated to a single regressor value by *difference* Δ, *composite* Σ, *distribution* Ξ (triads only) or as specified otherwise.

[a] $S^{\sim}$ based on spectral analysis of individual pitches.
[b] S° based on pitch-generalized spectral-envelope estimate.
[c] ERB: equivalent rectangular bandwidth (Moore and Glasberg, 1983).
[d] For triads, the ratio corresponds to Ξ of the individual amplitudes.

Table 3.3. The global descriptors (S) are commonly reported and discussed in detail in Peeters et al. (2011), whereas the local, formant-related descriptors (F) require some elaboration (see also Section A.2). Two descriptor frequencies describing the main formant, F_{max} and F_{3dB}[2], characterize its frequency at maximum magnitude and at the upper bound at which the magnitude has decreased by 3 dB, respectively, with the latter appearing to be more perceptually relevant (Chapter 2). Two related measures, F_{sl} and $F_{\Delta mag}$, evaluate the relative importance of the main formant compared to the spectral-envelope regions lying above it. The former evaluates the spectral slope above the main formant, whereas the latter quantifies the level difference between main formant and the averaged magnitude of the spectral envelope above it. Furthermore, the degree to which wind instruments are characterized by formant structure varies (e.g., strongest for oboe, much weaker for clarinet and flute), with F_{promi} quantifying the presence of up to two formants and their prominence (e.g., spectral maximum of pronounced frequency and magnitude extent). It is based on a cumulative score that increases with the number of prominent features, whose weights were determined heuristically, resulting in F_{promi} being greatest for oboe and lowest for clarinet (see Section A.2.4). Two measures, F_{freq} and F_{mag}, relate frequency and magnitude differences relative to formants between the constituent instruments. More specifically, for every formant of a constituent instrument, its differences to corresponding frequencies or magnitudes of other instruments is quantified and furthermore weighted by F_{promi}, after which all differences across instruments and formants are summed into an aggregate value.

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

Spectro-temporal descriptors account for spectral variation across time, which the (static) spectral descriptors leave unaddressed. Although previous research on blend has not reported a relevance for spectro-temporal (ST) features, in the interest of using a comprehensive set of timbre-related descriptors, two are included that involve the commonly reported *spectral flux* (ST_{fl}, Peeters et al., 2011) and the alternative measure *spectral in-*

[2]For simplicity, as of Chapter 3, F_{3dB} signifies the upper bound $F_{3dB}^{\rightarrow}$ as initially introduced in Chapter 2, because descriptors for lower bounds are no longer discussed explicitly.

coherence (ST_{in}), which quantifies the aggregate deviations of spectral magnitude between time frames (Horner et al., 2009).

Other descriptors and variables

The experimental designs involved factors that were likely to explain variance in median blend ratings, but were not related to or not reliably measured through timbre features. Their relevance as potential regressors is assessed through several categorical variables (C), in addition to acoustical descriptors that could serve as equivalent predictors in application scenarios lacking a priori knowledge of categorical distinctions, by quantifying pitch relationships or the loudness balance between combined sounds. The categorical variables make binary or ternary distinctions and for the use with PLSR have to be expressed as *dummy variables*[3](Martens et al., 2001). For triads, a strong distinction was expected beforehand for the presence versus absence of plucked string sounds ($C_{PizzNon}$), as they are highly impulsive. Similarly, the distinction into unison or non-unison dyads was also expected to yield higher ratings for the former (C_{UniNon} and Δf_0). Additional regressors account for the pitch level (C_{PiLoHi} and $f_{0|ERB}$), interval type (C_{In012}), and instrument position (C_{PosLR}). In addition, the production of dyads and triads also involved determining relative mix or scaling ratios between the amplitudes of the constituent sounds forming dyads or triads, which are also quantified to assess their possible influence on the blend ratings (x_{mix}).

3.3 Results

As mentioned in Section 3.2.1, PLSR analysis of a particular data set involves three stages, beginning with the regressor set in X_{orig}, then restricting the selection to X_{Q50} and finally X_{ortho}. Although statistics for all three stages are reported in Tables 3.4 and 3.5, only the results for final stage X_{ortho} are presented in detail. In the following visualizations, data points representing dyads or triads distinguish themselves based on pre-selected, exemplary instruments they involve, which helps to assess how, for example, an acoustical descriptor separates these instruments. The selected instruments concern those that exhibit the

[3] A categorical variable is represented by as many dummy variables as there are categories, with each category's dummy variable set to 1 for cases matching the category and 0 if not. As a result, these regressors yield multiple loadings. For example, a binary categorical variable yields two loadings in opposing orientations, their variable names appended by "-D1" or "-D2", symbolizing the categorical values "1" and "2", respectively.

highest or lowest aggregate blend ratings across the dyads or triads in which they occur, based on the median blend ratings for the corresponding dyads or triads. For dyads, the instruments clarinet, bassoon, and horn lead to the highest blend ratings of comparable magnitude, whereas the trombone leads to the highest blend ratings for triads. Allowing a better comparison between data sets, the horn and trombone represent instruments that blend well with others (colored green) in the dyad and triad sets, respectively, as both brass instruments' spectral descriptions also resemble each other (see Appendix B), whereas the oboe represents the exemplary instrument leading to bad blend (colored grey) in both sets.

3.3.1 Dyads (Experiment 3)

PLSR models attempting to predict median blend ratings for dyads initially involved 46 regressors (X_{orig}), which after elimination of loadings T_{orig} falling below the median threshold yielded 23 regressors in X_{Q50}. As listed in Table 3.4, a three-PC model explains 93% of the variance for X_{Q50}. Refining the regressors to an approximately orthogonal set, the resulting X_{ortho} consists of 14 regressors, again, leading to a three-PC model explaining 93% of the variance. The model fit in y for X_{ortho}, displayed in Figure 3.1, shows the variation in median blend ratings to be represented well. However, the blend ratings (x-axis) already exhibit two distinct groups of data points, corresponding to unison dyads (circles) leading to substantially more blend than non-unison dyads (diamonds). Furthermore, the non-unison dyads exhibit the trend of dyads involving horn (green) yielding more blend than those with oboe (grey), for overlapping sub-groups, whereas no such clear distinction is observable for unison dyads.

Figure 3.2 visualizes the loadings T_{ortho} and the scores P_{ortho} across the first two PCs. Larger symbols for scores correspond to higher degrees of blend. Loadings are visualized as vectors, with the black squares symbolizing the coordinates for the reported PLSR models, whereas the vector tips and ellipsoids surrounding the tips (often small and barely visible) represent the median and interquartile ranges obtained from the resampling technique with artificial noise, respectively. The resampling also adds two loadings that always appear around the origin of the coordinate system, representing the magnitude of variation introduced by noise with *normal* (N_n) and *uniform* (N_u) distributions. Reflecting the main distinctions in median blend ratings, the scores P_{ortho} also form two distinct groups for unison and non-unison dyads, with the corresponding categorical variable C_{UniNon} describing

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

Table 3.4 Dyad PLSR-model performance (R^2) and predictive power (Q^2) as well as component-wise contribution along up to three PCs. Three stages X_{orig}, X_{Q50}, X_{ortho} involve a sequential reduction of the number of regressors m.

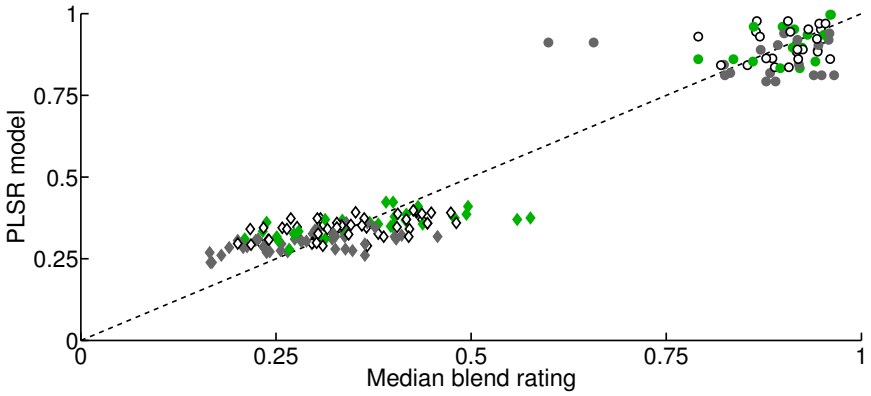

Fig. 3.1 Dyad model fit of y variables for X_{ortho}. Legend: circles, unison; diamonds, non-unison; grey involves oboe; green involves horn (excl. HO).

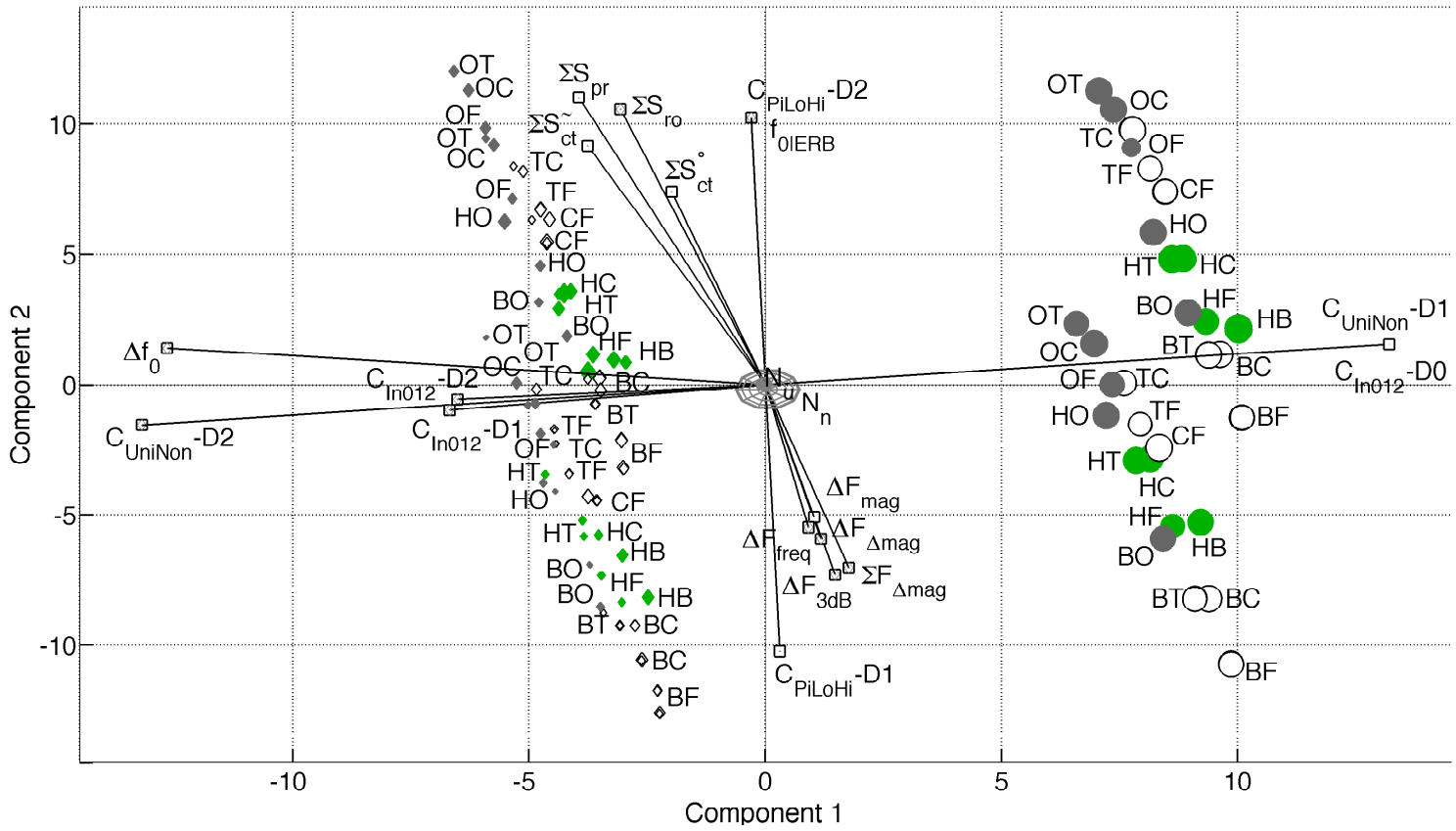

Fig. 3.2 Dyad PLSR loadings T_{ortho} (vectors) and scores P_{ortho} (points) for PCs 1 and 2. Legend: circles, unison; diamonds, non-unison; their size represents relative degree of blend; grey involves oboe; green involves horn (excl. HO); grey ellipsoids illustrate interquartile ranges from the added-noise resampling technique, e.g., N_n and N_u.

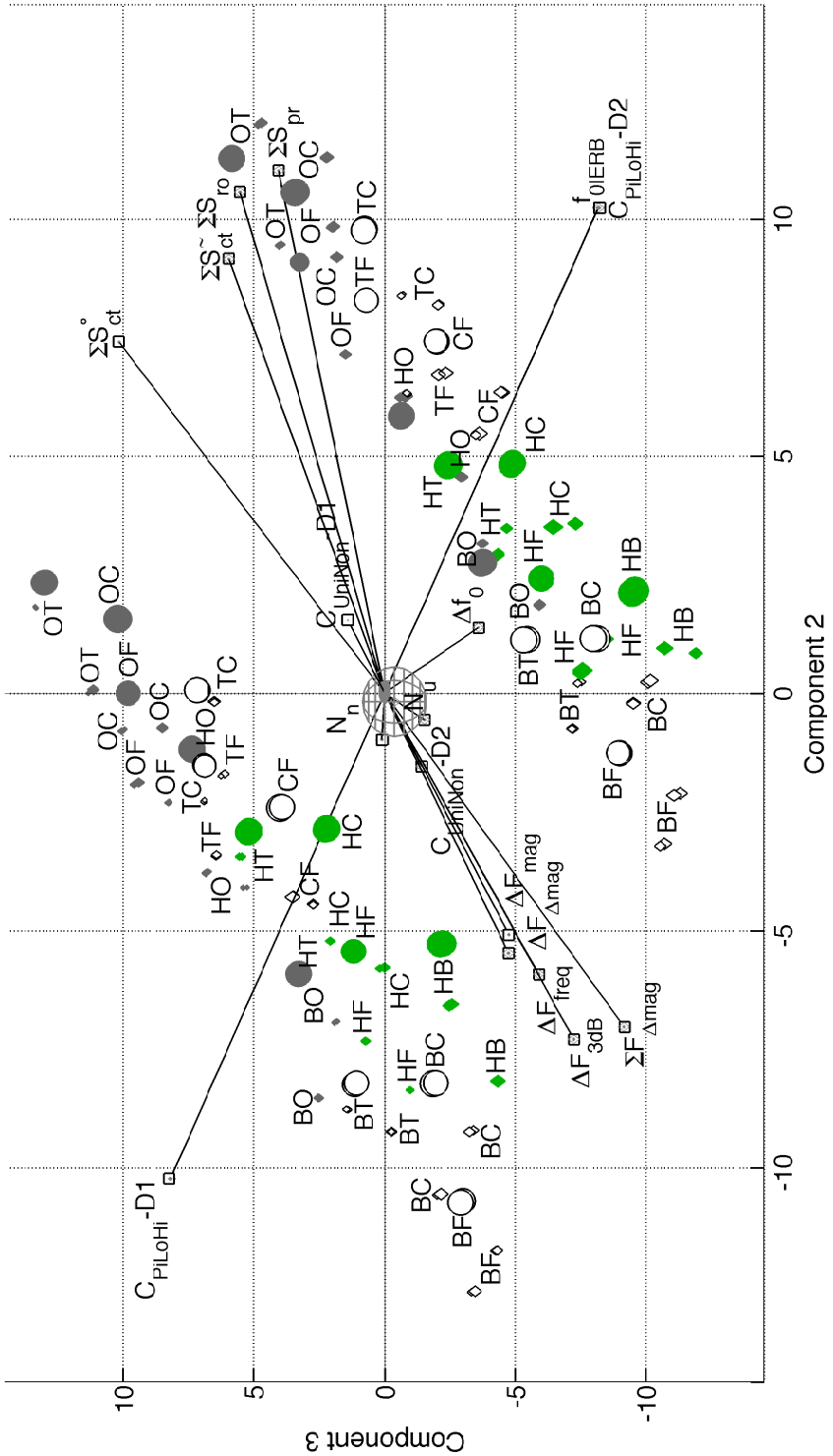

Fig. 3.3 Dyad T_{ortho} and P_{ortho} for PCs 2 and 3. See Figure 3.2 for legend.

this distinction most accurately along PC 1 and the acoustical descriptor Δf_0 predicting the same distinction comparably well. PC 2 appears to be influenced by two factors: 1) an additional grouping of dyads based on low and high pitch levels, described by the categorical variable C_{PiLoHi} and the acoustical descriptor $f_{0|ERB}$, and 2) a collinear set of spectral descriptors, falling slightly oblique to the PC axis, along which the dyads align in the four subgroups. Within the sub-groupings of interval type by pitch level (horizontal × vertical), the influence of spectral features appears to lead to similar dyad constellations. Furthermore, Figure 3.3 suggests the spectral and pitch influence to be independent (orthogonal) on the plane spanning PCs 2 and 3. The spectral regressors involve several composite (Σ) as well as difference (Δ) measures for S°_{ct} and formant-related descriptors. With regard to the resulting scores, P_{ortho} yields a grouping of dyads into those containing either horn or oboe (green/low-left vs. grey/top-right), interestingly, applying to both unison and non-unison dyads. Overall, the dyad data exhibit a complex structure of underlying factors, involving interval type, pitch level, and spectral features. Across all investigated models, their performance (R^2) is remarkably well matched by their predictive power (Q^2). Given the relatively large number of cases, $n = 180$, further PLSR analyses on subsets separated by interval type, yielding $n = 60$ for unison and $n = 120$ for non-unison dyads, are considered, as separate analyses allow an assessment of whether certain spectral and pitch trends are specific to only one of the interval types.

Unison

A three-PC model on X_{Q50} involving 22 regressors leads to 46% explained variance in median blend ratings for unison dyads, however, exhibiting a substantially lower predictive power of only 17% explained variance. Due to a fairly wide variation in T_{Q50} orientations, the angular threshold ϕ_i determining X_{ortho} had to be increased to $|\phi_i| < 30°$ to ensure that the reduction to an approximately orthogonal set would lead to a meaningful number of contributing regressors. The resulting model with nine regressors yields a two-PC model explaining 27% of the variance, which appears a more realistic estimate of the true predictive relationship between median blend ratings and X_{ortho}, as the divergence between model performance and the predictive power is substantially reduced, avoiding the risk of over-fitting to noise. As shown in Figure 3.4, the y_{unison} fit appears better than for the complete dyad data (Figure 3.1), but the blend ratings only span a relatively narrow scale

range, likely related to the dominant distinction between unison and non-unison dyads reducing the perceptual resolution among the unison dyads. The reduced resolution also makes it more likely for the variation in median blend ratings to contain increased noise levels, supported by the large divergence between R^2 and Q^2 in the initial models.

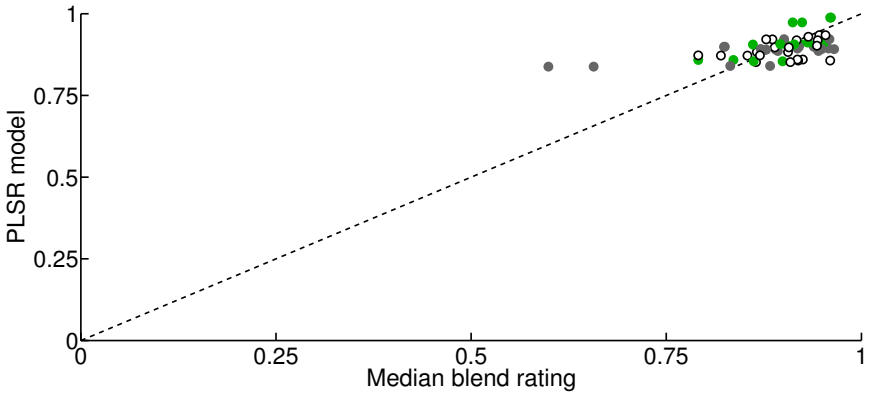

Fig. 3.4 Unison-dyad model fit of y variables for X_{ortho}. See Figure 3.1 for legend.

In Figure 3.5, the loadings T_{ortho} along PC 1 and PC 2 are quite stable and remain largely unaffected by noise, as the ellipsoids delineating the interquartile ranges for the noise variables (e.g., N_u) remain much smaller than the other loadings. With regard to the regressor relationships, PC 1 explains 22% of the variance, appearing to be linked to spectral composite (Σ) descriptors for main formant location (e.g., F_{max}, F_{3dB}) as well as centroid (e.g., S°_{ct}), which also achieves a distinction between low register and high register instrument dyads (e.g., HB vs. OF). PC 2 accounts for another 5% of the variance, involving a distinction between instrument dyads with similar formant structure and those with divergent structures (e.g., HB vs. BF and HF), explained by the formant-related descriptors ΔF_{sl} and ΔF_{freq}.

Non-unison

Twenty-three regressors in X_{Q50} yield a three-PC model explaining 55% of the variance in median blend ratings for non-unison dyads, with the predictive power corresponding to 47% of the variance explained. The reduction to X_{ortho} yields 11 regressors and a three-PC model explaining 48% of the variance, with a lower predictive power accounting for 35% of the variance. The model fit in $y_{non-unison}$ for X_{ortho}, shown in Figure 3.6, improved

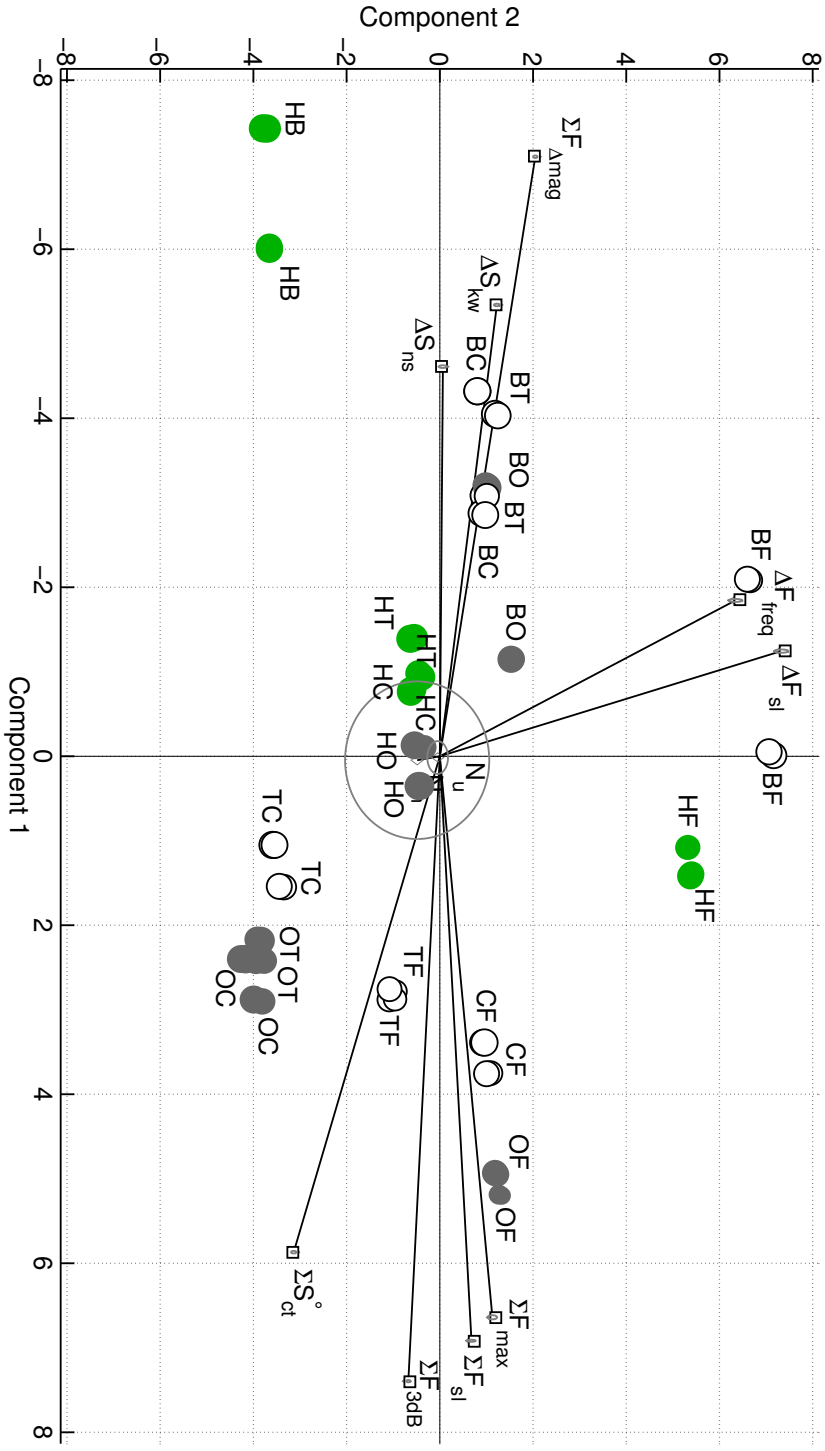

Fig. 3.5 Unison-dyad T_{ortho} and P_{ortho} for PCs 1 and 2. See Figure 3.2 for legend.

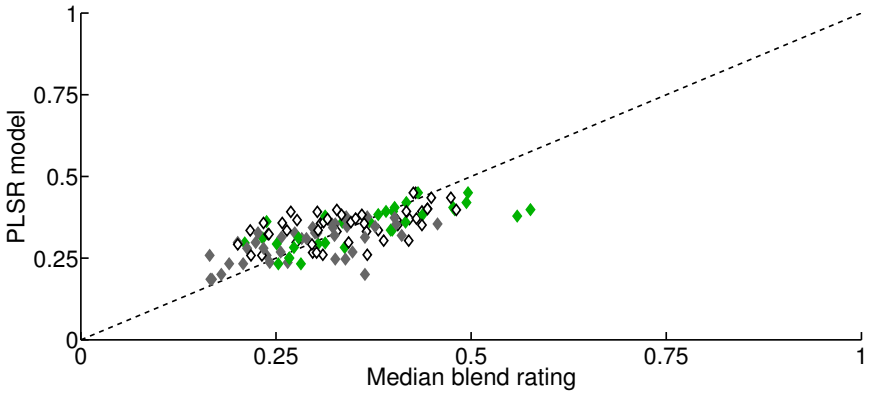

Fig. 3.6 Non-unison-dyad model fit of y variables for X_{ortho}. See Figure 3.1 for legend.

compared to the one for the complete dyad set (Figure 3.1), showing a better approximation to the ideal fit (dashed line).

As shown in Figure 3.7, PC 1 clearly reflects a grouping of dyads based on pitch level (C_{PiLoHi} and $f_{0|ERB}$), accounting for 33% of the explained variance. At the same time, the composite of the spectral slope $S_{lp}^{\sim}$ appears to covary with pitch change. All remaining spectral regressors appear relatively independent (orthogonal) to the pitch influence. Figure 3.8 illustrates that across the plane spanning PCs 2 and 3, two seemingly independent contributions of spectral regressors occur: 1) an implied triangle between the composite (Σ) regressors F_{3dB}, S_{ct}°, and the difference (Δ) descriptor F_{3dB} distinguishes dyads into those containing horn (bottom-left) and those involving oboe (top-right); 2) perpendicular to this orientation, difference in spectral slope $S_{lp}^{\sim}$ and composite in noisiness S_{ns} contribute somewhat more weakly. Together, PCs 2 and 3 account for 8% and 7% of the variance, respectively. Although in X_{Q50} a solitary loading of the composite spectral spread S_{pr} defines PC 3, its utility disappears in X_{ortho}.

3.3.2 Triads (Experiment 4)

The PLSR analysis of triads first involved 61 regressors, which reduced to 30 regressors in X_{Q50}, leading to a three-PC model explaining 88% of the variance in median blend ratings and with a predictive power explaining 71% of the variance. The subsequent reduction to X_{ortho} yields a two-PC model with seven regressors that still explains 81% of the variance, notably, without loss in predictive power compared to the previous models. As shown

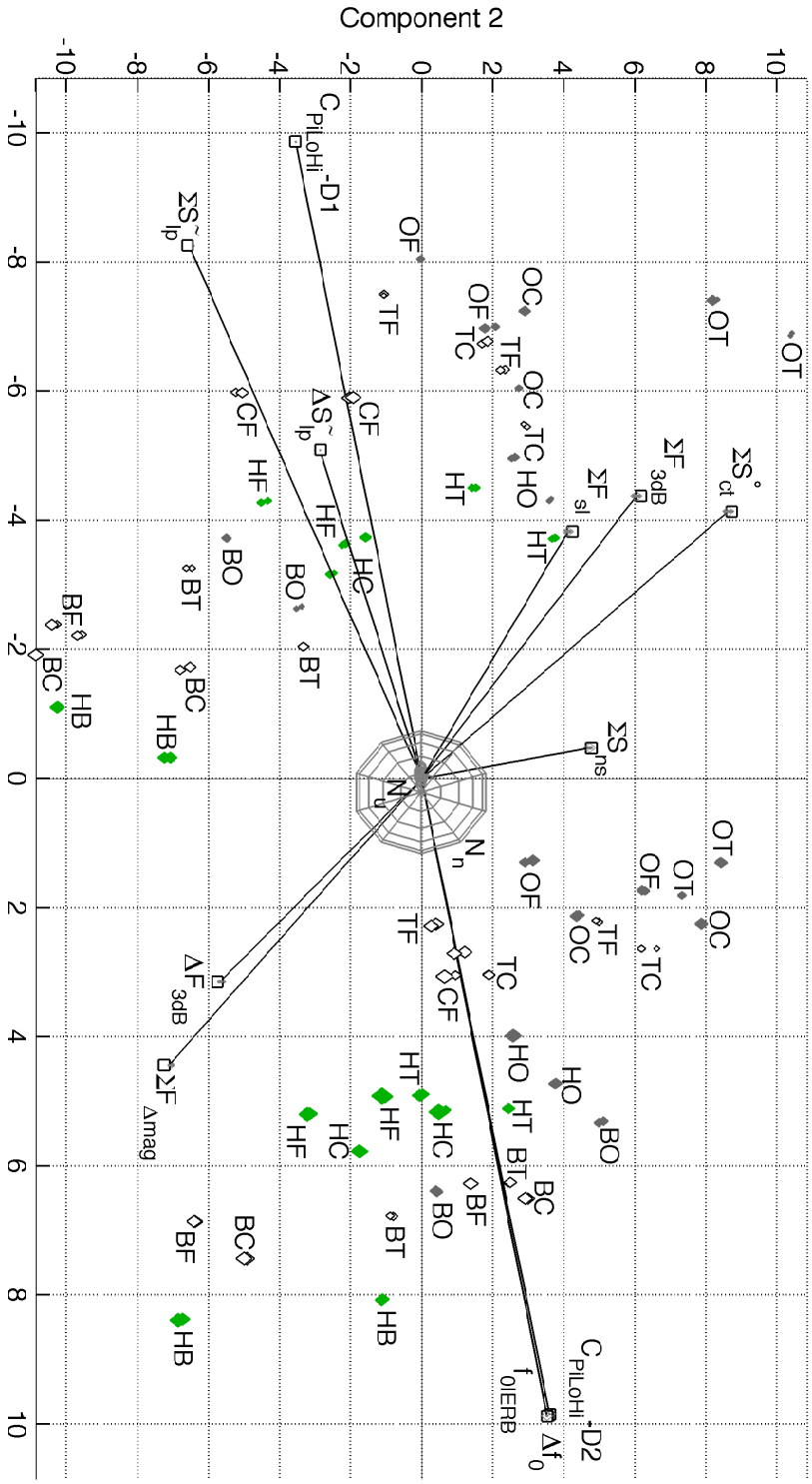

Fig. 3.7 Non-unison-dyad T_{ortho} and P_{ortho} for PCs 1 and 2. See Figure 3.2 for legend.

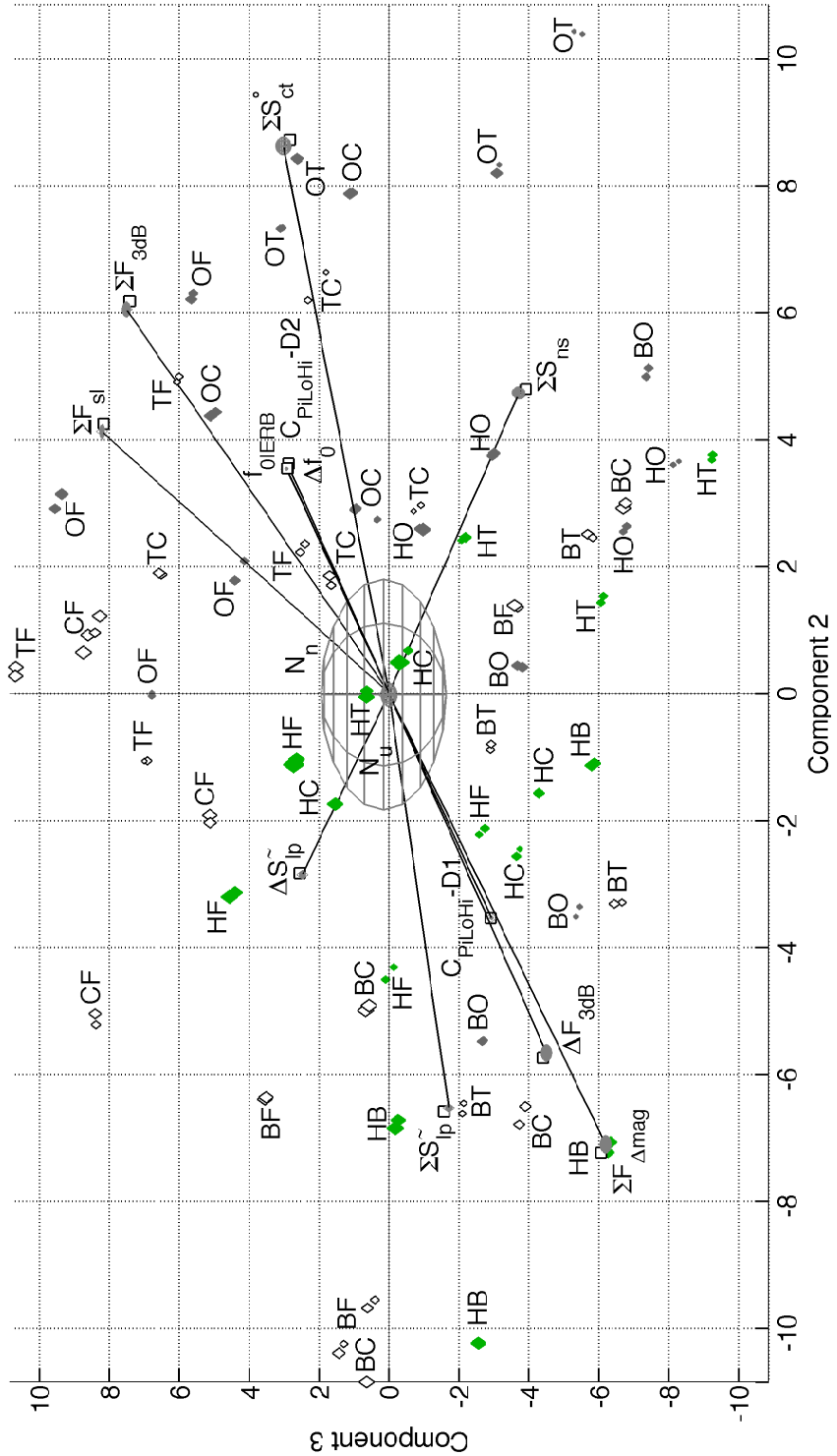

Fig. 3.8 Non-unison-dyad T_{ortho} and P_{ortho} for PCs 2 and 3. See Figure 3.2 for legend.

y triads	X regressors	m	R^2	Q^2	PC 1	PC 2	PC 3
all	X_{orig}	61	.94	.70	.86	.05	.03
	X_{Q50}	30	.88	.71	.80	.05	.03
	X_{ortho}	7	.81	.71	.78	.03	-

Table 3.5 Triad PLSR model performance (R^2) and predictive power (Q^2) as well as component-wise contribution along up to three PCs. Three stages X_{orig}, X_{Q50}, X_{ortho} involve a sequential reduction of the number of regressors m.

in Figure 3.9, the model fit for y appears satisfactory, given the smaller number of cases for triads ($n = 20$). Still, a compact cluster involving plucked cello (squares, bottom-left) stands in contrast to more spread out ratings for sounds lacking them (circles, right half). As with dyads, a trend for sounds involving trombone (green) being more blended than those containing oboe (grey) is apparent in each subgroup.

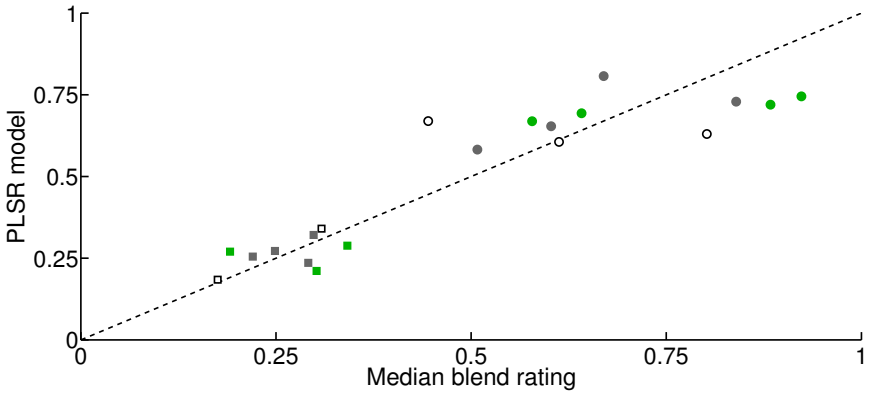

Fig. 3.9 Triad model fit of y variables for X_{ortho}. Legend: squares, incl. *pizz.*; circles, excl. *pizz.*; grey involves oboe; green involves trombone (excl. PTO, TTO, TCO).

Whereas for X_{Q50} a diverse range of spectral difference (Δ), composite (Σ) and also distribution (Ξ) regressors assume various orientations along all three PCs, the scores P_{ortho} and loadings T_{ortho} deliver a simplified picture. The main distinction found in Figure 3.10 along PC 1, which accounts for 78% of the variance, concerns the occurrence of plucked cello sounds or not, i.e., the categorical variable $C_{PizzNon}$, with the acoustical description of attack-slope difference A_{sl} performing similarly well. Apart from A_{sl}, the composite and difference descriptors for spectro-temporal incoherence ST_{in}, as well as the difference descriptor for noisiness S_{ns}, are somewhat correlated with $C_{PizzNon}$. This could result from

both the attack transient and rapid decay of the temporal envelope of plucked sounds contributing to both more noise and more spectral change over time. PC 2 explains the remaining 3% of the variance, appearing to relate to the difference in spectral spread S_{pr} and the distribution (Ξ) of the pitch-generalized spectral centroid S°_{ct}. However, PC 2 clearly shows an increased sensitivity to noise, illustrated by the enlarged radius for added normally distributed noise (N_n, large grey ellipsoid), which extends to about the same magnitude as S_{pr}, rendering this particular regressor less reliable. Even S°_{ct} exhibits an increased range of variability along PC 2 (narrow grey ellipsoid). However, its loading extends clearly beyond the interquartile range of N_n.

3.4 Discussion

Previous research had associated blend with acoustical measures describing spectral features and, under certain circumstances, also temporal features like the attacks or onsets of sounds. The current investigation pursued a correlational analysis by use of PLSR, modeling two perceptual data sets involving dyads and triads. PLSR loadings T allow the differentiated evaluation of relationships between regressors as being closely related (collinear) or independent of each other, which aids the identification of the most effective regressors. Applied to the complete data sets for both dyads and triads, the final models based on X_{ortho} explain around 80 to 90% of the variance in median blend ratings, which leaves only a marginal portion of the variance unaccounted for. The variation in both data sets is best explained by a dominant, primary factor, however, unrelated to spectral features.

For dyads, the distinction between unison and non-unison intervals explains 91% of the variance, with the fundamental-frequency difference Δf_0 representing a reliable acoustical predictor. That unison dyads would lead to higher blend than for non-unison had been anticipated, given that similar effects had been found in other studies (Kendall and Carterette, 1993; Chapter 4). The pronounced difference obtained in the current results, however, seems to exceed those previously reported, which could be related to the current study being the only one in which unison and non-unison were presented in a common stimulus set, whereas in other studies both interval types had been grouped into separate experimental blocks (Kendall and Carterette, 1993; Chapter 4) or had even been tested in separate experiments (Sandell, 1995). In addition, even the second-most important factor

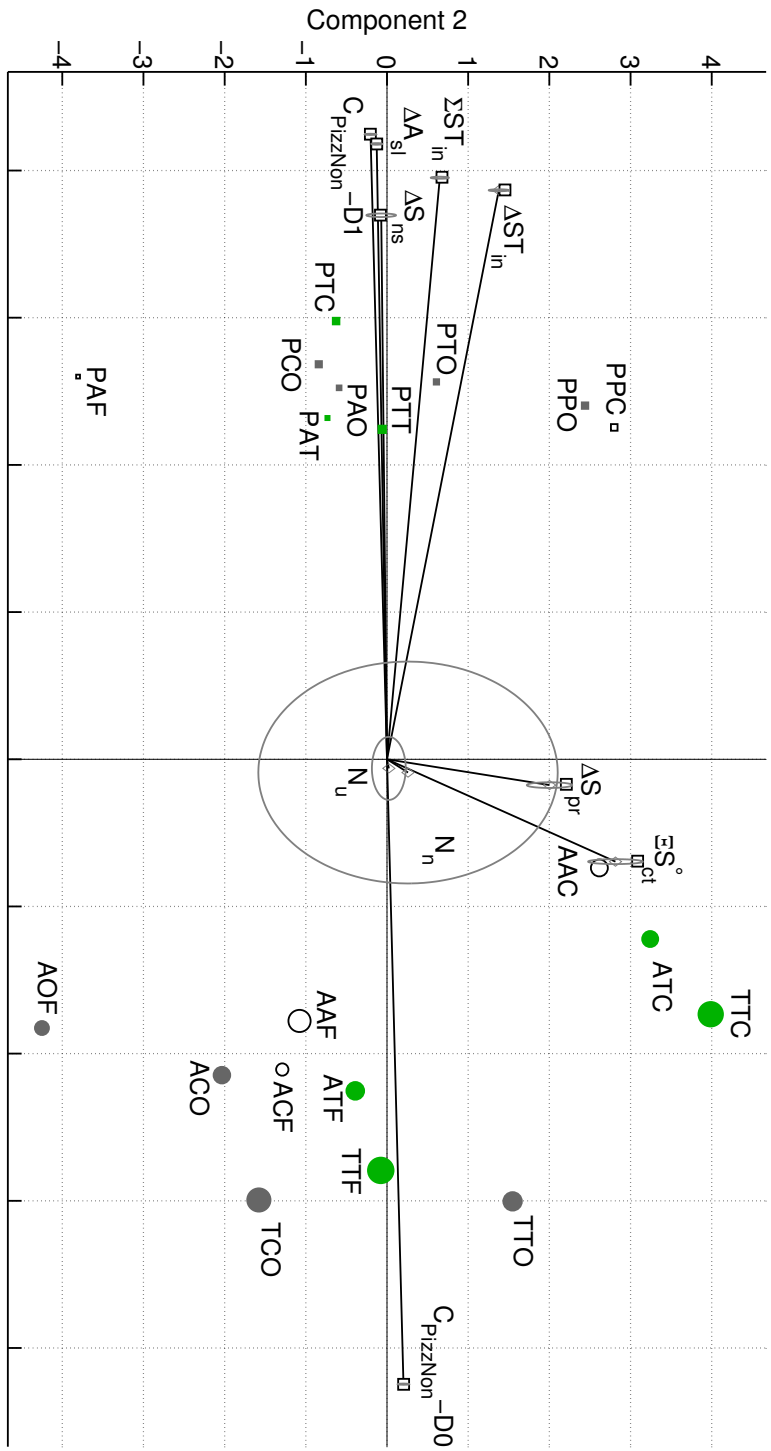

Fig. 3.10 Triad PLSR loadings T_{ortho} (vectors) and scores P_{ortho} (points) for PCs 1 and 2. Legend: squares, incl. *pizz.*; circles, excl. *pizz.*; their size represents relative degree of blend; grey involves oboe; green involves trombone (excl. PTO, TTO, TCO); grey ellipsoids illustrate interquartile ranges from the added-noise resampling technique, e.g., N_n and N_u.

in explaining the variation among dyads, $f_{0|ERB}$, is unrelated to spectral features, as it reflects differences between pitch levels, accounting for 2% and 33% in all dyads or just the non-unison dyads, respectively. The limitation to only the non-unison dyads (in the final models) implies that pitch may not influence unison dyads to the same extent. For non-unison dyads, it is also worth noting that inverting the assignment of instruments to either the upper or lower pitch had no effect on blend ratings. This negative finding provides another indication that inversion does not appear to influence blend (Sandell, 1995); only a single finding argues in its favor (Kendall and Carterette, 1993).

With regard to triads, the presence of a plucked (*pizz.*) cello evokes a strong decrease in blend ratings, whereas even triads including cello sounds excited by a single, brisk bow stroke lead to substantially more blend. Again, this distinction had been anticipated, given that increasingly impulsive sounds have been associated with comparable decreases in blend (Tardieu and McAdams, 2012). In line with the description of attacks, the difference in attack slopes A_{sl} bears a strong collinearity with the categorical distinction $C_{PizzNon}$, explaining about 80% of the variance; additional collinearity with spectro-temporal or noise features is assumed to co-occur as byproducts of the abrupt changes in temporal envelopes.

With both data sets being dominantly influenced by either factors related less to timbre (e.g., pitch) or to temporal features (e.g., attack time), spectral descriptors only occur as secondary or even tertiary sources of variation in the modeled median blend ratings, also, at lower magnitudes than the dominant ones. In perceptual tasks comparable to those employed in these experiments, participants may focus their attention on the dominant distinctions across stimuli at the cost of perceptual resolution for the less pronounced differences. This motivated a rigorous assessment of data reliability especially for the spectral regressors, for which two indicators were considered. First, a clear divergence between model performance R^2 and predictive power Q^2 indicates that models are likely over-fitting to noise artifacts instead of systematic factors of variation. For example, stripped of the dominant factor, the unison and non-unison datasets account for no more than 50% of the variance (R^2), with the unison-dyad data suggesting the true performance to be substantially lower as the predictive power is generally quite low. Explaining only 15-25% of the variance essentially means that a large portion of the variation is unaccounted for and likely reflects random variation or factors not captured by the tested regressors. Second, the degree of variation introduced by the noise-resampling technique along individual PCs provides another indicator of reliability, which shows smaller contributions by spectral

regressors and smaller sample sizes ($n = 20$) to be more sensitive to noise, as shown along PC 2 in Figure 3.10. This result presents one of the reasons for selecting two-PC models in some cases, because additional PCs were more clearly affected by noise. In summary, the identified tendencies for spectral regressors can be assumed valid for the obtained proportions of explained variance, but they should be considered preliminary until confirmed in additional datasets yielding greater resolution in the perceptual ratings.

Three spectral descriptors stand out in explaining the PLSR models for both data sets, namely, the pitch-generalized centroid S°_{ct} and the two main-formant descriptors F_{max} and F_{3dB}, notably representing spectral features that have previously been found to be relevant (Sandell, 1995; Reuter, 1996; Tardieu and McAdams, 2012; Chapters 2 and 4). Differences exist concerning the types of association between descriptor values of the instruments constituting dyads or triads. For unison dyads, the composite (Σ) measures for all three descriptors become relevant in explaining 22% of the variance, which is in agreement with the same association explaining other perceptual results for unison dyads (Sandell, 1995; Tardieu and McAdams, 2012).

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

The results for triads expand previous knowledge beyond dyadic contexts. Even if spectral features only account for 3% of the variance, the distribution of the sound carrying the intermediate value in spectral centroid S°_{ct} relative to the extremes serves as the strongest predictor, suggesting that this association (Ξ) may indeed be useful in describing instrument combinations with more than two instruments.

Overall, the global descriptor S°_{ct} and the main-formant location F_{3dB} indicate that prominent spectral-envelope properties represent reliable correlates to blend across vari-

ous instruments, pitches, and polyphonic combinations. Being the first investigation to test for the relevance of global and local spectral descriptors jointly, both domains seem equally helpful as regressors in a predictive application. Across all datasets, the loadings T confirm that most spectral descriptors are partially correlated, at the same time, allowing the identification of descriptors that appear independent of S°_{ct} and F_{3dB}, namely, $S^{\sim}_{lp}$ and S_{ns} (Figure 3.8), as well as the formant-based F_{sl} and F_{freq} (Figure 3.5). These additional descriptors could be of special interest in achieving more complete prediction models, although their relevance seems to depend on the stimulus context. A similar analysis approach on a wider data set is needed to confirm the obtained trends, and possibly even give further insight into the role of associations (Σ, Δ, Ξ) relevant for different musical scenarios. Furthermore, the apparent utility of pitch-generalized descriptors, i.e., all F descriptors and S°_{ct} as opposed to $S^{\sim}_{ct}$, implies that a case-by-case signal analysis on individual pitches may not be necessary, but instead, a prediction application could rely on a comprehensive, offline database of pitch-generalized instrument descriptions alone.

When taking into account the relative locations of instrument combinations along the PCs that correlate with spectral features, a recurring pattern of dyads or triads including oboe (grey), on one side, opposed to combinations involving horn or trombone (green), on the other, becomes apparent. In other words, dyads or triads containing oboe are often less blended, whereas combinations with horn or trombone (e.g., bassoon and horn, clarinet and horn, trombone and trombone) are among the most blended ones. Employing the notion of *blendability* of a particular instrument, the oboe should be considered a poor 'blender', which can be explained spectrally by its prominent and unique formant structure. Similar observations linking oboe to poor blend have been made in previous perceptual investigations (Kendall and Carterette, 1993; Sandell, 1995; Reuter, 1996; Tardieu and McAdams, 2012) as well as 'prescriptions' found in orchestration treatises (Koechlin, 1959; Reuter, 2002). On the other hand, the horn is generally considered an easily blendable instrument, again reflected in perceptual results (Sandell, 1995; Reuter, 1996). The relatively 'dark' timbre of the horn could support a general hypothesis of lower centroids leading to more blend (Sandell, 1995), at the same time, supporting the argument that similar main-formant locations explain the good blend obtained between horn and bassoon (see Chapter 4). In addition, Figure 3.10 illustrates that the distribution (Ξ) along S°_{ct} distinguishes triads with two identical instruments (e.g., TTC, PPC, AAC) from more diverse combinations, without, however, directly reflecting analogous relationships with respect to

the degree of blend (visualized size of the symbols for scores P). However, it does imply that timbral similarity, if not identity, aids blending. In summary, spectral features do seem to represent the main underlying factor governing whether instrument combinations blend or not, with pitch-generalized spectral descriptions conveying the timbral signature traits of instruments.

3.5 Conclusion

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

One main challenge for future research is determining the effective weighting between these different factors of influence. Whether the clear dominance of interval type or impulsiveness of attacks over spectral features, which became apparent in the current investigation, would extend to more complex musical contexts remains to be explored. It can be assumed that the growing complexity that a listening scenario involving musical contexts would present, given the simultaneous presence of other musical parameters, could significantly alter the relative importance of factors found in listening experiments employing isolated dyadic or triadic stimuli. For instance, a composer may assign a unison blend between two instruments to a melodic voice while juxtaposing this against a chordal, non-unison accompaniment whose instruments are chosen to blend amongst themselves into a homogeneous timbre. On another level, the melody may become more distinct from the accompaniment due to the distinction between unison and non-unison, which may also be desired. This case scenario illustrates that blend-related factors need not stand in compe-

tition with each other like they do in the investigated perceptual data, but instead could operate on independent levels, fulfilling separate functions within the musical context. For the composer, working with blend is not a matter of favoring unison intervals over non-unison intervals, but being able to employ it at individual levels of the musical scene (e.g., melody, accompaniment, or contrasting the two). Within each level, blend is achieved by relying on the same principles, i.e., similarity in spectral description as well as articulatory features (e.g., note attacks). This hypothetical scenario encourages future work on blend-prediction models that relies on perceptual data obtained from stimuli involving musical contexts (Kendall and Carterette, 1993; Reuter, 1996; Chapter 4), as it provides a more realistic setting from which weights between blend-related factors could be estimated. We thus propose the need of a meta-analytical investigation into a diverse range of perceptual blend data, in an attempt to move toward generally applicable blend-prediction tools.

Chapter 4

Blend-related timbral adjustments during musical performance

This chapter presents an investigation into performance-related factors related to timbre blending, by considering practical scenarios involving musical and acoustical factors. A production experiment (Experiment 5) investigates the interactive dependencies between horn and bassoon players attempting to attain blend during musical performance. The content is based on the following research article:

Lembke, S.-A., Levine, S. and McAdams, S. (In preparation). Blending between bassoon and horn players: an analysis of timbral adjustments during musical performance. *Music Perception*.

4.1 Introduction

Among the many aims of orchestration, the combination of instruments into a blended timbre is one that is most relevant perceptually. Although decisions concerning orchestration can be primarily guided by personal preference, blend relies on a set of perceptual factors. It is commonly assumed to concern the auditory fusion of concurrent sounds into a single timbre, with the individual sounds losing their distinctness, and, furthermore, it is thought to span a perceptual continuum from complete blend to distinct perception of individual timbres (Sandell, 1991; Kendall and Carterette, 1993; Sandell, 1995; Reuter, 1996; Tardieu and McAdams, 2012; and Chapter 2). Perceptual cues that are favorable to blend range from synchronous note onsets and pitch relationships emphasizing the harmonic series to instrument-specific acoustical traits. With respect to the latter, previous

studies have shown spectral properties to have the strongest effect on blend between sounds from sustained instruments. The spectra of many wind instruments have been shown to be largely invariant with respect to pitch and may also bear prominent features such as spectral maxima. These maxima are also termed *formants*, in direct analogy to the pitch-invariant spectral maxima found in human voice production. Previous explanations for blend being related to spectral features are either based on global spectral characterization or focus on local, prominent spectral traits. The global and more general hypothesis was established from studies for instrument dyads, in which the spectral centroids of individual instruments were evaluated, representing the global, amplitude-weighted frequency average of a spectrum. It was shown that a lower frequency sum of individual spectral centroids correlated with higher degrees of blend (Sandell, 1995; Tardieu and McAdams, 2012). The alternative hypothesis argues for localized spectral features to influence blend, more specifically, concerning formant relationships between instruments: when two instruments exhibit coincident formant locations, high blend is achieved, whereas increasingly divergent formant locations decrease blend, as the individual identities of instruments are thought to become more distinct (Reuter, 1996).

Chapter 2 followed up on the formant hypothesis by studying frequency relationships between the most prominent *main* formants. The investigation considered dyads of recorded and synthesized instrument sounds. While the former remained a static reference, the latter was varied parametrically with respect to its formant frequency. For the instruments with prominent formant structure, namely bassoon, (French) horn, trumpet, and oboe, blend was found to decrease markedly when the synthesized main formant exceeded that of the reference, whereas comparably high degrees of blend were achieved if the synthesized formant remained at or below the reference. This rule proved to be robust across different pitches, with the exception of the highest instrument registers, and even applied to non-unison pitch intervals. Yet, this rule relies on one instrument serving as a reference, which raises the conundrum of which of two instruments in an arbitrary combination would function as the reference. The answer may lie in musical practice: either the instrument leading the joint performance or the one with a more dominant timbre could assume this function. In musical practice, achieving blended timbres involves two stages: its conception and its realization. Blend is first conceived by composers and orchestrators, who lay out the foundations by providing necessary perceptual cues, i.e., ensuring that musical parts have synchronous note onsets and pitch relationships favorable to blend, with the parts

being assigned to suitable instrument combinations. The successful realization of blend as perceived by listeners still depends on musical performance, which necessitates precise execution by several performers with respect to intonation, timing, and likely also coordination of timbre. Previous research precluded the influence of performance by relying on stimuli that were mixed from instrument sounds that had been recorded in isolation, with there being only a single exception (Kendall and Carterette, 1993) in which dyad stimuli had been recorded in a joint performance (Kendall and Carterette, 1991). The interaction between performers may in fact affect blend in a way that previous research has not considered. For instance, differences between performer roles could provide answers to the question of a certain instrument serving as a reference as intimated in Chapter 2.

4.1.1 Musical performance

Psychological research on musical performance has primarily investigated temporal properties. Although past investigations have focused on note synchronization and timing between performers (Rasch, 1988; Goebl and Palmer, 2009; Keller and Appel, 2010) as well as related motion cues (Goebl and Palmer, 2009; Keller and Appel, 2010; D'Ausilio et al., 2012), performer coordination with respect to timbral properties remains largely unexplored. Rasch (1988) established that a certain degree of asynchrony between performers is common and practically unavoidable, whereas perceptual simultaneity between musical notes is still conveyed. For example, typical asynchronies between wind instruments (e.g., single and double reed) performing in non-unison are reported as falling within 30-40 ms. Moreover, the asynchronies relate to different roles assumed by musical voices, e.g., the melody generally precedes bass and middle voices.

Two studies investigated the relationship between two pianists being assigned performer roles as either *leader* or *follower*. In one study, followers exhibited delayed note onsets relative to leaders (Keller and Appel, 2010), whereas in the other, followers displayed a higher temporal variability, thought to be linked to a strategy of error correction relative to leaders (Goebl and Palmer, 2009). In addition, the second study showed that under impaired acoustical feedback, performers increasingly relied on visual cues to maintain synchrony, which motivates investigations of performance-related factors involving auditory properties alone to prevent visual communication.

Role dependencies between performers are indeed common to performance practice.

They have been investigated for larger ensembles (D'Ausilio et al., 2012) and have been discussed in terms of *joint action* (Keller, 2008), in which they may modulate how performers rely on cognitive functions such as anticipatory imagery, integrative attention, and adaptive coordination. In terms of musical interpretation, leaders commonly assume charge of phrasing, articulation, intonation, and timing, whereas followers "adapt their own expressive intentions to accommodate or blend with another part" (Goodman, 2002; p. 158). It therefore appears plausible that the performance of blended timbre may similarly rely on role assignments between musicians. For instance, when two instruments are doubled in unison, one of them assumes the leadership in performance, toward which followers may orient their timbral and timing adjustments.

The current study explores what timbral adjustments are employed in achieving blend and how these interact in a performance scenario with two musicians. A set of acoustical measures monitors the spectral change and potential covariates that are assumed to be related to timbral adjustments. In addition, performances are also evaluated through musicians' self-assessment. Besides timbral adjustments, performances naturally also involve aspects related to timing, intonation, and adjustment of dynamics. Intonation has not been previously discussed as relating to blend, likely due to past research having precluded performance-related aspects, but reports from performers argue that correct intonation aids blending. Given the emphasis on timbre, however, performer coordination with respect to synchronization and intonation remains outside the focus of the current study. Moreover, they represent aspects that are important to accurate delivery of musical performance in general, which greatly limits the extent to which they can be varied independently to affect blend. Furthermore, a high frequency resolution is desirable for spectral analyses, which, due to an inherent inverse dependency, comes at the expense of temporal resolution. As the adequate measurement of asynchronies between note onsets would call for the relatively high time resolution of 5 ms (Rasch, 1988), which would require a separate series of acoustical analyses, it was considered prohibitive, given the focus on timbral adjustments.

The investigation attempts to feature a realistic account of factors encountered in musical practice and situates musicians in an approximation to the ecologically valid setting of a concert hall, realized through controlled and reproducible virtual performance environments. The *coloration* of instrument timbre as a function of relative position in a concert hall has been reported to be perceptible (Goad and Keefe, 1992) and would similarly extend to differences between rooms. Furthermore, an impairment of the acoustical

communication between musicians (Goebl and Palmer, 2009) may be relevant to the performance of blended timbre as well. Because the investigation considers a potential effect of performer roles, an instrument combination should be chosen that allows for sufficient timbral coordination, i.e., by avoiding situations in which one instrument's timbre dominates the other so strongly that a change in role assignments is unlikely to overcome the timbral mismatch. An instrument combination that finds widespread use in the orchestral repertoire is bassoon and horn. Orchestration treatises discuss these two instruments as forming a common blended pairing (Rimsky-Korsakov, 1964; Koechlin, 1959), with these observations reflected in findings of high degrees of blend in perceptual investigations (Sandell, 1995; Reuter, 1996). The horn is often considered an unofficial member of the woodwind section, bearing a timbral versatility that succeeds in blending with woodwinds, brasses, and even strings, which suggests that it, at the very least, should succeed in bridging timbral differences with the bassoon.

4.1.2 Acoustical measures for timbre adjustments

With regard to acoustics, bassoon and horn bear a high resemblance in their spectral envelopes. Figure 4.1 illustrates their global spectral envelopes for the dynamic marking *piano*. This envelope approximates the spectral traits found to be invariant across their pitch ranges, as estimated from spectra across all playable pitches (see Chapter 2 and Appendix A). As their most prominent traits, main formants are located around 500 Hz and can be characterized by the frequency corresponding to the maximum magnitude F_{max} and the upper frequency bound F_{3dB} at which the magnitude has decreased by 3 dB. Both instruments' main formants are quite similar, with their F_{max} differing by about 80 Hz, whereas their F_{3dB} lie much closer. In addition, the spectral centroids S_{ct} are located in the vicinity of the main formants, showing that even the global spectral distribution focuses on the main formants. Figure 4.1 provides a generalized description approximating the instruments' structural invariants, i.e., related to what informs orchestrators in their choice of instruments. In practice, these structural constraints still allow for a certain degree of timbral variation that musicians can exploit. Because wind instruments act as acoustical systems in which all sound originates from common structural elements (e.g., mouthpiece, resonator tube), timbral adjustments are expected to be inherently linked to the primary parameters of sound excitation performers focus on, namely, pitch and dynamic intensity.

For both instruments, blend-related adjustments of timbre can still be assumed to relate to spectral changes, which can be quantified through the measures F_{max}, F_{3dB}, and S_{ct}. From a preliminary qualitative investigation with bassoon and horn players, the timbre variability at the players' control was found to be greater for horn than for bassoon. For the latter, the location and shape of the main formant is relatively fixed, with spectral changes primarily affecting the magnitudes of higher frequency regions relative to the main formant, whereas the structural constraints of the horn allow for greater changes to main-formant location and shape.

Musicians reported that during performance, the greatest timbre change could be achieved by varying dynamics, which suggests a dependency between them. The identification of perceived dynamic markings has been shown to be mediated by both timbre and sound level (Fabiani and Friberg, 2011), which argues that when performers adjust dynamics, both timbre and the sound level (L_{rms}) are affected.

Apart from dynamics, pitch presents another source of covariation with spectral measures, with pitch being expressed through the fundamental frequency (f_0). Figure 4.2 shows a horn playing an ascending A-major scale over two octaves. All spectral measures show some variation as pitch ascends, quantified descriptively by the linear-correlation coefficient (Pearson's r): The strongest covariation with f_0 is apparent for S_{ct}, $r = .92$, whereas the correlation with main-formant measures is less pronounced, $r < .40$, with F_{max} and F_{3dB} meandering around idealized average values. Given these differences in covariation with f_0, the two types of spectral measures seem to capture independent contributions of timbral change. It is important to note that even f_0 and L_{rms} yield a clear degree of correlation, $r = .72$, with about 10 dB of level change across the two octaves. In orchestration practice, this correlation could correspond to the notion of *pitch-driven dynamics*, with experimental evidence showing that ascending pitch contour can enhance the identification of changes in dynamics, e.g., *crescendo* (Nakamura, 1987). In summary, the preliminary investigation argues for timbral adjustments to be jointly evaluated through measures of spectral variation as well as potential factors of covariation, namely, pitch and dynamics.

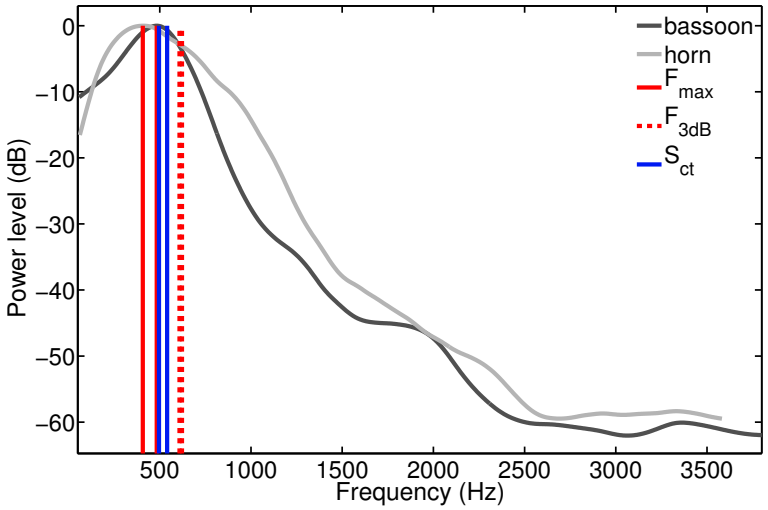

Fig. 4.1 Spectral-envelope descriptions for bassoon and horn at dynamic marking *piano*. Spectral descriptors F_{max}, F_{3dB}, and S_{ct} exhibit clear commonalities between the two instruments.

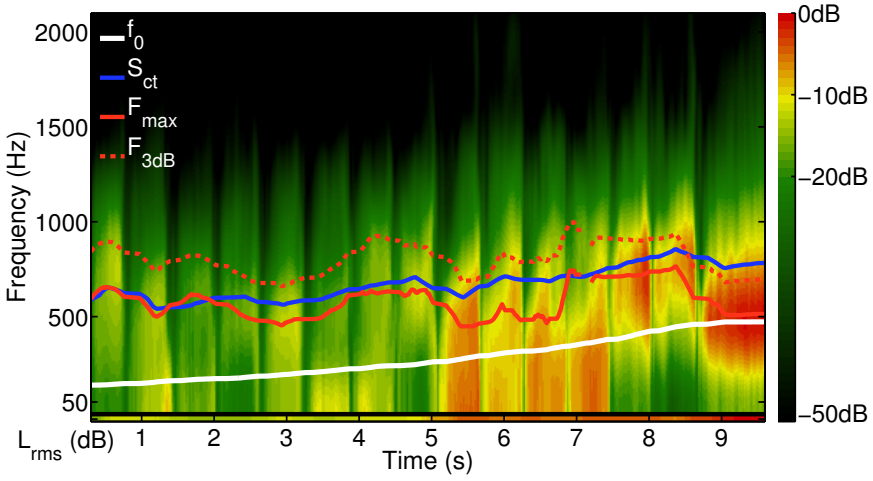

Fig. 4.2 Horn playing A-major scale from A2 to A4. Time course of spectral envelopes (magnitude in color; legend, far-right), with corresponding measures for spectral properties and pitch (curves) as well as dynamics (horizontal strip, bottom).

4.2 Method (Experiment 5)

4.2.1 Participants

Sixteen musicians were recruited primarily from the Schulich School of Music at McGill University and the music faculty of the Université de Montréal. The bassoonists, three female and five male, had a median age of 21 years (range 18-31). The hornists, six female and two male, had a median age of 20 years (range 17-44). Across both instruments, 10 participants considered themselves professional musicians, and overall, the musicians reported playing or practicing their respective instruments for the median duration of 21 hours per week (range: 5-35). All musicians were paid for their participation.

4.2.2 Stimuli

Three musical excerpts were investigated, all taken from Mendelssohn-Bartholdy's *A Midsummer Night's Dream*, Op. 61, No. 7 (measures 1-16). The chosen instrument combination is featured prominently in this musical passage. In a thin orchestral texture, low strings, second horn, and clarinet establish the harmonic structure through long, separated notes, while two bassoons accompany a solo horn melodically. In the absence of other salient voices, the combination of bassoons with horn can therefore be thought to aim for a homogeneous, *blended* timbre. All phrases were transposed by a fifth down to A major from the original key of E major, to reduce the impact of player fatigue through repeated performances in high instrument registers, at the same time ensuring little change in key signature. The transposed excerpts are shown in Figure 4.3. The melody, voice A, is used for unison performances, whereas voices B and C served as non-unison material.

Although the musicians played in separate rooms in order to record their individual sounds, they heard themselves and the other player over headphones in a simulated virtual-acoustics environment, which allowed the control over acoustical factors (see Section 4.2.3). The simulation was achieved through binaural reproduction (Paul, 2009) using real-time convolution of the instruments' source signals with individualized binaural *room impulse responses* (RIRs). Each musician's performance was captured through an omnidirectional microphone (DPA *4003-TL*). Both microphone signals were routed to a control room, where preamplification gain was digitally matched for both performers. The analog signals were converted to 96 kHz / 24-bit PCM digital data, recorded at full resolution for later

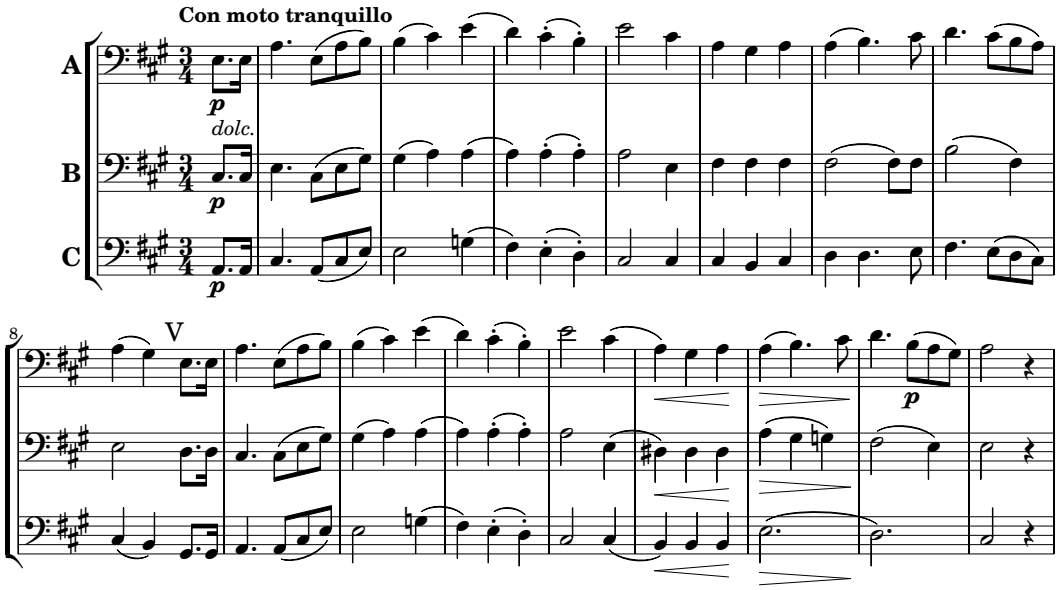

Fig. 4.3 Investigated musical excerpts A, B, and C, in A-major transposition, based on Mendelssohn-Bartholdy's *A Midsummer Night's Dream*. The 'V' marks the separation into the first and second phrases (see *Musical factors* under Section 4.2.3).

acoustical analysis and at the same time fed into separate convolution engines that processed the source signals with customized RIRs, based on the manipulations of acoustical factors. Individualized binaural signals were then fed to headphones for each performer. Headphone amplifier volume was held constant, as were the circumaural closed-ear headphones (Beyerdynamic *DT770*). A latency inherent to the convolution delayed the arrival of the simulated room feedback by about 8.4 ms, affecting both performers equally. The RIRs had been previously collected in real concert venues and were measured with a binaural head-and-torso system (Brüel & Kjaer *Type 4100*), excited by a loudspeaker (JBL *LSR6328P*) positioned to emulate the instruments' main sound-radiation directivity (Meyer, 2009). In the simulated environment, musicians would hear themselves and the other musician in a common performance space, which provided realistic room-acoustical cues (e.g., room size, its reverberation characteristics, relative spatial positions of players). The instrument locations were based on a typical orchestral setup (see Figure C.3): horns on the conductor's left front side and bassoons on the conductor's right front. For instance, hornists heard themselves in direct proximity and the bassoonist towards their left, at a distance of 3.6 m, whereas the bassoonists' viewpoint was reversed in orientation. In or-

der to take these individual viewpoints into account, i.e., as performers heard themselves (*self*) and the other musician (*other*), the acoustical analyses of performances consider the individualized binaural signals. Although four possible binaural signal paths result from one performer having two ears and the two viewpoints *self* and *other*, only two are considered for simplicity: *self* considers the ear facing away from the other performer, and *other* considers the ear closer to the other performer.

4.2.3 Design

Performances are studied as a function of musical and acoustical factors using a repeated-measures design to rule out confounding individual differences for instruments and playing technique or style with the investigated effects.

Musical factors

Three independent variables consider the performer role, the influence of different musical voice contexts, and performance differences across time. For the Role factor, one instrumentalist is assigned the role of *leader*, and the other performer acts as *follower*, i.e., takes on an accompanying role. According to the Interval factor, musicians either perform a melodic phrase in *unison* (voice A in Figure 4.3) or a two-voice phrase in *non-unison* (B and C); in non-unison, the top voice (B) is assigned to the leader. The Phrase factor divides the musical excerpts into two, with the separation occurring right before beat three of measure eight (see the 'V' in Figure 4.3). This separation yields two musical phrases of identical length consisting of similar musical material, more so for unison than for non-unison excerpts.

Acoustical factors

Two other variables investigate effects for communication directivity between performers and the room-acoustical properties of performance venues. The Communication factor assesses the influence of whether both performers are able to hear each other or only the follower hears the leader, denoted *two-way* or *one-way*, respectively. For the Room factor, the influence of acoustics is assessed for two different performance spaces: musicians are simulated as performing in either a large, multipurpose performance space ($RT_{60} = 2.1$ s,

time for reverberation to decrease by 60 dB) or in a mid-sized recital hall ($RT_{60} = 1.3$ s).[1]

4.2.4 Procedure

The experiment was conducted in two research laboratories at the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) at McGill University. Separate laboratory spaces were called for in order to create individual acoustical environments for each participant, ensuring the capture of separate source signals as well as preventing visual cues between performers. Each performance laboratory was treated to be relatively non-reverberant, with $RT_{60} < 0.5$ s. Performers received instructions and provided feedback through dedicated computer interfaces. Musical notation for all three excerpts was provided on a music stand, while performances were temporally coordinated by a silent video of a conductor. With both performers seated on chairs, the stand was positioned to allow the performer's field of view to cover both the musical notation and the conductor, arranged similarly to the binaurally simulated orchestra situation, i.e., the stand slightly to the right of the conductor as seen from a hornist and vice versa for a bassoonist. The video was recorded in advance by having an experienced conductor (with baton) outline the metrical structure of the musical excerpts, including gestures related to phrasing and articulation. He used a constant reference tempo of 58 beats per minute.

A pair of bassoon and horn players was tested in a single experimental session, being instructed to perform together to achieve the highest degree of blend possible. They performed three repetitions of 16 different experimental conditions (four factors by two treatment levels, excluding Phrase), leading to a total of 48 experimental trials. The experiment lasted around two hours in total, including a break scheduled after half of the trials. To avoid disorientation of musicians through strongly varying performer-role and voice assignments, the musical factors were blocked. Participants assumed the role of either leader or follower throughout the first or second half of the experiment. Furthermore, shorter eight-trial blocks grouped conditions based on voice assignment (e.g., four unison trials, another four non-unison), with the repetitions occurring after each block. For instance, a given participant would begin as leader for 24 trials, performing the first repetition of four unison trials, then proceed to four non-unison trials, followed by the second repetition of the same four unison trials, etc. The four possible block-ordering schemes were counter-

[1] The performance venues correspond to the *Music Multimedia Room* and *Tanna Schulich Hall*, respectively; both are located at the Schulich School of Music, McGill University.

balanced across all participants and instruments. The acoustical-factor combinations were nested in sub-blocks of four trials and randomly ordered. Three practice trials were conducted under the guidance of two experimenters, presenting the experimental conditions encountered at the beginning of individual block-ordering schemes.

A single experimental trial consisted of three stages: preparation, performance, and ratings. During preparation, musicians were asked to prepare the assigned musical excerpts and individual performer roles, while being able to hear themselves in the current simulated room environment. After both participants indicated being prepared, the actual performance commenced and once it ended, each participant judged their individual experience of the performance by providing two ratings. The first rating assessed how well they thought they had individually *performed* given their assigned role on a continuous scale with the verbal anchors *very badly* and *very well*. The second rating concerned the perceived degree of achieved *blend* with the other performer on a continuous scale with the verbal anchors *low blend* and *high blend*.

4.2.5 Acoustical measures

In addition to the behavioral ratings, several acoustical measures that account for timbre features related to blend were derived from time-series analyses of the musical performances. Timbral adjustments were evaluated through spectral descriptors and also monitored through the covariate measures pitch and dynamics. Likewise, two additional cues important to blend, namely, intonation and synchrony, were initially considered to allow their influence to be filtered out subsequently. Performances were analyzed with respect to the time-averaged magnitude of a measure, its temporal variability during performance, and its temporal coordination between performers. Therefore, each measure yielded three corresponding dependent variables (DVs).

All acoustical measures were based on spectral analyses across the time course of performances, for which short-time Fourier transforms (STFT) and further derived representations were computed using dedicated software (AudioSculpt/SuperVP, IRCAM, Paris). STFT is based on the fast Fourier transform (FFT), using Hann-windowed analysis frames consisting of 7620 samples, FFT length of 8192 bins, and an overlap of 25% between successive frames. This corresponds to a frequency and time resolution of 11.7 Hz and 19.8 ms, respectively. Pitch detection employed harmonic analysis of the STFT spectra (Doval and

Rodet, 1991), with the identified fundamental frequency f_0 configured to fall within the possible range $f_0 \in [92.5, 370]$ Hz, which reflects the pitch range across all excerpts expanded by a whole tone on each end. The f_0 estimates provided by AudioSculpt were complemented by corresponding confidence scores, i.e., the likelihood for identified harmonics to be linked to f_0, which in turn were used to discard time frames falling below 80% confidence from further analysis for all measures. This elimination improved the reliability of both f_0 and the spectral measures. Based on the remaining STFT frames, spectral envelopes were obtained through *True Envelope* (TE) estimation (Villavicencio et al., 2006). The TE algorithm applies iterative cepstral smoothing on STFT-magnitude spectra, with the computed estimates using a constant cepstral order oriented at $f_0 \leq 300$ Hz. A formant-analysis algorithm (see Section A.2) evaluated the spectral envelopes, identifying main formants (F_1), which were quantified in terms of frequencies characterizing their maximum F_{max} and the upper bound F_{3dB}, as well as computing the spectral centroid S_{ct} (Peeters et al., 2011). The spectral envelopes also served to quantify dynamics by determining relative, root-mean-square (RMS) power levels L_{rms}.

As the raw time-series data for the measures exhibited some fine temporal variation and occasional outliers, some prior data treatment was needed. All measures were smoothed by a weighted moving-average filter. Weights were based on the f_0-confidence scores, assuming that higher confidence reflects a more robust and reliable parameter estimate. Smoothing used a sliding-window duration of 475 ms, which corresponds to an eighth note at the performed tempo. Especially for horn signals, the automated formant detection at times led to erroneous estimates, which could be identified and eliminated. Prior to smoothing, the main-formant descriptors F_{max} and F_{3dB} were filtered for outlying values that lay beyond an octave below and two-thirds of an octave above their time-averaged median value, because unrelated spectral features beyond these frequencies were occasionally classified as the main formant. Deemed an artifact of cepstral smoothing, the TE estimates for horns sometimes also exhibited spectral-envelope maxima at 0 Hz, in which case formant identification failed. Therefore, resulting gaps for F_{max} greater than two metrical beats were replaced by f_0 values, serving as the lowest tonal signal components. The corresponding F_{3dB} values were determined from the replaced F_{max}. The final step of data treatment ensured that the measures yielded values across all analysis frames of a performance, allowing comparisons between performers across all time points. This was achieved through linear interpolation of all remaining gaps to a reference time grid. Extrapolation was applied for values missing

at the edges, which rarely exceeded a quarter-note duration (e.g., delayed entry of the first note or the final note not being held for its entire duration).

The investigation focuses on timbral adjustments as reflected in spectral changes. However, not all changes are necessarily related to the intent to achieve blend. Performer actions related to errors in intonation or timing also create a certain degree of spectral change. Therefore, the performances were filtered for cases in which bad intonation and/or synchrony were apparent. Intonation is measured by comparing f_0 between performers, expressed as the relative deviation in cents. For unison, this characterizes deviations from a f_0 ratio of unity; for non-unison, deviation considers f_0 ratios of the corresponding intervals in equal temperament. Asynchrony can also be assessed through the intonation measure, because asynchronous note entries also evoke substantial deviations from perfect intonation for the duration by which they are offset from synchrony. The time series for all measures retained only values falling within the intonation range of ± 25 cents, which corresponds to musically acceptable intonation (Rakowski, 1990). Pitch (f_0) and dynamics (L_{rms}) are intrinsically related to the spectral measures and cannot be directly excluded from further analysis, but will instead be monitored for similar trends in the time-series spectral measures. The influence of f_0 is twofold: First, systematic differences in f_0 between the musical excerpts are likely reflected in deviations between unison and non-unison performances. Second, f_0 also varies over time, and all spectral measures covary with f_0 to some extent. By taking residuals (ϵ) from the linear regression of the f_0 time series onto the time series of each of the three spectral measures and adding the residual scores to the spectral time-series means, the linear covariation with f_0 over the excerpts can be removed. This procedure yields the *residual* measures ϵF_{max}, ϵF_{3dB}, and ϵS_{ct}.

The performance analysis focuses on individual performers and evaluates each acoustical measure with three DVs. The first DV quantifies the acoustical measure's average magnitude, using the *median* across time values. The second DV assesses the temporal variability along a measure, expressed as a robust *coefficient of variation* (CV): the ratio between interquartile range and median. The third DV assesses the temporal coordination between performers, evaluating the maximum *cross-correlation coefficient* (XC) for their time series.[2] Due to the expected covariation with f_0, the XCs for the spectral measures

[2] Although cross-correlation time lags were also evaluated, no evidence for relative delays in coordination was found across all measures. For instance, L_{rms} displayed a median lag of 0 ms across all conditions and both instruments, with the interquartile range also being 0 ms, showing hardly any variation along this

were assumed to be inflated by the inherent similarity in f_0 profiles between excerpts A and A, and even B and C. Therefore, this DV considers the residual measures (ϵ), whereas the remaining DVs are based on the original acoustical measures. Furthermore, in considering the individual viewpoints of performers within the binaural simulation, the DVs evaluating median and CV are based on time series for the binaural signal *self*, whereas the DV evaluating XC compares *self* with *other*.

4.3 Results (Experiment 5)

The results will focus on several hypotheses, tested by several experimental variables (set in *italics*). It is expected that musicians will perform differently as leaders than as followers, with those in the *Role* of followers adjusting their timbre to that of the leader. Unison *Intervals* are hypothesized to yield higher perceived blend than the non-unison case as well as showing more coordination between instrumentalists. At the same time, the known differences in f_0 register between excerpts are likely to be accompanied by substantial covariation with the spectral measures. Furthermore, the coordination between performers is predicted to increase throughout a performance, i.e., it should be higher in the second musical *Phrase* than in the first. With respect to the acoustical factors, differences between *Rooms* may affect the degree of coordination between performers to some extent, although it is not clear in what way. Finally, given an assumed stronger dependency of followers on leaders than vice versa, performances in which leaders lack acoustical feedback from followers are not expected to differ from the case with unimpaired *Communication*.

Performances across the 16 factorial combinations (excluding Phrase) were repeated three times. The actual analysis retains only two repetitions per participant pair that yield the highest self-assessed performance ratings, which need to reflect agreement between the two participants performing together. Out of three repetitions, at least one finds mutual agreement between both performers as to being rated among the highest two. For no mutual agreement on the remaining one, the repetition yielding the higher average rating across performers is taken. Some unforeseen technical issues during two experimental sessions rendered data for a total of five trials unusable. Fortunately, this affected only one repetition per experimental condition, allowing the remaining two repetitions to be used. In

measure. S_{ct} exhibited a median lag of 0 ms with an extremely wide interquartile range of 871 ms, which reflects little agreement across participants.

the analyses, separate performances are considered as independent cases, i.e., corresponding to a total number of 16 cases (eight performers × two repetitions) per instrument.

Analyses of variance (ANOVAs) tested effects across the within-participants musical and acoustical factors. The within-participant residuals yield slight departures from a normal distribution (Shapiro-Wilk test). Based on the known robustness of ANOVA to violations of normality for equal sample sizes (Harwell et al., 1992), its use is considered justified for DVs exhibiting less than 10 violations over all 32 factor cells, which all reported statistical effects fulfill. Furthermore, the two *Instrument* groups can be implemented as a between-participants factor if both groups exhibit similar variances. This is fulfilled for the behavioral ratings, as both groups of players use identical rating scales and do not exhibit systematic differences in their ratings. The acoustical measures, however, exhibit clear violations (Levene's test), brought about by consistent differences in their acoustical characterization. As a result, the acoustical measures involve separate ANOVAs by instrument.

4.3.1 Behavioral ratings

Participants provided two ratings quantifying their perception of *blend* and assessment of their own *performance* given their assigned role. As these apply to entire performances, only the four within-participants factors Role, Interval, Room, and Communication are analyzed, with Instrument forming a between-participants factor. Also, for the impaired acoustical feedback, performers acting as leaders actually did not provide *blend* ratings, as they were unable to hear the follower. This resulted in missing data in 4 out of 16 conditions, which were substituted with within-participant medians. The obtained effects are visualized in Figure 4.4.[3] As the substituted values are balanced evenly across the two voice levels, the obtained moderate main effect, $F(1,30)=12.02$, $p<.01$, $\eta_p^2=.29$, is assumed valid (far-left panel), confirming that performances in unison are perceived as blending more than in non-unison. However, two-way interaction effects with Communication and Role are considered unreliable, because the comparisons along both factors lead to half of the

[3]Figure 4.4 displays group medians and interquartile ranges, indicating the usage of the rating scale across all participants. Although this does not reflect the within-participants differences evaluated through ANOVA, the obtained trends still become apparent.

data points for one of the levels (e.g. *one-way*, *leader*) being determined by the substituted values.

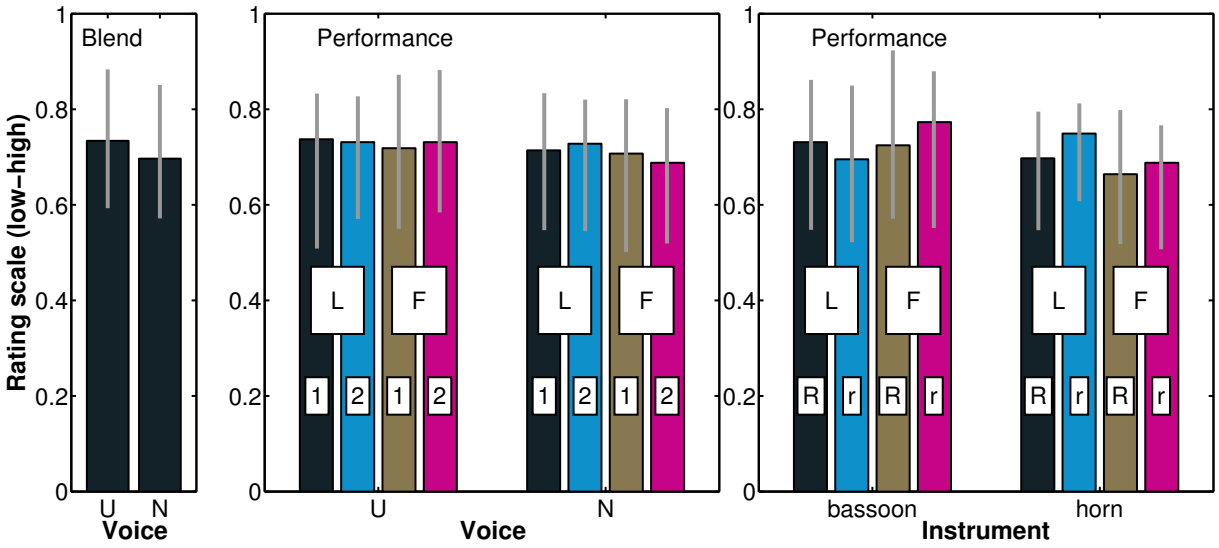

Fig. 4.4 Medians and interquartile ranges of ratings across all participants illustrating main effects for *blend* (left) and interaction effects for *performance* (center and right). Factor abbreviations: Role: leader (L), follower (F); Interval: unison (U), non-unison (N); Room: large (R), small (r); Communication: one-way (1), two-way (2).

Despite the *performance* ratings only leading to a marginally significant main effect for Interval, $F(1,30) = 3.90$, $p = .06$, $\eta_p^2 = .12$, this factor still yields two-way interactions (middle panel) with Role, $F(1,30) = 6.43$, $p = .02$, $\eta_p^2 = .18$, and Communication, $F(1,30) = 4.70$, $p = .04$, $\eta_p^2 = .14$. The first interaction involves musicians rating themselves as having performed their role better as followers than as leaders in unison conditions, with the inverse relationship holding for non-unison performances. The second suggests that in unison performances, musicians rate their performances higher for unimpaired, two-way communication, whereas the ratings for non-unison performances appear to be unaffected by communication directivity. Two additional interactions involve differences between instruments (right panel). A two-way interaction with Role, $F(1,30) = 6.49$, $p = .02$, $\eta_p^2 = .18$, yields higher performance ratings for bassoons than horns in the role of followers, whereas no difference between instruments is found for leaders. The same interaction suggests that bassoonists provide higher ratings as followers than as leaders, with the opposite applying to horns. A related three-way interaction adds the influence of the Room factor,

$F(1, 30) = 4.22$, $p = .05$, $\eta_p^2 = .12$. For bassoons, the difference between roles becomes larger in the smaller room, whereas for horns, the role difference appears to be limited to just the smaller room. Overall, these interdependencies suggest that communication impairment has a stronger effect on unison performances and that followers are more satisfied with their performances than are leaders. Differences between instruments and across roles could be related to instrument-specific issues concerning playability of the corresponding excerpts. Furthermore, the less reverberant acoustics of the small room seem to affect performances (or their evaluation) more critically.

4.3.2 Acoustical measures

Figure 4.5 presents the spectral measures and f_0 as a function of time, with a separate horizontal strip at the bottom for L_{rms}. In this example, the unison excerpt is performed under normal, two-way communication in the larger room, with the bassoon acting as leader and also considering its viewpoint, i.e., based on its binaural signals for bassoon (*self*) and horn (*other*). Three DVs are derived from each measure — median, CV, and XC — and are analyzed in repeated-measures ANOVAs investigating the factors Role, Interval, Room, Communication, and Phrase.

Since the acoustical measures and associated DVs are quantified along physical scales or quantities derived from them, statistical effects are also evaluated against psychoacoustically meaningful thresholds. For median L_{rms}, differences need to exceed 1 dB, as this value estimates the just-noticeable difference (JND) for amplitude (Zwicker and Fastl, 1999). For all spectral measures (F_{max}, F_{3dB}, S_{ct}), differences below 5 Hz are disregarded, as this corresponds to 1% frequency variation relative to the investigated main formants, falling slightly above the JND for frequency (Zwicker and Fastl, 1999; Moore and Moore, 2003). For CV, differences below 10% are considered negligible, because even confounding variables can be shown to introduce greater variability (see Covariates). Lastly, XC differences below 1% (e.g., 0.3% improved temporal coordination) are considered of too little value to be reported. The threshold for XC is expressed in terms of explained variance, i.e., differences between r^2 values.

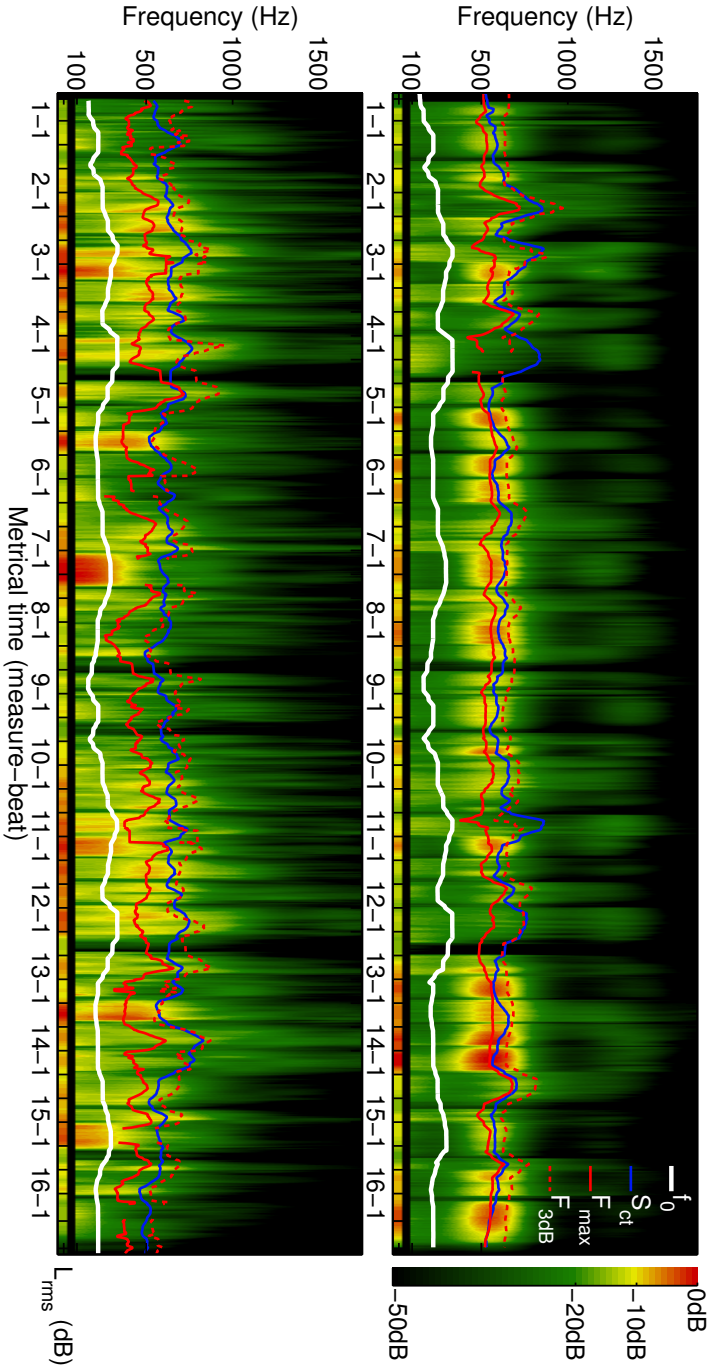

Fig. 4.5 Single performance of the unison excerpt by a bassoon (top) and a horn (bottom) player. TE spectrogram and time series of (smoothed) acoustical measures (compare to Figure 4.2).

Covariates

As the acoustical measures are based on real-life signals, they may contain and introduce some covariate influence across factor distinctions, which could be unrelated to deliberate timbre adjustments by performers. For instance, different rooms typically impose a characteristic *coloration*, i.e., frequency filter, that may induce shifts in the spectral measures. Likewise, the apparent differences in f_0 register between excerpts likely impose spectral shifts that lie beyond performers' control.

The assessment of potential room effects compares fixed reference performances simulated at the *self* positions in the small vs. large rooms. For greater representativeness, this procedure is applied to two selected performances per participant, for excerpts A and C, yielding 2×16 cases. The comparison of group medians yields shifts for all spectral measures and L_{rms}: identical horn performances exhibit slightly stronger dynamics in the large than in the small room, with the opposite applying to bassoon. Likewise, the spectral measures vary by about 1% in main-formant frequency between rooms. In terms of CV, the spectral measures exhibit up to 30% more temporal variability in the large room, whereas the same room decreases variability in L_{rms} by about 20%. It appears that higher reverberation introduces greater spectral variation, whereas it smoothes out temporal variability in dynamics. At the same time, the increased reverberation in the larger room would be expected to even affect XC, which compares time series for *self* against *other* as heard by a single performer. As apparent in Figure 4.5 (bottom compared to top panel), the performance for *other* yields more variability than the *self* position, i.e., signals heard from afar are more reverberated. An additional change in reverberation between rooms could therefore modulate the XC further, as a greater change can again be expected for *other*. Unfortunately, these considerations suggest that pre-existing, systematic differences in room acoustics introduce a confounding influence on all measures and across all DVs, compromising the ability to quantify differences in performer adjustments between rooms separately. As a result, obtained ANOVA effects will be measured against the thresholds quantified above, serving as baselines for the systematic variation. The baselines for median DV between rooms are visualized in Figure 4.6 (horizontal lines matching brown bars).

Spectral covariation with f_0 between excerpts is quantified on the actual performer data. The comparison considers separate group medians by excerpt, with the spectral shifts expressed relative to excerpt A, which has the highest f_0. Spectral shifts can also be

	f_0		Bassoon			Horn		
x/A	Hz	%	F_{max}	F_{3dB}	S_{ct}	F_{max}	F_{3dB}	S_{ct}
B	-62	-25	-4	-2	-6	-19	-12	-13
C	-104	-42	-13	-7	-13	-24	-12	-21

Table 4.1 Covariation of spectral measures with f_0 for excerpts B and C relative to A (in % if not indicated otherwise), quantified as medians across all performances of an excerpt. f_0 per excerpt corresponds to the median across pitches, weighted by their relative durations.

compared to corresponding changes in f_0 itself, represented by the median across pitches per excerpt, which is weighted by the relative duration of individual pitches. Table 4.1 displays these comparisons: While f_0 varies as much as -42%, the spectral shifts are less pronounced, nonetheless exhibiting a monotonic decrease by excerpt, i.e., C is lower than B, which is lower than A. Bassoons exhibit only up to -13% of covariation, whereas horns show decreases up to -24%. The averaged frequency shifts for B and C are taken as the baselines for spectral shifts induced from f_0 changes alone and are also visualized in Figure 4.6 (horizontal lines matching blue bars).

Given the covariate influence of rooms and f_0, the presentation of results for the factors Room and Interval precede the three remaining factors. Figure 4.6 visualizes potential main effects for median DV across all original acoustical measures, i.e., F_{max}, F_{3dB}, S_{ct}, and L_{rms} (individual panels from left to right, respectively). The bars and error lines symbolize medians and interquartile ranges for within-participants differences between factor levels, respectively, for the factors Role (black), Interval (blue), Room (brown), Communication (pink), and Phrase (orange). The factor-level abbreviations above and below the x-axis indicate the orientation of a difference. For instance, for positive values in S_{ct}, the spectral centroid is higher for unison (U) than non-unison (N); the reverse applies for negative values.

Room

ANOVAs on the median DVs yield differences for the spectral measures and for L_{rms} that directly mirror the expected covariate baselines between rooms alone, as illustrated in Figure 4.6 by comparing the brown bars to the corresponding horizontal lines. All spectral

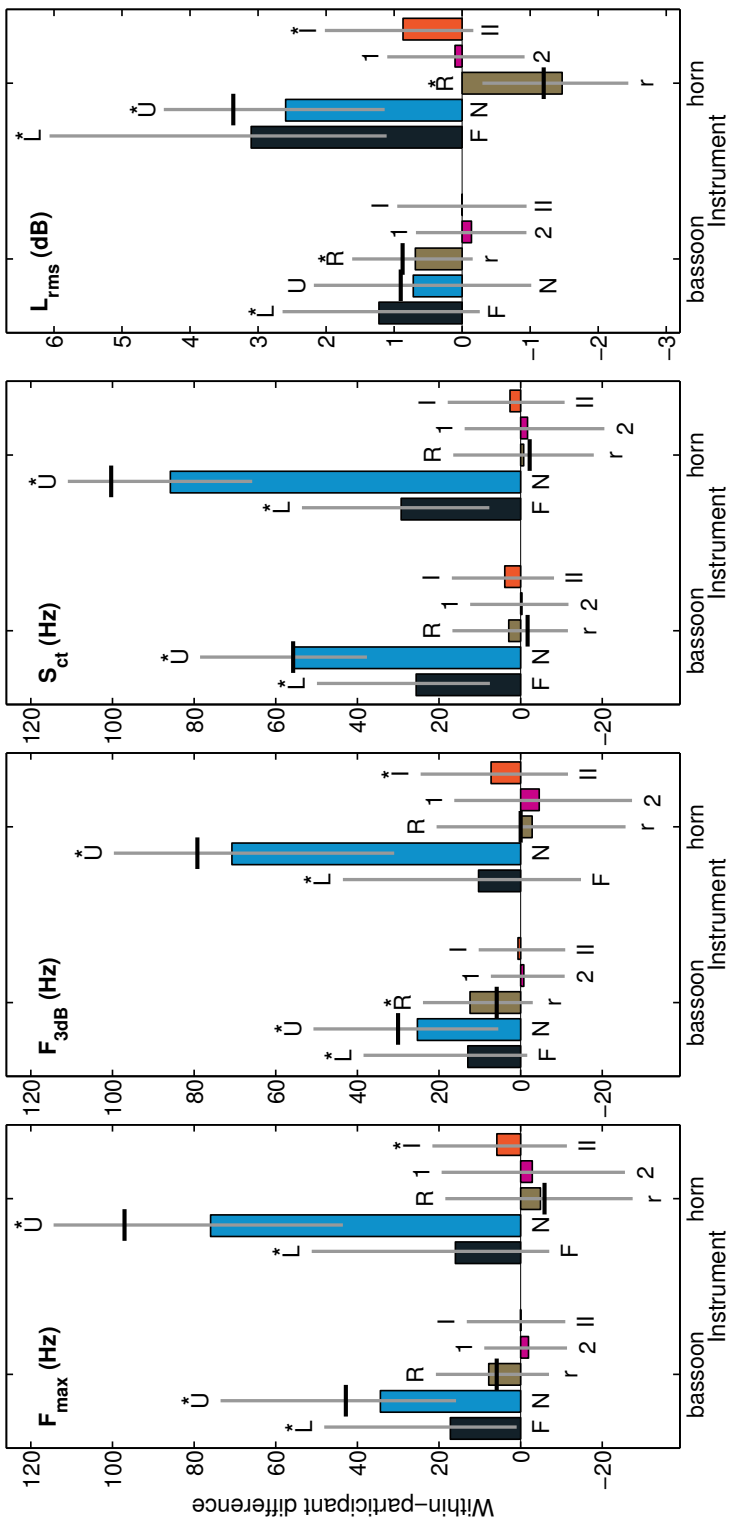

Fig. 4.6 Medians and interquartile ranges of within-participants differences for all acoustical measures (each panel) as a function of instrument (left and right parts in each panel) for the five independent variables. Factor levels are abbreviated and labelled above and below the x-axis. For instance, positive differences signify $U > N$; negative ones $N > U$. Abbreviations: Role: leader (L), follower (F); Interval: unison (U), non-unison (N); Room: large room (R), small room (r); Communication: one-way (1), two-way (2); Phrase: first (I), second (II). Asterisks (*) indicate significant ANOVA findings falling above the predefined thresholds. Black horizontal lines for Interval and Room indicate the expected differences arising from f_0-register and room-acoustical variability alone, respectively (see Covariates).

measures show slight positive or negative trends between rooms for bassoon and horn, respectively. F_{3dB} differences are significant only for bassoon, $F(1,15) = 22.86$, $p < .01$, $\eta_p^2 = .60$. In addition, L_{rms} yields differences for both bassoon, $F(1,15) = 24.02$, $p < .01$, $\eta_p^2 = .62$, and horn, $F(1,15) = 164.48$, $p < .01$, $\eta_p^2 = .92$. Also the CV exhibits greater temporal variability in the larger room. All spectral measures yield differences for the horn, $F(1,15) \geq 7.74$, $p < .02$, $\eta_p^2 \geq .34$, whereas differences are limited to both main-formant measures for bassoon, $F(1,15) \geq 5.29$, $p < .04$, $\eta_p^2 \geq .26$. These differences again reflect the expected trends for room acoustical variation alone. As a result, the obtained findings appear to be influenced primarily by pre-existing, systematic differences between room acoustics and do not seem to be changed by deliberate performer actions.

Interval

All spectral measures exhibit higher median-DV frequencies in unison than in non-unison, for both bassoon, $F(1,15) \geq 60.41$, $p < .01$, $\eta_p^2 \geq .80$, and horn, $F(1,15) = 106.45$, $p < .01$, $\eta_p^2 \geq .88$. Moreover, these differences closely match the covariate baselines for f_0 register, as illustrated in Figure 4.6 when comparing the blue bars against the horizontal lines. For horn, unison also yields higher L_{rms}, $F(1,15) = 124.79$, $p < .01$, $\eta_p^2 = .89$. Again, these effects cannot be assumed to correspond to blend-rated performer actions, as they were dictated by the musical notation. The pronounced influence of Interval, however, is still relevant to the interpretation of effects along the remaining factors.

In addition, bassoonists show greater temporal coordination in unison, with XC increasing by 4% for ϵS_{ct}, $F(1,15) = 4.82$, $p < .05$, $\eta_p^2 = .24$, although the difference is explained mainly by the smaller room, Interval×Room: $F(1,15) = 5.69$, $p = .03$, $\eta_p^2 = .28$. By contrast, horns exhibit 8% greater coordination in L_{rms} in non-unison performances, $F(1,15) = 12.00$, $p < .01$, $\eta_p^2 = .44$, with the difference being only half as pronounced in the second phrase, Interval×Phrase: $F(1,15) = 7.76$, $p = .01$, $\eta_p^2 = .34$.

Role

The clearest indication for timbre adjustments by performers concerns differences between *leader* and *follower* roles. For the median DVs, role-based differences across spectral features and also dynamics become apparent in Figure 4.6, considering the black bars. Musicians yield higher spectral frequencies and increased sound levels as leaders than when

performing as followers. Figure 4.7 illustrates these trends even more clearly as equivalent spectral-envelope changes. These spectral envelopes (curves) and the indicated acoustical measures (vertical lines) represent medians taken across all performances, collapsed across the remaining factors. Although these aggregate differences do not compare to within-partipant differences, they still show how the effects influence the entire spectrum. As suggested by the pairs of spectral envelopes, the main formants of followers (light grey) recede in frequency and level compared to the leaders' (darker grey). This is reflected in analogous differences across the acoustical measures, although the detailed analysis reveals distinctions between instruments. For bassoon, the main-formant measures are larger for leaders, $F(1,15) \geq 33.02$, $p < .01$, $\eta_p^2 \geq .69$, but it appears to be limited to non-unison, which is likely related to the f_0 differences between excerpts B and C, Role×Interval: $F(1,15) \geq 34.76$, $p < .01$, $\eta_p^2 \geq .70$. Whereas S_{ct} decreases across all performances, $F(1,15) = 60.24$, $p < .01$, $\eta_p^2 = .80$, however, more so in non-unison, for similar reasons as before, Role×Interval: $F(1,15) = 76.50$, $p < .01$, $\eta_p^2 = .84$. At the same time, L_{rms} exhibits a slight decrease for followers, $F(1,15) = 14.49$, $p < .01$, $\eta_p^2 = .49$. Overall, the bassoons' main formants in unison performances remain fixed, whereas the change in S_{ct} suggests spectral adjustments relative to the main formant, which co-occurs with a slight decrease in L_{rms}. The differences obtained for horn are more pronounced. Both F_{max} and F_{3dB} yield higher frequencies for leaders, $F(1,15) \geq 9.45$, $p < .01$, $\eta_p^2 \geq .39$, with the difference for F_{3dB} appearing to be limited to unison performances, Role×Interval: $F(1,15) = 10.19$, $p < .01$, $\eta_p^2 = .40$. Also S_{ct} yields a difference between performer roles, with higher frequencies for leaders, $F(1,15) = 45.91$, $p < .01$, $\eta_p^2 = .75$, being again more pronounced for non-unison performances, Role×Interval: $F(1,15) = 6.43$, $p = .02$, $\eta_p^2 = .30$. Analogous differences concern leaders yielding higher L_{rms}, $F(1,15) = 22.84$, $p < .01$, $\eta_p^2 = .60$, and more so in the non-unison situation, Role×Interval: L_{rms}, $F(1,15) = 30.23$, $p < .01$, $\eta_p^2 = .67$. In summary, these findings argue that in the attempt to blend with leaders, followers adjust to 'darker' timbres and, interestingly, spectral features and dynamics change in a coherent way. For both instruments, S_{ct} drops by about 30 Hz and L_{rms} decreases by 1-3 dB for followers.

With regard to temporal variation, the DVs quantifying the CV exhibit instrument-specific effects. Leading hornists vary more than followers along F_{3dB} and S_{ct}, $F(1,15) \geq 9.15$, $p < .01$, $\eta_p^2 \geq .38$, whereas the contrary applies to bassoonists across all spectral measures, $F(1,15) \geq 22.42$, $p < .01$, $\eta_p^2 \geq .60$. For both instruments, these effects are limited to non-unison performances, which suggests that they arise from instrument-specific issues

between the excerpts B and C, Role×Interval: $F(1,15) \geq 5.93$, $p < .03$, $\eta_p^2 \geq .28$. For instance, the low registral range of excerpt C posed more playing difficulty to hornists than to bassoonists. Another role-dependent difference is specific to horns, in which the temporal coordination as quantified by XC is up to 3% higher for leaders concerning ϵF_{3dB} and ϵS_{ct}, $F(1,15) \geq 5.68$, $p \leq .03$, $\eta_p^2 \geq .28$.

Phrase

The comparison of acoustical measures between the first and the second phrase indicates that both players adapt throughout performances toward an assumedly improved configuration. With regard to median DV, leading bassoonists lower S_{ct} by about 12 Hz towards the second phrase, whereas followers increase by 10 Hz, still, remaining below leaders, Phrase×Role: $F(1,15) = 25.63$, $p < .01$, $\eta_p^2 = .63$. The effect for followers appears limited to non-unison, whereas in unison, followers do not vary S_{ct} throughout performances, Phrase×Role×Interval: $F(1,15) = 31.22$, $p < .01$, $\eta_p^2 = .68$. This notable interaction reveals that even leaders attempt to close larger gaps in S_{ct}, while followers fulfill the same objective by remaining stable or closing gaps in the opposite direction. Hornists show similar effects, although without interactions with other factors, as illustrated in Figure 4.6 (orange bars). The formant measures decrease by about 5 Hz in the second phrase, $F(1,15) \geq 6.69$, $p \leq .02$, $\eta_p^2 \geq .31$. Likewise, L_{rms} also decreases by about 1 dB throughout performances, $F(1,15) = 28.22$, $p < .01$, $\eta_p^2 = .65$. Similar effects for temporal coordination support the previous findings. For L_{rms}, the second phrase yields 6% and 8% higher XC for bassoon, $F(1,15) = 37.93$, $p < .01$, $\eta_p^2 = .72$, and horn, $F(1,15) = 125.05$, $p < .01$, $\eta_p^2 = .89$, respectively. Similarly, the coordination in ϵS_{ct} also increases in the later phrase by 3% for bassoon, $F(1,15) = 9.86$, $p < .01$, $\eta_p^2 = .40$, and 5% for horn, $F(1,15) = 19.14$, $p < .01$, $\eta_p^2 = .56$.

Communication

Among the acoustical measures, no clear indications were obtained that the absence of auditory feedback from the follower affected performances differently than in the unimpaired case. Of the few statistically significant findings, all fall below the pre-defined thresholds for psychoacoustically meaningful differences.

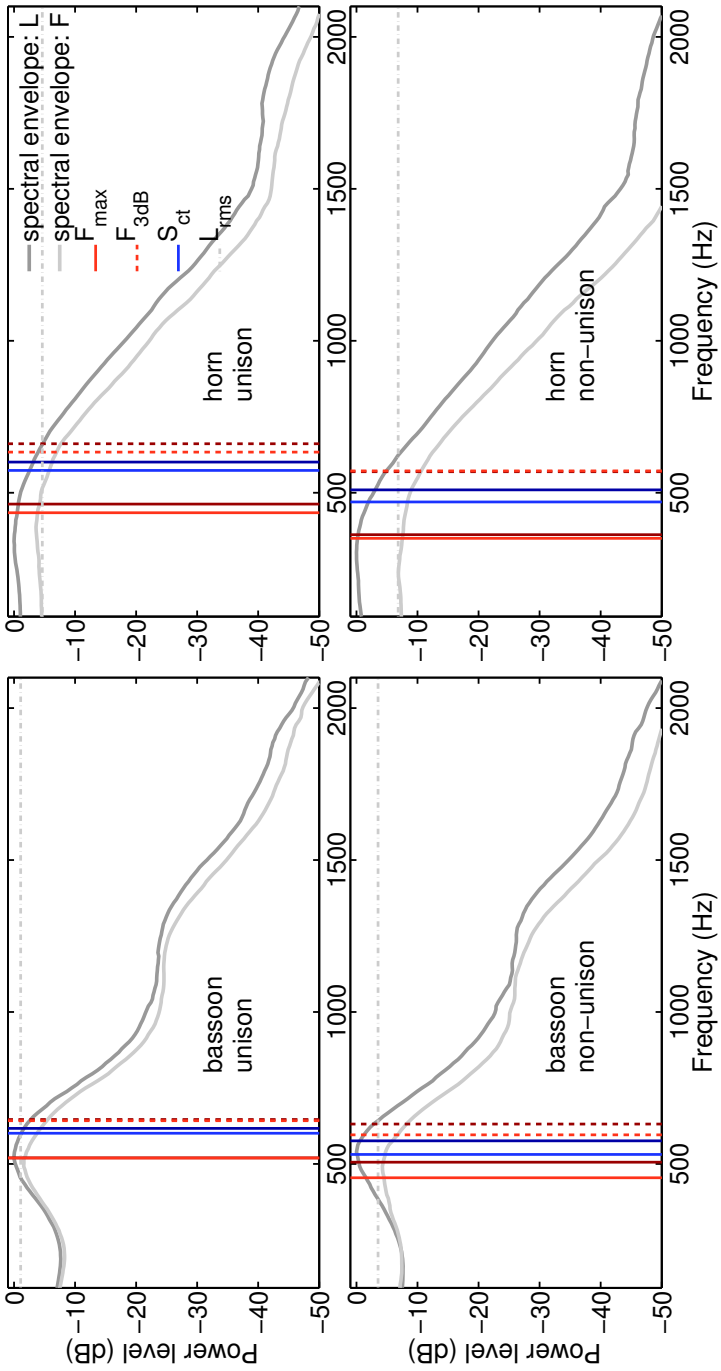

Fig. 4.7 Spectrum and level variations as a function of performer role, unison vs. non-unison, and instrument. Followers (F, shaded lighter) exhibit lower spectral frequencies (F_{max}, F_{3dB}, S_{ct}) and dynamic level (L_{rms}) relative to leaders (L, shaded darker).

4.4 Discussion

When two musicians aim to achieve a blended timbre during performance they coordinate their playing in a certain way. Both performers aim for the idealized timbre the musical score conveys, which usually also implies the instrument that should lead in performance. The leading musician determines timing, intonation, and phrasing, providing reference cues that accompanying musicians closely follow, who likely also adjust their timbre to ensure blend. The employed strategies of performer coordination may or may not be influenced by whether they are playing in unison or non-unison, whether they perform in different venues, or whether the leading instrument is unable to hear the other musician. These factors were studied for pairs of bassoon and horn players, focusing on the timbral adjustments they employed. Performances were evaluated over their time courses through a set of acoustical measures, complemented by self-assessment from the performers, delivering a differentiated picture how performers adjust timbre in achieving blend.

Measuring timbre adjustments as they occur in the realistic setting of musical performance yields a high degree of complexity. These adjustments were evaluated through spectral features, which, however, seem inseparable from covariation with pitch and dynamics. These covariates are what a musical score essentially communicates to performers and although timbre is implied through instrumentation and articulation markings, in reality it occurs more as a byproduct of notated pitches and dynamics. The covariates lastly also determine how performers excite their instruments' acoustic system, in turn, establishing inherent links on the resulting spectral properties. Although correlation analyses on their own do not prove causal relationships, the inherent coupling of pitch, dynamics, and spectral properties in wind instruments has been established physically (Benade, 1976) and should, hence, allow their association to be justified. Correlations between spectral measures and the covariates f_0 and L_{rms} are visualized in Figure 4.8. As individual differences across performers and their instruments are to be expected, the evaluation considers correlations across all performances of individual players and then summarizes these as medians and interquartile ranges for bassoon and horn separately. The impact of pitch variation between excerpts is strong, reflected in fairly high positive correlations, $r \approx .75$, between f_0 and all spectral measures. This applies to both instruments, with S_{ct} being most affected and, moreover, there being little variability among players. Dynamics also exerts an influence on the spectral measures, however, to a different degree across instruments.

Half of the bassoonists show moderate positive correlations with L_{rms}, $r \approx .40$, with there being pronounced variance for the other half, while hornists exhibit higher correlations, $r \approx .60$, and less variability among players. In addition, there is also a clear correlation between pitch and dynamics, which differs between instruments in magnitude and variance, similar to the differences obtained between L_{rms} and the spectral measures. In summary, pitch induces substantial spectral change and due to it being dictated by musical notation, these changes lie beyond performers' control. Although both instruments also show tendencies for increases in dynamics to be associated with increases in spectral frequencies, the covariation is greater for horns; not all investigated bassoons exhibit clearly positive correlation with L_{rms}. Regardless of these differences, dynamics afford performers of both instruments greater liberty in timbral control. Subtle changes in dynamics that fall within the notated dynamic markings could thus be used for slight timbre adjustments and may be more easily achieved than adjustments independent of both dynamics and pitch. Experienced orchestrators likely have internalized the inherent links between pitch, dynamics, and timbral properties in their instrumentation knowledge (e.g., *pitch-driven dynamics*), whereas the current findings argue that research on timbre perception aiming to situate it in musical practice should abandon its definition as that residual quality alongside pitch and dynamics, instead accepting the notion of it being closely entwined with the other musical parameters.

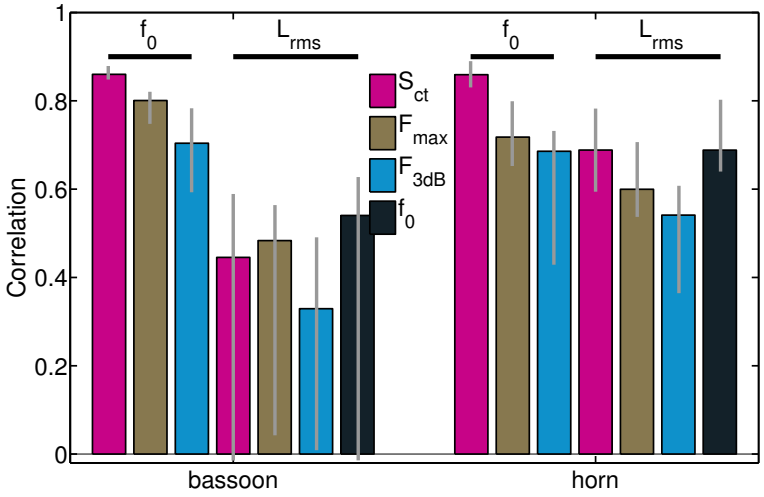

Fig. 4.8 Covariation introduced by pitch (f_0) and dynamics (L_{rms}) per instrument. Median and interquartile range for bassoon or horn of players' within-participants correlations across all factor cells and repetitions (32×2).

Assigning roles between performers yields the clearest effects for timbral adjustments related to blend. Players acting as leaders indeed function as a reference toward which followers orient their playing. In order to achieve blend, followers adjust towards *darker* timbres compared to when they perform as leaders. For both instruments, the darker timbre corresponds to shifts of S_{ct} by about 30 Hz towards lower frequencies, whereas, only for the horn, the main formant shifts as well. At the same time, a *darker* timbre occurs together with *softer* dynamics, which suggests that performers may partially achieve the timbre change through subtle changes in dynamics in addition to potential changes in embouchure or the position of the right hand in the bell of the horn. The extent to which spectral change is employed varies between instruments, with the horn clearly producing more change and also known to be the timbrally more versatile instrument. Due to the nature of the within-participants design, these role comparisons consider how the same musicians perform differently as followers than as leaders, i.e., they do not assess how bassoonist followers darken their timbre relative to hornist leaders and vice versa. At the least, Figure 4.7 suggests that as followers, hornists lower their F_{3dB} to be about the same as that of the bassoonists. With regard to the magnitude of changes in dynamics, differences in L_{rms} (e.g., 1-3 dB) are not as pronounced as signifying a departure from the notated dynamic marking *piano*. From interviewing players of both instruments, musicians appear to consciously consider adjustments of both dynamics and timbre as strategies to achieve blend. For instance, in accompanying a leading instrument, a hornist described his goal as achieving a "rounder" or less brilliant timbre, at the same time reporting that playing with woodwinds, he would need to avoid 'overpowering' the other instrument in dynamics. Likewise, a bassoonist reports the importance of loudness balance to blend, also clarifying that to her, dynamics and timbre are not independent.

Unison performances were indeed perceived as yielding higher blend than their non-unison counterparts. However, the employed range of the rating scale does not imply a large difference, which can be explained in a number of ways: Listening experiments conducted in the past obtained clearer differences in blend ratings between unison and non-unison. In the current experiment, however, participants provided ratings alongside the more demanding performance task, with the ratings also being well separated in time, which did not allow immediate comparisons of unison vs. non-unison performances. Furthermore, performers were asked to use the rating scale based on their previous musical experience, i.e., judging performances and blend relative to what they had learned was

achievable in musical practice. In addition, blending could have also been understood as how 'coupled' the musicians' performance was, i.e., related to additional factors such as synchrony and intonation. Lastly, the musicians' own playing could have partially masked their perception of the other player (e.g., hearing their instrument in greater proximity and via bone conductance), which does not compare to conventional listening experiments, where participants are presented a comparatively balanced representation of two instruments. Together, these factors could have led to the less pronounced rating differences between the voice conditions. Nonetheless, higher blend may still relate to unison performances influencing player coordination more critically. In unison, the performance ratings suggest that followers gave higher ratings than did leaders, which could imply that leaders were generally less satisfied with their performance, given their more important role and responsibility for its success. By contrast, non-unison performances yield higher ratings by leaders than followers, which could be related to excerpt C, located in a low register, having led to some noticeable playing difficulty with a few players. While communication directivity does not appear to affect performances as measured acoustically, the only time it does become relevant concerns unison performances, as impaired communication was judged to be detrimental to musicians' performance. In a similar way, although room-acoustical effects related to blend cannot be deduced from the acoustical measures, some indications are obtained in the performance ratings, which suggest that the smaller, less reverberant room yielded clearer differences between ratings based on performer role. These effects also show that performer coordination between instruments is more critical in the room exhibiting less reverberation, which may allow more subtle differences to become audible. Indeed, temporal coordination for one spectral measure is found to be higher for unison and the smaller room, although this remains limited to ϵS_{ct} and bassoons.

Several indications suggest that musicians improved their coordination throughout a performance. The temporal coordination for both instruments improved in the later phrase for both dynamics (L_{rms}) and global spectral change (ϵS_{ct}) by up to 5% and 8%, respectively. It should be noted that while medians computed on XC across both measures and instruments are comparable, $r \approx .24$, they indicate a fairly weak positive correlation, which suggests that timbre-related performer coordination does not operate at a fine time resolution, but only appears to apply to larger time segments, such as first vs. second phrase. Furthermore, the assessment of temporal change suggests that even leaders adjust their timbre. For instance, regardless of assigned role, horn players slightly reduced their

main-formant frequencies and dynamic level in the second phrase. Although these changes are of considerably smaller magnitude than the ones between performer roles (e.g., 5 vs. 30 Hz), performer coordination appears to motivate adjustments by both musicians to a limited degree. Overall, this confirms that performer coordination adapts over time, ideally leading to an improvement, and that the reference function of leaders still allows for a certain degree of bilateral adjustment between performers. As there is no indication for performer coordination to be modulated by both communication impairment or performance venue among the acoustical measures, the strategies musicians employ in achieving blend appear to be fairly robust to acoustical factors.

This investigation represents a case study by featuring two instruments that commonly form a blended timbre in the orchestral literature. Given the high timbral similarity between bassoon and horn, an effect for performer roles was obtained across both instruments, i.e., regardless of which was leading in performance, whereas obtaining a role-based effect would become less likely when there are starker differences between instrument timbres. In the latter scenario, the more dominant timbre would seem predisposed to assume the lead and serve as the reference, into which the other instrument would either succeed or fail to blend. This case concerns what Sandell (1995) referred to as the *augmented timbre*, in which a dominant instrument is timbrally enriched by another instrument. With this case being a common goal in orchestration, its success depends on the ability of the other instrument to blend into the context defined by the reference. Either its spectral envelope lacks any prominent features that would otherwise 'challenge' the dominant instrument or it bears a sufficiently high resemblance to the latter. In the current investigation, both instrument timbres are similar, yet, its greater timbral versatility allows the horn to blend into a bassoon sound (see Figure 4.7), whereas the bassoon would in return not succeed in adjusting towards a more brilliant or 'brassy' timbre. This imbalance in timbral adjustments, paired with instrument-specific issues related to the playability of excerpts, could explain the obtained differences in performance ratings between instruments. For example, hornists generally gave higher ratings of their performances as leaders than as followers, which could be linked to the greater ease of playing in their default timbre as leaders, as opposed to having to adjust to a substantially darker timbre as followers. This implies that even in this common pairing, the horn may generally assume the more dominant role over bassoons, which also becomes apparent in the orchestral repertoire. Their combination in unison in fact is less common, likely explained by their high similarity not adding

much timbral enrichment, whereas their combination in non-unison is widespread. In the latter cases, bassoons are often found substituting in for missing horns, because up to the mid-nineteenth century, orchestras generally only included two horns, with the addition of bassoons overcoming this limitation, as is also the case in the investigated orchestral passage by Mendelssohn-Bartholdy. In practice, bassoonists more often find themselves blending into the horn timbre than vice versa. Despite the various scenarios concerning instrument combinations as well as dominance or role relationships, a common rule of blending seems to apply to all: The accompanying instrument darkens its timbre in order to avoid 'outshining' the leading, dominant instrument. In other words, when an accompanying instrument blends into the leading instrument, it adopts a strategy of remaining subdued and low-key, very similar to how it subordinates itself to the lead instrument's cues for intonation, timing, and phrasing.

4.4.1 Conclusion

The current investigation showcases how the orchestration goal of achieving blended timbres is mediated by factors related to musical performance. For instrument combinations exhibiting similar timbres (e.g., bassoon and horn), the assignment of performer roles determines which instrument serves as a reference toward which accompanying musicians adapt their timbre to be darker. In an arbitrary combination of instruments, a possible dominance of one timbre likely biases that instrument toward assuming the reference and leading role, requiring that another instrument be able to blend in, otherwise resulting in a heterogeneous timbre. With respect to previous research on musical performance, the current findings illustrate a case in which performer coordination, as related to concepts like *joint action* and *leadership*, directly applies to performers' control of timbre. Achieving a blended timbre requires coordinated action in which an orchestrator's intention becomes the common aim of two or more performers, involving strategies based on relative performer roles that ensure the idealized goal is realized. Standing in the limelight of performance, leading musicians assume the responsibility over the accurate and expressive delivery of musical ideas, whereas the accompanist's primary concern is to blend in, and if successful, remain somewhat obscured in the lead instrument's timbral shadow.

Chapter 5

Conclusion

The perceptual phenomenon of timbre blending was investigated in three independent studies. With the common aim of establishing relationships between acoustical factors and the perception of blend, the investigations comprised a scope that would allow them to be of utility to musical practice, i.e., considering orchestration during its conception and its realization. Chapters 2 and 3 are dedicated to the investigation of acoustical factors linked to blend for dyadic and triadic instrument combinations, whereas Chapter 4 ventured into the novel (scientific) terrain of timbre blending during musical performance. Chapter 5 now attempts to situate the findings obtained and individual conclusions into the larger context of musical practice by also tying in relevant knowledge from orchestration practice and taking into account treatises and other relevant literature on orchestration.

In Section 5.1, all factors relevant to blend are summarized and discussed in terms of the implications of the current research. Section 5.2 situates the findings within musical practice by considering how blend is employed in music and orchestration and how the findings for certain instruments and instrument combinations compare to their discussion in orchestration treatises. Finally, Section 5.3 sketches out a blueprint for a general model for blending timbres in music that draws on the dependencies of musical context, the relevant spectral factors characterizing instruments, and the application and disposition of all blend-related factors at different stages of musical practice.

5.1 Factors influencing blend

In Section 1.3.2, potential factors contributing to blend were introduced. At this stage, an attempt is made to define broader categories with respect to which acoustical properties

they concern, namely, *temporal*, *pitch*, and *spectral* relationships, with the latter one being thought to involve the acoustical 'signature traits' of instruments. These factors are discussed with respect to findings from the research reported in Chapters 2, 3, and 4 as well as previous research.

5.1.1 Temporal factors

Synchronous note onsets are considered an important factor contributing to blend. In musical practice, composers provide the basis by specifying to-be-blended timbres to occur simultaneously, whereas musicians ensure their precise execution during performance. In the conducted investigations, Experiments 1 to 4 comprised dyad or triad stimuli that had been adjusted for favorable synchrony previously, presuming that this shifted the focus to other blend-related factors, such as spectral ones. By contrast, Experiment 5 situated participating musicians in the realistic scenario of performance, allowing them to vary across all performance-related factors. Therefore, synchrony should have played a role, which likely also affected the obtained blend ratings (see Section 4.3.1). However, the focus on timbral adjustments and the lack of a sufficiently high degree of temporal resolution prevented any valid evaluation of the effect of onset synchrony, but the potential negative influence of asynchronous note onsets could nonetheless be removed from the time-series data considered for the evaluation of timbral adjustments. Even in the absence of empirical evidence, it is still reasonable to assume onset synchrony to directly relate to fundamental principles of auditory scene analysis (ASA) concerning spectral fusion and synchrony among tones (see Section 1.3.1) and, therefore, it is expected to affect blending between established timbral identities in a similar way.

Apart from synchrony, the relevance of differences in the onset or attack characteristics between the instruments forming dyads or triads are also relevant (Tardieu and McAdams, 2012). In music, these correspond to attack articulations, and composers generally define them in the notation (e.g., *legato*, *tenuto*, *staccato*, *sforzando*), although instrumentalists may add their contribution to still improve or work against blend, given the implied intention. This of course could also be related fundamentally to instrument-specific differences in the onset times. For instance, double-reed instruments (e.g., oboe, bassoon) are known for their short attacks compared to the more gradual onsets of flutes (Reuter, 1995). In the investigations conducted, Experiment 4 provides clear indications that differences in attack

slopes (e.g., *pizzicato* vs. *arco*) assume a dominant role in the obtained blend ratings. At the same time, this experiment considered isolated note contexts, whereas the relevance of temporal properties has been argued to diminish when embedded within musical contexts (see Sections 1.2.2 and 1.3.2), although more impulsive articulations (e.g., *staccato*, *sforzando*) could still be expected to counteract this trend.

5.1.2 Pitch-related factors

Pitch relationships involve a set of musically relevant parameters, which, moreover, may relate to blend in several ways. From the perspective of composers or orchestrators, the musical material governs pitch relationships, fulfilling roles or functions with respect to counterpoint and harmony, e.g., melodies, accompanying voices, chord progressions. These constraints still offer degrees of freedom that can be exploited to affect blend between the timbres involved, such as by the replication of melodic lines through doublings as well as the relative positions among instrument timbres in chords, however, with both options not being specifically investigated in the conducted experiments. For instance, doublings could achieve greater blend by being spaced along the harmonic series with respect to their individual pitches, which, for a serious investigation, would have required preferably more than three voices as well as a musical context. On the other hand, the role of the instrument in chords does become relevant to Experiment 4, without, however, being investigated as a dedicated independent variable, i.e., not including all possible combinations of instrument permutations. Regarding musical performance, interviewed musicians have reported the role of intonation as an additional factor, which is seen as important in achieving blend. Experiment 5 precluded the study of intonation, only quantifying it to validate the investigated timbral adjustments as not being related to faults in intonation. Still being of informational value, let it be mentioned that a trend could be observed for greater intonation accuracy for performances in unison compared to non-unison, which directly corresponds to findings that tolerance ranges for musically acceptable intonation are clearly smaller for unison than for non-unison (Rakowski, 1990).

Distinctions between unison and non-unison intervals yield clear effects on blend, confirming previous findings (Kendall and Carterette, 1993). In Experiment 3, unison dyads led to much higher blend ratings than their non-unison counterparts, whereas the same trend in Experiment 5 is clearly less pronounced. The disparity between findings could

again be related to the dyadic stimuli representing either isolated or musical contexts, respectively, with the latter experiment, moreover, involving additional tasks. A realistic estimate for the general impact of unison vs. non-unison distinctions would probably fall somewhere in between the effects observed across experiments, and, furthermore, musical contexts may also mediate the importance of interval distinctions, as blend may be operating at different, independent levels of the musical 'scene' (see Section 5.3.1), i.e., not necessarily interfering with their respective individual aims.

Non-unison intervals can be further distinguished in terms of consonance or dissonance, where dissonant intervals could be thought to lead to less blend (see Section 1.3.2). In musical reality, dissonances are hardly avoidable, as varying degrees of consonance simply result from higher-level musical considerations, not allowing special attention to blend. Experiment 1 investigated the blend for consonant and dissonant non-unison intervals. It did not, however, compare the degree of blend between the two conditions, but instead assessed whether the categories would lead to differences in the spectral configurations associated to blend. The results yield no such indication for the investigated instruments (horn, bassoon, clarinet, trumpet), showing that the observed spectral relationships favorable to blend (see Section 5.1.3) apply to non-unison intervals regardless of their degree of consonance. Furthermore, Experiment 2 also showed no difference for the spectral relationships between non-unison and unison for four of six investigated instruments. All four instruments (horn, bassoon, oboe, trumpet) exhibit pronounced formant structure and could be classified as *pitch-invariant*, because they maintained the *plateau*-shaped blend-rating profile across all interval conditions (see Figures 2.6 and 2.7). By contrast, the two remaining instruments (clarinet, flute) yield much weaker formant traits (see Figure 2.1) and have also been shown to be sensitive to interval differences. In summary, strong formant prominence appears to allow the same rules concerning spectral relationships and blend to apply even to non-unison intervals, whereas its lack results in an increased influence of pitch.

Apart from interval-related factors, pitch height may also affect blend. The extreme registers of instruments, often associated with rather atypical timbres, may be thought to depart from conveying the typical timbral signature that can be assumed to be relatively pitch-invariant in the normal registers. Experiment 1 found the spectral relationships explaining blend at unison intervals to no longer apply to high registers of the three investigated instruments (bassoon, clarinet, trumpet). Smaller differences in pitch than those between registers require a more differentiated discussion. Experiment 2 delivers two

different trends, again, based on the prominence of formant structure. The same four pitch-invariant instruments mentioned earlier lead to blend-rating profiles remaining unaffected across the two or three investigated pitch levels. By contrast, weakly pronounced formant structure again leads to the corresponding instruments exhibiting greater sensitivity to pitch variation. Unlike Experiment 2, which investigated local variations of formant structure, Experiment 3 investigated arbitrary instrument combinations for the same six instruments, where the mixed instrument pairs exhibited larger differences in formant structure and also did not involve an experimental task in which one variable sound formed a dyad with a constant reference. The obtained blend ratings seem to be affected by variation of pitch height, notably, even for combinations involving the pitch-invariant instruments; however, a clear contribution to partial-least-squares-regression (PLSR) models is only found for non-unison intervals (see Section 3.3.1). From the evaluation of correlation matrices for auditory-modeling representations, i.e., *stabilized auditory images* (SAIs, see Section 2.2.2), it can be argued that SAIs begin varying as a function of pitch above D4 (horn, bassoon) or A4 (oboe, trumpet). Whereas the pitches in Experiment 2 were chosen to fall in regions below these boundaries, Experiment 3 employed pitches beginning at C4 and reaching as far up as B♭4, which exceeded the boundaries and thereby could have introduced a pitch influence to all instruments. In summary, pitch height can be shown to affect blend and, more specifically, it may bear more weight on non-unison than unison contexts, possibly due to diverging locations of the resulting partial tones.

5.1.3 Spectral factors

Spectral features serve as dominant factors in both the perception of timbre (see Section 1.2.2) and blending between multiple timbres (see Section 1.3.3). For musical applications, spectral features would furthermore be of importance if they are shown to generalize across extended pitch regions (see Section 2.2 and Appendix B), based on which prominent spectral traits could in fact convey the signature traits of an instrument's timbre. Experienced composers and orchestrators have likely internalized the structurally invariant traits these instruments bear, with it being just as likely for musicians to have acquired implicit knowledge of it through frequent interaction with their instruments' acoustic systems.

After establishing that certain spectral features, such as main formants, can be generalized to wind instruments, the series of experiments yields findings confirming the relevance

of both previously reported spectral factors, namely, the global descriptor *spectral centroid* (Sandell, 1995; Tardieu and McAdams, 2012) and the role of formant structure (Reuter, 1996). Specific relationships along these spectral features are shown to be important, suggesting that spectral similarity as opposed to divergence serves as the general principle governing high degrees of blend. Experiments 1 and 2 investigated local variations of main-formant frequency of a synthesized sound forming a dyad with a constant recorded instrument sound, with the synthesis leading to frequency alignment between both main formants for the zero-deviation case ($\Delta F = 0$). Both experiments revealed two typical response patterns, which grouped the instruments based on pitch-invariant or pitch-variant performance (see Section 5.1.2). For the pitch-invariant instruments (horn, bassoon, oboe, trumpet), similarly high degrees of blend were obtained for $\Delta F \leq 0$, whereas when the synthesized formant exceeded the other ($\Delta F > 0$), blend decreased markedly, which overall resembled the profile of a *plateau* (Figure 2.6, left panel). The remaining two instruments (clarinet, flute) exhibited monotonic increase in blend as ΔF decreased, resembling a *linear* blend profile (Figure 2.6, right panel). As these profiles are defined relative to a constant anchor or reference, i.e., the recorded sound, these findings left unanswered what such a reference would correspond to in practice, although in musical scenarios, a leading instrument or the more dominant timbre (e.g., the one being 'augmented' or 'highlighted') could fulfill this function. Results from Experiment 5 can be interpreted as confirming the first hypothesis, showing that performers leading in performance indeed function as a reference toward which followers adjust their timbre. With regard to spectral adjustments, followers 'darken' their timbre compared to when playing as leaders, which agrees with the rule established relative to ΔF from Experiments 1 and 2. Horn players acting as followers in fact lower their main-formant frequencies to avoid exceeding those of leading bassoon players, as Figure 4.7 suggests. Yet, it also becomes apparent that although hornists are able to vary their formant frequencies, bassoonists are more constrained in their timbral control, achieving mainly decreases in spectral centroid relative to the almost stationary main formants. As a result, more drastic divergences in relative formant positions may not be overcome by performance-related factors alone and, moreover, a combination of spectral descriptors may become necessary in comprehensively explaining blend through acoustical properties.

Experiments 3 and 4 investigated arbitrary combinations of instruments in dyads or triads, respectively, in contrast to the pairing of relatively similar instruments (horn, bassoon)

in Experiment 5, as well as local variations from 'maximum similarity' ($\Delta F = 0$) in Experiments 1 and 2. For one, this addresses the case of no particular instrument being assumed to serve as a reference based on performance-related factors. In addition, the divergence in spectral envelopes and formant structure is extended to the entire frequency range and is not just limited to main formants. Despite a broad list of spectral and spectro-temporal descriptors being considered for the PLSR models, two spectral measures appear most relevant, again corresponding to spectral centroid (S°_{ct}) and the main formants (F_{3dB}). The way in which the descriptor values for the individual sounds forming dyads or triads are best associated seems to vary as a function of interval type. In previous investigations, both the centroid *composite* (sum) for unison intervals and the absolute centroid *difference* for non-unison intervals were found to be most relevant (Sandell, 1995). Experiment 3 supports these trends by showing that the composite (Σ) of both S°_{ct} and F_{3dB} are the most useful spectral regressors for unison, whereas for non-unison, greatest utility seems related to a combination of the composite for S°_{ct} and F_{3dB} and the difference for only the latter. As Experiment 4 dealt with triadic combinations, the role of the instrument taking the intermediate value along a descriptor defined as the *distribution* (Ξ) becomes relevant, which for S°_{ct} serves as the most useful regressor. The regression analysis of Experiment 2 employing multiple linear regression (MLR) leads to similar conclusions, however, showing the relevance of F_{3dB} to be more useful when expressed as equivalent level differences ($\Delta L^{\rightarrow}_{3dB}$) and also contributing clearly more than the absolute difference in spectral centroids ($|\Delta S_{centroid}|$). Notably, across all the regression analyses (Experiments 2, 3, and 4), the spectral descriptors in question can all be formulated from the pitch-generalized spectral-envelope estimates, as opposed to those determined for individual pitches, which shows that these generalized descriptions have great utility in serving as the signature traits of the instruments. However, this point has only been sufficiently investigated for wind instruments to date.

In conclusion, matching spectral features between instruments appears to be a general strategy to achieve blend (see Section 5.3.2). In addition, there are several indications that a relative 'darkening' of timbre is also understood as a general strategy, reflected in orchestrators' choice of instruments and dynamic markings, and, similarly, also adopted in terms of timbral adjustments during musical performance. Darker timbres result from an effective reduction of the global centroid, which can result from a more selective adjustment of spectral features in frequency. The role of dynamics becomes especially relevant,

as spectral envelopes of instruments recede in their extent towards higher frequencies with decreasing dynamic markings (see Appendix B). Orchestrators may therefore opt for softer dynamics, with the orchestral repertoire reflecting this in numerous examples of blended instrument combinations, e.g., the investigated excerpt by Mendelssohn-Bartholdy in Experiment 5 being marked *piano* (see Figure 4.3). At the same time, performers employ the same strategy to minimize the potentially problematic effect of salient spectral features at higher frequencies.

5.1.4 Blend prediction through acoustical factors

At this stage, the successful prediction of perceived blend relying on acoustic measures to describe all relevant factors, which furthermore would apply to any musical scenario, still seems an insurpassable obstacle. As a simplified alternative, predictive models that rely only on the signature traits of instruments could still be of utility to orchestrators, as they would provide an objective tool for the selection of suitable instruments, which would complement any aesthetic considerations influenced by individual preference or general convention. The use of mathematical predictive models, such as those employing linear combinations, requires the knowledge of the relative weights for the combination of factors. Through the use of MLR and PLSR, the relative contributions of factors could be assessed, but these weights may be restricted to the particular stimulus context. In addition, greater complexity in the datasets also exhibited more noise artifacts in modeling the behavioral blend ratings, which reduces the reliability of the identified correlational relationships. The data from Experiment 2 originated from a relatively controlled stimulus set, from which spectral properties alone could be found to explain up to 87% of the variance in MLR. Although the greater diversity of instrument combinations from Experiments 3 and 4 still leads to strong overall regression performance (80-90% explained variance), the individual contribution of spectral features is comparably low, explaining no more than 50% of the variance and showing increased sensitivity to noise artifacts. This indicates a trade-off in the utility of diverse stimulus sets. For one, they allow the effect of multiple factors to be weighed against one another; on the other hand, this has been shown to make the reliability of behavioral measures somewhat problematic for the less dominant factors. As a compromise, this motivates future investigations to adopt meta-analytical approaches, to associate separate analyses at their points of intersection, in the hope of attaining more

generalizable weights for the complete set of factors, all the while controlling for data reliability. Unfortunately, an insufficient number of datasets has been considered to allow such an undertaking to be of value at the current stage.

5.2 Contributions to musical practice

5.2.1 The use of blend in music

The notions of *blend* and *contrast* were initially established as sonic goals in Chapter 1.1, which, moreover, were suggested to fulfill functional roles for larger aims within orchestration. Indeed, composers seem to vary in their aesthetic visions concerning orchestration, but its more elementary function makes blend and contrast apply universally. For instance, orchestration can aim to achieve unity of the orchestral texture, argued to be a matter of selecting instrument combinations that blend into a background (French: *fond*) against which other musical ideas unfold (Koechlin, 1959). Another ideal could aim for transparency, which is achieved through heterogeneity rather than blend, to assist the elucidation of independent musical ideas (Schoenberg, 2006). Up until the end of the nineteenth century, approaches to orchestration were also divided into a German tradition which sought to emphasize the musical *line*, relying on blend to achieve it, and the French school, which emphasized *color*, thus striving for heterogeneity (Mathews, 2004). Being even more specific in its implications, the composer Schnittke (2006) equates his notions of *timbre consonance* or *dissonance* with those of blend and its opposite. Hence, it can be seen that blend and contrast act as elementary techniques serving higher functions or even philosophies with regard to orchestration, furthermore showing some objective basis in its reliance on a common perceptual process. Two of the main perceptual investigations on blend arrive at similar conclusions:

> For most composers, 'obtaining a blend' is rarely the primary objective of selecting instruments for a chord or melodic doubling; more likely, 'blend' is used as a *means* for insuring the success of some other effect. (Sandell, 1991; p. 319)

> [E]ven a cursory examination of orchestral music leads to the conclusion that the degree of blend is a variable being manipulated by the composer according to the demands of the musical context. (Kendall and Carterette, 1993; p. 56)

At the same time, blend is not detached from other musical parameters, as shown by the relevance of the temporal and pitch-related factors, which suggests its inherent dependency on the musical material and context, bearing relevance to both compositional and performance-related factors.

> Le plus ou moins de fondu résulte de la manière dont vous combinez les timbres et, dans une certaine mesure aussi, de la disposition des accords. (Koechlin, 1959; vol. 3, p. 3)

> Unless the principles of good voice-leading, spacing, and doubling are applied in an arrangement, no amount of clever orchestration will produce satisfactory results; and without an understanding of harmonic content and form, intelligent scoring is impossible. In orchestrating, it is of the greatest importance to think in terms of lines rather than in terms of isolated notes. (Kennan and Grantham, 1990; p. 2)

Depending on the nature of the musical material, blend and contrast can in the end be exploited at multiple independent levels. For instance, a non-unison blend could involve a quartet of two horns and two bassoons, whereas it serves as a contrasting backdrop to a unison doubling of two blended melody instruments, thereby fulfilling blend or contrast at multiple layers of the orchestral texture. In another scenario, temporal factors could also be interpreted variably. For example, it is at the liberty of a composer to decide at what time blend is desired. Given an impulsive sound at the outset that would naturally contribute towards contrasting it from the rest, blend could still establish itself throughout the course of a long, sustained note following the initial abrupt attack. Similarly, one could imagine there being more 'vertically' oriented blends, based on synchrony of otherwise heterogeneous sounds, or conversely, more 'horizontally' oriented blends of sounds bearing little synchrony in their occurrences but a high similarity concerning their spectral features. As a result, blend may be exploited in a variety of ways that may not be easily predictable without a 'musical' intelligence. In addition, these factors also vary as a function of performance-related contributions. Unlike the fairly controlled investigation in Experiment 5, in practice, performers 'read' the music by interpreting role assignments from it; they know how to adapt specifically to a particular instrument (e.g., intonation, balance), relying on a substantial degree of expertise in ensemble performance, as well as

hearing and adjusting the result in rehearsal under the guidance of the conductor. These factors are not only limited to the previously addressed factors during performance (e.g., intonation, timbral adjustments, articulation such as note onsets, vibrato), but also premeditated preparations such as optimizing fingerings, reeds, or even instrument choice (e.g., bore width) to achieve blend in a given context.

5.2.2 Orchestration and instrumentation

Orchestrators benefit from knowledge concerning which instruments are particularly useful in blending with others, addressing a notion of individual blendability. At the same time, particular instrument combinations are also known to blend well compared to others. Orchestration and instrumentation form important components in the training of composers aspiring to write orchestral music, which at the university level can involve several course modules to cover the subject in depth. In orchestration treatises and courses, the instrument families are usually discussed separately at first, namely, for strings, woodwinds, brass, and percussion. Students may first be familiarized with the blend relationships among the string sections, which can generally be dealt with quite briefly, as there appears to be general agreement that blend among strings poses little problem (Piston, 1955; Kennan and Grantham, 1990). Even the brass instruments are seen as less problematic to blend, whereas the woodwind section is deemed "the most quarrelsome" group within the orchestra (Adler, 2002; p. 164), requiring careful consideration in attempting to blend their diverse range of timbres. Concerning the timbral signature traits of instruments, these varying degrees of complexity in combining instruments could be based on the less 'problematic' instruments bearing less prominent spectral characteristics than some strongly pronounced formant structure known of wind instruments. It could also be that some instruments are more versatile in their timbral variations, allowing them to blend with more instruments than others. In an attempt to address these questions with regard to orchestration treatises and the orchestral repertoire, the following paragraphs are dedicated to a discussion of the 'quarrelsome' members of the wind quintet (flute, oboe, clarinet, bassoon, horn).

The oboe finds little support for being suited to blend with many other instruments. In selecting the three most blendable woodwind members to form a trio, Koechlin (1959) favors bassoon, clarinet, and flute over the oboe. Reuter (2002) summarizes its verbal descriptions across various instrumentation treatises as being judged 'nasal' and 'pierc-

ing', supported by similar judgments from perceptual evaluations (Kendall and Carterette, 1991), which already implies that its utility in orchestration could be more towards contrast than blend. Not surprisingly, it also happens to exhibit the most prominent formant structure among the woodwinds (see Figure 2.1 and Table A.1), bearing pronounced main and secondary formants, which, moreover, are higher in frequency than the comparably prominent spectral features of other instruments (e.g., horn, bassoon). This spectral uniqueness appears to serve as a valid explanation for why the oboe finds itself among the least blended combinations in both Experiments 3 and 4 as well as in previous studies (see Section 3.4).

With respect to more strongly blending instruments, a distinction has to be made between those that still exhibit a prominent formant structure as opposed to the ones that lack such prominence. The horn is seen as the instrument most capable of blending with other instrument groups (Berlioz and Strauss, 1905), is considered equally at home among the woodwinds (Piston, 1955), and exhibits a wide utility in mediating blend within and across instrument families (Koechlin, 1959). Its formant structure is quite pronounced, but it affords a timbral versatility that allows players a substantial range of control (see Chapter 4). At the same time, the lower frequency location of the main formant (e.g., 500 Hz) could make the horn benefit from an apparent limitation of the auditory system, as modeled by the *auditory image model* (AIM), which favors blend (see Sections 2.6 and 5.3.2). Given a strong similarity in terms of their main formants, the bassoon also bears a similar potential to the horn. In combination with the horn, the bassoon seems to form the prime example of a blended pairing (Rimsky-Korsakov, 1964; Koechlin, 1959; Piston, 1955; Adler, 2002; Reuter, 2002), but it also blends well with other instruments, although its timbral versatility is shown to be clearly more limited than that of the horn (see Chapter 4). Nonetheless, both instruments find themselves among the best blended dyadic combinations of Experiment 3.

The clarinet is attributed a similar potential as the horn by being deemed a good candidate for doublings, due to its unobtrusive timbre (Berlioz and Strauss, 1905). With its pitch-generalized spectral envelope lacking strongly pronounced formant structure, which similarly applies to the flute, this could explain why both instruments easily go unnoticed in instrument combinations (Koechlin, 1959). It is important to note, however, that the clarinet's large timbral versatility strongly depends on the timbral evolution across its distinct registers (e.g., low *chalumeau*, middle, high *clarion*, extreme high). As a result, the versatility or blendability with certain instruments may in fact be specific to one of

these registers, instead of applying to its entire pitch range (Reuter, 2002). In this regard, its lower and middle registers may blend well with horn and bassoon (Koechlin, 1959), whereas in its high register, it can be used to *augment* the dominant timbre of the oboe, as is well illustrated in the example of their unison doubling for the main theme of the first movement of Franz Schubert's *Unfinished* Symphony No. 8 in B minor (measure 13 onwards).

From this discussion, *blend* as discussed in orchestration treatises does seem to address the same perceptual phenomenon, as it finds explicit mention across various treatises (e.g., English: *blend*, French: *fusion* and *fondu*, German: *Verschmelzung*) and, furthermore, agreement is found with respect to instruments and instrument combinations suitable to blend. Moreover, the individual utility of instruments appears to match their predisposition as concerns their spectral characteristics.

5.3 Perceptual model for timbre blend in musical practice

In view of the discussed set of factors relevant to blend, an attempt is made to discuss how they would come into play in a perceptual model for timbre blending. This model aims to situate the perception of blend within realistic scenarios encountered in musical practice, i.e., involving independent stages during the conception and realization of blend. Unlike simpler perceptual phenomena (e.g., loudness, pitch), where models trace the initial physical stimulus across different stages of auditory processing (e.g., peripheral, central), the case of blend is much more complex in nature, because it deals with timbre's multi-dimensionality as well as the various degrees of freedom musical contexts offer, for which a distributed theory seems more appropriate. Also, in the absence of more compelling sensorineural evidence than the model representations generated by AIM, the model will be more conceptual and qualitative. Therefore, the discussion considers a more schematic model, presenting blend as being related to factors of musical context and spectral relationships, before situating all factors within the greater framework of musical practice, i.e., considering orchestration and performance.

5.3.1 Layers within the musical scene

As the musical material or context effectively motivates the usage of blend (or contrast) within orchestration, a perceptual model for blend should also consider its relation to the

musical scene, i.e., falling within the framework of ASA (see Section 1.3.1). Picturing the case of a simplified musical scene, instead of time and frequency being the elementary dimensions for ASA with *tones* prior to fusion, we now deal with higher-level dimensions such as metrical time and pitch applied to already established timbral identities. Figure 5.1 shows the case of a musical scene involving two independent musical 'streams' or layers. The most apparent factor contributing to the distinction between the two layers is the pronounced separation in pitch (Δf_0). The top layer involves a coupled melody (e.g., in parallel thirds) consisting of a regular and relatively dense succession of notes. Given the exact replication of the top voice just a small pitch interval lower, the pitch contour of both voices is identical, as is their rhythmic profile. With regard to blend-related factors, this case exhibits highly pronounced note-onset synchrony (red lines) and involves the Gestalt principle of *common fate* (dashed contour), when the similarities in rhythmic and pitch profiles are taken into account. In contrast, the lower layer resembles a triadic chordal accompaniment, with individual chords being well separated in time, but their individual voices occur in synchrony (red lines) and thus contribute to blend. If the chord progression is further based on parsimonious voice leading, blend may be further strengthened based on the Gestalt principle of *good continuity* (connecting dashed lines).

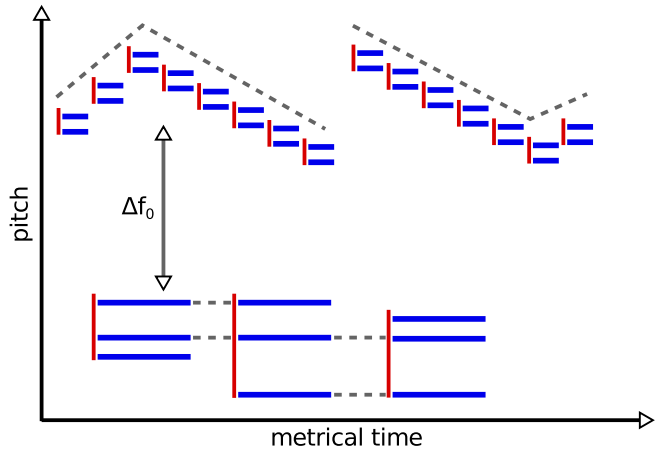

Fig. 5.1 Schematic of independent layers in a musical scene. Red vertical lines mark synchronous note onsets. Dashed lines trace Gestalt principles (*common fate*, top; *good continuity*, bottom).

This musical context already establishes blend-related factors without still taking into account the spectral relationships among individual voices, which provides orchestrators

with additional means of elucidating musical ideas. The matching of such spectral traits can then create additional blend or contrast within and between the two levels. Such conceptual distinctions into independent musical or timbral layers within the orchestral texture are summarized under the term *stratification* in music theory, distinguishable into different hierarchical levels of the texture, such as the *micro*, *meso*, and *macro* levels. In the current example, relationships among voices and instruments within and between levels would operate at the *micro* and *macro* levels, respectively. The important implication taken from this example is that in practice, the expected degree of blend always depends on the musical context, as there are many factors that unfold across pitch and, importantly, also time. Attaining a general perceptual model for blend in these musical cases would therefore almost require as much of a general model for musical scene analysis as one based merely on spectral relationships, i.e., matching instruments according to their timbral signatures.

Musical scenarios should eventually also be discussed with respect to recent trends in ASA research (Micheyl and Oxenham, 2010; Shamma and Micheyl, 2010). The stratification of the orchestral texture would similarly apply to *figure*-and-*ground* relations. Neurocognitive findings suggest auditory figures to lead to neural activation traces, which are modulated by attentional focus (Elhilali et al., 2009), and even exist at the pre-attentive level (Teki et al., 2011), with both cases being relevant to normal scenarios of music listening and production.

5.3.2 A spectral model to blend

The two main viewpoints of past research have argued for either 'darker' timbres leading to more blend (Sandell, 1995) or the necessity of favorable formant relationships (Reuter, 1996). From the discussion of spectral factors, these two viewpoints seem to serve independent contributions in achieving blend: 1) a general strategy of limiting spectra to lower, i.e., 'darker', frequencies, and 2) matching relationships of spectral features in frequency. Both relate to the conception and realization stages of musical practice and, moreover, even exhibit interdependencies, with the role of dynamics also becoming important. Especially with instruments exhibiting strong formant structure, the preference of softer over louder dynamics, i.e., yielding darker timbres, can in fact facilitate the matching of spectral features, as higher-order formants become less pronounced (see Appendix B) and thus less problematic to blend.

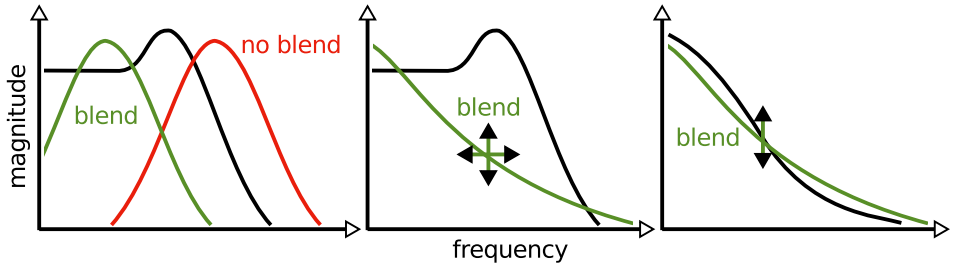

Fig. 5.2 Three blend scenarios as a function of spectral/formant prominence, descending in importance from left to right. Black spectral envelopes serve as the reference. Left: two formants require careful frequency matching. Center: one formant and a less pronounced envelope can lead to blend given amplitude and frequency matching. Right: two less pronounced envelopes yield blend mainly as a function of amplitude matching.

Whereas the general strategy of darkening timbre is conceptually simple, the frequency relationships require a more differentiated discussion. Figure 5.2 summarizes three principal scenarios that could apply across arbitrary instrument combinations and govern blend, arranged in descending order of critical importance to blend. The left panel considers the presence of two prominent *formants*, representing the most problematic case. Assuming the instrument with the more dominant timbre or the one leading in performance serving as a reference (black), the formant of the other instrument has to stay at or below it to ensure blend (green) or otherwise, stand out and contrast itself (red) from the reference. With our investigation only considering the relationships between main formants, it is still assumed that comparably pronounced higher-order formants would require similar treatment to ensure blend. For instance, the oboe exhibits a pronounced secondary formant (see Figure 2.1, $\approx$ 3 kHz), which may serve as one of the reasons for its difficulty in blending with other instruments and why it is used to tune the orchestra, because it can be heard through the din of tens of musicians tuning up. The middle panel considers the intermediate case of the presence of one prominent formant juxtaposed against a spectral-envelope region lacking pronounced formant structure, i.e., exhibiting a quasi-monotonic decrease in magnitude toward higher frequencies. In this case (e.g., oboe with clarinet), blend can be assumed if the latter spectral envelope remains 'unobtrusive', with more blend expected as it recedes further in frequency. The right panel concerns the case of only instruments with less pronounced formant structure (e.g., string section, flute, clarinet), where relative frequency location becomes less important, given the absence of prominent

spectral features that require matching. In this case, a suitable level balance between two instruments may already suffice to achieve high degrees of blend, with level adjustments possibly being of utility to the intermediate case as well.

In addition, there are two mediating factors that could bear an important role on blend. The first was observed through the use of auditory-modeling representations employing the AIM. Identified as the *high-pass characteristic* (see Section 2.6), spectral-envelope regions below 500 Hz appear less important to blend, due to their SAI magnitudes being strongly attenuated. This limitation would actually prove favorable to instruments exhibiting formants at or below that frequency (e.g., main formants of horn, bassoon), with blend becoming harder to achieve for higher reference formants (e.g., oboe's main formant). If in Figure 5.2 (left panel), for instance, the 'red' formant were centered on 500 Hz, one could assume the divergence of spectral envelopes below its maximum to be negligible and thus actually lead to blend. Second, a wider spacing of partial tones with increasing pitch will lead to notches in the SAI profiles, outlining individual partial tones, which could facilitate individual timbral identities permeating through the notches and thus becoming more distinct, especially in non-unison combinations, i.e., when partial tones are likely to diverge in frequency.

In conclusion, it becomes apparent that in the presence of prominent spectral features of one instrument, blend depends critically on whether: 1) other instruments bearing comparably prominent traits will be sufficiently similar, or 2) other instruments bear much less pronounced spectral traits. In other words, given a dominant timbre, blend is achieved when other instruments do not 'challenge' its timbral identity, whereas blend becomes easier to achieve when pairing relatively 'unobtrusive' timbres, with less careful attention in matching required.

5.3.3 Map of blend-related factors in musical practice

In musical practice, blend relates to a conception and a realization phase (as does communicating musical ideas in general: Kendall and Carterette, 1990), with all discussed factors occurring either exclusively in one or the other, or applying to both similarly. Figure 5.3 illustrates a map of the temporal, pitch-related, and spectral factors relative to where they occur, i.e., during conception (left region), realization (right region) or shared between the two stages (intersection of both regions). Orchestrators may conceive musical ideas

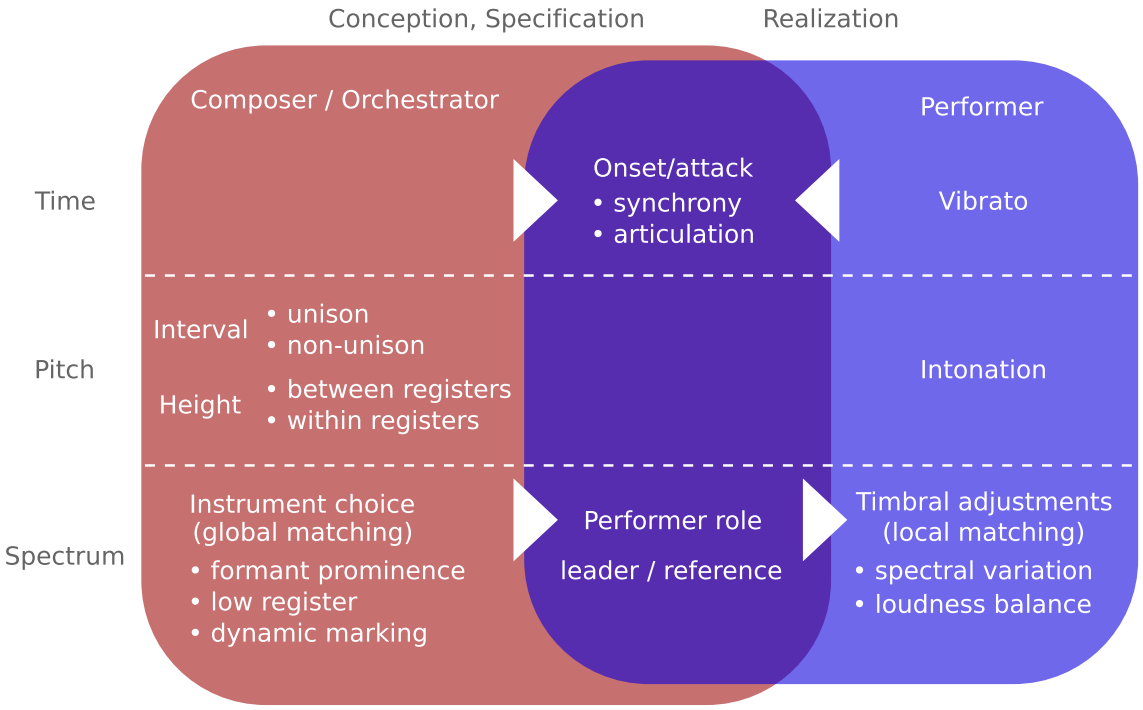

Fig. 5.3 Blend-related factors mapped onto musical practice.

by exploiting blend or contrast, which then becomes a *specification* once laid down in the musical notation. Thereupon, musicians interpret these specifications and furthermore apply blend-related adjustments as a function of the interactive relationship with other performers. With regard to temporal factors (top section), note-onset synchrony and articulation are part of the notation but equally also require precise execution during musical performance and, therefore, apply to each stage independently. By contrast, the usage or avoidance of vibrato has been reported to be based on musicians' expertise in ensemble performance alone, e.g., performing without vibrato when accompanying instruments that do not generally employ it. Pitch-related factors (middle section) group exclusively to one stage or the other, based on the predisposition of notes (e.g., pitch height and interval type) during composition and orchestration and correct intonation during musical performance. Lastly, spectral matching (bottom section) applies on a global and local level, corresponding to the conception and realization stages, respectively. In the former, composers and orchestrators select instruments based on whether they match, also exploiting the potential of low-register instruments, pitch-invariant formant structure, and softer dynamic

markings. On the other side, performers use the versatility of their particular instruments for timbral and dynamics adjustments. As a linking factor, the implied dominance relationships from the notation and instrumentation inform performers of their individual roles on which spectral and dynamics adjustments will be based.

5.3.4 Current limitations and future directions

Concerning the investigation of auditory-model representations, i.e., AIM and its specific reliance on a dynamic, compressive gammachirp (DCGC) basilar-membrane model, the existence and frequency limits of the *high-pass characteristic* (see Section 2.6 and Section 5.3.2) need to be validated and more precisely defined.[1] In addition, AIM should also be further investigated with respect to how wider spacing of partial tones with increasing pitch may introduce notches in the SAI representations, with these expected to be detrimental to blend. These investigations of auditory-modeling representations could provide valuable information for future attempts at blend prediction, and it is hoped that they would reveal potential redundancies in the acoustical description of blend-related factors. In the envisaged meta-analytical approach to attaining more comprehensive and generalizable blend-prediction models (see Section 5.1.4), AIM representations may again be valuable in overcoming potential discrepancies in correlating acoustic with perceptual measures. In addition, studying the evolution of SAIs across pitch and various spectral-envelope types could aid in consolidating previous perceptual blend theories based on 'darker' timbres or formant coincidence. Lastly, perceptual models may also be beneficial to a special scenario of timbral contrasts, i.e., the notion of timbre *salience*, which has only recently received special attention (Chon, 2013).

In Chapter 2, the rule explaining blend as a function of variations in main-formant frequency requires a reference formant, with Chapter 4 showing this reference to correspond to performer roles between musicians in practice. However, this finding, based on a case study of paired bassoon and horn players, will still have to be extended to other instrumental and musical contexts in order to allow greater generalizability. With regard to other factors related to musical performance, the role of room acoustics should not yet be ruled out even if evidence for its relevance is weak. For one, timbral adjustments could not be independently

[1] It should be noted that these observations are specific to the DCGC auditory filterbank and do not manifest themselves using the more basic and dated *gammatone* filters, which do not account for changes in auditory-filter shape as a function of sound level.

associated with changes between rooms in Experiment 5. However, the self-assessment of performers does suggest an influence of room acoustics. One could assume that although timbral adjustments may be less dependent on room acoustical changes because performers may indeed be able to 'extrapolate' the instrument's source timbre from the room-induced coloration (see Section 1.3.2), differences in reverberation could still affect performances to some degree. Overall, the study of timbral adjustments between musicians opens interesting avenues into practical aspects concerning orchestration and instrumentation. Especially for instrumentation, the study of instrument-specific dependencies and interactions between musicians seems to offer a potential for many intriguing research projects. Even beyond the context of timbre blending, instruments bearing a greater timbral variability and affording performers with more timbral control may in turn also expand their expressive capabilities, which may add to the performance-related expressivity achieved from variation in timing and dynamics. The quantification of the extent to which timbral adjustments may be independent from other covariates (e.g., pitch, dynamics, articulation), which should not be assumed as generalizable across instruments, could bring research closer to the ultimate aim of determining what timbre really represents and how it operates with respect to musical practice.

Finally, loudness balance between instruments still needs to be addressed in a more systematic manner as it presents a critical component in the matching of spectral-envelope features. For instance, a much higher sound level of one instrument over another will likely mask the weaker sound altogether, i.e., not really relating to the notion of blend, as only a single timbre is effectively heard. In Experiment 1, loudness balance was determined by participants, while in Experiments 2, 3, and 4, it was predefined as part of the stimulus design. For the latter cases, no indication of a systematic influence of measures quantifying the relative mix ratios between sounds constituting dyads or triads was found in the regression analyses (see Section 2.5.2 and Section 3.3). Yet, the role of loudness balance should still be considered relevant to musical practice, as the notion of *balance* between instruments and instrument sections is considered important, e.g., in setting the orchestral background or *fond* and in achieving *équilibre des sonorités* (Koechlin, 1959), which similarly applies to the perspective of a conductor, who aims to mediate the loudness balance between the orchestral groups. Likewise, musicians also generally acknowledge the importance of loudness balance to blending with other instrumentalists. Therefore, it will be of interest to assess the extent to which loudness balance alone, which essentially involves *global*

spectral-envelope matching, may potentially mediate the influence of strong formant structure (see Section 5.3.2). For instance, Reuter (1996) has argued that blend between formant-based wind instruments and bowed strings is only achieved for sufficiently high levels of the latter. In addition, the work of sound engineers also relies a great deal on achieving blend in the loudspeaker reproduction through level adjustments between individual microphone feeds, although subtle use of frequency equalization may also be employed. In terms of orchestration and its room-acoustical implications, the influence of loudness balance has been investigated and quantified earlier (Burghauser and Špelda, 1971), without, however, specifically addressing perceived blend. Furthermore, loudness is similarly entwined with the notion of dynamics as is timbre or spectral variation (Fabiani and Friberg, 2011), representing yet another set of interdependencies to account for in future investigations of blend in increasingly realistic musical scenarios.

5.4 Concluding remarks

Blend and contrast serve elementary functions in fulfilling sonic goals that are in fact relevant to many potential, higher-order aims of orchestration. With regard to perceptual and cognitive processes involved in music listening, both appear to be important to musical scene analysis where they may mediate ASA processes. They should therefore be considered as central to the long-term aim of establishing a perceptually informed theory of orchestration. Blend's wide usage within orchestration practice also relates to musicianship, with performers having acquired and internalized their own strategies to blend, which show instrument-specific dependencies based on the limitations and affordances of particular instruments or even other factors of performance practice. Research on performance factors aids in understanding what timbre essentially represents in music, with regard to its non-trivial entwinement with other parameters. The challenges involved in measuring timbral adjustments during performances taking place in real-life acoustic environments require interdisciplinary approaches and tools, which I have been able to acquire and contribute to this research and will gladly build on in the future.

Timbre instilled a personal interest in me from an early age. As a native of Berlin, the immersion in concert experiences at its *Philharmonie* crucially affected my musical enculturation as did the diversity of sonic arts this city offers in general. The sounding nature of things had since become a focus of my creative and academic work, with the effects

of timbre perception often generating streams of curiosity, which proved a reliable driving force in helping me overcome the many obstacles encountered during years of doctoral research. Timbre is a challenging and not-seldom puzzling affair, but it also represents the musical parameter most anchored in reality, i.e., as its *sounding* manifestation, which warrants its continued study despite its complex interdependencies.

Appendix A

Estimation and description of spectral envelopes

Section 2.2.1 introduced an estimation method for instruments' pitch-generalized spectral envelopes, allowing the identification of pitch-invariant properties, such as formant structure. These empirically derived estimates are then used to describe the instruments investigated in Chapters 2, 3, and 4. As Chapter 2 does not cover all aspects of this procedure, Appendix A presents the necessary details.

A.1 Empirical, pitch-generalized estimation

Empirical estimations of spectral envelopes yield pitch-generalized descriptions of instruments. They evaluate composite distributions of partial-tone spectra, ideally compiled across each instrument's entire pitch range. As spectral envelopes exhibit substantial change across dynamic markings (see Appendix B), separate estimates should be acquired for different dynamic markings.

At first, time-averaged power-density spectra for individual pitches are computed, evaluating only the continuous, steady-state portions of sounds, i.e., excluding their attack and decay phases, as these likely also include broadband transients as opposed to tonal content. For individual spectra, a partial-tone-detection routine determines the frequencies and amplitudes of partials.[1] Across all analyzed pitches, a global, composite partial-tone distribution is compiled, with individual partial-tone spectra normalized to their mean power. Partials above 5 kHz are disregarded, as no occurrence of formants has been reported above that frequency (Reuter, 2002). An example of the composite distribution (black dots) is shown in Figure A.1 for the contrabass trombone.

[1] The MATLAB function `PeakPicker.m` by Bertrand Scherrer was used.

Spectral envelopes are obtained by searching for a general trend within the composite distribution, with a comparable approach having applied a sliding moving-average filter window across the frequency dimension (Luce and Clark, 1967). In the current case, a curve-fitting procedure is applied based on a *cubic-smoothing-spline* algorithm in MATLAB[2]. A smoothing parameter p controls the relative weight between the contrary aims of obtaining a detailed spline fit to the data and a smooth linear trend. In order to obtain an objective indicator for determining the ideal smoothing coefficient p, a measure of fit F between the power levels of partial tones $L_{partials}$ and the spectral envelope L_{cfit} of length K was derived. As shown in Equations A.1 to A.3, the fit measure F assesses the 'smoothness' of the measure D across frequency indices k. $D[k]$ quantifies power-level deviations between the estimated spectral envelope at frequency $f_{cfit}[k]$ and partial frequencies $f_{partials}[i]$ located within an octave bandwidth of the former. The curve-fitting procedure is conducted across a set of logarithmically scaled, decreasing values of p, with the optimal p coefficient being attained when the improvement in F for successive values of p falls below 1%.

$$F = \sum_{k=1}^{K-1} \left(D[k+1] - D[k] \right) \tag{A.1}$$

$$D[k] = \sum_{i[k]} \left(L_{partials}[i] - L_{cfit}[k] \right) \tag{A.2}$$

$$i[k] = \text{all } i \text{ located within } f_{cfit}[k] \cdot 2^{-1/2} \leq f_{partials}[i] < f_{cfit}[k] \cdot 2^{+1/2} \tag{A.3}$$

Figure A.1 provides an example of a spectral envelope estimated from a composite partial-tone distribution (black dots), based on spectral analyses of a contrabass trombone. The analyzed audio samples, taken from the Vienna Symphonic Library (VSL), were recorded at *mezzoforte* across 37 pitches, in the range from E♭1 to E♭4. The optimal smoothing coefficient is $p = 2{\cdot}10^{-7}$. A single, main formant can be identified, characterizing the instrument's spectral-envelope shape at around 500 Hz.

[2] `csaps.m`-function, MATLAB Spline Toolbox

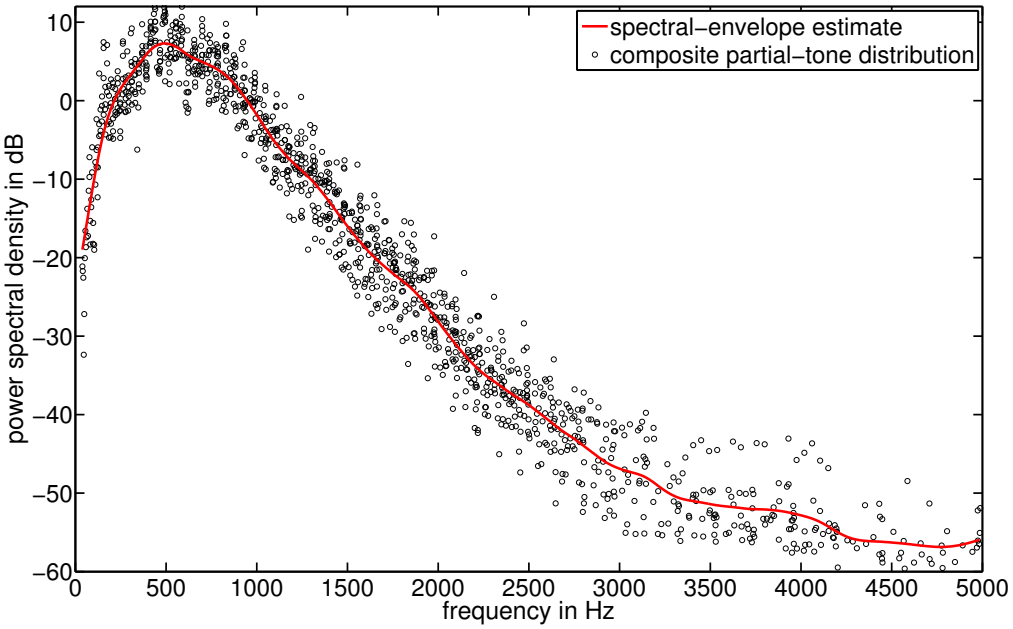

Fig. A.1 Estimated pitch-generalized spectral envelope for contrabass trombone based on a composite distribution of partial tones across 37 pitches.

A.2 Description of formant structure

Spectral-envelope estimates serve as the basis for the identification and description of prominent features, such as formants, with respect to their frequency location and extent, and relative contribution to the remaining spectral-envelope regions. Whereas during the initial work on spectral-envelope description and during the design of the spectral-envelope synthesis filters (see Appendix D) formants were identified and characterized qualitatively, the long-term aim involved the development of reliable algorithms for automated formant identification and description, which could be applied to pitch-generalized spectral-envelope estimates (Chapters 2 and 3) as well as time series of spectral-envelope/TE estimates (Chapter 4). Furthermore, the algorithm also supplies formant descriptors to be tested in prediction models for timbre blending.

A.2.1 Identification and classification of formants

Formants are classified as features that prominently characterize a given spectral envelope. *Prominent* in this regard more specifically addresses a sufficiently large extent in the

frequency dimension, a significant difference in magnitude to adjoining spectral-envelope regions, and/or a clear global contribution to the spectral-envelope magnitude. The chosen approach of formant identification and classification evaluates the spectral-envelope magnitude function and associates it to its corresponding frequencies. The algorithm involves the following steps:

1. **Numerical derivatives** of the spectral-envelope magnitude function are computed in order to evaluate spectral shape characteristics (see lower panel in Figure A.2), with the first derivative representing its slope. In addition, the second and third derivatives aid in identifying and distinguishing different kinds of formant classes.

2. **Local spectral maxima** are identified at frequencies yielding zero-crossings in the first derivative for which the second derivative also yields negative values.

3. **Spectral plateaus** are locations along the spectral envelope clearly approaching but not attaining zero-slope. These are identified as local maxima in the first derivative, confirmed additionally by occurring together with zero-crossings and negative polarities in the second and third derivatives, respectively. As shown in the top panel of Figure A.2, a plateau is identified slightly above 1000 Hz (grey solid line), whereas the point identified just below that frequency is a local maximum.

4. **Classification into formants** considers all identified local maxima and plateaus, which are compared in their magnitudes to the global spectral-envelope maximum, with those falling beyond the range of 50 dB below the latter being discarded. Among the remaining maxima and plateaus, ones exhibiting sufficiently high proximity in frequency and/or magnitude are grouped, with the most prominent of that group being assigned as the formant maximum. Finally, only the three most prominent formants are retained, denoted *main*, *secondary*, and *tertiary* formants, with the first being lowest in frequency and for all investigated (real-life) cases also strongest in magnitude. In Figure A.2, the plateau and maximum (solid grey lines) mentioned earlier are grouped together with the classified formant maximum (solid red line) located just above 500 Hz.

5. **Global maxima** of spectral envelopes can either be represented by the identified main formants or not. In the latter case (e.g., flute, clarinet), the main formant

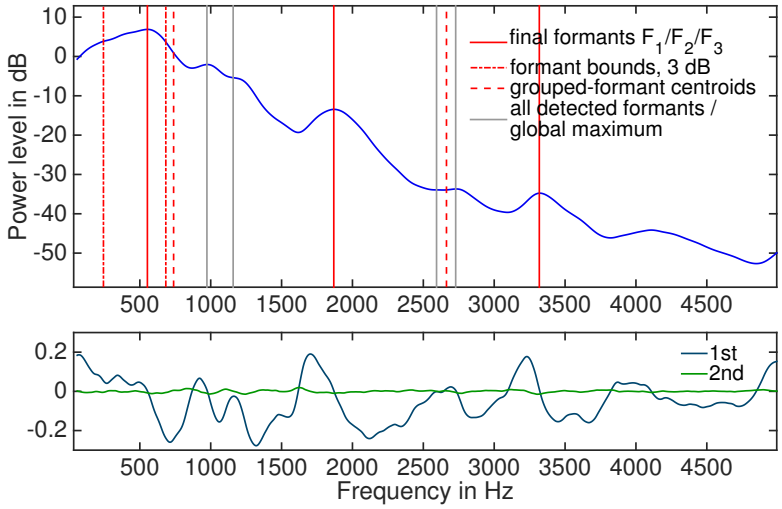

Fig. A.2 Output from the spectral-envelope description algorithm for an empirical spectral-envelope estimate of bassoon at a *mezzoforte* dynamic. Top panel: spectral-envelope description; bottom panel: derivatives.

assumes a less prominent role in the entire spectral envelope and, accordingly, the description acknowledges the global spectral maximum as the most prominent feature. As the latter in most cases lies at the lower frequency extreme, it is not considered a formant.

A.2.2 Characterization of classified formants

Individual formants can be characterized by several descriptors with respect to frequency location and extent. F_{max} denotes the frequency yielding the maximum magnitude. Two pairs of frequencies characterize *upper* and and *lower* bounds at which the power magnitude has decreased by either 3 dB or 6 dB relative to F_{max}, i.e., $F_{3dB}^{\rightarrow}$ or $F_{6dB}^{\rightarrow}$ and $F_{3dB}^{\leftarrow}$ or $F_{6dB}^{\leftarrow}$, respectively. For less pronounced maxima such as spectral plateaus, lower bounds are often not available. In addition, as formant shapes are seldom symmetric, the central tendency for formant-frequency location can alternatively also be expressed through the arithmetic or geometric means of the frequency bounds, provided that both frequency bounds are available.

A.2.3 Characterization of relationships among formants

The power-level difference $F_{\Delta mag-1|2}$ between the main and secondary formants serves as a measure of relative prominence of the main formant to the remaining formant(s). As secondary formants are not always available for certain instruments, another more general expression, $F_{\Delta mag}$, considers the relationship between the main formant and the remaining spectral-envelope regions. The power level for the remaining spectral-envelope regions are derived from the average computed on all inverse-logarithmic magnitude values above $F^{\rightarrow}_{6dB}$. Another descriptor for the relative contribution of the main formant to the global spectrum, F_{sl}, computes the *spectral slope* (Peeters et al., 2011) selectively for the spectral-envelope region above $F^{\rightarrow}_{3dB}$.

A.2.4 Formant prominence

Formant prominence assesses the extent to which the identified formant structure assumes a pronounced role in characterizing the global spectral envelope. It currently is based on a cumulative score F_{promi}, evaluating features that contribute to formant prominence. More specifically, the main and secondary formants are quantified with respect to their frequency extent as well as the degree of protrusion in the formant magnitude (e.g., presence of upper and lower bounds). Table A.1 lists formant-prominence scores for the six wind instruments whose spectral-envelope estimates are presented in Figure 2.1. For instance, the scores reflect the tendency of the oboe to exhibit a much more prominent formant structure than that of the flute, providing a potential utility to predict dominance relationships between instrumental timbres (see Section 3.2.3).

Instrument	oboe	horn	bassoon	trumpet	flute	clarinet
Score	13.7	10.2	9.1	6.9	5.7	5.6

Table A.1 Formant-prominence scores for six wind instruments based on spectral-envelope estimates for *mezzoforte* dynamic marking. Compare to Figure 2.1.

Appendix B

Spectral envelopes of orchestral instruments across dynamic markings

Pitch-generalized spectral-envelope estimates appear to convey the timbral signature traits that are relevant to the blending of instrument sounds. The estimates should be obtained for separate dynamic markings, as spectral properties can be expected to vary as a function of dynamics (see Sections 1.3.2 and 4.4). Appendix B aims to quantify the extent to which spectral envelopes vary as a function of dynamics and, furthermore, serve as a source of reference for the main instruments of the orchestra, covering the principal three sections woodwinds, brass, and strings.

All instrument descriptions are based on audio samples taken from the Vienna Symphonic Library (VSL), processed and analyzed in the same way as discussed in Section 2.2 and Appendix A. For all instruments, the comparison considers three dynamic markings, which, given their availability, correspond to *forte* (f), *mezzoforte* (mf), and *piano* (p) or, if not available, dynamic markings that most closely approximate them. For the wind instruments, formant characterization (see Section A.2) is employed to identify the maxima of main (F_1) and secondary (F_2) formants, as well as the 3 dB upper bound for the former ($F_1^{\rightarrow}$). In the visualizations, the indicated arrows trace the evolution of the main-formant descriptors from the softest to the loudest dynamic marking, whereas F_2 indicates its approximate location across dynamic markings, e.g., illustrating how secondary formants become more pronounced with increasing dynamic marking. For all instruments, the visualized spectral-envelope estimates can be taken as a reference in estimating the perceived degree of blend from spectral relationships that take the prominence of spectral traits into account (see Section 5.3.2). Overall, a tendency of flattening spectral slope

with increasing dynamic markings becomes apparent and, therefore, it appears that softer dynamics may indeed be less problematic to blend, as 1) they reduce the global spectral extent, and 2) they attenuate the presence of pronounced spectral features.

B.1 Woodwinds

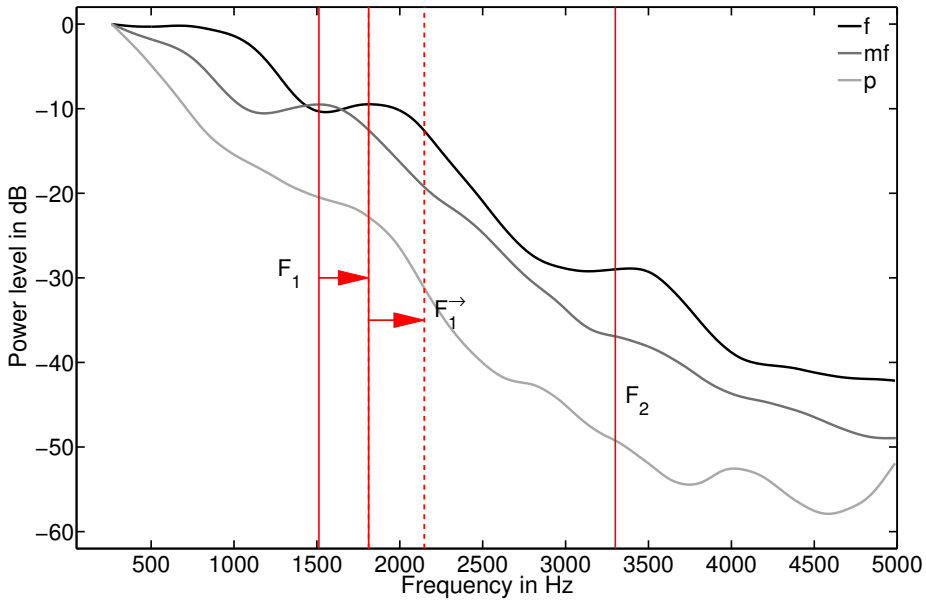

Fig. B.1 Spectral-envelope estimates for flute across dynamic markings *forte*, *mezzoforte*, and *piano*.

Chapter 2 established a distinction among woodwind instruments into those exhibiting less pronounced formant structure (flute, clarinet) and those whose spectral envelopes are clearly dominated by formants (oboe, bassoon). As illustrated for flute in Figure B.1 and clarinet in Figure B.2, all spectral envelopes do not exhibit strong formant structure and present rather 'unobtrusive' shapes. However, the identified formants do become more pronounced as the dynamic marking increases, e.g., they evolve from spectral *plateaus* into *maxima* (see Section 2.2.1). As main formants become more pronounced, their locations also shift slightly toward higher frequencies.

Figures B.3 and B.4 show the spectral envelopes for the double-reed instruments, i.e., oboe and bassoon, respectively. Discussed as the least blendable member of the woodwinds

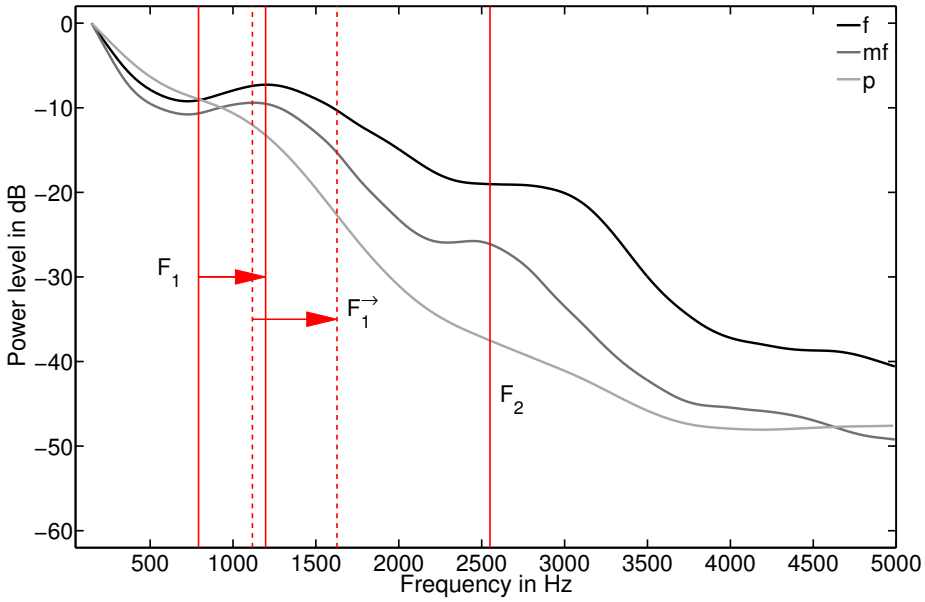

Fig. B.2 Spectral-envelope estimates for B♭ clarinet across dynamic markings *forte*, *mezzoforte*, and *piano*.

(see Section 5.2.2), which also exhibits the highest degree of formant prominence (see Section A.2.4), the spectral envelopes of the oboe yield pronounced main and secondary formants at all dynamic markings. Still, with increasing dynamics, the secondary formant becomes more pronounced in magnitude and frequency extent, whereas the main-formant maximum F_1 shifts toward higher frequencies, while its upper bound exhibits a similar but weaker shift. The low-register instrument bassoon similarly exhibits two pronounced formants, which barely change in position across dynamic markings. However, for dynamic markings beyond *piano*, the spectral slope above the main-formant upper bound $F_1^{\rightarrow}$ flattens and also the secondary formant gains substantially in magnitude.

B.2 Brass

As shown in Figure B.5, the horn's main formant gains some frequency extent with increasing dynamic marking, but it is the secondary formant that becomes substantially more pronounced. The degree of spectral variation across dynamics may be linked to the known timbral versatility of the horn. For instance, its status as an unofficial member

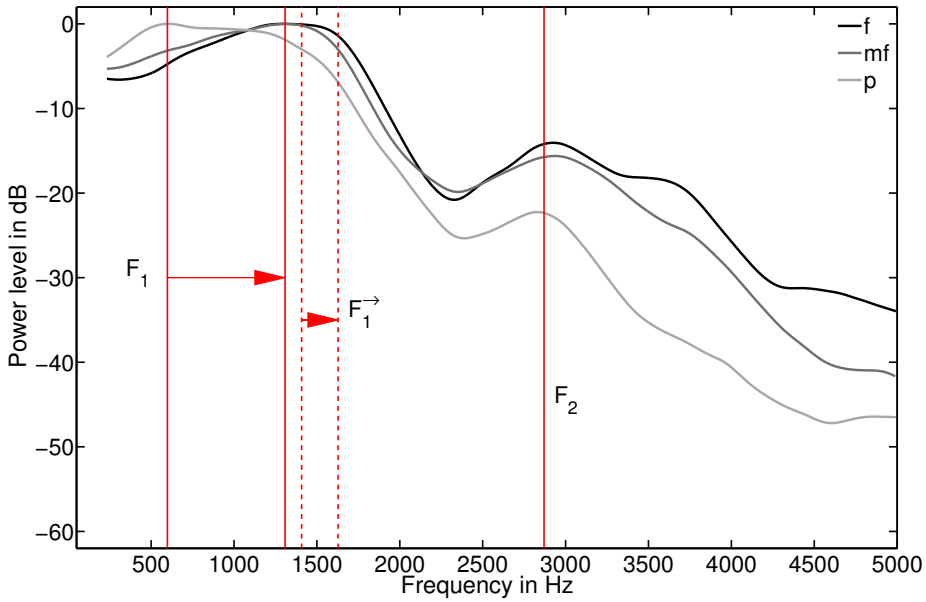

Fig. B.3 Spectral-envelope estimates for oboe across dynamic markings *forte*, *mezzoforte*, and *piano*.

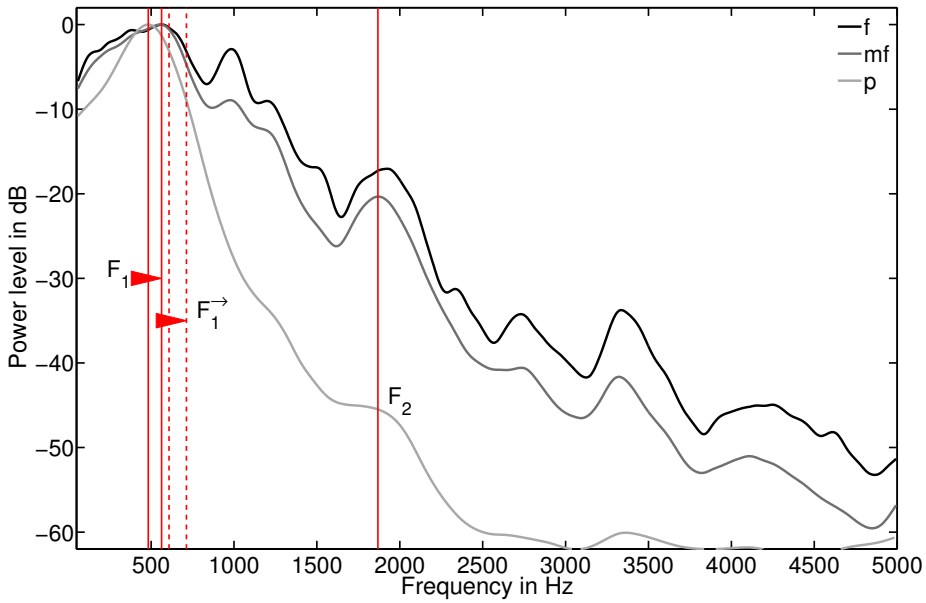

Fig. B.4 Spectral-envelope estimates for bassoon across dynamic markings *forte*, *mezzoforte*, and *piano*.

of the woodwinds (see Section 5.2.2) may be achieved with its steeper and lower spectral extent at lower dynamics, whereas at higher dynamics it resembles other brass instruments more. Furthermore, it bears a striking resemblance to the spectral envelopes of the bassoon (Figure B.4) at all dynamic markings, which reflects their frequent use as a blended pairing.

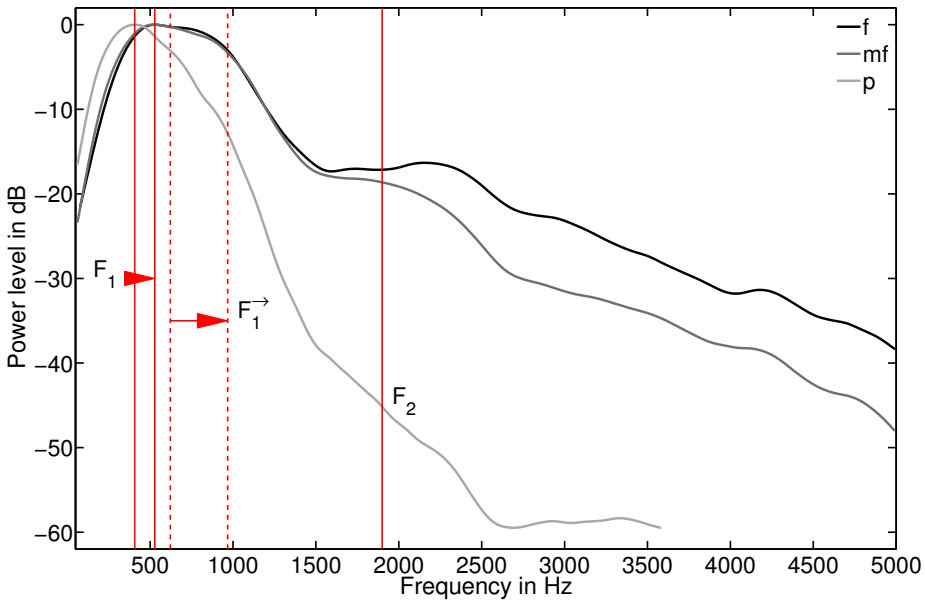

Fig. B.5 Spectral-envelope estimates for (French) horn across dynamic markings *forte*, *mezzoforte*, and *piano*.

Figure B.6 shows that the tenor trombone is similar to the horn's spectral envelope at the lower dynamic markings, whereas at *forte* its main formant expands toward higher frequencies, resembling that of the trumpet (Figure B.7). This feature may allow it to fulfill a role as a 'bridging' instrument within a *forte* brass section. As illustrated in Figure B.7, the broad main-formant region of the trumpet is apparent at all dynamic markings, although it becomes more pronounced as dynamics increase. Likewise, the secondary formant also gains prominence.

B.3 Strings

String instruments from the violin family exhibit spectral maxima that result from resonances stemming from their body plates and the air cavity contained therein, rendering

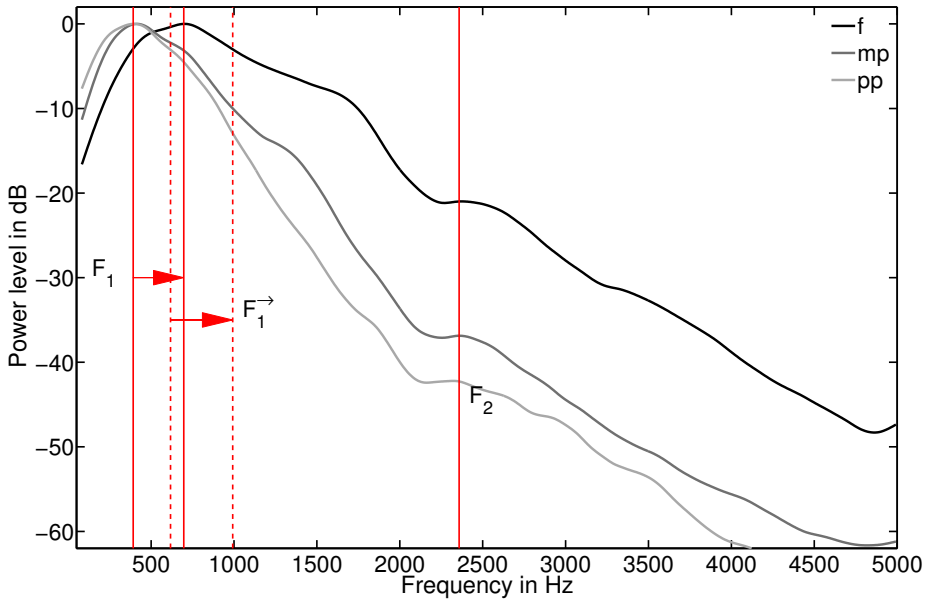

Fig. B.6 Spectral-envelope estimates for tenor trombone across dynamic markings *forte*, *mezzopiano*, and *pianissimo*.

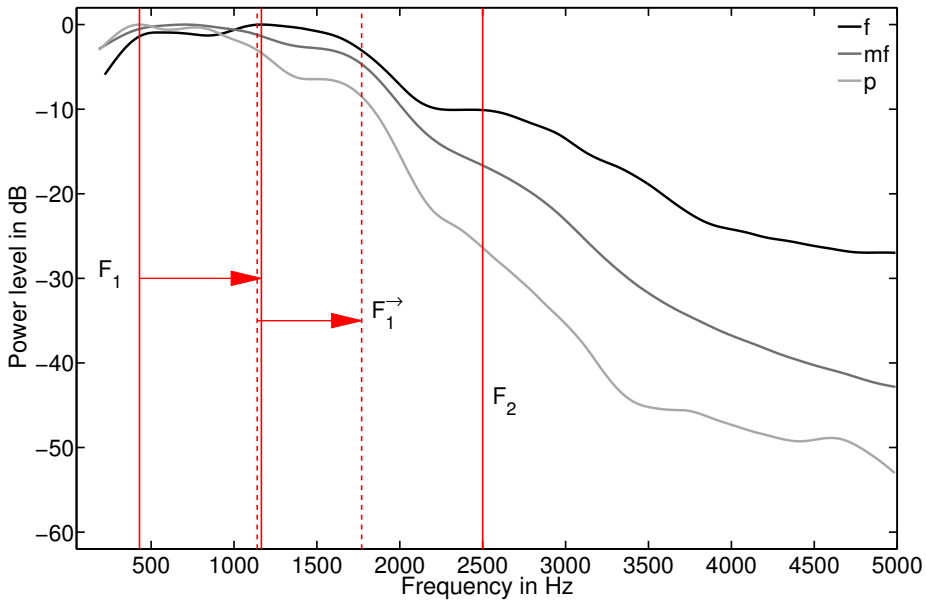

Fig. B.7 Spectral-envelope estimates for C trumpet across dynamic markings *forte*, *mezzoforte*, and *piano*.

their spectra highly individual. However, in orchestras, they are mainly used in a choric way, in which case their individual characteristics can be assumed to average into an aggregate, pitch-generalized spectral envelope. Therefore, the following estimates consider string sections, also involving vibrato, as this represents the most common way strings are played, further complemented by other sources of incoherent variation such as slight deviations in intonation and timing, which together add to the typically 'thicker' sound of string sections. The string instruments considered comprise the violin and violoncello sections, representing high- and low-register instruments, respectively. Figure B.8 displays spectral-envelope estimates for a section of 14 violinists, which bear quite similar spectral slopes across all dynamic markings. With increasing dynamic marking, the spectral envelopes become slightly undulated, which, however, is noticeably less pronounced than even for the weak formant structure of flute (Figure B.1) and clarinet (Figure B.2). The violoncello section, made up of eight players, exhibits slightly steeper spectral slopes, with *forte* dynamics leading to a more pronounced undulation of the spectral envelope than for violins, as shown in Figure B.9. Overall, the string instruments exhibit the least prominent spectral features, arguing for their general utility to form blended combinations amongst themselves (see Section 5.2.2) and also with the other instrument families.

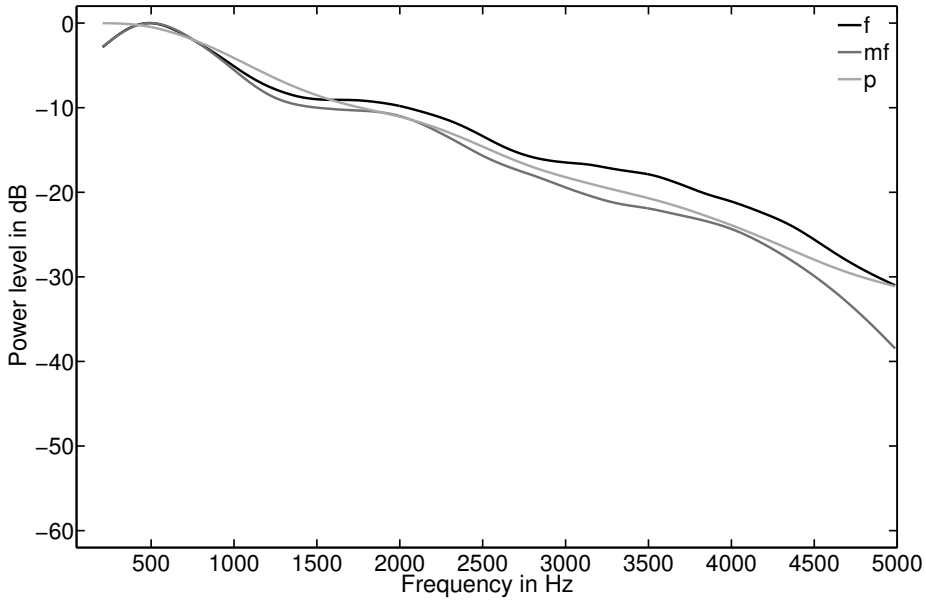

Fig. B.8 Spectral-envelope estimates for violin section (14 players) across dynamic markings *forte*, *mezzoforte*, and *piano*.

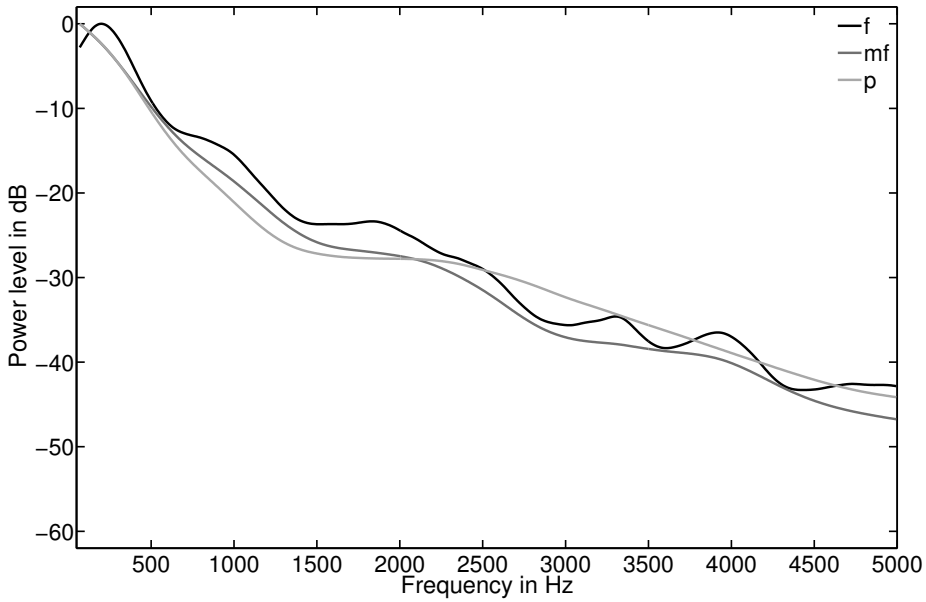

Fig. B.9 Spectral-envelope estimates for violoncello section (8 players) across dynamic markings *forte*, *mezzoforte*, and *piano*.

Appendix C

Stimulus presentation and spatialization

Since timbre blending in the context of orchestration and musical practice would involve the occurrence of instruments at spatially distinct locations as well as in relatively expansive and reverberant venues, stimuli were presented in a way that would recreate a listening environment likely encountered in practice. All conducted experiments presented stimuli either via two-channel, stereophonic loudspeakers or binaural reproduction, which corresponds to individual instrument sounds being spatialized differently. Only the stimuli of Experiment 4[1] relied on the original stereo recordings from the Vienna Symphonic Library (VSL); Experiments 1 to 3 utilized only their left-channel data. Experiment 5 processed audio signals performed in real time. Both the stereophonic and binaural spatialization rendered individual instruments at distinct locations in a simulated virtual space.

In a scenario with several instruments acting as acoustic sound sources, listeners act as corresponding receivers and hear the instrument sources as a function of spatial disposition and room-acoustical characteristics. These factors can be accurately modeled by convolution of the source signals with room impulse responses (RIRs) that describe the frequency transfer functions between all sources and receivers. The highest realism would be achieved by taking the source directivity of instruments and the binaural hearing of listeners into account. The importance of frequency-dependent directivity patterns of instruments has been studied to some extent, but it has unfortunately never led to substantial findings with respect to perceptual difference thresholds that would inform the level of spectral and spatial resolution needed (Meyer, 2009; Otondo and Rindel, 2004). In simulations or *auralizations*, instrument directivity is commonly based on octave-filter resolution. Despite

[1]The stimuli of Experiment 4 were presented in the same setting as for Experiments 1 to 3, illustrated in Figure C.2, without, however, using the spatialization environment.

this spectral simplification, both computational room-acoustical modeling (e.g., Otondo and Rindel, 2004) and multi-loudspeaker diffusion strategies (e.g., Pasqual et al., 2010) exhibit a high degree of technical sophistication and requirements, while at the same time still falling short of accurately modeling physical reality.

C.1 Experiments 1, 2, and 3

All stimuli of Experiments 1 to 3 were simulated in a virtual room environment with its proportions and absorptive properties resembling a mid-sized, slightly reverberant concert setting, as illustrated in Figure C.1. The two instrument sources were 4 m apart at an elevation of 1.2 m. Two receiver locations represent two microphones with omni-directional directivity spaced 0.6 m apart (e.g., *AB* main microphone commonly used in recordings of orchestral music), facing the instruments at an axial distance of 8 m. Both pairs of sources and receivers are centered in the room width, while their mid-distance reference plane is centered in the room length. The distance between the microphones determines the recording angle[2], which leads to the perception of the direct sound of sources falling within that angle as *phantom sources* along the stereo image between two loudspeakers, as indicated by the red crosses in Figure C.2.

Inside an IAC double-walled, isolated sound booth, participants were seated in the *sweet spot* defined by the standard two-channel stereo setup, situating listeners in an equilateral triangle with the loudspeakers, with the latter oriented towards the listeners, as illustrated in Figure C.2.[3] Two active, near-field studio monitors (Dynaudio Acoustics *BM 6A*) mounted on stands at ear height were used as loudspeakers spaced 1.67 m apart with the main level on the monitor controller (Grace Design *m904*) held constant. The locations for the two phantom sources (red crosses) resulting from the spatial disposition of sources relative to the recording angle are ±46% off-center, with 100% representing localization at the loudspeakers themselves.

[2]For the spacing of 0.6 m, the recording angle corresponds to about ± 60° off the normal axis.

[3]Exact execution of the described listener position was only achieved for Experiments 2, 3, and 4, as participants were instructed to remain aligned to visual markers on the front and side walls throughout the experiment. Experiment 1 lacked such visual guides. However, participants were still asked to remain seated in the center of the two loudspeakers. As Experiments 2, 3, and 4 relied on pre-determined values for the loudness balance between instruments, and the stereophonic presentation varied as a function of spatial location, an exact execution was necessary, whereas in Experiment 1 it was negligible as participants themselves had direct control over the loudness balance.

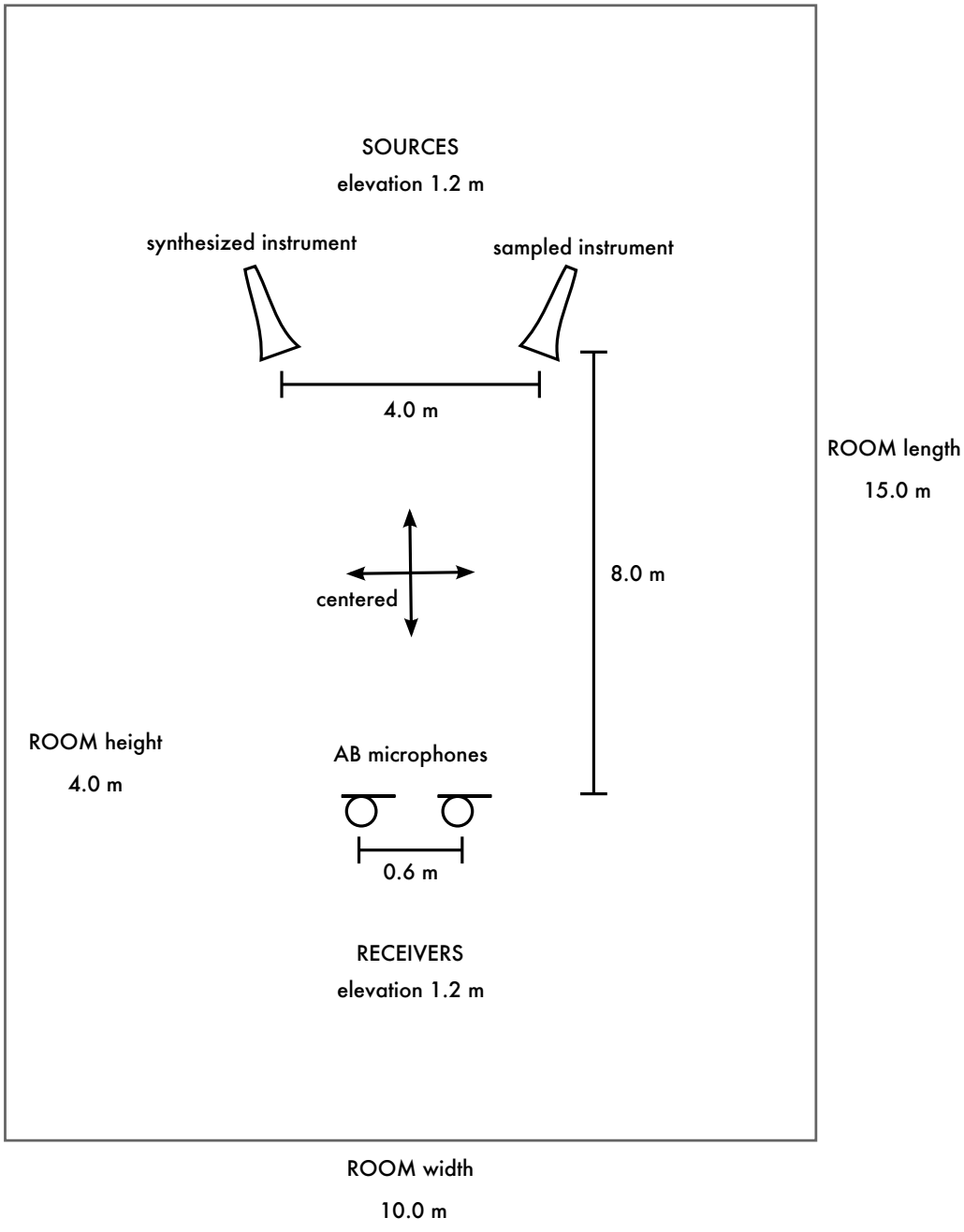

Fig. C.1 Source and receiver disposition and room dimensions used for spatialization in Experiments 1, 2, and 3. For Experiment 3, the *synthesized* instrument is substituted by the second *recorded* one.

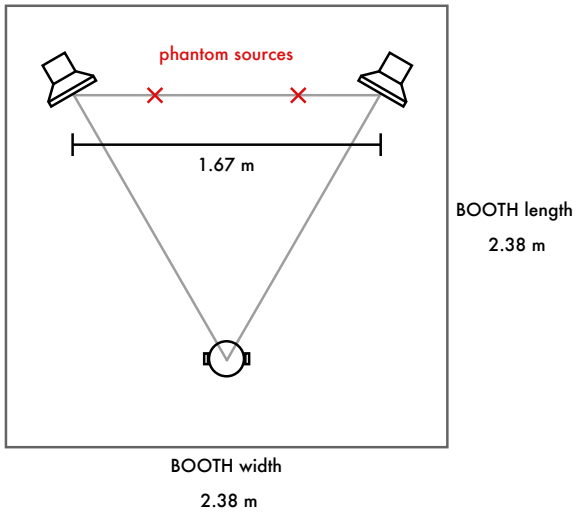

Fig. C.2 Loudspeaker and listener disposition in the sound booth for Experiments 1 to 4. For Experiments 1 to 3, the spatialization outlined in Figure C.1 corresponds to the indicated phantom sources (red crosses).

The spatialization was achieved through convolution of the sources' audio signals with the corresponding RIRs at the receiver locations, with four parallel convolutions being required and the resultant signals at each receiver (from the two sources) being summed. The four RIRs for all combinations between sources and receivers were generated with a MATLAB-based room-acoustical simulation[4] employing a mirror-source model with up to 12th-order reflections. All surrounding surfaces were given the same absorptive properties for all frequencies, leading to a reverberation time of approximately 0.33 s. The real-time spatialization was implemented in Max/MSP 5, accomplished by the use of four parallel instances of an external object[5], which employs an FFT-based fast-convolution algorithm.

The implemented spatialization model considers several simplifications compared to physical reality. The sources are modeled as radiating sound omni-directionally. Similarly, receivers are also modeled to be omni-directional. In theory, this agrees with the usage of two omni-directional microphones, but in practice, omni-directional microphone directivity is only given below frequencies whose wavelengths are significantly larger than the size of the microphone membrane and enclosure. For higher frequencies, microphones became

[4] `rir.m`, version 3.2, by Stephen G. McGovern, 2003
[5] `Tconvolution~`, version 0.1, by Thomas Resch, 2006

increasingly directional towards sound incident from the front. Furthermore, modeling all surrounding surfaces as having the same absorptive properties and as being frequency-independent does also not correspond to physical reality. As a result, the modeled sound radiation from instruments contributes to more reflections from surrounding surfaces than usual. Likewise, reflections from the back wall are over-represented at high frequencies compared to what would be expected from real omni-directional microphones.[6] Hence, the achieved sound image may exhibit less realism than possible, but given that a satisfactory degree of realism is still achieved subjectively, these limitations were acceptable for the research purposes.

C.2 Experiment 5

For Experiment 5, binaural reproduction was employed to simulate musicians as hearing themselves and another musician in a common performance venue. Over the last century, research on binaural hearing and reproduction has made substantial advances as well as finding a wide range of applications, e.g., artificial-head systems (Paul, 2009). Characterizing binaural perception, head-related transfer functions (HRTFs) involve the human torso and head, with notably the pinnae strongly influencing sound localization. HRTFs have been found to be highly individual and weaknesses of artificial-head systems, e.g., front/back or up/down localization confusions, are due to the use of HRTFs that have been averaged across individual humans (Møller et al., 1999). Knowledge of individual HRTFs can allow the correction of these potential sources of errors, which, however, would have required their measurement for each performer. As the aim was to assess a general influence of room acoustics and not, for instance, more specific perceptual thresholds for spatial hearing, the effort that would have been necessary to acquire individualized HRTFs and to implement complex simulations of instrument directivity would have been prohibitive. As a result, binaural reproduction employed an artificial-head system, using a single broadband loudspeaker system as excitation that was positioned so as to emulate a simplified instrument directivity. The arrangement between performers was held constant across both rooms. A fixed distance separated both performers spatially, with both centered relative to the room dimensions along the lateral plane defined by them, as schematized in Figure C.3.

[6] Given a typical microphone-membrane diameter of 2 cm, this would only apply to frequencies $\geq$17 kHz.

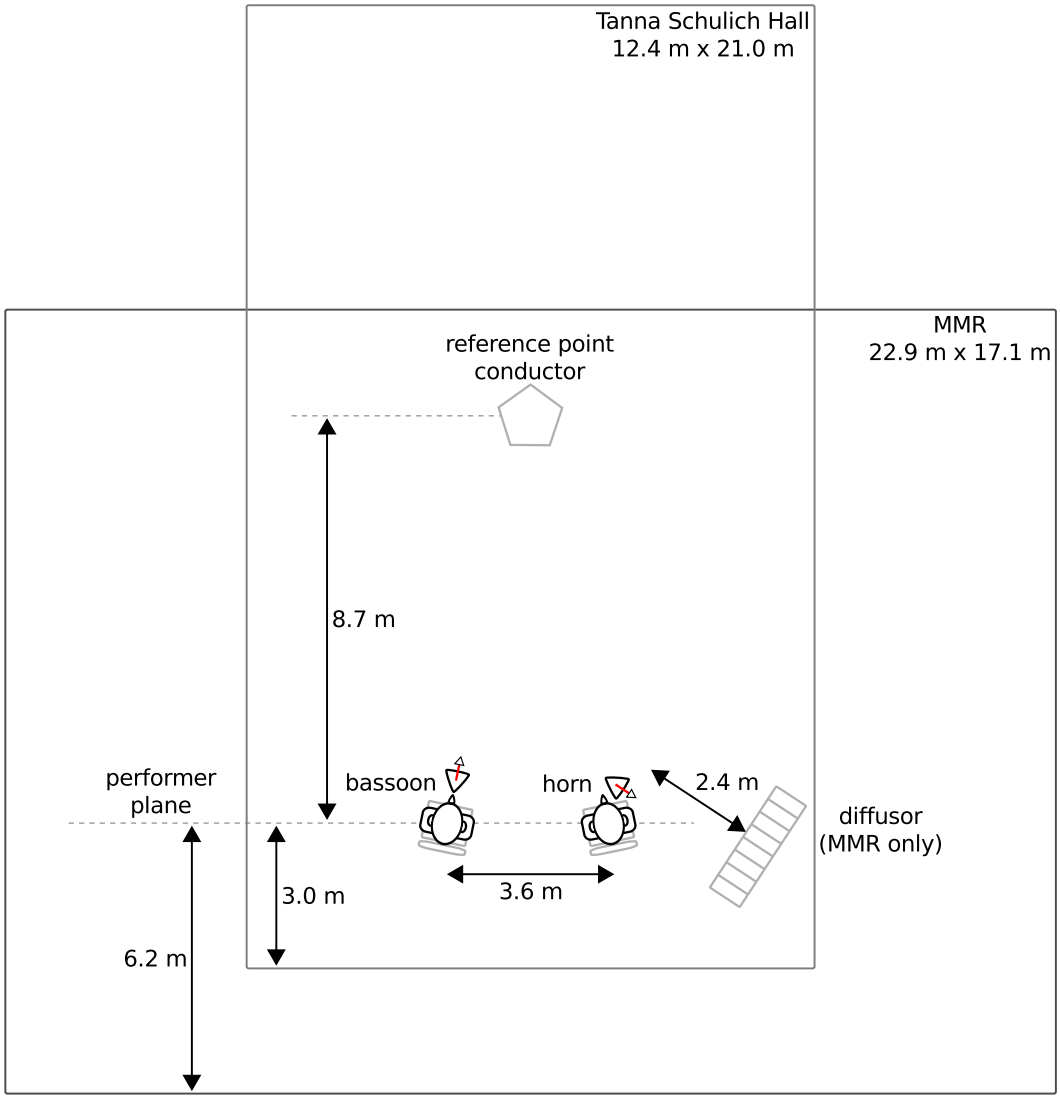

Fig. C.3 Floor plan of simulated positions between performers inside Tanna Schulich Hall and the MMR. Rounded triangles represent instrument sources, with red arrows indicating their main directivity; the seated manikins act as receivers, facing a central conductor location. Distances and room dimensions (simplified to rectangular geometry) are to scale, whereas objects are disproportionately magnified.

Four independent, binaural RIRs for different receiver-source constellations were captured in each of the investigated rooms, i.e., *Tanna Schulich Hall* and the *Multimedia Music Room* (MMR), corresponding to both performers listening to themselves and both performers listening to the other performer. In order to simulate the performers as being seated, the head-and-torso receiver was placed on chairs, whereas the loudspeaker source was placed on a pedestal (de-coupled by sandbags) in front of or to the side of the chair. Based on the usual orchestral seating arrangements oriented towards a conductor, the chairs were slightly angled inwards, aimed at a central reference point. Instrument-specific source directivity was simulated by loudspeaker orientation, pointing the main axis of loudspeaker radiation, i.e., perpendicular to its driver cones, to a corresponding direction for which each instrument radiates the broadest range of frequencies (Meyer, 2009). For the bassoon, the loudspeaker was aimed frontwards as seen from the performer, whereas for the horn, the loudspeaker orientation was chosen to be on-axis with its bell, i.e., aimed towards the right-back side as seen from the performer (approx. 135° angle). Due to the relatively large distance to the back wall in the MMR, a sound diffusor was placed several meters behind the hornist, simulating the influence of a stage wall or shell occurring at a similar distance.

For improved signal-to-noise ratio in measuring RIRs, a sine-tone sweep was used as excitation, generated and converted to impulse responses by using a convolution-based reverberation plug-in (Altiverb) hosted by a digital-audio workstation (ProTools) and AD/DA-converted through an audio interface (RME *Fireface UFX*). The same software and similar hardware components were used for the real-time convolution during the experiment. This technical implementation achieved reasonably realistic room simulations. Shortcomings of binaural reproduction via artificial-head systems, e.g., front/back confusion, were minimized through the lateral disposition of the performers. Also, the simplified usage of a loudspeaker source radiating into a real room, as opposed to employing a computational room model, achieved a satisfactory degree of realism.

Appendix D

Spectral-envelope synthesis

The interest in investigating the contribution of main formants to timbre blending motivated the development of a sound-synthesis model that allowed for parametric variations of spectral-envelope characteristics. This model was used in Experiments 1 and 2 reported in Chapter 2. Inspired by previous formant synthesis approaches, which had mainly focused on voice synthesis (Rodet et al., 1984; Sundberg, 1991), a source-filter model was adopted in which a composite filter structure describes the pitch-generalized spectral envelope and is grouped into two independent filters, one corresponding to a main formant, the other modeling the remaining spectral envelope. During synthesis, the filter structure is fed a broadband, harmonic source signal that can be varied in fundamental frequency. In order to fulfill the requirements for its subsequent use in perceptual tests, the synthesis had to meet several criteria. The independent filters were controllable with respect to frequency location and relative magnitude or gain. Furthermore, a real-time functionality was sought that exhibited instantaneous response to parameter changes and could handle discontinuous parameter value changes. The implementation was made in Max/MSP 5, which fulfilled all requirements and provided the flexibility of modeling the required digital source signals and filter structures.

D.1 Source signal

As the motivation behind the creation of controlled stimuli focused on partial tones outlining the spectral envelope in a region relevant to the occurrence of formants, the excitation source signal was implemented as being limited to 5 kHz and not containing any noise components. As a result, the source signal $s[n]$ comprised harmonics of the fundamental

frequency f_0 of equal amplitude, as formalized in Equation D.1 for the sampling period T_s.[1] The number of harmonics H was chosen to limit the bandwidth based on f_0, as illustrated in Equation D.2.

$$s[n] = a[n] \cdot \sum_{h=1}^{H} sin\left(2\pi\, n\, h f_0\, T_s\right) \tag{D.1}$$

$$H = \lfloor \frac{5000\, Hz}{f_0} \rfloor \tag{D.2}$$

With regard to the temporal amplitude envelopes $a[n]$ for isolated notes, the attack and decay portions were modeled as linear ramps of 100 ms duration. Although this by no means represents an accurate modeling of instrument-specific attack and decay properties, the equality of temporal envelope characteristics across different synthesized instruments aided the desired primary focus on spectral properties.

D.2 Spectral-envelope filters

Each of the two spectral-envelope filters (index i) was modeled as two cascaded second-order all-pole filters (index j), with both spectral-envelope filters implemented as a parallel structure. The composite filter transfer-function $H(z)$ is defined in Equations D.3 to D.6.[2] Each component all-pole filter is defined by a set of coefficients for their individual bandwidths B_{ij}, center frequencies f_{ij}, and gains g_{ij}.

$$H(z) = \sum_{i=1}^{2} \left[\prod_{j=1}^{2} \frac{G_{ij}}{1 - 2\, R_{ij}\, cos\left(\theta_{ij}\right)\, z^{-1} + R_{ij}^2\, z^{-2}} \right] \tag{D.3}$$

$$R_{ij} = e^{-\pi T_s B_{ij}} \tag{D.4}$$

$$\theta_{ij} = 2\,\pi\, T_s\left(f_{ij} + \Delta F_i\right) \tag{D.5}$$

[1] Max/MSP external `oscil~`, version 2.0, by Eric Lyon, 2006

[2] Despite the parallel implementation of the two filter structures, their individual contributions to $H(z)$ are not independent. As a result, relative magnitude differences are greater than the individual parameter variations suggest. Since no quantification of exact magnitude differences (e.g., determination of perceptual thresholds) was sought, this did not compromise our investigation.

$$G_{ij} = 10^{\,\Delta L_i/20} \cdot \left(1 + \frac{\Delta F_i}{f_{ij}}\right) g_{ij} \tag{D.6}$$

The independent control parameters for each filter were implemented as absolute deviations from a pre-defined origin (zero) for frequency ΔF_i in Hz and gain ΔL_i in dB.[3] Audible glitches due to discontinuous parameter value changes were avoided by insertion of linearly interpolated transitional values over durations of up to 200 ms.

D.3 Modeling of instruments

Each spectral-envelope filter was matched to the identified main formant of a particular instrument or the remaining spectral-envelope regions, as shown in Figure D.1 and discussed in Section 2.2.3. More specifically, the modeling involved manual adjustments of the sets of component-filter coefficients B_{ij}, f_{ij}, and g_{ij}, with the result defined as the original filter response, i.e., the case for which the filter-parameter deviations ΔF_i and ΔL_i are zero. The reported findings only investigated variations in main-formant frequency ΔF_1, with all other parameters remaining constant, i.e., $\Delta F_2 = \Delta L_1 = \Delta L_2 = 0$.

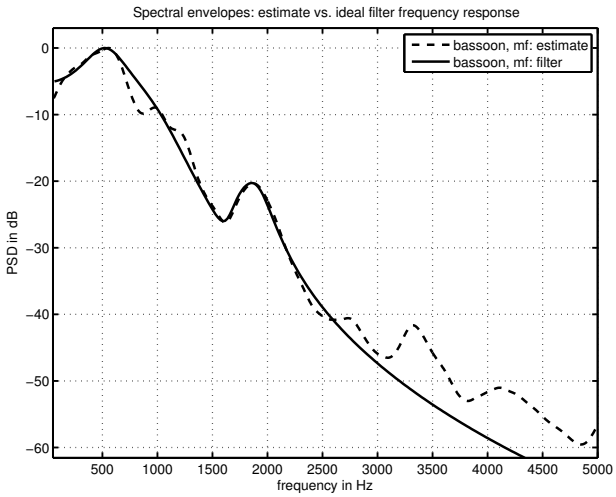

Fig. D.1 Modeled filter frequency response (solid) and spectral-envelope estimate (dashed) for bassoon.

[3] In Equation D.6, the ΔF_i-dependent weighting of gains g_{ij} becomes necessary to achieve a quasi-constant gain across variations of ΔF_i.

References

Adler, S. (2002). *The study of orchestration*. 2nd edition. New York: Norton.

ANSI (1973). *Psychoacoustical terminology*. New York: American National Standards Institute.

Benade, A. H. (1976). *Fundamentals of musical acoustics*. New York: Oxford University Press.

Berlioz, H. and Strauss, R. (1905). *Instrumentationslehre von Hector Berlioz*. Leipzig: C.F. Peters.

Bregman, A. S. (1990). *Auditory scene analysis: the perceptual organization of sound*. Cambridge, MA: MIT Press.

Brown, J. C., Houix, O., and McAdams, S. (2001). Feature dependence in the automatic identification of musical woodwind instruments. *Journal of the Acoustical Society of America*, 109(3):1064–1072.

Burghauser, J. and Špelda, A. (1971). *Akustische Grundlagen des Orchestrierens*. Regensburg: G. Bosse.

Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, 118(1):471–482.

Chon, S. H. (2013). *Timbre saliency, the attention-capturing quality of timbre*. PhD thesis, McGill University.

Conover, W. J. and Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129.

D'Ausilio, A., Badino, L., Li, Y., Tokay, S., Craighero, L., Canto, R., Aloimonos, Y., and Fadiga, L. (2012). Leadership in orchestra emerges from the causal relationships of movement kinematics. *PLoS One*, 7(5):e35757.

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

Doval, B. and Rodet, X. (1991). Estimation of fundamental frequency of musical sound signals. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. V–3657–V–3660, Toronto.

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

Elhilali, M., Xiang, J., Shamma, S. A., and Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology*, 7(6):e1000129.

Fabiani, M. and Friberg, A. (2011). Influence of pitch, loudness, and timbre on the perception of instrument dynamics. *Journal of the Acoustical Society of America*, 130(4):EL193–199.

Flanagan, S. and Moore, B. C. J. (2000). The influence of loudspeaker type on timbre perception. In *Audio Engineering Society Convention 109*, pp. 1–10, Los Angeles, CA.

Fletcher, N. H. and Rossing, T. D. (1998). *The physics of musical instruments*. 2nd edition. New York: Springer.

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

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

Goad, P. J. and Keefe, D. H. (1992). Timbre discrimination of musical instruments in a concert hall. *Music Perception*, 10(1):43–62.

Goebl, W. and Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Perception*, 26(5):427–438.

Goodman, E. (2002). Ensemble performance. In Rink, J., ed., *Musical performance: a guide to understanding*, chap. 11, pp. 153–167. Cambridge: Cambridge University Press.

Goodwin, A. W. (1980). An acoustical study of individual voices in choral blend. *Journal of Research in Music Education*, 28(2):119–128.

Grey, J. M. (1978). Timbre discrimination in musical patterns. *Journal of the Acoustical Society of America*, 64(2):467–472.

Grey, J. M. and Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, 63(5):1493–1500.

Hajda, J. M. (2007). The effect of dynamic acoustical features on musical timbre. In Beauchamp, J. W., ed., *Analysis, synthesis, and perception of musical sounds: the sound of music*, chap. 7, pp. 250–271. New York: Springer.

Hajda, J. M., Kendall, R. A., Carterette, E. C., and Harshberger, M. L. (1997). Methodological issues in timbre research. In Deliège, I. and Sloboda, J., eds., *Perception and cognition of music*, chap. 12, pp. 253–306. Hove: Psychology Press.

Hall, M. D., Pastore, R. E., Acker, B. E., and Huang, W. (2000). Evidence for auditory feature integration with spatially distributed items. *Perception & Psychophysics*, 62(6):1243–1257.

Handel, S. (1995). Timbre perception and auditory object identification. In Moore, B. C., ed., *Hearing*, chap. 12, pp. 425–461. San Diego, CA: Academic Press.

Handel, S. and Erickson, M. L. (2004). Sound source identification: the possible role of timbre transformations. *Music Perception*, 21(4):587–610.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., and Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics*, 17(4):315–339.

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

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

Irino, T. and Patterson, R. D. (2006). A dynamic compressive gammachirp auditory filterbank. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2222–2232.

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. Technical report, International Organization for Standardization, Geneva.

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(4):751–763.

Keller, P. E. (2008). Joint action in music performance. In Morganti, F., Carassa, A., and Riva, G., eds., *Enacting intersubjectivity*, vol. 10 of *Emerging communication: studies in new technologies and practices in communication*, chap. 14, pp. 205–221. Amsterdam: IOS Press.

Keller, P. E. and Appel, M. (2010). Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles. *Music Perception*, 28(1):27–46.

Kendall, R. A. (1986). The role of acoustic signal partitions in listener categorization of musical phrases. *Music Perception*, 4(2):185–213.

Kendall, R. A. (2004). Musical timbre in triadic contexts. In *Proc. 8th International Conference on Music Perception and Cognition*, vol. 1, pp. 600–602, Evanston, IL.

Kendall, R. A. and Carterette, E. C. (1990). The communication of musical expression. *Music Perception*, 8(2):129–163.

Kendall, R. A. and Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, 8(4):369–404.

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

Kendall, R. A. and Vassilakis, P. (2006). Perceptual acoustics of consonance and dissso-nance in multitimbral triads. *Journal of the Acoustical Society of America*, 120(5):3276.

Kendall, R. A. and Vassilakis, P. N. (2010). Perception and acoustical analyses of tradi-tionally orchestrated musical structures versus nontraditional counterparts. *Journal of the Acoustical Society of America*, 128(4):2344.

Kennan, K. W. and Grantham, D. (1990). *The technique of orchestration*. Englewood Cliffs, NJ: Prentice Hall.

Koechlin, C. (1954–1959). *Traité de l'orchestration: en quatre volumes*. Paris: M. Eschig.

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

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

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

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

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

Mathews, P. (2004). Delius and the joining of French & German orchestration. In *Proc. Music Theory Society of the Mid-Atlantic Annual Meeting*, pp. 1–24, Philadelphia, PA.

McAdams, S. (1984). *Spectral fusion, spectral parsing and the formation of auditory images*. PhD thesis, Stanford University.

McAdams, S. (1993). Recognition of auditory sound sources and events. In McAdams, S. and Bigand, E., eds., *Thinking in sound: the cognitive psychology of human audition*, chap. 6, pp. 146–198. New York: Oxford University Press.

McAdams, S. (2013). Musical timbre perception. In Deutsch, D., ed., *Psychology of music*, 3rd edition, chap. 2, pp. 35–67. San Diego, CA: Academic Press.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192.

Mertens, P.-H. (1975). *Die Schumannschen Klangfarbengesetze und ihre Bedeutung für die Übertragung von Sprache und Musik*. Frankfurt am Main: E. Bochinsky.

Meyer, J. (2009). *Acoustics and the performance of music*. 5th edition. New York: Springer.

Micheyl, C. and Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. *Hearing Research*, 266(1-2):36–51.

Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F. (1999). Evaluation of artificial heads in listening tests. *Journal of the Audio Engineering Society*, 47(3):83–100.

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

Moore, B. C. and Moore, G. A. (2003). Discrimination of the fundamental frequency of complex tones with fixed and shifting spectral envelopes by normally hearing and hearing-impaired subjects. *Hearing Research*, 182(1-2):153–163.

Nakamura, T. (1987). The communication of dynamics between musicians and listeners through musical performance. *Perception & Psychophysics*, 41(6):525–533.

Otondo, F. and Rindel, J. H. (2004). The influence of the directivity of musical instruments in a room. *Acta Acustica united with Acustica*, 90(6):1178–1184.

Pasqual, A. M., de França Arruda, J. R., and Herzog, P. (2010). Application of acoustic radiation modes in the directivity control by a spherical loudspeaker array. *Acta Acustica united with Acustica*, 96(1):32–42.

Patterson, R. D., Gaudrain, E., and Walters, T. C. (2010). The perception of family and register in musical tones. In Riess Jones, M., Fay, R. R., and Popper, A. N., eds., *Music perception*, vol. 36 of *Springer handbook of auditory research*, chap. 2, pp. 13–50. New York: Springer.

Paul, S. (2009). Binaural recording technology: a historical review and possible future developments. *Acta Acustica united with Acustica*, 95(5):767–788.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The Timbre Toolbox: extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, 130(5):2902–2916.

Piston, W. (1955). *Orchestration*. New York: Norton.

Rakowski, A. (1990). Intonation variants of musical intervals in isolation and in musical contexts. *Psychology of Music*, 18(1):60–72.

Rasch, R. A. (1988). Timing and synchronization in ensemble performance. In Sloboda, J., ed., *Generative processes in music: the psychology of performance, improvisation, and composition*, chap. 4, pp. 70–90. Oxford: Clarendon Press.

Reuter, C. (1995). *Der Einschwingvorgang nichtperkussiver Musikinstrumente: Auswertung physikalischer und psychoakustischer Messungen*. Frankfurt am Main: P. Lang.

Reuter, C. (1996). *Die auditive Diskrimination von Orchesterinstrumenten - Verschmelzung und Heraushörbarkeit von Instrumentalklangfarben im Ensemblespiel*. Frankfurt am Main: P. Lang.

Reuter, C. (2002). *Klangfarbe und Instrumentation: Geschichte–Ursachen–Wirkung*. Frankfurt am Main: P. Lang.

Reuter, C. (2003). Stream segregation and formant areas. In Kopiez, R., Lehmann, A. C., Wolther, I., and Wolf, C., eds., *Proc. 5th Triennial ESCOM Conference*, pp. 329–331, Hannover.

Rimsky-Korsakov, N. (1964). *Principles of orchestration*. New York: Dover Publications.

Rodet, X., Potard, Y., and Barrière, J.-B. (1984). The CHANT project: from the synthesis of the singing voice to synthesis in general. *Computer Music Journal*, 8(3):15–31.

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

Saldanha, E. L. and Corso, J. F. (1964). Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, 36(11):2021–2026.

Sandell, G. J. (1991). *Concurrent timbres in orchestration: a perceptual study of factors determining blend*. PhD thesis, Northwestern University.

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

Schnittke, A. (2006). Timbral relationships and their functional use. In Mathews, P., ed., *Orchestration - an anthology of writings*, pp. 162–177. New York: Routledge.

Schoenberg, A. (2006). Instrumentation. In Mathews, P., ed., *Orchestration - an anthology of writings*, pp. 133–138. New York: Routledge.

Schumann, K. E. (1929). *Physik der Klangfarben*. Professorial dissertation, Universität Berlin.

Shamma, S. A. and Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, 20(3):361–366.

Song, H. and Beilharz, K. (2007). Spatialization and timbre for effective auditory graphing. In *Proc. 8th WSEAS International Conference on Acoustics & Music: Theory & Applications*, pp. 18–26, Vancouver.

Strong, W. and Clark, M. (1967a). Perturbations of synthetic orchestral wind-instrument tones. *Journal of the Acoustical Society of America*, 41(2):277–285.

Strong, W. and Clark, M. (1967b). Synthesis of wind-instrument tones. *Journal of the Acoustical Society of America*, 41(1):39–52.

Stumpf, C. (1926). *Die Sprachlaute: experimentell-phonetische Untersuchungen nebst einem Anhang über Instrumentalklänge*. Berlin: Springer.

Sundberg, J. (1991). Synthesizing singing. In de Poli, G., Piccialli, A., and Roads, C., eds., *Representations of musical signals*, chap. 9, pp. 299–324. Cambridge, MA: MIT Press.

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

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., and Griffiths, T. D. (2011). Brain bases for auditory stimulus-driven figure-ground segregation. *Journal of Neuroscience*, 31(1):164–171.

van Dinther, R. and Patterson, R. D. (2006). Perception of acoustic scale and size in musical instrument sounds. *Journal of the Acoustical Society of America*, 120(4):2158–2176.

Villavicencio, F., Röbel, A., and Rodet, X. (2006). Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I–869–I–872, Toulouse.

Wedin, L. and Goude, G. (1972). Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psychology*, 13(1):228–240.

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

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics: facts and models*. 2nd edition. Berlin: Springer.