A spectral-envelope synthesis model to study perceptual blend between wind instruments

Sven-Amin Lembke, Stephen Mcadams

ACOUSTICS 2012

A spectral-envelope synthesis model to study perceptual blend between wind instruments

S.-A. Lembke and S. McAdams

CIRMMT, McGill University, Schulich School of Music, 555 Sherbrooke St. W., Montreal, Canada H3A 1E3
sven-amin.lembke@mail.mcgill.ca

Wind instrument sounds can be shown to be characterized by pitch-invariant spectral maxima or formants. An acoustical signal-analysis approach is pursued to obtain spectral-envelope descriptions that reveal these pitch-invariant spectral traits. Spectral envelopes are estimated empirically by applying a curve-fitting procedure to a composite distribution of partial-tone frequencies and amplitudes obtained across an instrument's pitch range. A source-filter synthesis model is designed based on two independent formant filters with their frequency responses matched to the spectral envelope estimates. This is then used in perceptual experiments in which parameter variations of the synthesis filter are manipulated systematically to investigate their contribution to the degree of perceived blend between the synthesized sound and a recorded instrument sound. The perceptual relevance is assessed through two tasks in which participants either produce the best attainable blend by directly controlling synthesis parameters or rate the degree of blend for 5 parameter presets. Behavioral data from both experiments suggest the utility of this formant-based model for correlating pitch-invariant acoustical description with perceptual relevance, as both formant frequency and magnitude appear to affect perceived blend.

1 Introduction

Research in auditory perception has for a long time relied on sound synthesis methods for the creation of controlled stimuli that allow a systematic investigation of parametrized acoustical factors. In the perceptual study of timbre blending between instruments, agreement of pitch-invariant spectral traits has been argued to contribute to perceived blend between concurrent instrument sounds [1]. Previous acoustical investigations have confirmed the existence of these pitch-invariant spectral maxima [2, 3], with these maxima being termed *formants*, by analogy with the acoustic properties of the human voice.

With regard to our investigation of perceived timbre blend between two concurrently sounding instruments, a capability was sought to parametrically vary the spectral shape of a synthesized instrument. The spectral shape is based on a pitch-invariant spectral envelope representation, which was operationalized in terms of being expressed as a combination of formant regions whose perceptual relevance could be tested.

The following sections describe the development of a stimulus production environment enabling concurrent presentation of a synthesized and recorded sampled instrument and how the technical infrastructure was used in the perceptual investigation.

2 Stimulus production

An essential requirement for the design of a synthesis method to model wind instruments was its reliance on a spectral envelope representation derived from estimates based on real instruments. Therefore the following section will first present the technique for obtaining the spectral envelopes and establish concepts related to their description, before discussing the technical infrastructure of the synthesis model and stimulus presentation environment.

2.1 Spectral-envelope description

Past considerations of the description of formants or other pitch-invariant spectral traits for instruments have called for a comprehensive assessment to encompass a whole range of pitches of an instrument [4]. In order to validate previous claims of pitch invariance as well as confirm their relevance to our set of instruments (e.g. bass trombone, horn, trumpet, bassoon, clarinet, oboe, flute), an acoustical characterization aimed at obtaining spectral-envelope descriptions of our instrument sample database was conducted.

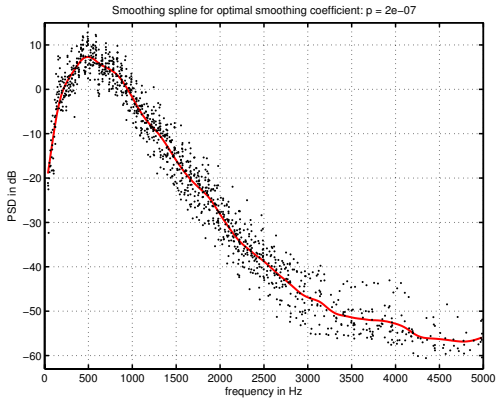

Figure 1: Spectral envelope estimate for bass trombone (line) and distribution of partial tones (dots).

Similar to past approaches [3, 5], an empirical estimation of the pitch-invariant spectral envelope of wind instruments was pursued. This involved the initial computation of power density spectra for sustained portions of sounds for up to 40 pitches per instrument, followed by a partial tone detection routine. A curve-fitting procedure employing a so-called *smoothing spline* applied to the composite distribution of partial tones yielded the spectral envelope estimates, as shown in Figure 1.

The obtained spectral envelope estimates served as the basis for qualitative identification and categorization of two formants that were implemented in the synthesis model of the instruments. The main formant represented the most prominent spectral maximum with decreasing magnitude towards both lower and higher frequencies or if not available, the most significant plateau along the magnitude decrease towards high frequencies.[1] The secondary formant was the next most prominent spectral maximum or plateau. Furthermore, pitch-invariant descriptors for the main formant were formulated that described the frequencies of the formant maximum f_{max} as well as upper and lower bounds at which the power magnitude decreased by either 3 dB or 6 dB relative to the maximum.

2.2 Spectral-envelope synthesis model

Inspired by previous formant synthesis approaches which had mainly focused on voice synthesis [6, 7], a source-filter

[1] The latter case only applied to the spectral envelope estimate of the flute.

model was adopted in which a composite filter structure describes the pitch-invariant spectral envelope and is grouped into two independent *formant filters*. 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 formants were required to be 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 modelling the required digital source signals and filter structures.

2.2.1 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 and equal amplitudes as shown in Equation 1 for the sampling period T_s. The number of harmonics H was chosen to limit the bandwidth based on f_0 as illustrated in Equation 2.

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

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

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

2.2.2 Formant filters

Each of the two formant filters (index i) was modelled as two cascaded second-order all-pole filters (index j), with both formant filters implemented as a parallel structure. The composite filter transfer-function $H(z)$ is defined in Equations 3 to 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}.

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

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

[2]Despite the parallel implementation of the formant filters, 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) is sought, this does not compromise our investigation.

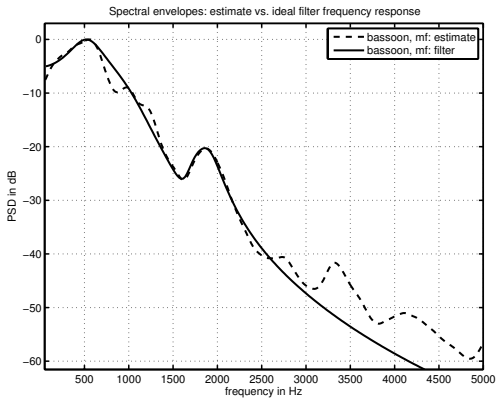

Figure 2: Modelled filter frequency response (solid) and spectral envelope estimate (dashed) for bassoon.

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

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

The independent control parameters for each formant filter were implemented as absolute deviations from the ideal (zero) for frequency ΔF_i in Hz and gain ΔL_i in dB.[3]

2.2.3 Modelling of instruments

Each of the instruments was modelled by using their respective spectral envelope estimates as a reference and matching each formant filter to the identified formants, as shown in Figure 2. The modelling involved manual adjustments of the sets of component-filter coefficients B_{ij}, f_{ij} and g_{ij}, with the result being termed the *ideal* filter response, i.e. the case for which the control parameter deviations ΔF_i and ΔL_i are zero. The achieved closeness in spectral shape between models and estimates was not meant to deliver realistic emulations of the instruments per se but instead to reduce spectral differences not associated with identified formants and as a result to improve a selective evaluation of their perceptual relevance.

Most of the instruments considered are well-characterized by the formant representation, with the flute and clarinet representing the least appropriate cases. For instruments for which only a single formant appeared to characterize the spectral envelope significantly, the other formant filter served an entirely technical function of complementing the main formant's frequency response to adequately model the entire spectral envelope estimate. Furthermore, since the spectrum of the clarinet is characterized by the well-known attenuation of the lowest even-order partials, which notably also varies as a function of pitch, the modelled filter structure for the clarinet intentionally diverged from the obtained estimate. As our study aimed at finding a relevance of pitch-invariant properties, the clarinet was modelled to describe only the identified formant (located above low-order partials) and the remaining spectral envelope towards higher frequencies, thus excluding its pitch-variant frequency region.

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

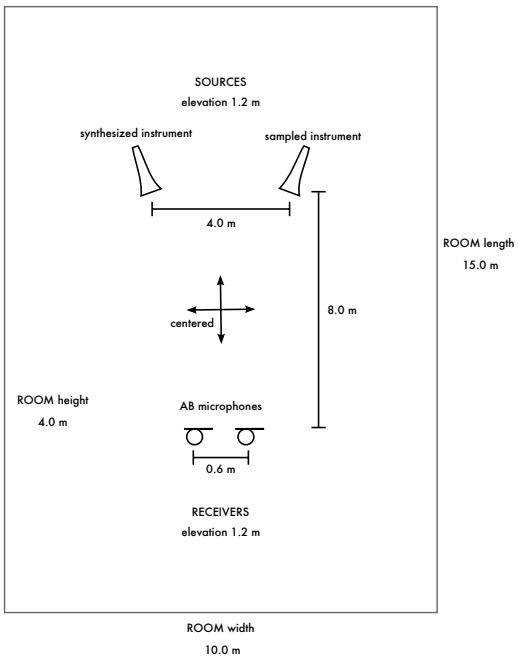

Figure 3: Sources and receivers in acoustical room simulation model.

2.3 Stimulus presentation environment

Although the synthesis presented a central part of the stimulus production infrastructure, the perceptual investigation of blend still required it to be paired with the concurrent presentation of a recorded instrument sample. A stimulus presentation was chosen that would recreate a listening environment likely encountered in listening to instrumental music. As a result, the *synthesized* and *sampled* instruments were simulated as spatially distinct sources in an acoustical room simulation model, shown in Figure 3, employing real-time convolution. Two receiver locations simulate a common time-delay-based stereophonic AB main-microphone setup to be presented over a standard stereo loudspeaker listening configuration. The loudness balance between the instruments presented another control parameter that concerned the investigation, being implemented as a linear crossfade between the amplitudes of the two sources prior to convolution with the room model. Sounds for both instruments were triggered synchronously and due to the instrument samples' limited duration of about 5 s, repeated throughout an experimental trial.

3 Perceptual investigation

Given the developed synthesis model and the overarching stimulus presentation environment, the subsequent perceptual investigation aimed to investigate the possible perceptual relevance of the pitch-invariant formants. Applied to the spectral-envelope synthesis model, this concerned studying if and how the formant control parameters ΔF_i or ΔL_i were related to perceived blend. Furthermore, pitch invariance would only be deducible if a certain trend could be confirmed across several different pitches. The original hypothesis of a perceptual relevance of formants argued for perceived timbre blend to be related to a coincidence of formant frequencies between instruments [1], which in our case would correspond to finding no significant deviations from the *ideal* formant parameter values. Two behavioral experiments were conducted to investigate these assumptions.

3.1 Experimental design

The two experiments differed in the experimental tasks that were employed, with the second also aiming to provide further validation and clarification of findings from the first experiment. The synthesized instruments were paired with recorded samples of the same instruments at selected pitches. All instruments except for the bass trombone were included as stimuli in the main experiments. With respect to multifactorial statistical hypothesis tests, both experiments adopted a within-participants design.

3.1.1 Experiment A: Blend production

The first experiment employed a production task and was conducted with 17 participants, recruited as musically experienced listeners. Across 66 trials (22 conditions $\times$ 4 repetitions) participants were given the task to adjust either ΔF_i or ΔL_i directly in order to achieve the maximum attainable blend. User control of the stimulus production environment was provided via a two-dimensional graphical interface, with controls for the investigated formant parameter and the loudness balance between instruments. ΔF_i was investigated for the main formant of all instruments, with the secondary formant only being tested for oboe. ΔL_i for the main formant was only tested for instruments prominently characterized by formants, namely, horn, bassoon and oboe. The latter also served as the only case the secondary formant gain was investigated.

3.1.2 Experiment B: Blend rating

The second experiment was based on a simplified and less time-consuming rating task and involved 20 participants, again recruited as experienced listeners. Across 120 trials (30 conditions $\times$ 4 repetitions) participants were asked to rate the relative degree of blend for a total of 5 sound dyads per condition. A continuous relative blend rating scale was used, spanning from *most blended* to *least blended*. Across 5 dyads the same instrument sample formed pairs with varying formant parameter value presets for ΔF_1 or ΔL_1, with only the main formant ($i = 1$) being considered.[4] For both parameters one of the presets presented the zero-deviation *ideal* case. The remaining 4 presets comprised moderate deviations below (*-mod*) and above (*+mod*) the ideal and likewise, a pair of extreme deviations (*-ext* and *+ext*). The presets were based on generalizable formant properties which allowed comparisons between instruments to be made on a common scale of spectral-envelope description.

For ΔF_1 the 4 non-ideal preset values were derived from formant descriptors (see Section 2.1) and defined as the difference between a frequency f_{preset}, expressed in terms of formant bounds, and the identified formant maximum, as shown in Equation 7. More specifically, moderate deviations $f_{preset}(\pm mod)$ fell 10% within both 3 dB-bounds relative to their frequency width. $f_{preset}(-ext)$ corresponded to

[4] The presets included predetermined values for the loudness balance between instruments and also had been equalized for loudness between presets.

the higher of the two values between 80% of the lower 6 dB-bound or 150 Hz. Whereas $f_{preset}(+ext)$ was identical to the upper 6 dB-bound. All instruments were considered for this formant parameter.

$$\Delta F_1 = f_{preset} - f_{max} \quad (7)$$

For ΔL_1 the moderate deviations represented values obtained from the behavioral findings of Experiment A paired with values mirrored relative to the ideal. The extreme deviations were defined as being 60% more extreme than the moderate ones. The instruments included the same ones in Experiment A plus trumpet, with the missing behavioral reference value for the latter being substituted by the average behavioral values of the other instruments.

Apart from the direct blend ratings, additional measures were also formulated to quantify the preferred parameter deviations. These were derived from rating-weighted measures of parameter values based on the two highest ratings per trial, and were defined as weighted average and standard deviation, serving as descriptors for central tendency and spread of parameter values across within-participant repetitions.

3.2 Behavioral findings

Both experiments motivated a broad range of statistical tests, generating a large quantity of reportable data. Due to space constraints, only the most meaningful behavioral findings with respect to the synthesis model are reported for both experiments. This limitation concerns only reporting results for the main formant for which clear indications have become apparent, arguing for its dominant role in explaining perceptual blend.

Experiment A yielded results for the scenario in which participants themselves determined the parameter values leading to the best perceived blend. For relative parameter deviations $\Delta F_1/f_{max}$, a common trend to slightly underestimate the ideal by about 10% was found, as is shown in Figure 4. For 4 instruments, the underestimations were statistically significant (*), determined through a single-sample t-test against a sample mean of zero.[5] Notably, the horn and bassoon did not differ significantly from the ideal formant frequency. The absolute deviations ΔL_1 showed a clear trend to relative amplification of the main formant contributing to best blend, results for all considered instruments being significantly different from the ideal.

Experiment B aimed to confirm tendencies found in Experiment A and investigate whether they exhibited pitch invariance across a set of representative pitches. Each formant parameter was tested at 2-4 pitches per instrument, including the original conditions from the previous experiment. Instead of finding the best blend along a continuum of parameter deviations as in Experiment A, participants compared the relative degree of perceived blend between presets, which could, and in fact did, lead to some differences in the results. With regard to frequency deviations ΔF_1, the preferred (i.e. highest-rated) presets were not only oriented toward the ideal value and moderate underestimations (*-mod*), but included the extreme underestimation (*-ext*) as well. Conversely, the lowest ratings were obtained for overestimations of the ideal value (*+mod* and *+ext*), which agrees with the general trend

[5] All reported statistically significant results are based on a significance level: $\alpha = .05$.

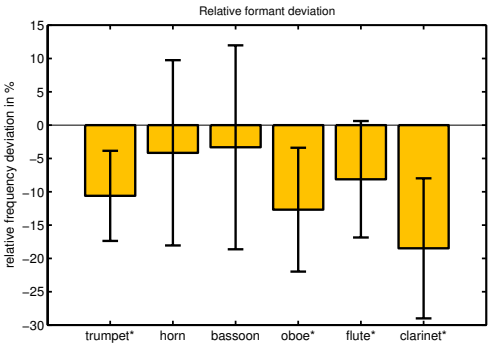

Figure 4: Mean behavioral $\Delta F_1/f_{max}$ (error: std. deviation).

	instrument					
effect	flute	clarinet	oboe	trumpet	horn	bassoon
preset	s	s	s	s	s	s
pitch	ns	ns	ns	ns	ns	ns
preset x pitch	s	s	ns	ns	ns	ns

Figure 5: Main and interaction effects for the factors 'preset' and 'pitch' (s: significant, ns: not sig.).

of underestimation found in Experiment A. Given a similar preference of frequency deviations at and below the ideal, the relative rating-weighted measures showed clearer underestimations than in Experiment A, on the order of 20%. Yet, even in this case the bassoon and horn only yielded significantly different results from the ideal in about half the comparisons. For gain deviations ΔL_i, amplification of the main formant could again be confirmed for the same instruments as in Experiment A, with nearly all comparisons being significantly different from the ideal. However, the trumpet did not show a clear trend for main formant amplification.

With the reported tendencies for both formant parameters being in strong agreement across the direct rating and rating-weighted measures, several multifactorial Friedman tests and ANOVAs were conducted to investigate whether the findings argue for a pitch-invariant constancy of ΔF_1-ratings. The analysis rationale involved testing for main effects for the factors 'preset' and 'pitch'. Significant main effects for 'preset' would confirm that ratings could be considered as a reliable indicator of perceptual differences. Furthermore, the finding of significant interaction effects between the factors 'preset × pitch' would argue against pitch invariance, as the profile of blend ratings across presets would be shown to vary as a function of pitch.[6] As summarized in the table shown in Figure 5, all instruments yield statistically significant main effects for 'preset' but not for 'pitch'. Significant divergence from pitch-invariant performance was only found for the flute and the clarinet, which interestingly are also the instruments that are the least-well represented by the formant representation.

Although both experiments display somewhat different results concerning the perceptual relevance of exact overlap of the formants, they both support the hypothesis that perceived blend is achieved around and below the ideal formant

[6] Due to violations of the assumption of normality for about half the presets, main effects were tested with a non-parametric 2-way Friedman test. As no non-parametric test was available to test for interaction effects, this was done through a repeated-measures 2-way ANOVA, after ensuring that both tests yielded similar tendencies for the main effects.

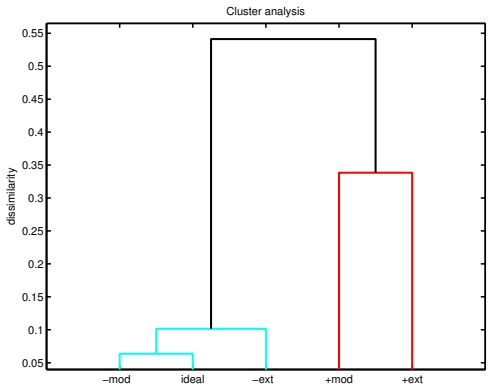

Figure 6: Dendrogram displaying clustering based on effect size from post-hoc analyses for 'preset'.

location and is clearly reduced above this value. To further elucidate this tendency across instruments for which pitch invariance could be assumed, a cluster analysis was conducted with the rating differences between preset levels being interpreted as a dissimilarity measure. This measure considered squared effect sizes (r) of statistically significant non-parametric post-hoc analyses for pairwise comparisons between presets (Wilcoxon signed-rank test).[7] The *complete*-linkage clustering algorithm considered dissimilarity data averaged across 15 independent sets of effect sizes for oboe, trumpet, horn and bassoon. As shown in Figure 6, the overestimations of ΔF_1 (*+mod* and *+ext*) are maximally dissimilar to a compact cluster associating deviations centered on and below the ideal formant location (*ideal*, *-mod* and *-ext*).

4 Conclusion

The development of the spectral-envelope synthesis model has been an essential tool in allowing a selective and focused investigation of the acoustical spectral-envelope properties related to the perception of blend between wind instruments.

We have shown that localized formant regions are perceptually relevant to blend for the main formant parameters describing relative magnitude and frequency location. As concerns the latter, the theory of formant coincidence [1] does not appear to hold across both investigated experimental tasks. Instead, it becomes clearly apparent that the role of formants in the perception of blend may function as a critical frequency boundary. The degree of perceived blend decreases markedly whenever the relative location of formants exceeds the frequency boundary of a *reference* formant. As the reference formant in our investigation was predetermined by the static sampled instrument, it remains to be studied how this would apply to musical practice, in which musicians perform blend in an interactive relationship.

Pitch invariance is suggested by both the acoustical description and perceptual findings for most of the investigated wind instruments, its perceptual relevance being meaningful to the development of realistic renditions of wind-instrument synthesis. It may be assumed that an accurate physical modelling synthesis of these wind instruments would exhibit the same pitch-invariant relationships described here. Given the perceptual role attributed to formants and their apparent pitch-invariant relevance, it can be assumed that pitch-invariant descriptors describing the frequency boundary may be able to serve as acoustical predictors of perceived blend.

As to the utility of the model to individual instruments, the flute and clarinet have been found to deviate from a pitch-invariant behavior and would thus need to be treated as special cases. This would likely limit their musical usage to not serving as candidates for non-unison blended combinations. By contrast, the bassoon and horn display a strong robustness across pitch and are centered on the ideal formant location. Apart from being commonly used in orchestration practice to achieve blend, their lower pitch ranges, could furthermore support a hypothesis of 'darker' timbres generally leading to more blend [8]. With this hypothesis having being derived from an acoustic description based on a global spectral average (e.g., spectral centroid), our investigation has contributed further by delivering more differentiated explanations based on a more local spectral origin.

[7] Dissimilarity was assumed zero for non-significant differences.

Acknowledgments

The authors would like to thank Bennett Smith for assistance in the setup of perceptual testing hardware. This work was supported by a Schulich School of Music scholarship to SAL and a Canadian Natural Sciences and Engineering Research Council grant to SM.

References

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

[2] Schumann, K. E. *Physik der Klangfarben - Vol. 2*. professorial dissertation, Universität Berlin, Berlin (1929).

[3] Luce, D. & Clark, J. Physical Correlates of Brass-Instrument Tones. *The Journal of the Acoustical Society of America* **42**, 1232–1243 (1967).

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

[5] Luce, D. A. Dynamic Spectrum Changes of Orchestral Instruments. *Journal of the Audio Engineering Society* **23**, 565–568 (1975).

[6] Rodet, X., Potard, Y. & Barrière, J.-B. The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General. *Computer Music Journal* **8**, 15–31 (1984).

[7] Sundberg, J. Synthesizing singing. In *Representations of musical signals*, chap. 9, 299–324 (MIT Press, Cambridge, MA, 1991).

[8] Sandell, G. J. Roles for Spectral Centroid and Other Factors in Determining "Blended" Instrument Pairings in Orchestration. *Music Perception* **13**, 209–246 (1995).