PERCEPTUALLY CORRELATED PARAMETERS OF MUSICAL INSTRUMENT TONES

James W. Beauchamp

School of Music and Dept. of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, Illinois USA
jwbeauch@illinois.edu

ABSTRACT

In Western music culture instruments have been developed according to unique instrument acoustical features based on types of excitation, resonance, and radiation. These include the woodwind, brass, bowed and plucked string, and percussion families of instruments. On the other hand, instrument performance depends on musical training, and music listening depends on perception of instrument output. Since musical signals are easier to understand in the frequency domain than the time domain, much effort has been made to perform spectral analysis and extract salient parameters, such as spectral centroid changes, in order to create simplified synthesis models for musical instrument sound synthesis. Moreover, perceptual tests have been made to determine the relative importance of various parameters, such as spectral centroid variation, spectral incoherence, and spectral irregularity. It turns out that importance of particular parameters depend on both their strengths within musical sounds as well as the robustness of their effect on perception. Methods that the author and his colleagues have used to explore timbre perception are: 1) discrimination of parameter reduction or elimination; 2) dissimilarity judgments together with multidimensional scaling; 3) informal listening to sound morphing examples. Ramifications of this work for sound synthesis and timbre transposition will be discussed and demonstrated.

1. INTRODUCTION

The principal long-term goal of this study is to achieve a synthesis system where a minimal set of independent but perceptually meaningful parameters are used to control and synthesize musically useful sounds, including sounds of traditional musical instruments. Basic steps for accomplishing this goal are a) using spectral analysis to obtain static and time-varying parameters; b) building synthesis models to utilize these parameters; c) conducting formal listening tests on single sounds to test the efficacy of the models; and d) conducting informal listening tests using synthesis of extended musical passages.

Although it might seem that this goal could be achieved in a few weeks or months, in practice, timbre has been studied using a series of less ambitious steps. Typically the first step is to select a group of musical sounds to study. The parameters to be identified from the sounds are first of all the time-varying amplitudes and frequencies obtained from a spectral analysis. Then, more detailed, possibly perceptually important parameters can be inferred such as attack and decay times, spectral envelope features (such as spectral centroid spectral irregularity, and spectral flux), vibrato characteristics, and inharmonicity.

Conducting a formal listening test for timbre requires the following steps:

- Stimuli preparation
- Psychoacoustic testing (the actual listening test)
- Data processing and presentation
- Interpretation of results

Either synthetic or recorded acoustic ("real") sounds can be used as stimuli, but in either case they should be normalized to eliminate sonic attributes that are not part of timbre, namely, loudness, pitch, and duration. The latter two are not a problem for synthetic sounds (sounds consisting solely of harmonically related frequencies), but for either sound type loudness equalization through gain factor adjustment must be achieved by additional loudness testing or a special program [1] or alternatively by randomizing the levels [2]. For pitch normalization of harmonic sounds it is generally acceptable to make certain that the fundamental frequencies are the same. Duration is a bit more complicated, but a method is given in [3], where attack and decay structures are retained and the total duration is reduced to a standard 2 s.

If a test is to only compare different acoustic sounds, as in the case of a dissimilarity test, no further stimuli modifications may be necessary. Physical (spectral) differences between sounds can be measured in terms of specific parameters and correlated with the measured perceptual differences. However, for a discrimination test the experimenter will often want to modify specific acoustic parameters of the sounds and then examine how discrimination ability varies with each parameter that is changed or, in more detail, the amount of change of each parameter.

Important questions are: 1) What specific parameters should be varied? 2) Why do we choose these particular parameters? How do we measure them? How do we vary them? For the studies reviewed in this paper, the specific parameters are spectral irregularity, spectral flux, spectral centroid variation, amplitude and frequency microvariations, and inharmonicity. Reasons for choosing these parameters are discussed in the timbre literature. (See [4] and [5] for reviews.) Methods for measuring and varying them are given in [3] and [6].

Also, it should be remarked that two specific parameters, average spectral centroid and attack time have proved to be so salient that they are sometimes factored out (normalized) of the stimuli. Such was the case with [6] and [7] and is a method utilized in the second study covered in this paper.

Preparation of a psychoacoustic test requires the selection of listener subjects and the design of the test. Chosen listeners are generally young people with good hearing and divided between those with extensive and those with meager musical experience. The two formal tests described in this review paper use either timbre discrimination or dissimilarity judgments. With discrimination the subjects are generally asked to judge whether pairs of sounds are same or different. With a dissimilarity task

they are asked to judge the amount of difference on a scale of say 0 to 10.

Once a test is completed and the data is collected, the data can be processed in various ways. Discrimination averages can be simply presented or graphs of discrimination vs. a particular parameter can be illustrated. In the case of dissimilarity judgments, the method of multidimensional scaling (MDS) is commonly used to display the positions of the sound stimuli in a 2- or 3-dimensional space [8] [9]. To show trends within the space the dimensions can be correlated with parametric measures of various spectrotemporal parameters of the stimuli.

Finally, an interpretation of the data is usually given. One of the biggest problems is estimating the scope of validity of the results. Scope is usually limited because it is difficult to design tests that cover a wide range of cases and can still be conducted over a reasonable time period.

2. THREE TIMBRE STUDIES

Three studies will be described, a 1999 timbre discrimination study, a 2006 timbre dissimilarity judgment study with MDS solution, and a 2008 timbre transposition study.

2.1 Timbre discrimination study (1999)

The objective of this study [3] was to investigate the relative importance of some different spectrotemporal parameters by simplifying musical sounds with respect to these parameters. The stimuli prototypes (reference sounds) consisted of tones performed on seven different instruments: clarinet, flute, oboe, trumpet, violin, harpsichord, and marimba at pitch $E^{\flat}_4$ (311 Hz). Loudnesses were equalized using a brief test, and durations were equalized to 2 s using a method described in [3]. The sound signals were analyzed using a pitch-synchronous short-time Fourier transform program [10], and the resulting partial amplitude and frequency data was simplified as follows:

1) partial amplitude-vs.-time envelopes smoothed
2) spectral envelope smoothed (irregularity reduced)
3) spectral flux (aka *incoherence*) eliminated
4) partial frequency-vs.-time envelopes smoothed
5) partial frequencies locked to time-varying harmonic
6) partial frequencies flattened to harmonic

and then the sounds were resynthesized to the time domain by additive synthesis.

Note that the word *partial* is used here instead of *harmonic* because even though the frequencies of these tones are close to harmonic, departures from harmonicity are possible.

Although there was considerable variation with instrument (see [3] for details), the discrimination results averaged over the seven instruments were:

a) spectral envelope smoothed	96%
b) spectral flux eliminated	91%
c) frequencies flattened	71%
d) frequency envelopes smoothed	70%
e) frequencies locked harmonically	69%
f) amplitude envelopes smoothed	66%

An interpretation of these results is that the spectral parameters *irregularity* (i.e., *jaggedness*) and *flux* (change of spectrum shape over time) are, for this set of instruments, most salient. Smoothing the amplitude and frequency envelopes (using a 10 Hz cutoff low-pass filter) eliminates fine grained temporal detail, but this elimination is relatively unnoticeable. So is locking the frequencies harmonically or removing any trace of frequency variation.

However, when an error metric (similar to those discussed in [7]) was constructed based on the difference between reference and modified partial amplitudes, and a regression line was constructed to fit discrimination (given in terms of d') against the log of this error, it was found that the regression straight line explained 77% of the discrimination variance (88% if one outlier point was removed). Since the modifications done could also change the spectral centroid, d' was also plotted against the log of *normalized spectral centroid difference* between the reference and modified sounds. In this case a regression straight line explained only 54% of variance, but when the spectral centroid difference was combined into a total formula with the partial amplitude error metric, 83% of variance was explained (with no outliers removed).

A final interpretation from these results is that, yes, spectral irregularity and flux are important specific parameters, but discrimination is also strongly correlated with a total metric difference between two sounds which takes into account all of the frequency components of the sounds.

2.2 Timbre dissimilarity study (2006)

With this study, originally presented as a talk in 2006 [11], subjects had the task of judging the dissimilarity between musical sounds. The original stimuli consisted of tones performed on ten sustained-tone instruments: bassoon, cello, clarinet, flute, horn, oboe, recorder, alto sax, trumpet, and violin. Two types of tones were constructed from these: *dynamic* (with flux) and *static* (without flux). The tones were also normalized with respect to pitch (F_0 = 311 Hz), attack time (.05 s), decay time (.05 s static, .15 s dynamic), total duration (0.5 s static, 2.0 s dynamic), loudness [1], and average normalized spectral centroid (3.7). Average centroids were normalized by applying a filter with response k^p to each harmonic k's amplitude, where p was varied to achieve the desired centroid value, as described in [6].

The listening test employed ten musically experienced subjects to judge dissimilarity between tone pairs using a method of triadic comparison [12]. While dissimilarity scores theoretically could vary from 0 to 17, actual scores varied from about 4 to 13. The scores were placed in a 10×10 dissimilarity matrix which was processed by two different classical MDS programs (SPSS and Matlab). For the static tones, only 2D solutions were made, whereas both 2D and 3D solutions were made for the dynamic tones. Stresses (average normalized difference between inter-timbre distances given by the dissimilarity matrix and those given by the MDS solution) for the 2D solutions were both 0.12 for the static case and 0.15–0.17 for the dynamic case; for the dynamic 3D solutions they were both 0.095.

(It was somewhat of a surprise for this author to discover the degree to which the distances between pairs of timbres in an MDS solution do not exactly match the values given by the dissimilarity matrix and that *stress* is commonly given by MDS programs as an important measure of their average agreement. Stress generally decreases as the number of dimensions increases, but for visualization 2 or 3 dimensional solutions are preferred. Stress is useful for estimating the accuracy of an MDS solution. Unfortunately, in reading music several MDS papers, I could not find a single mention of the word *stress*, even though it is a very basic concept in the theory of MDS.)

Meanwhile, static tone solutions were correlated with two parameters measured from the sound signals, *even/odd harmonic ratio* (ratio of the average rms amplitude of the even harmonic amplitudes to that of the odd harmonics) and *spectral irregularity*. The dynamic tone solutions were correlated with those parameters plus two others: *spectral flux* (aka *incoherence*) and *normalized spectral centroid variation* (spectral centroid standard deviation divided by its average value). All MDS solutions were rotated so that the best possible even/odd correlation aligned with the horizontal axis. For the other parameters, best-fit straight lines of highest correlation to the various parameters were computed.

Details of the corresponding SPSS and Matlab solutions were different. However, for the 2D static case instrument groupings were quite similar. The most obvious groupings were {recorder, clarinet, cello} and {trumpet, oboe, violin}. R^2 correspondences with the even/odd and spectral irregularity parameters were 78-79% and 69-75%, respectively, for the two solutions. For the 2D dynamic case the correspondences were 71-69% for even/odd, 68-68% for spectral centroid variation, 56-53% for spectral incoherence, and 39-40% for spectral irregularity. Also, the spectral centroid variation and spectral incoherence straight lines were close together, indicating that these variables were highly correlated.

For the 3D dynamic case the correspondences for even/odd, spectral centroid variation, spectral incoherence, and spectral irregularity were 82-68%, 83-82%, 53-83%, and 82-71%, respectively, indicating rather strong disagreement between the SPSS and Matlab solutions as to the saliency of 3 out of 4 of the parameters. Averaging over the two solutions gives 82.5% for spectral centroid variation, 76.5% for spectral irregularity, 75% for even/odd, and 68% for spectral incoherence. Therefore, assuming that 3D solutions are best for the dynamic case because of their relatively low stress, it appears that spectral centroid variation is the parameter with the highest and most consistent saliency (beyond average centroid and attack/decay) for dynamic tones. On the other hand, any of the four parameters corresponds as well as the others for at least one of the two solutions. Also, it is curious that the average correspondence for the four parameters is about the same for the SPSS solution (75%) as for the Matlab solution (76%), which means it would be difficult to conclude that one solution is *better* than the other, but they certainly are significantly different (average correspondence difference equals 14%).

After making all of these computations one might ask: What is the advantage of using MDS? Why not just correlate with the original dissimilarity data? Certainly MDS yields some pretty pictures, showing the relative positions of timbres relative to one another, but as the two 3D solutions for dynamic tones show, different solutions with the same stress can result in timbres in very different positions and can yield quite different correlations. At least with the original dissimilarity matrix there is only one set of data to correlate with, and it has no stress.

2.3 Timbre transposition study (2008)

This study was presented as a talk in 2008 [13]. The point of the study was to explore synthesis using a small set of time-variable control parameters and a family of spectral envelopes [10] [14], which represent a particular instrument, but then switch the spectral envelope family to a different instrument and see what happens. Either the spectral envelopes will dominate, or the temporal data will dominate, or a hybrid instrument that shares characteristics will be produced. The instrument supplying the time-varying parameters is called the *source instrument* and the one supplying the spectral envelope family is called the *target instrument*.

In an earlier project it was discovered that using the time-varying parameters $A_{rms}(t)$, $f_0(t)$, and $f_c(t)$ (i.e., rms amplitude, fundamental frequency, and spectral centroid), combined with a spectral envelope family based on spectral centroid clustering, could produce trumpet tones that were quite realistic. The spectral envelope family was derived from a training set of trumpet tones that covered a wide gamut of pitches and dynamics (i.e., intensity levels). Every frame of every tone was analyzed (using the pitch-synchronous analyzer) and sorted into different "bins" based on ranges of centroid values, 0-200, 200-400, etc. The spectra in each bin were normalized and then sorted into critical bands, and finally the amplitudes within each band were averaged to give a single value that represents the band amplitude for that bin. These amplitudes as a function of the band center frequencies formed a spectral envelope, and the collection of spectral envelopes for the various centroid ranges formed a family of spectral envelopes for the trumpet.

Synthesis was done by first deriving representative time-varying parameters $A_{rms}(t)$, $f_0(t)$, and $f_c(t)$ from a trumpet solo recording. $f_c(t)$ was used to compute the instantaneous spectral envelope by interpolation from the spectral envelope family, and harmonic amplitudes were obtained from the spectral envelope by sampling it at frequencies $k\ f_0(t)$, where k is the harmonic number. These amplitudes can be easily adjusted to match the total amplitude $A_{rms}(t)$. Then the sound is synthesized using additive synthesis. A demonstration of this method using a restricted parametric model for the temporal variations is given on this author's website [15]. This includes the addition of low frequency noise microvariations to the pitch and amplitude controls to make the synthesis sound more realistic.

There is a question of whether the same family of spectral envelopes is adequate for all pitches (F_0's) or whether the family has to change as a function of F_0 . It seems to be the case (but not proven) that single families are adequate for brass instruments but perhaps not for woodwinds or strings. However, with the abundance of memory available in computers these days, it is entirely reasonable to compute and store a different family for each F_0. Thus, spectral envelope becomes a function of both spectral centroid (f_c) and pitch (f_0).

It is fairly obvious that there needs to be a match between the source instrument and the target instrument in terms of the ranges of pitches and centroids. Thus, if the source and target pitch ranges and centroid ranges don't overlap sufficiently, timbre transposition won't work. However, the control ranges coming from the source can be easily mapped to correspond to the best ranges for the target using simple linear equations.

Findings for this study were informal. It appears that if the temporal data of the source instrument is similar to that of the target instrument, the result will likely be identified as the target instrument. On the other hand, if the target instrument's spectral envelope family is similar to that of the source, the temporal information of the source may dominate and the result may still be identified as the source instrument. We have found this to be true if the source is a bassoon playing a series of low-pitched short-duration notes is the source and a horn is the target. In between these cases are cases where neither the temporal information or spectral data are similar for the two instruments, so a true hybrid is produced. If horn is the source and clarinet is the target, we have a situation where the resulting hybrid sound can be recognized as a horn in terms its temporal envelopes but as a clarinet in terms of its unique spectrum with emphasis on odd harmonics.

Gradual morphing or interpolations between the instruments could be produced by cross-fading the temporal controls or the spectral envelopes or both. We actually haven't tried this yet, but there is no reason why the method shouldn't produce interesting results.

Another possibility to investigate is the addition of additional external controls designed to modify parameters such as those shown to be salient in the 1999 and 2006 timbre studies discussed above.

3. REMARKS AND CONCLUSIONS

The 1999 timbre study, which used parameter simplification and discrimination, indicated that spectral irregularity and spectral flux were more important than amplitude and frequency microvariations and inharmonicity. However, this author would take that result with a grain of salt because it is well known that temporal details and inharmonicity are important for instrument recognition and for warmth and realism.

The 2006 dissimilarity study, which used multidimensional scaling to summarize relative perceptual distances between instrument timbres, yielded some interesting results and raised some nontrivial issues. One issue was the unexpected importance of the concept of stress (see above). Another was the usefulness of rotation for comparing solutions using different MDS programs. Still another was that solutions from different MDS programs can be quite different, although for the same number of dimensions their stresses tend to be in approximate agreement. Still another, was that best-fit straight lines that don't normally correspond to dimensional axis lines can be used to maximize R^2 correspondence. Finally, different MDS programs can yield different correlations with acoustic parameters, making exact conclusions about the saliency of these parameters problemmatic. Nonetheless, our conclusions from the MDS solutions can be summarized thusly: For static tones (those without flux) for two different 2D solutions with stresses of 12%, even/odd harmonic ratio correlated quite high (78-79%) and better than spectral irregularity (69-75%). For dynamic tones (those with flux) for two different 2D solutions with stresses of 15 and 17%, even/odd correlated best (69-71%), followed by spectral centroid variation, spectral flux, and spectral irregularity. With the 3D solution for these tones, the stresses dropped to 9.5% and some correlations increased 82-83%, but there was very significant disagreement between the solutions, except for spectral centroid variation (both solutions close to 82%).

The 2008 timbre transposition study showed that combining some time-variant parameters with fixed spectral envelopes can not only allow the formation of a compact resynthesis model for a given instrument, but it can also serve as a method for applying the temporal characteristics of one instrument to the spectral characteristics of another. In some cases the resulting sounds demonstrate one of the two characteristics dominating the other. When the differences between the corresponding characteristics of the two instruments are both pronounced, a true hybrid is generally produced, where the temporal (articulatory) characteristic can be recognized as coming from one instrument and the spectral (tone color) characteristic as coming from the other.

4. REFERENCES

[1] Moore, B. C. J.; Glasberg, B. R.; and Baer, T., "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness", *J. Audio Eng. Soc.,* vol. 45, pp. 224–240 (1997).

[2] Dai, H., "On suppressing unwanted cues via randomization", *Perception & Psychophysics*, Vol. 70, No.7, pp. 1379–1382 (2008).

[3] McAdams, S.; Beauchamp, J. W.; and Meneguzzi, S., "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters", *J. Acoust. Soc. Am.*, Vol. 105, No. 2, Pt. 1, pp. 882–897 (1999).

[4] Hajda, J. M., "The Effect of Dynamic Acoustical Features on Musical Timbre", in *Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music*, J. W. Beauchamp, ed (Springer), pp. 250–271 (2007).

[5] Donnadieu, S., "Mental Representation of the Timbre of Complex Sounds", in *Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music*, J. W. Beauchamp, ed (Springer), pp. 272–319 (2007).

[6] Beauchamp, J. W. and Lakatos, S., "New spectro-temporal measures of musical instrument sounds used for a study of timbral similarity of rise-time- and centroid-normalized musical sounds", *Proc. 7th Int. Conf. on Music Perception and Cognition* (ICMPC 7), Sydney, Australia, pp. 592–595 (2002).

[7] Horner, A. B.; Beauchamp, J. W. and So, R. H. Y., "A Search for Best Error Metrics to Predict Discrimination of Original and Spectrally Altered Musical Instrument Sounds", *J. Audio Eng. Soc.*, Vol. 54, No. 3, pp. 140–156 (2006).

[8] Miller, J.R. and Carterette, E. C., "Perceptual space for musical structures", *J. Acoust. Soc. Am.*, Vol. 58, No. 3, pp. 711–720 (1975).

[9] Grey, J. M., "Multidimensional perceptual scaling of musical timbres", *J. Acoust. Soc. Am.*, Vol. 61, No. 5, pp. 1270–1277 (1977).

[10] Beauchamp, J. W., "Analysis and Synthesis of Musical Instrument Sounds", in *Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music*, J. W. Beauchamp, ed (Springer), pp. 1–89 (2007).

[11] Beauchamp, J. W.; Horner, A. B.; Koehn, H.-F.; and Bay, M., "Multidimensional scaling analysis of centroid- and attack/decay normalized musical instrument sounds" (abstract), *J. Acoust. Soc. Am.*, Vol. 120, No. 5, Pt. 2, p. 3276.

[12] Plomp, R., "Timbre as a multidimensional attribute of complex tones", in *Frequency Analysis and Periodicity Detection in Hearing*, R. Plomp and G. F. Smoorenburg, eds. (A. W. Sijtohoff, Leiden), pp. 397–414 (1970).

[13] J. W. Beauchamp and M. Bay, "Timbre transposition based on time-varying spectral analysis of continuous monophonic audio and precomputed spectral libraries" (A), *J. Acoust. Soc. Am.*, Vol. 123, No. 5, Pt. 2, p. 3805 (2008).

[14] Luce, D. and Clark, M. Jr., "Physical Correlates of Brass-Instrument Tones", *J. Acoust. Soc. Am.*, Vol. 42, No. 6, pp. 1232–1243 (1967).

[15] http://ems.music.uiuc.edu/beaucham/