Analysis of Percussion for Timbre Measurement and Synthesis

yeoOn lo and dan hitt (e-mail: acoustics.netcom.com)

Abstract
This paper examines a marimba tone and the ear's response to it and then uses the results to build a more suitable framework for measurement and synthesis of timbre.

1 Introduction
Timbre is a percept that pertains to a family of tones. For instance, the 88 notes of a modern Steinway collectively define a particular piano timbre which might be distinguishable from that of a Yamaha piano and which certainly differs from the tones of a clarinet. It is not that all 88 notes of any piano project exactly the same timbre but that there is a sense of continuity in that as we make discernible change in pitch, the change in timbre is imperceptible, even though notes far apart may in fact make slightly different tones1. One of the challenges for synthetists of timbre is to discover the information that makes all the notes of the 'same' timbre exhibit this sort of continuous behavior while at the same time notes of 'same' pitch but 'different' timbres, say a Yamaha and a Steinway, pair-wise distinct.

Timbre is not restricted to that of musical instruments, although the latter are distinguished engineering specimens designed to maintain for all their notes a given tone (when articulated under like conditions). A sound may not have a pitch but it certainly has some kind of a timbre or some superposition of a number of them. Timbre, as a percept, is at once immediate and universal; yet as a notion, it is complicated to articulate and even more difficult to formulate for fruitful measurement and synthesis. Although it is widely believed to be linked to some overtone structure of the sound, what, precisely, is timbre? What is its behavior and what are its universal characteristics?

2 Statement of Purpose
Why are we interested in timbre measurement and what does it entail? To this day, research in computer music and psychoacoustics continues to treat the temporal aspects of acoustic signals with benign neglect. Measurements of timbre typically concern only frequency distribution. Control of temporal behavior is done exclusively with amplitude envelopes, on bins of frequency or on the overall waveform. Unfortunately, detailed amplitude control on frequency bins works quite poorly when the update mechanism is rigid and/or when bandwidth is restricted. A consequence is that we do not accurately capture what is in the signal that the ear might pick up and perceive as information, something relevant to forming a timbre. Resynthesis based on incomplete information2 leads to perceptual difference (from the original). Ways to minimize the latter by adding a component from the analysis residue may not be too illuminating unless this component provides also for high fidelity timbre-preserving syntheses (a term we reserve for those of which some other perceptual dimensions other than timbre may vary). The relevance of such a provision should be clear to those who appreciate the familial notion of timbre.

It is well-established that the ear as an observer of acoustic signals in the audible range has extraordinary capabilities in following temporal events [Schubert]. Thus if a signal's pattern of vibration changes, the ear can keep track of these changes as well as when they materialize. Therefore, timbre measurement pertains not only to frequency, but time also; when do perceptually significant changes occur in a signal? A suitable notion of timbre must exhibit a suitable update mechanism for changes and their measurements. Our purpose here then is to present a way to think about and deal with (in the sense of measurement and synthesis of) timbre, first by observing the behavior of a marimba tone and then by devising a framework to quantitatively describe and analyze this and other timbres in general.

3 Some Criteria a Suitable Timbre Framework Must Meet

1. It (the framework) must have a temporal component which closely tracks that which the ear operates with. In particular, it must account for what we hear in some of the most stringently articulated percussion tones such as what we are dealing with here.

1We sometimes use 'tone' for 'timbre'; our meaning will be clear from context.
2Complete signal information such as that expressed through the Fourier isometry is also uninformative even though it yields, apart from round-off errors, perfect results because the information, e.g., that of the phase, is not presented in a sufficiently distilled form to provide for high fidelity timbre-preserving syntheses of modified tones.
2. It must enjoy optimal data reduction, that is, satisfy the principle of Occam's Razor.

3. It must asymptotically reduce to the Helmholtz Model for static signals, also an expression of (2).

4. It must have an analytical component which is able to strongly capture a timbre under study in the sense that the measurements are fully relevant for and can thus be, with suitable transformations, conveyed to the entire family of sounds projecting the same timbre. This is a corollary to (2) and constitutes a principle of completeness integral to a modern theory of timbre—a point missing among many contemporary writings on the subject.

5. It must satisfy the principle of universality so that the treatment applies to a wide range of timbre and enables the filling or populating of its subspaces (again an expression of (2)).

6. Synthesis must be a driving component in such a treatment because not only is it an obvious modern means for hypothesis verification about any aspect of a sound but also is it a reliable measure of the precision of our knowledge about the matter at hand.

7. In particular, it ought to abstract and encapsulate all spectral and temporal features that need to be captured and conveyed for the entire family of sounds. Abstraction and encapsulation enable the satisfaction of optimal data reduction. Furthermore, analysis and synthesis form mirror images of each other in light of both temporal and spectral measurements.

8. In regard to timbre-preserving syntheses (4), e.g., over a range of several octaves, a challenge for us is this: What is that collection of information in the signals which, through appropriate transformation, gives rise to that perceived sameness of timbre, as their pitch, or more typically pitch contour, move up and down the scale? We know from experience that the information available in the shape of a single magnitude spectrum is not sufficient. Furthermore, the notion of 'sameness' is really a manifestation of continuity rather than invariance. Therefore, when we research a timbre, a key criterion is how well the analytical information satisfies timbre-preserving syntheses, according to the idea of continuity. More precisely, we say that a perceptual structure $\sigma$ is conveyable (over a sound set $S$ with respect to pitch) if

$$||P_{x}(x) - P_{x}(x')|| \ll ||P_{x}(x) - P_{x}(x')|| \quad \text{for } x, x' \in S$$

Here, $P_{x}$ is the projection onto the pitch dimension, i.e., $P_{x}(x)$ is the pitch of $x$, and $P_{x}$ is the projection onto the parameter space of $\sigma$; distances indicated by the delimiting pairs of vertical double bars are normalized to the number of just noticeable differences in the respective dimensions.

The idea is that we want to say that we've captured a particular timbre weakly if we can only resynthesize that particular sound itself; but we've captured it strongly if we can resynthesize the timbre, i.e., a whole group of sounds with the 'same' timbre, along some range of pitches.

These criteria lay the foundation for our work on timbre.

### 4 Signals and Observers

![Poincaré portraits (marimba data)](image)

![Waveform associated with Poincaré portraits (marimba data)](image)

Due to space considerations, we refer the reader to our earlier paper [Lo and Hitt] for detailed discussion.

---

3Now we can see, for example, that the Short-Time Fourier Transform, a commonly used analysis-synthesis method, fails Occam's Razor, even though it studiously attempts to track temporal behaviors, and thus makes for a poor timbre measurement and synthesis tool.

These criteria also imply that a method which exposes a single spectrum, an overall spectral envelope, or some variant of these would not do because the picture of the "partials" of a marimba tone [Lo and Hitt] would not naturally admit such a framework. In the marimba example, partials rise and fall at different times, have different life spans, and exhibit different shapes.

An analysis algorithm which succeeds only in resynthesis but is not capable of translating the analysis to sounds of other pitches, durations, or manners of articulation, fails our criterion (4) of capture and convey.
of the experiments. We examined an anechoic recording of a hard-mallet-excited marimba tone whose waveform is displayed in juxtaposition with its Poincaré trajectory.4

Most of us are familiar with how it sounds, at least on some superficial level. In the literature, it is customary to talk about “the” spectrum when one thinks of its timbre. Here the Poincaré portraits suggest that signal characteristics are changing and the ear is following the change with its very refined local-in-time observations [Boomsmit]. Hence, from the observer’s view, the marimba timbre correlates with not just one but a series of spectra. In the reference cited above, the reader can find very clear evidence of spectral evolution in the sound’s picture of the “partials”, or sonogram, which is most illuminating when intensity variation at a time-frequency cell is represented by indexed colors.

Furthermore, unlike the sound of a vowel or a gently excited clarinet tone, we find no easily discernible synchrony in the behavior of the partials. The sonagram5 [Lo and Hitt] reveals primary bands of energy at three spots: the fundamental, 2 octaves above, and around the 10th harmonic. They have different extents, onsets, peak locations, and decay rates. Finally, they present two paradoxes:

- What we hear does not seem to correspond to what is indicated in the sonagram. For example, in the last few tenths of a second, there is little energy except in the fundamental and yet we continue to hear a bright timbre orginated by the energy surrounding the 10th harmonic.

- What we hear in the sound seems inexplicably different from the sound obtained by reversing the sampled sequence. Notice that the Fourier Transforms of the two sounds have identical magnitude spectra, differing only in the phase information. Yet the sounds are puzzlingly different.

Elsewhere [Lo and Hitt] we accounted for these paradoxes based on understanding the relationship between the temporal characteristics of the signal and the manner the human ear perceives or responds to it. Here we shall focus on how this relationship dictates the way timbre should naturally be treated.

5 Color, Innovation, Motion, & Interpolation

The ear’s temporal resolution characteristics dictate the length (in a signal) within which amplitude variation cannot be resolved in sequence. As a result, each tiny waveform segment in a signal is not perceived as a development in time but rather as a set of simultaneous outputs of resonator banks (the basilar membrane), each responding to the same tiny waveform input. A natural way to think about this is: The ear parses an incoming acoustic signal like a sample-and-hold device with a finite buffer. Contents of successive buffers reveal a temporal development of the signal’s vibration patterns.

Thus we view a sound as a file or train of wave segments or columns of frequency distribution. Each column or segment may be regarded as a sound color, following Slawson [Slawson]. As the ear-brain seeks patterns, it keeps track of regularity while always staying ready for and reacting to any noticeable change [Boomsmit]. The locations of changes, or innovations [Lo, 1986] form a motion [Lo, 1986]. The timbre of a sound is then sound color in motion. A difference between Slawson’s colors and those defined in this framework is that the former are always those of the vowels which are immutable while the latter vary from sound to sound and must be detected by analytical means. Consequently, our framework introduces an analytical component which is not present in Slawson’s. Furthermore, within the context of continuous speech synthesis, this framework attempts to capture all the innovations or crucial patterns of vibration which form a superset of the vowel colors.

In other words, it is the dynamic placement, or motion of the vowels and consonants, suitably modified by the individual speaker’s voice quality, intonations, and inflections, which fully describe what we hear in a spoken specimen, not just a series of quasi-static vowels alone. In this perspective, our framework attempts to build a bridge between naturally occurring specimens and Slawson’s sound colors. This perspective is important because with the advent of text-to-speech applications in modern communication environment, often it remains very desirable to have efficient means to synthesize speech which has a natural quality. Naturalness (versus perceived “machine” quality) facilitates recognition and is most likely accomplished by a synthesis model which most closely approximates the speech as we hear it.

4A Poincaré trajectory is a sort of map in phase space. In physics, it is sometimes a plot of derivative against value; here, we plot value at a running time against value at a fixed offset from the running time. The underlying evolution from something very complex to something very simple is clearly more illuminating with the trajectory than its waveform counterpart.

5A sonagram presents for an acoustic signal, such as a marimba tone, a collection of narrow-bandpass filter output signals. As such, they are time functions, albeit sinusoid-like functions.
The column-in-file representation may seem at first less elegant than the horizontal representation of additive, or Fourier synthesis. Such an appearance is deceiving because the smooth evolution of change in a non-pathological, e.g., a typical, musically interesting sound, except in the transient regimes, requires rather sparse or infrequent updates by the ear-brain. That means that, from a signal point of view, a fully precise formulation of timbre consists of a sequence of columns of crucial acoustical information together with their interpolations. Interpolation here is used primarily in the perceptual sense; its mathematical nature may well vary from class to class of sound sources. Knowledge about the specifics is to be gained from analysis, which then encodes it into the scheme of synthesis. The locations of these crucial vibration patterns, collectively called the motion, are variable, just as the patterns themselves are variable; they are both salient features of a given timbre. The model thus accounts for temporal behaviors which the ear can resolve (and follow) as well as the vibration patterns which the ear cannot resolve and which constitute local spectral information.

6 Conclusions

The marimba signal composition and the ear's response modality suggest a broadly general alternative to the single spectrum description, some of its haphazardly formulated derivatives, or the frame-by-frame STFT description, the first being too simple to be fully informative—failing criterion (1), the second, holds little promise for satisfying (4) because of its arbitrary formulation, and the third, failing (2). Such an alternative automatically satisfies (1) & (3) and abstracts the physical correlates of timbre into a very compact but flexible format with maximal data reduction built in, a feature essential to any rational theory—Occam's razor (2). Within this formalism, data reduction is in a position to continue to exploit the ear's frequency response characteristics (critical band and frequency masking) but it can and usually goes further in the temporal dimension, depending on the dynamical properties of the signals themselves. The nature of the abstraction in this model provides for further data reduction. It does so by allowing other sounds of the 'same' timbre to be synthesized with high fidelity, through suitable transformation of information derived from the colors in motion of the one being studied (2).

This data reduction property is generally absent in contemporary treatment of timbre. The key to success, we believe, should involve following a suitable formalism such as ours and focusing on discoveries of the precise algorithms of transformation for classes of timbre. Our work, being synthesis-driven (6), has been positively tested on a variety of timbre, notably the 3-octave high-fidelity synthesis of violin based on the analysis of a $G^2$ (7), following the capture and convey paradigm and frequency resampling on certain colors in motion (8).

In general, the timbre model we introduce here has wide applications [Lo, 1986], notably in filling the space of each timbre along other perceptual dimensions, such as pitch and duration (4), and in filling timbre subspaces by interpolation (5), one of the most glorious promises computer music research has yet to deliver on. Our framework tries to correct this situation by providing a universal yet conceptually simple approach to synthesis of timbre, allowing room for refinements for individual timbre families, a formal structure not unlike which is associated with perceptual encoding standards in modern digital audio compression.

References


---

Footnote: Line-segment approximation in additive synthesis has the same aim, but it is inflexible about bin allocation. Interpolation begins with the nature of the signal and dutifully reflects itself in the ear's response.