

Perceptual Evaluation of Principal-Component-Based Synthesis of Musical Timbres*

GREGORY J. SANDELL, *AES Member*

Parmly Hearing Institute, Loyola University, Chicago, IL 60626, USA

AND

WILLIAM L. MARTENS, *AES Member*

Headspace, San Mateo, CA 94401, USA

Harmonic-based analysis and resynthesis of musical instrument tones, for example, using the phase vocoder method, is a valuable technique, but its data representation is very large. However, this data set is usually highly redundant. Principal components analysis (PCA) can be used to encode such data into a smaller set of orthogonal basis vectors with minimal loss of information. Techniques for applying PCA to such data are explored, and the aural impact of the method on three tones (cello, trombone, and clarinet) are studied in two perception experiments. Results show that nearly identical resyntheses can be produced with a 40–70% data reduction. A preprocessing step called variable-duration temporal partitioning (VDTP) is introduced, which also affords a natural-sounding method for time expansion and contraction of tones. An extension of the PCA technique is also introduced that implements a “timbre space,” or coordinate system for interpolation among a group of musical instruments.

0 INTRODUCTION

Harmonic-based analysis and resynthesis (additive synthesis) of musical instrument tones has been a staple of timbre research since the 1960s [1]–[3]. It remains a popular tool for researchers and musicians because of the good balance it obtains between effective parametrization of timbre features and faithfulness of resynthesis. Two frequently cited shortcomings of the technique are its computational costs at resynthesis time and its consumption of computer storage space (usually more than the raw samples of the original source sound). Advances in digital signal processing and computer hardware are gradually improving the former problem [4], [5], and the outlook for additive synthesis as a practical replacement for the more restrictive control of sampling synthesizers now seems promising. The problem of storing such a representation in memory remains, however, as

well as the difficulty of delivering such data to a real-time algorithm swiftly enough (throughput). An effective method of keeping additive synthesis data sets as small as possible is therefore desirable.¹

Fig. 1 provides an illustration of the amplitude portion of an additive synthesis data set, using a small set of values invented for the purpose of demonstration. We can observe in these data a feature characteristic of most additive synthesis data sets, the fact that the amplitude envelopes are highly correlated. Less easy to appreciate in this graphic but nonetheless also characteristic is the correlation of spectra over time, that is, the profile of amplitudes for the 15 partials at a given point in time is highly correlated to many other such profiles in the event. Both kinds of correlations indicate that the data are highly redundant, that is, that the same information is

* Manuscript received 1995 February 13; revised 1995 October 5.

¹ We are assuming here that the advantages of preserving fundamental representation of additive synthesis outweigh the fact that a greater data reduction may be achieved with alternative techniques.

represented multiple times in the event with only minor variations. Attempts to discover just what can be eliminated, while retaining the most desirable aspects of the sound, has been a focus of a number of research efforts.

1 DATA REDUCTION OF ADDITIVE SYNTHESIS DATA

1.1 Previous Research

A number of approaches have been taken to reduce the size of additive synthesis data sets, although most of them can be categorized as either *envelope-reduction* or *wavetable-interpolation* techniques. Both approaches seek to identify where actual values of the original data can be substituted with much simpler, algorithmically generated data such that the resulting loss in detail is perceptually negligible. Envelope reduction techniques treat the data as a set of partials with amplitude envelopes, and approximate the functions of each partial with a series of straight lines and break points. Since only the break points are saved, a great deal of data reduction is obtained. Pioneering research in this area was done by Risset and Mathews [6], who demonstrated that a trumpet tone could be represented in this fashion with no appreciable loss in quality. Later Grey [7] produced a large set of very realistic sounding instrument tones in this manner. However, both projects involved the fitting of line segments completely "by hand" and trial and error. Subsequent research has gone in the direction of seeking a more automated process. Schindler [8] suggested a more hierarchic organization of the problem, whereas Strawn [9] provided a detailed algorithm for arriving at the segments computationally. Charbonneau [10] suggested a reduction by averaging envelopes together, whereas Kleczkowski [11] offered a more formalized version of a similar idea called *group additive synthesis*. Kleczkowski's idea was further expanded by Eaglestone and Oates [12] by the addition of hierarchical clustering analysis on the set of amplitude envelopes.

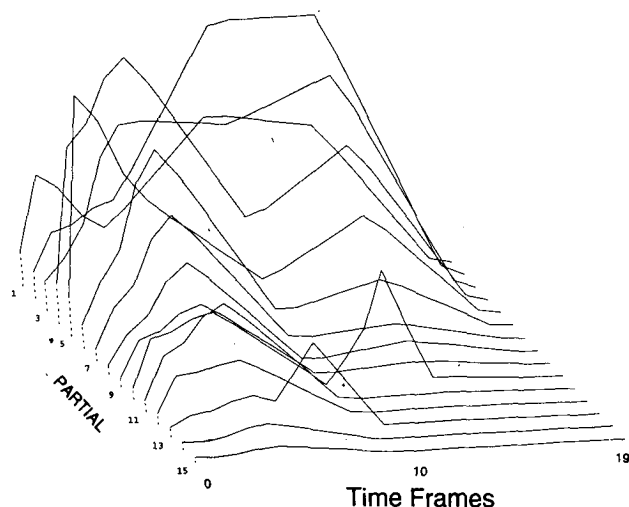


Fig. 1. Amplitude portion of additive synthesis data set (artificially constructed for illustration). Data consist of 15 time-varying amplitude envelopes, one for each partial, each with 20 time frames.

Wavetable-interpolation techniques address the additive synthesis data from another orientation, viewing the data as a set of spectral envelopes, one at each time frame. Here the event is broken up into a series of spectral slices where actual changes can be replaced by smooth transitions from one spectrum to another. Such systems are described in brief reports by Sasaki and Smith [13], Bowler [14], and Schwartz [15], whereas a more extensive and carefully explored wavetable-interpolation scheme is described in Serra, Rubine, and Dannenberg [16], [17]. Horner, Beauchamp, and Haken [18] present a method of reducing the number of wavetables by a genetic algorithm.

While some of these methods take into account the redundancy of the data in the process of selecting what information to discard, none of them identify or act on the redundancy in an ideal manner. Principal components analysis (PCA), a procedure of multivariate statistics [19], is a tool that systematically identifies redundancy and offers a method for redistributing the data such that unnecessary replication is eliminated. In this paper we describe how PCA can be applied to additive synthesis data sets to obtain a data reduction in this manner. We also explore the quality of resynthesis obtainable following such data reduction in two formal perception experiments, and we note some of the possible side benefits that PCA offers for timbre researchers.

1.2 Principal Components Analysis

PCA as a statistical technique was first described in Hotelling [20], who applied it to the scoring of intelligence tests. We shall describe it heuristically here in the context of additive synthesis data. Details for implementing PCA computationally are given in the Appendix. PCA's essential task is to reduce a correlated matrix of data to a set of orthonormal basis functions, which is done by calculating the eigenvectors of the matrix of the correlation coefficients of the original data. This encodes the original data into a form in which the correlated input is redistributed more efficiently into a smaller number of orthogonally related functions. From this data set a likeness of the original data can be constructed. The reconstruction is such that the sum of the mean-square differences between the original and reconstructed data set is systematically minimized, meaning that only a small amount of the original variance is lost. The amount of variance that can be restored is in inverse proportion to the amount of data reduction that is achieved, so a balance may be struck in order to meet the needs of data reduction and accuracy.

Consider a spectrotemporal amplitude matrix $X_{[1:M][1:N]}$ consisting of N partials (in columns) and M time points (in rows). At each time m the N partials can be considered as the coordinates of a point in N -dimensional space. Consider all the times m for only three partials N plotted against each other in a three-dimensional space U the axes of which we call u_1 , u_2 , and u_3 . As time progresses, the point will trace out a pattern, and because the signals are well correlated, this pattern will tend to concentrate in some particular area. We project this

pattern onto the original data, that is, for each time m we calculate $(u_1 * X_{[m][1]}) + (u_2 * X_{[m][2]}) + (u_3 * X_{[m][3]})$. This amounts to a rotation of coordinate axes to a new set of axes Y with columns $y_1, y_2,$ and y_3 . If the rotation is chosen such that the y_1 axis (which is called the principal axis) lies in the direction along which most of the variance in the data matrix is observed, then the x_1 signal will contain most of the information about $u_1, u_2,$ and u_3 . The axes are subsequently rotated to yield y_2 , which accounts for the residual variance remaining from the previous stage; this is continued again for y_3 . The three y values eventually account, in decreasing amounts, for 100% of the variance of X . If X is highly correlated, then y_1 will account for a very large proportion of this variance. Consequently, discarding the x_2 and x_3 signals and reconstructing the data set from x_1 alone will result in a minimum error with respect to the original data set. Thus a likeness of the original data is reproduced from a representation that is a little over one-third the size of the original data set.

PCA yields two matrices of data, called *scores* (or *basis vectors*) and *weights*; the scores and the weights together comprise the *principal components* (PCs). The relationship between the scores, weights, and input data and the values $Y, X,$ and U in the previous explanation is shown in Fig. 2. The scores are the result of a matrix multiplication between the input data and the PC weights. In this example the scores will resemble amplitude trajectories and the weights spectral envelopes. The weights specify by what magnitudes each of the scores must be multiplied in order to arrive at the reconstruction. The principal components are ordered so that the first captures the largest portion of the variance of the population of envelope shapes, whereas subsequent scores capture decreasing amounts of variance. The original data are reproduced perfectly only when all principal components are used (100% of the variance is accounted for). It is assumed, however, that for additive synthesis

data sets, less than 100% variance is necessary to make the tone perceptually indistinguishable from the original, or at least, to make the differences musically acceptable. Later in the paper we shall show just how many PCs are required to achieve these two goals.

At this point it is necessary to introduce a number of terms, summarized in Table 1, which will be used throughout the paper. The fact that the additive synthesis matrix can be in two possible orientations creates two important possibilities for the way in which PCA is applied to it. So far we have described additive synthesis matrices in *temporal orientation*, that is, N partials in columns and M time frames in rows, as in the preceding example. PCA treats the N partial amplitude envelopes as the variables to be measured, or *variates*, and their M time-varying amplitudes as cases, individuals, or *observations* on those measures.² PCA gives as output (1) a set of scores representing the variance of the envelopes as a set of orthogonally related basis functions, and (2) a set of weights giving the multiplying factors to shape the envelopes to the proper spectrum envelope. We call this *temporal PCA*. When the PVA³ matrix is in *spectral orientation*, on the other hand, the M time frames are in columns, and their N partial amplitudes are in rows. Now PCA treats the time frames as the variates and their partial amplitudes as observations. PCA gives as output (1) a set of scores representing the variance of the spectral profiles as a set of orthogonally related basis functions, and (2) a set of weights giving the multiplying factors to shape the spectra to the appropriate amplitude at each point in time. We call this *spectral PCA*.

Figs. 3 and 4 illustrate both types of PCA as applied to the data shown in Fig. 1. The left column of Fig. 3 simply redisplay the amplitude envelopes of each of the partials (here limited to eight to simplify the illustration) of the event shown in Fig. 1. To the right of this, in four columns, are four different reconstructions of the event employing 1, 2, 3, and 4 principal components, respectively. We refer to the practice of reconstructing

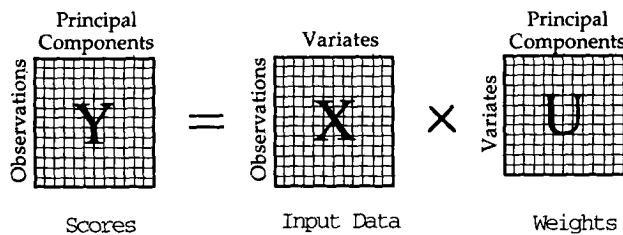


Fig. 2. Matrix-algebra illustration of relationship between input data, scores, and weights.

² The terms *variates* and *observations* are in common currency in statistics as it is practiced in the behavioral sciences. We adopt them here as a syntax for explaining PCA independent of a particular rows and columns orientation. When we do refer to rows and columns, we follow most statisticians and software implementations in associating columns with variates and rows with observations; readers accustomed to the opposite association should read with caution.

³ PVA is an abbreviation of phase vocoder analysis, a form of additive synthesis data.

Table 1. Terms used in application of PCA to additive synthesis data.

Input Matrix	Columns	Rows
PCA input	Variates	Observations
PVA matrix orientation		
Temporal	Partials	Time frames
Spectral	Time frames	Partials
Type of PCA		
Temporal PCA		
(input matrix in temporal orientation)	Scores resemble partial amplitude envelopes	Weights resemble spectra
Spectral PCA		
(input matrix in spectral orientation)	Scores resemble spectra	Weights resemble partial amplitude envelopes

a data set from principal components as *PC resynthesis*. Above each column are the principal components themselves—the scores above and the weights below. The scores consist of 20 discrete points (the number of time frames) and the weights comprise eight points (the number of partials). Consider the first column, the resynthesis with one principal component. The score for this principal component seems to have captured the two “bumps” (one near the attack, one near the decay) that are most apparent in partials 4 through 8. The eight amplitude envelopes are in fact this shape duplicated eight times at various magnitudes according to the eight values given in the weights. We see from the weights that this shape is only weakly weighted for partials 1 through 3, but strongly weighted for partials 4 through 8. The score for the second principal component appears to capture the broader bump (halfway through the tone), characteristic of partials 1 through 3, and we see from its weights that it is emphasized for those three partials. The solution for the second principal component consists, again, of the score duplicated eight times over according to the magnitude of the weights, but added to the solution arrived at for the first principal component. Each successive PC resynthesis is in fact cumulative in this manner. A given principal component can only be factored in once all the principal components below it in number have been factored in. We see that by the third PC already the eight partials have come to resemble the original data set. PCs 3 and 4 add further refinements, perhaps to account for idiosyncrasies of particular partials. For example, principal component 3 seems devoted to sharpening up the early spike in partial 5, whereas principal component 4 seems devoted to putting an indentation midway through partial 3.

Fig. 4 shows the way PCA is applied to the spectral orientation of the data set. Again the original data are shown running down the left column, as a series of spectral slices over time. Each of four resyntheses, employing 1, 2, 3, or 4 principal components, is shown in separate columns to the right. Above this group are the set of PCs associated with each solution. Each of the scores has 15 points (the number of partials) whereas

each of the weights has 20 points (the number of time frames). Consider the first column of the four resyntheses, a resynthesis with one principal component. The score seems to have captured the average spectral shape of the first five partials or so, and neglected partials six and above. The 20 spectra of the resynthesis are in fact this shape duplicated 20 times at various magnitudes according to the 20 values given in the weights. We see from the weights that this shape is strongly weighted at the beginning one-third of the duration of the event, and weakly weighted thereafter, reflecting the fact that partials 1 through 5 rise and fall over this time period. The score for the second principal component appears to capture a similar trend, but applied to partials 3 through 5, and over a different time course (that is, an early, broad bump, as seen in the corresponding weights). The score for PC 3 seems devoted to adjustments to the first and fifth partials during the attack portion of the tone. We see again that by the third PC already the 20 partials have come to resemble the original data set.

We note in these illustrations that PCA calls attention to certain “organizational” features of the tones, which may be less easy to appreciate in their original additive synthesis form. For example, Fig. 3 showed that the event could be characterized by one trend (PC 1) with narrow bumps at the beginning and end of the tones, and another trend (PC 2) with a broad bump in the middle. The correspondence of principle components to such acoustic features suggests that PCA obtains an analysis similar to that provided by Kleczkowski [11]. Strictly speaking, however, the correspondence is fortuitous. For musical tones containing large amounts of spectro-temporal flux the principal components will have to account for the variance of such a large number of features that its organizational properties will be less easily appreciated by eye.

The value of the PCA procedure as a data-reduction tool is perhaps best appreciated by viewing the *eigenvalues* for the two solutions we have just observed (see

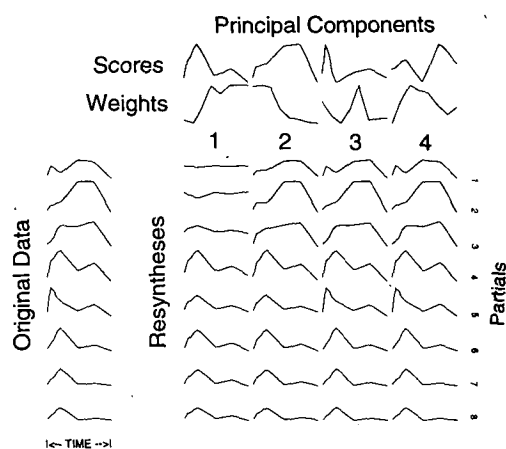


Fig. 3. Illustration of temporal PCA using data set shown in Fig. 1.

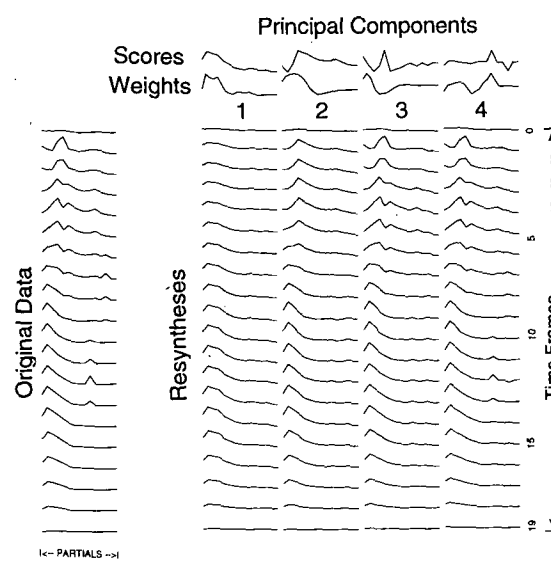


Fig. 4. Illustration of spectral PCA using data set shown in Fig. 1.

Fig. 5), which show the proportion of variance captured by each of the principal components. The fact that, in both, the proportions fall rapidly to small values reflects the fact that the input data are strongly correlated. Eigenvalues that decreased in a gradual linear fashion, on the other hand, would indicate little correlation (and thus little benefit is to be gained by PCA). Both analyses yield 15 principal components, but show that over 99% of the variance of the original event is captured with a representation of roughly one-third the size, that is, the first five principal components (out of 15). We note that spectral PCA captures a larger proportion of variance in the lower numbered principal components. As we shall see later, we generally find this to be the case with musical tones.

Both examples considered only the analysis of the amplitude portion of the additive synthesis data set. Including the frequency matrix in our analysis regime consists simply of running a separate PCA analysis in parallel with the amplitude analysis (they cannot be combined into a single analysis). We will discuss the application of PCA to frequency data later in this paper.

1.3 Previous Uses of PCA

The first application of PCA for audio data reduction appears to have been by Kramer and Mathews [21], who applied it to data from a channel vocoder analysis of speech. A more recent application of PCA for data reduction of speech signals is given in Zahorian and Rotherberg [22]. Beyerbach and Nawab [23] describe a Fourier analysis technique that they term principal short-term Fourier transform (PSTFT). Stautner [24] describes an audio analysis system based on critical-band-spaced auditory filters, which is subsequently data-reduced with PCA. A technique related to PCA, Gramm-Schmidt orthogonalization, has been used in adaptive filtering [25]. Stapleton and Bass [26] modeled musical instruments with another related technique, the Karhunen-Loève transform, although their approach is differ-

ent from ours in that they apply it directly to the time-domain signal rather than a channel-based analysis. PCA has also been used as a statistical tool for musical research in contexts other than the direct representation of audio signals, such as in Li, Hughes, and House [27], Pols, van der Kamp, and Plomp [28], Martens [29], and Kistler and Wightman [30].

To our knowledge the first study to apply PCA to musical tones in additive synthesis form is Laughlin and coworkers [31], [32]. Laughlin applied what we call temporal PCA to an additive synthesislike data matrix (peak-picked partials from a short-term Fourier analysis). Of primary interest was his practice of combining multiple instruments into single analyses by PCA (a practice we also consider later in this paper), and postulating family membership attributes from an examination of the principal components. Laughlin also provided informal evaluations of the quality of tones that were obtained by PCA resynthesis. Analysis and resynthesis of trombones and guitars yielded sounds "surprisingly characteristic of the instruments" [31, p. 86]. For most other instruments, however, he noted "obvious difference[s] in sound quality" and that they "tended to lack subtle attack characteristics of the original digitized sounds" [31, p. 85]. Piano sounds lacked important decay features, and most wind instruments lack critical microfluctuations. As Laughlin noted, however, much of the outcome could be attributed to simplifications his computing setup obliged him to make in the process of sampling and analyzing the tones.⁴ Furthermore, Laug-

⁴ Laughlin's process of sampling the instruments yielded a poor signal-to-noise ratio (48 dB). To make the data sufficiently compact for PCA, the spacing between time frames in the analysis was a rather crude 36 ms, which certainly resulted in the loss of much perceptually critical timbral detail. For similar reasons, all instruments were standardized in length by truncation (hence decays were missing in many cases). Furthermore, the maximum number of harmonics considered was 20, which for some of the lowest fundamentals (86 Hz) resulted in a very narrow frequency range.

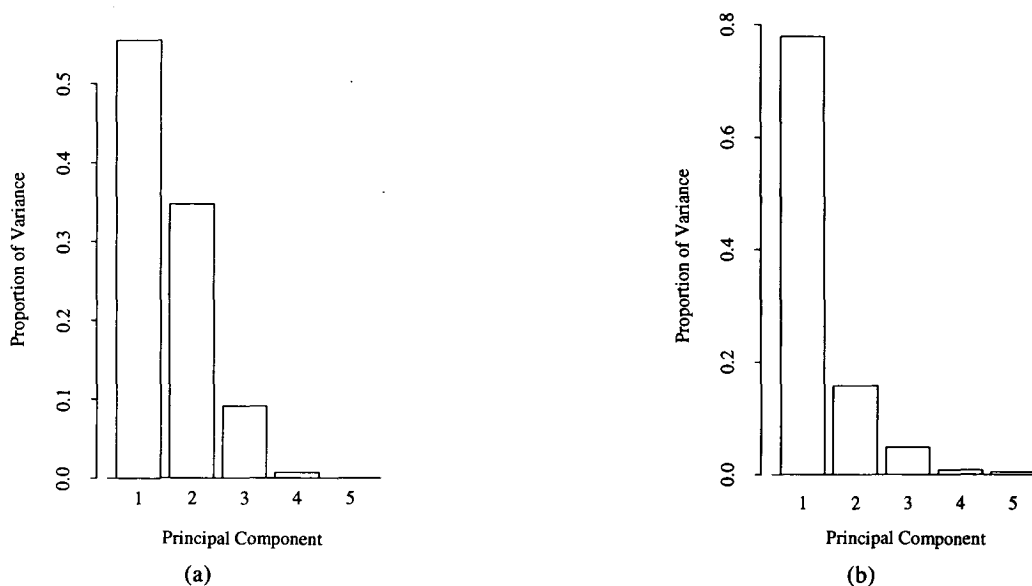


Fig. 5. Eigenvalues for PCA analyses given in Figs. 3 and 4. (a) Temporal PCA. (b) Spectral PCA.

hlin's PCA resyntheses never involved more than five principal components. As the present study will show, five principal components yield tones that are quite easily distinguishable from the original tones.

Sandell and Martens [33] reported on a use of PCA on additive synthesis data sets for data reduction and timbre interpolation, but this research is more extensively reported in the present paper. More recently, Horner, Beauchamp, and Haken [18] compared PCA (in the form that we call *spectral PCA*) to a genetic algorithm (GA) for spectral matching. They describe a number of flaws in the sound of PC-resynthesized tones, although their study reported the use of up to only five principal components, and employed static frequency ratios in the frequency portion of the additive synthesis data.

2 DOWNSAMPLING

We have now introduced the basic features of PCA and how it may be applied to additive synthesis data. Real additive synthesis data sets tend to be much larger than in the illustration shown in Fig. 1. This poses some problem which we shall now describe, and it motivates us to do some preprocessing on additive synthesis data sets which we call *downsampling*.

The calculation of eigenvectors, which is at the heart of the PCA procedure, requires the construction of a square symmetric matrix with as many rows and columns as the number of variates. Since a sound of only a few seconds can have several thousand time frames,⁵ for spectral PCA this matrix may be as large as 20 Mbytes for a 5-s tone. The recursive nature of the eigenvector calculation can not only take very long, even on a very fast computer, but more or less requires that the entire matrix remain in active memory. With the current memory capacity of today's high-performance workstations tending to be between 16 and 32 Mbytes, this can make computational demands impractical. It is therefore desirable to preprocess the additive synthesis data set prior to PC analysis in order to reduce its size.

The goal is to use a method that eliminates any obviously unnecessary detail, such as identifying places where actual data can be replaced by algorithmic processes, so we can leave PCA the sole task of doing what it is good at, eliminating redundancy. We began by applying a simple *wavetable-interpolation* technique to the data—selectively removing groups of frames at points where they could be replaced by interpolation without noticeable injury to the tone. Recognizing that the perceptual relevance of spectral change is of course greatest at attack and decay times and less so during steady-state portions, our approach consists of sampling the additive synthesis data more densely during transient times than steady-state times. To restore the data set to

its original size, for resynthesis we use a cubic spline between frames. We call the two procedures downsampling and upsampling, respectively.

Specifically the procedure works as follows. Suppose we have a PVA data set in spectral orientation with 1500 frames, which we want to reduce to 200 partitions. We find that most instruments require roughly 80 of these 200 partitions to be dedicated to capturing attack transients at the original PVA frame rate (80 was typical for many of the tones we studied, but a more suitable number can be settled upon by examining the tone). During the relatively slowly varying bulk of the tone, partition durations gradually increase to a maximum of roughly 24 frames, and then decrease toward the end of the tone in order to capture the decay at the original frame rate. All partials are reduced according to the same scheme, with the break points occurring at the same times. Fig. 6(a) shows the number of additive synthesis time frames subsumed in each of the partitions, whereas Fig. 6(b) graphically depicts the temporal mapping of the 200 partitions in units of original additive synthesis time frames. A more descriptive name for our approach is variable-duration temporal partitioning (VDTP). In order to shift a downsampled file back to its original temporal resolution, we use a spline to interpolate the necessary number of points between the amplitude values from one partition to another for each partial.

We are of course adopting one data-reduction procedure in order to make it possible to investigate another. This presents no problem so long as we ensure that appropriate events are being compared at evaluation time; that is, if PCA is being performed on downsampled

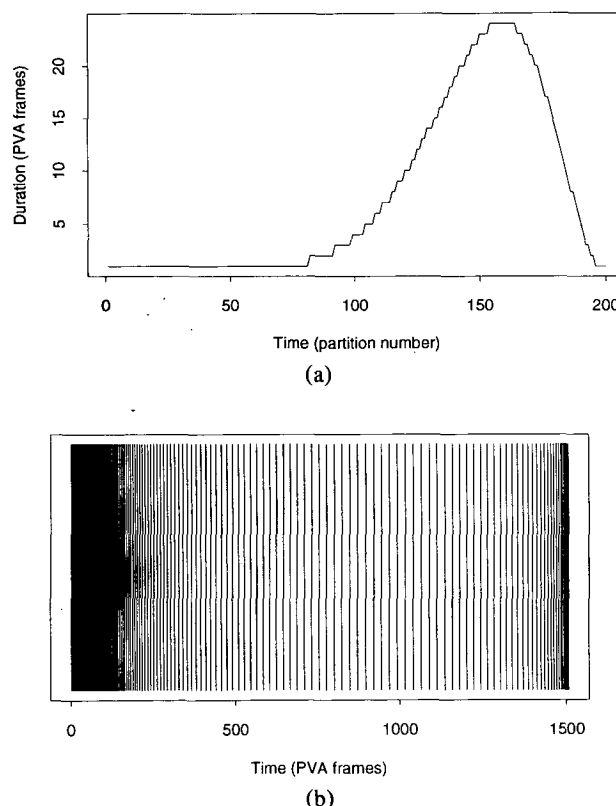


Fig. 6. Illustration of downsampling.

⁵ The usual convention in phase vocoder analysis is to have one time frame for each half-period of the (user-estimated) fundamental of the tone. Thus a phase vocoder analysis of a 5-s note on the pitch of c4 will yield 7110 frames, for example.

data, the perceptual consequences of PCA should be evaluated with respect to resynthesized downsampled sounds rather than the original recorded sounds. Second, to ensure that PCA is being applied to suitably "real" sounds, we want our downsampling procedure to minimally alter the audio quality of the original recording. The quality of resynthesized downsampled tones will be evaluated later in this paper.

3 EVALUATION OF PC-RESYNTHESED TONES

3.1 Instruments

Figs. 7 and 8 give the results of our processing on a musical instrument tone. (The instrument, a cornetto, is a Renaissance precursor to the trumpet.) Fig. 7 shows the original amplitude data for the instrument (first 22 partials). It was downsampled and analyzed with PCA, then reconstructed by PC resynthesis with six principal components and upsampling. Fig. 8 shows the reconstruction. For the first nine partials we see that the reconstruction preserves the shapes quite faithfully, with the

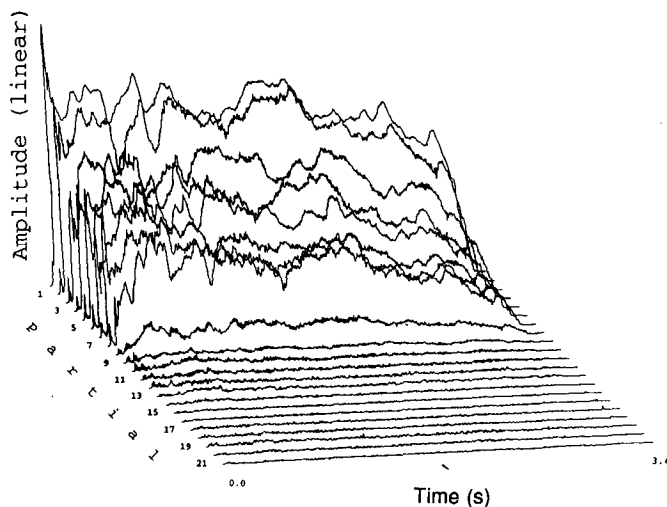


Fig. 7. Amplitude portion of additive synthesis data set for cornetto (Renaissance precursor to trumpet).

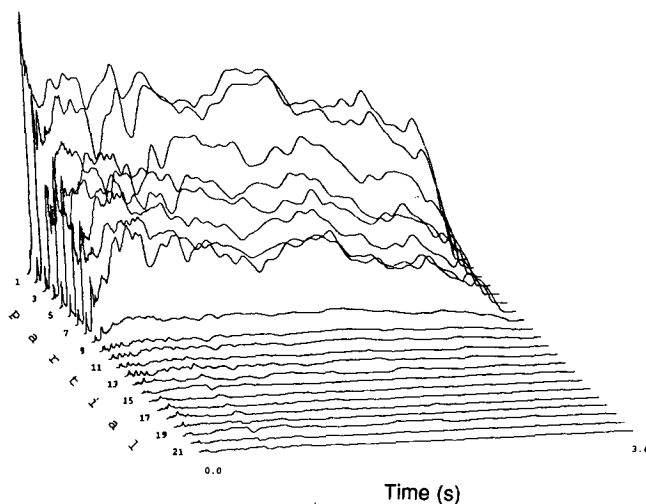


Fig. 8. Reconstruction of data set in Fig. 7 with six principal components.

exception of the frame-to-frame jitter. For the higher partials, jitter seems to be the primary attribute in the original data, and this is lost in the reconstruction. In addition some of the attributes of the lower partials have been inaccurately passed on to the reconstruction of the higher partials. Using more than six principal components would restore more of the original detail, but it is not clear how many are necessary.

We shall now present an evaluation of the aural impact of PCA processing on the downsampled amplitude portions of additive synthesis data sets. Three digitally sampled musical instrument tones, a cello, a clarinet, and a trombone, were analyzed by a phase vocoder. The cello (c3, 1.08 s) was bowed in martelé fashion (swift, sharp attack); it came from the McGill University Master Samples Compact Discs [34], vm. 1, track 15, index 14. The clarinet (g5, 0.764 s) was performed by John Bruce Yeh and recorded at IRCAM. The trombone (Bb3, 0.48 s) performed with a rapid change of a "wah-wah" mute; it came from the Prosonus Sound Library of compact discs [35], brass volume 1, track 15, index 18. Samples were converted from their original sampling rate of 44.1 kHz to 22.05 kHz for analysis. A phase vocoder analysis [36], [37] using the software described in Beauchamp [38] was applied to these samples in order to put them in additive synthesis form, yielding amplitude and frequency matrices for each. For the cello and trombone, the first 22 partials of the additive synthesis data were saved and the rest discarded (meaning a cutoff frequency of about 5.1 kHz), whereas all partials of the clarinet were saved (15 partials, cutoff of 11.76 kHz). Both matrices for each instrument were downsampled, and the downsampled amplitude data were then analyzed by the PCA process. For the reconstruction of a tone from the PCA data, the downsampled data were PCA resynthesized and then upsampled, whereas the (non-PCA-processed) downsampled frequency data were simply upsampled. At this point the full amplitude and frequency spectrotemporal matrices were ready for resynthesis into audio samples.

Our choice of whether to use spectral or temporal PCA has been guided by the fact that spectral PCA tends to obtain a more statistically effective reduction for musical instrument tones than temporal PCA. The number of principal components necessary to capture 99% (an arbitrarily selected value) of the variance for the cello, clarinet, and trombone are 6, 4, and 7 for spectral PCA, and 16, 13, and 11 for temporal PCA. This suggests that musical instrument tones are more correlated in spectrum over time than they are in amplitude shape across partials, and that they will be more effectively analyzed by spectral PCA than by temporal PCA. For this reason we have chosen spectral PCA. But if PCA is used as a means of studying timbre, one may be interested in the information one learns from temporal PCA as well. The question boils down to whether one wants to observe successive refinements in time or in frequency.

Relatively short durations of tones were selected because our experiments would be quite long and difficult if tones of lengthy duration were used. However, we

have used PCA on numerous tones of greater length (between 2 and 4 s) and find the results to be as good as for the short tones discussed here.

3.2 Experiment 1

All three instruments were analyzed with a phase vocoder, downsampled, analyzed with spectral PCA, reconstructed at 31 levels of PC resynthesis (1 PC, 2 PCs, . . . , 31 PCs), upsampled, and resynthesized into mono sound files with additive synthesis. We wished to see if PCA-resynthesized tones could produce sounds that were perceptually identical to their unprocessed counterparts, and what number of principal components attained such success.⁶ In order to test perceptual identity in as ideal a manner as possible, we used a three-interval, two-alternative forced-choice paradigm with feedback. For a given instrument *A*, the first interval always consisted of *A* in downsampled resynthesis form. The second and third intervals consisted of the same downsampled *A*, and *A* resynthesized with a particular number of PCs. However, the order of the latter two was scrambled from trial to trial. The subject identified the odd one out, that is, which of the latter two sounds was different from the first. Thus the answer was correct if they identified the PC-resynthesized tone, incorrect if they identified the downsampled tone. Intervals were separated from each other with 500-ms silence.

The sounds were played diotically to listeners using a Silicon Graphics Indy computer via a pair of Sennheiser HD-414 SL headphones inside a IAC sound isolation booth. Subjects faced the Indy display screen (placed outside the booth and viewed through a window), which displayed a simulated response box with response buttons corresponding to the second and third intervals. Positioned above each of the two response buttons were lights that flashed in synchrony with the presentation of the second and third sounds. Listeners gave their answers by pressing one of the two response buttons with the computer mouse. After each answer, the listener was given feedback on the correctness of the response—a light below the response buttons flashed green if the answer was correct, red if the answer was false. All the trials ($N = 50$) for a particular level of PC resynthesis were presented in a single block rather than scrambled together with other levels. By these means (feedback and blocking) the listener was able to quickly reach maximally correct performance. The blocks were in no particular order. (For example, the subject might rate all 50 trials for the clarinet with 10 principal components, followed by all 50 trials for the clarinet with 2 principal components.) The set of blocks of trials for each instrument were all presented together. For example, subjects heard all the clarinet blocks before going on to the trombone blocks.

⁶ We reiterate that the PCA resyntheses are of downsampled tones with limited numbers of harmonics, and in these experiments we compare them to equivalent tones without PCA processing. The original recordings of the instruments are not involved in the experiment.

Three subjects participated, RH, SH, and SA, all male and having normal hearing. None were musicians, although RH had training as an audio recording engineer.

Each instrument was downsampled to 200 time frames. This means that the number of variates in our spectral PCA analysis is 200. The maximum number of principal components that can be employed for clarinet, trombone, and cello while still achieving a data reduction relative to the original downsampled data are 12, 19, and 19, respectively.⁷ For the success of PCA we hope to find that indistinguishability is achieved with this number of principal components or fewer.

Fig. 9 shows the results. We regard tones as indistinguishable when the proportion of correct values fall within the area marked off by dotted lines (that is, $50\% \pm$ a standard deviation). For the clarinet and cello this point is obtained just at the border for data reduction (12 and 19 principal components, respectively), meaning that the amount of data required for resynthesis is of the same size as the original data. For the trombone the point is obtained where 50% data reduction can be achieved. However, we note that for the trombone the function is not declining steadily, and that for two of the subjects the tone paradoxically appears to be worse for a few higher order PC resyntheses.

The sorts of flaws that made the tones identifiable as resyntheses differed from instrument to instrument and changed over successive principal components. Generally the same flaw persisted in the instrument with each added principal component, becoming increasingly subtle until the flaw could no longer be detected. For the clarinet a repetitive "popping" sound would occur during the steady state. For the cello, a "boing" (similar to the sound of a spring) would appear for a brief moment during the decay of the tone. The trombone simply sounded harsh and distorted at its loudest point. These observations are consistent with the kinds of flaws noted by Horner, Beauchamp, and Haken [18].

The results show that only one of the three tones could meet the requirement of obtaining a resynthesis that was indistinguishable from its original while still obtaining a data reduction. We do not find this particularly surprising since listeners were tested in an unusually ideal listening situation, uncharacteristic of a reverberant, real-world musical situation with no feedback on the correctness of answers. In fact, for the higher orders of PC resynthesis the flaws that enabled detection of a difference were nearly unnoticeable, and according to subjects' reports, only through having learned "where to

⁷ Comparisons between sizes of original data and PCA data sets (using spectral PCA, and considering just the amplitude portion) are made using the following formulas. For a given PCA resynthesis of a tone with NPC principal components, only a portion of the full PCA data will be required; the number of values will be $2NO + (NPC * NV) + (NPC * NO)$, where NV is the number of variates, and NO the number of observations. The $2NO$ term refers to the means and standard deviations of the variates which are required for the reconstruction. The number of values of the additive synthesis data set is $NV * NO$. We assume the storage size of each value (such as float, double, long integer) to be the same for both data sets.

look” for them were they able to perform so well. Although an instrument’s characteristic flaw might persist into the higher numbered PC resyntheses, it shrunk to a barely detectable level much earlier. From an experimental point of view the sounds were distinguishable, but from a musical point of view the flaw enabling that was insignificant. For this reason, a second experiment was designed to make an evaluation relating more to real musical criteria.

3.3 Experiment 2

The same stimuli as in experiment 1 were used in a quality judgment task to discover how many principal components were necessary to obtain a synthesis without “noticeable” artifacts. The procedure scrambled trials and eliminated the feedback procedure that was used in experiment 1 in order to eliminate some of the cues that enhanced listener performance. Experiment 2 was performed six weeks after experiment 1, with the same three subjects.

For a given instrument *A* the presentation consisted of only two intervals. The first interval was always the downsampled version of *A*, and the second interval was a PC resynthesis. Listeners rated the second tone for its faithfulness according to the following five categories:

- 1) Identical
- 2) Suspect a difference
- 3) Subtle difference (barely noticeable)
- 4) Difference (but might be tolerable)
- 5) Serious flaw in tone

On each trial they were told to notice first of all whether they could detect a difference and to respond with category 1 if they were certain, with category 2 if less certain. If they did notice a difference, but it was subtle to the point where they would have failed to hear it unless they were “looking for it,” they were to select category 3. They were encouraged to be conservative in their criteria for selecting this category. For category 4 an example was given of what would be considered “tolerable”: if they felt the flaw would be rendered negligible or even undetectable in the context of normal musical presentation (such as over loudspeakers in a normally reverberant room). If the flaw was severe enough to injure the identity of the instrument in any context, they should answer with category 5.

Each level of PC resynthesis for each instrument was played and rated a total of 10 times. Instruments were presented in blocks (all clarinets before any trombones, and so on), but in contrast to experiment 1, the levels of PC resynthesis and their 10 repetitions were completely scrambled.

Fig. 10 shows the results. First note the considerable agreement among the listeners in spite of the subjectivity of the rating scale, which strengthens the interpretability of the categories. Second, by comparing equivalent data between the two experiments, a decreased sensitivity to flaws is found in experiment 2, which we attribute to the absence of both feedback and blocking of levels of PC resynthesis. For example, the highest order of PC resynthesis that is distinguishable from its “original” in

experiment 1 is 19 for cello, 13 for clarinet, and eight for trombone. In experiment 2 the equivalent PC levels (the highest order of PC resynthesis for which they were certain the tones were different, category 3) fall to 14, 11 and three, respectively.

We evaluate the success of PCA as a data-reduction technique in terms of the amount of data reduction achieved by the lowest order of principal components that attains a rating of category 3. With this as our criterion, PCA is successful at achieving a significant

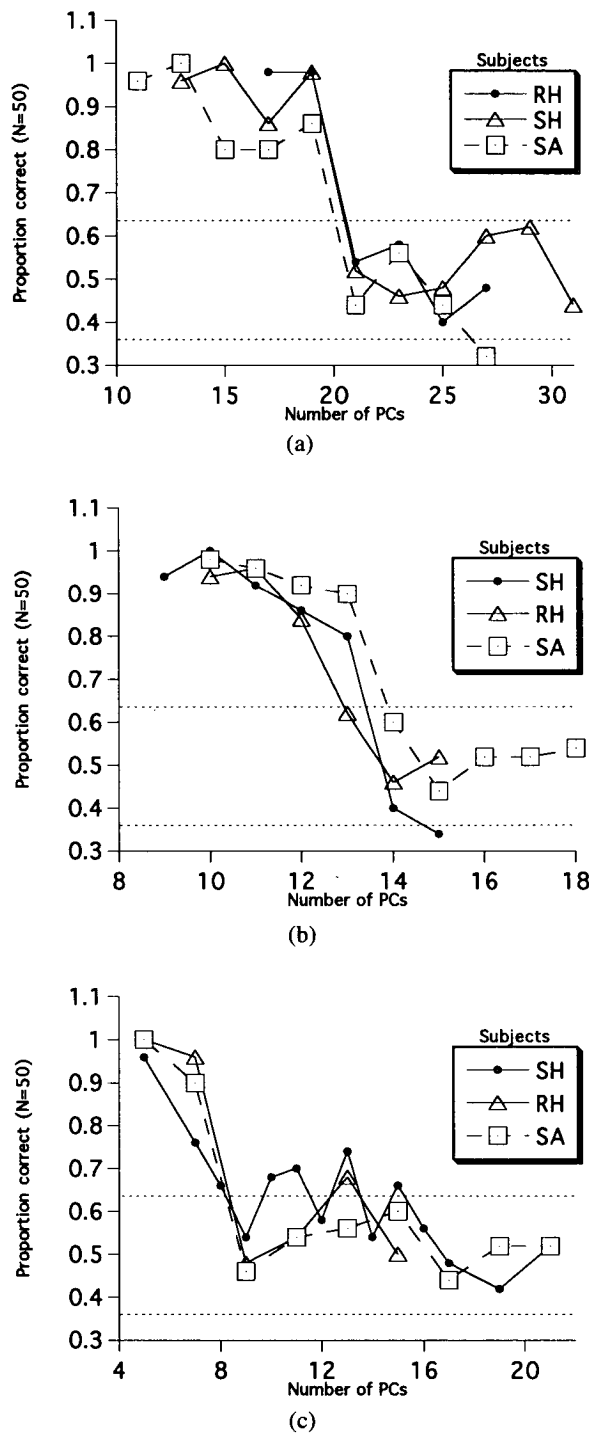


Fig. 9. Scores for distinguishing between downsampled/resynthesized tones and various levels of PC resynthesis. Data are for three listeners (SH, RH, SA). (a) Cello. (b) Clarinet. (c) Trombone.

data reduction. Fig. 11 shows for both experiments 1 and 2, averaged over the three listeners, the minimum number of principal components needed to obtain indistinguishable resyntheses ("diff") and the quality category 3 judgments. Fig. 12 shows the data reduction achieved for each of these data points. We distinguish between the contribution of PCA (comparing the size of the PCA data to the full downsampled data set) and the contribution of PCA and downsampling together (comparing the size of the PCA data to the original data prior to downsampling). With a distinguishability criterion, PCA alone achieves a data reduction only for the trom-

bone, but with a quality criterion, data reduction is achieved for all three instruments. Taking into account the contribution of downsampling, data reduction is achieved for all three instruments using both criteria.⁸

Ratings of both the quality and the distinguishability of the downsampled version of the tones (made during the course of experiments 1 and 2) are given in Fig. 13. In both cases they were compared to an equivalent tone: a 15- (clarinet) or 22-partial tone (cello and trombone) resynthesized from the additive synthesis data set at a sample rate of 22.05 kHz. Downsampled cello and trombone were indistinguishable from their nondownsampled counterparts in both experiments. The downsampled clarinet was distinguishable from its nondownsampled version, but it did receive a good quality rating.

4 FREQUENCY DATA AND PCA

Thus far we have only considered modeling the amplitude portions of additive synthesis data with PCA and resynthesizing sounds using the original frequency information. Our research efforts have shown that while additive synthesis frequency data sets survive downsampling rather well, the frequency data in general (that is, whether downsampled or not) pose special problems to our PCA approach. The problem seems to be rapid, wide frequency excursions with inharmonic ratios when partial amplitudes are low, or during attack portions of the tone, and seemingly random and inharmonic microjitter (about 0.5% of their nominal frequency values) during steady-state portions of the tone. Both seem attributable to noisy aspects of the sound, either in the instrument itself or in the recording environment, or to leakage between filters in the phase vocoder analysis process. This fluctuation cannot be removed without changing the sound noticeably, so the information is relevant. A glance at the eigenvalues for a particular tone (Fig. 14) supports the obvious conclusion that such

⁸ The reason the clarinet data reduction is so much higher than other instruments when PCA is combined with downsampling is because of the clarinet's higher pitch. Because its fundamental frequency was higher, there were more frames in the additive synthesis data set, so when these were reduced to 200 in downsampling, a significant data reduction was achieved.

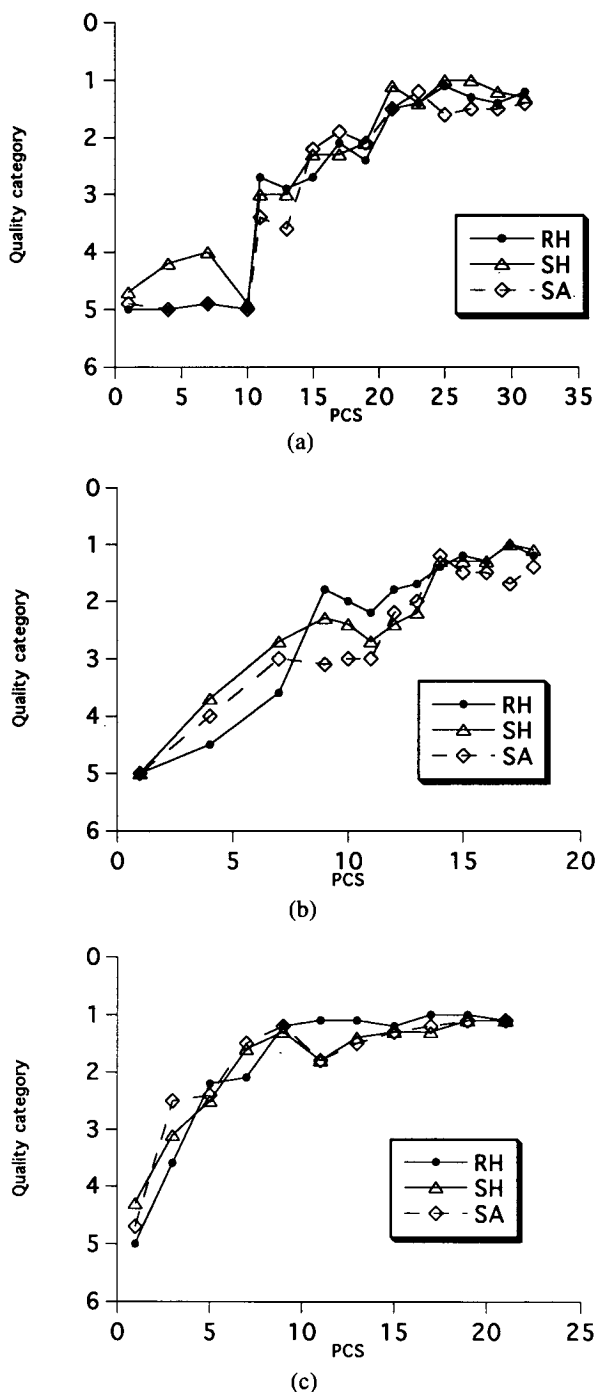


Fig. 10. Quality judgments for comparisons between downsampled/resynthesized tones and various levels of PC resynthesis. Data are for three listeners (SH, RH, SA). (a) Cello. (b) Clarinet. (c) Trombone.

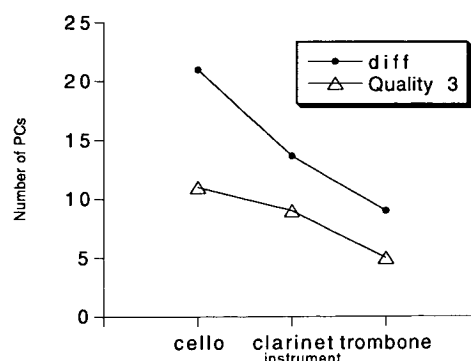


Fig. 11. Summary of two experiments, averaged over three listeners, with reference points for achieved data reduction.

features are highly uncorrelated, and thus resistant to the data reduction offered by PCA. That is, while PCA will still be able to produce satisfactory sounding resyntheses, it will be at the expense of having to include so many principal components that no data reduction is achieved. This problem is equally apparent in both spectral and temporal PCA.

We have experimented with a number of possible solutions to this problem. One possible solution is to coerce the frequency trajectories of higher frequency partials to parallel the trajectory of the fundamental, either using perfectly harmonic ratios or using a set of inharmonic frequency ratios derived statistically from the sound. We have found that both approaches yield sounds that are flat and artificial sounding, so we consider this strategy unacceptable.⁹ Rescaling the frequencies to values proportional to their nominal frequency (so that each variate had the same range of values) provided no improvement in sound quality. Another strategy involved viewing the microjitter as random, removing it prior to PCA analysis and then restoring it following PC resynthesis by inserting appropriately scaled Gaussian noise into the frequency functions. Some sounds resynthesized from such data sets were satisfactory, suggesting that further research in this area is warranted. We consider even more promising the notion of using an approach that systematically separates "sinusoidal" and "stochastic" parts of sounds, as in the work of Serra [39], and submitting this to PCA. We speculate that some solution of this sort can be found that can be used in parallel with the more successful PCA on amplitude data.

We note that the frequency issue has plagued many

⁹ Charbonneau's study [10] is often cited as showing that frequency data can be reduced by substituting a single frequency function for all the harmonics with no significant perceptual consequences. His data actually show that identical or near-identical judgments (when comparing normal and reduced tones) occurred on an average of only 36% of the time; moreover the sounds were presented over loudspeakers in a room. Our own experience with such transformations (including the same stimuli as Charbonneau's) confirm that the effect is quite noticeable.

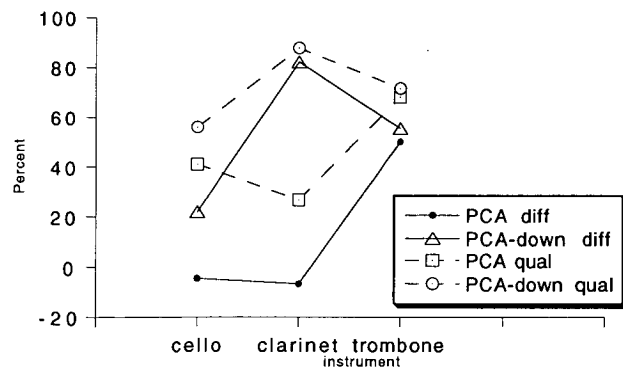


Fig. 12. Amount of data reduction achieved by number of principal components meeting minimum criteria for distinguishability (diff, experiment 1) and quality category 3 (qual, experiment 2). PCA-diff—measurement with respect to total reduction achieved by PCA and downsampling together; PCA—reduction obtained by PCA alone.

other researchers in the area of additive synthesis data reduction and that several have sidestepped it by adopting the strategy we rejected earlier, of coercing the frequencies to static integer ratios [10], [13]–[18], [26], [31]. Except in the case of a restricted group of instruments (certain wind instruments) only a "likeness" of the original tone will be achieved with static frequency ratios. We prefer instead to concentrate on PCA's effectiveness at modeling amplitude data and evaluate this in the context of tones that sound as natural as possible, so we have chosen to use relatively unaltered frequency data.

5 MULTIPLE-TONE PCA AND TIMBRE SPACES

One of the side benefits of downsampling is that it puts tones of widely varying duration into a common domain, a matrix of 200 columns in our case. This allows us to pursue a more advanced application in which multiple tones are analyzed with a single PCA. For example, three tones, each with 22 partials and 200 partitions,

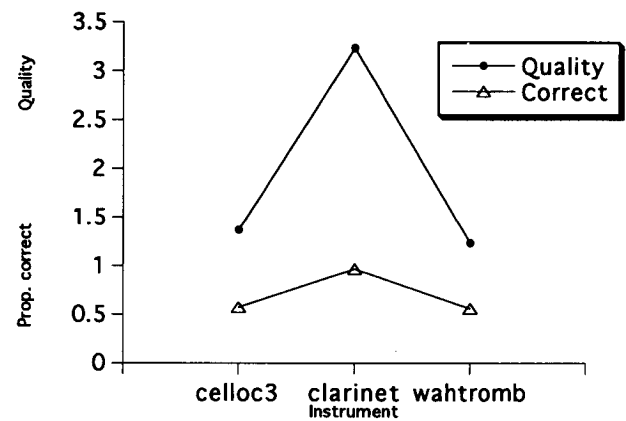


Fig. 13. Evaluation of downsampled tones using distinguishability and quality rating tasks of experiments 1 and 2. y axis has dual function, showing proportion correct (0.0 : 1.0) in lower half, quality scale (1 : 5) in upper half.

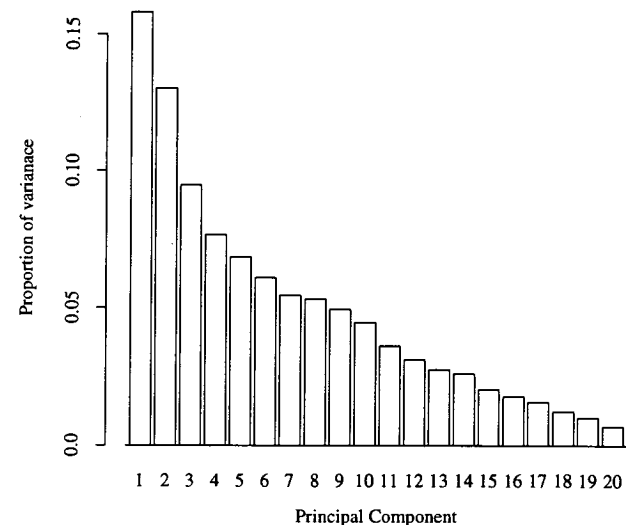


Fig. 14. Eigenvalues for frequency portion of additive synthesis data set for cello used in experiments 1 and 2.

could be concatenated to make 22 partials with 600 frames. Like any PC analysis, the original data (that is, all three tones) will be reproduced perfectly only when using all principal components. More interesting to consider, however, is the sounds resulting from using a smaller number of principal components. If three tones are concatenated, analyzed, and PC resynthesized with a small number of principal components, the three tones (when deconcatenated) will share features with one another. We can capitalize on this feature and use it to implement a *timbre space*, such as described in Grey [7] and Wessel [40]. In a timbre space we wish to make it possible (1) to portray visually the similarities between a set of tones in two- or three-dimensional space and (2) to extrapolate plausible sounding syntheses of tones that are inbetween points in the space, that is, make timbre interpolations.

We explored this by combining PCA with methods from multivariate analysis of variance (MANOVA; see Harris [19]). We begin by generating a prototypical timbre from a set of timbres using a PCA that finds weighting functions only for what the set of timbres has in common. The second part is the interpolation between those timbres using a PCA that finds weighting functions only for what distinguishes between the timbres. The prototype weighting functions were generated by finding the eigenvectors of the pooled within-group sums of squares and cross products (SSCP) matrix. The interpolation weighting functions were based on the between-group SSCP matrix. The within-group and between-group SSCP matrices sum to equal the total SSCP matrix for all the additive synthesis data sets submitted. Hence the information for complete reconstruction of the original data set is not lost in these operations.

By isolating what all the tones' data sets have in common, we can generate a new data set that includes aspects of information from all the data sets but emphasizes no single tone in particular. When resynthesized, the sound can be likened to a "prototype" for that group of instruments. When several different horns are analyzed, for example, the resulting prototype sounds like a bland, generic horn sound. This result is perhaps not musically useful, but if we subtract the prototype matrix from all the individual instrument matrices, it becomes the origin of a multidimensional coordinate system within which each instrument is located at a unique point relative to that origin. This has the musically useful result of setting up a control structure for timbral interpolations. To reduce the dimensionality of the interpolation coordinate system, the deviation matrices were scored on the interpolation weighting functions. Using only the first three principal components for the deviation scores creates a coordinate system that can be easily explored.

The primary source of information for interpolation is in the principal components for deviation scores. Once the prototype is generated, it becomes a "center of gravity" for the space within which interpolation takes place. Note that a prototypical timbre could have been synthesized from a simple average of all the additive synthesis data matrices in the set of timbres, but we found that

this result had objectionable idiosyncrasies that did not appear in the more idealized PC-based prototype. Since the interpolation space is defined by three principal components, there are multiple paths between each of the analyzed timbres. We can take the shortest path between two timbres, or follow a piecewise linear path that changes values on one dimension at a time.

Like any arithmetic scheme for timbre interpolation, what is obtained is only the approximation of a psychologically and musically real interpolation. It is hard to predict how plausible any interpolation (or prototype, for that matter) will sound, but a careful selection of instruments can help ensure better results. As the number and variety of instruments included in a single PC analysis grows larger, the less the principal components will relate to anything resembling any of the given tones, or even anything musically plausible. The syntheses begin to sound noisy, because the principal components themselves have become more noiselike. Thus the instruments for a particular timbre space should be selected with a bias toward optimizing commonality of features.

Hilmar Thordarson, a composer in residence with us at the Center for New Music and Audio Technologies (CNMAT) at the University of California at Berkeley, explored the sounds produced by this method with us, and used them in a composition that was played to an audience at the 1992 International Computer Music Conference. As expected, more musically satisfactory results were obtained with certain groups of instruments than with others. A group consisting of alto flute, cornetto, trumpet, trombone, and English horn yielded plausible and interesting sounding interpolations. A group of brass (French horn, trumpet, trombone, tuba), however, was too homogeneous, making all the points in the space sound too much like one another. When the contrast was too high and the common features were very few, the interpolations sounded highly distorted and implausible. A mixture of a timpani with some wind and string instruments gave very noisy results because the timpani's decay patterns were so unlike any of the sustaining instruments. It appears that the most successful combinations of instruments are those that strike a balance between variety and homogeneity.

6 DISCUSSION

6.1 Attractive Features of PCA

One of the most powerful features of PCA is the opportunity for highly accurate resynthesis while achieving a data reduction. Under stringent evaluative criteria, flaws arising from the process can be heard out. These flaws seem to arise from singular, idiosyncratic (and usually brief in time) features that are correlated with none of the other properties of the tone. Hence many principal components are needed to completely account for them. They are subtle, however. Using more musically realistic evaluative criteria, such that flaws are negligible, a considerable reduction can be achieved. For applications with much laxer criteria (low-cost syn-

thesizers, arcade games, or for applications with competing sounds present) tolerance for such flaws may be even higher, and an extremely high data reduction could be achieved. In either case, artifacts tend to occur in a fixed location of the tone, and preprocessing or cosmetic postprocessing could eliminate them entirely.

On the whole we are rather pleased with the variety of instruments that can be analyzed and resynthesized with PCA while maintaining a high quality. We have obtained very good results with nearly all wind instruments. Instruments that are somewhat resistant to efficient PC resynthesis, however, are strings and flutes. This is probably a consequence of their characteristic noise components during the steady state. In a partial-based analysis system all sound is represented as sinusoids, and noise is simulated by spectral flux, or uncorrelated behavior among partials. From a PCA perspective, this means that it will take a large number of principal components to capture an acceptable portion of the variance for quality resynthesis. Another problem occurs with freely decaying strings, such as in pizzicato or after the bow leaves the string. The absence of the periodic driving force also leads to incoherent spectral fluctuations.

The fact that the underlying representation of PCA is lossless (that is, identity resynthesis can be obtained when using all principal components) is a compelling and elegant feature of PCA. In an actual data-reduction context the degree of loss can be roughly quantified in terms of variance accounted for (although it must always be at least 99% for high-quality synthesis). Thus there is some degree of predictability between the quality obtained and the associated data reduction. Also compelling is that this whole process can be completely automated, without any user intervention. No burden need be placed on the user to decide what level of detail is important.

PCA also offers timbre researchers some insight into the makeup of musical tones. One measure that is needed by timbre researchers is a measure of spectral flux. The eigenvalues of a spectral PCA analysis provide a kind of estimate of this by showing the number of principal components that are required for a given proportion of variance. If 10 principal components are needed to reach 99% of the variance, the event is probably very complex. Similarly, temporal PCA would offer a measure of harmonic independence.

The scores sometimes contain features corresponding to primary timbral attributes such as the global amplitude envelope, the vibrato, or attack-time "blips." This can be very useful for separating essential acoustical features from one another and studying them in detail. This and the orthogonal nature of the principal components suggest the promise for attaining one of the central goals of an ideal analysis-by-synthesis system— independent controls for primary features of the sounds, or "perceptual knobs"—factors that manipulate brightness, vibrato, bite, and so on. Unfortunately, however, the essentially statistical basis of PCA precludes the possibility that the scores might correspond directly to

perceptually salient features of tones. Higher order principal components act as statistically "corrective" features to the features neglected by previous principal components, and the complexity of their function cannot be manipulated in any intuitive way without upsetting the delicate relationships between them. We experimented with manipulating the principal components in order to alter the quality of sound and were unable to produce anything other than distortions of unlikely musical value.

It is not clear why the data for the trombone in experiment 1 did not improve monotonically with the number of principal components for two of the subjects [see Fig. 9(c)]. The accuracy for the resyntheses in question (principal components 8–14) never exceeded 73%, so it would appear to be a phenomenon occurring at the fringe. Nonetheless we note that the quality judgments also show a dip from "identical" to "possibly identical" around the same principal component, 11 [see Fig. 10(c)]. It remains somewhat of a mystery how accounting for greater variance could introduce more inaccuracies.

Finally, we point out that our study only evaluated the perceptual consequences of spectral PCA. A side-by-side comparison of the perceptual consequences of both temporal and spectral PCA processing tones is warranted and would make a worthy topic for further research.

6.2 Downsampling

Although variable-duration temporal partitioning (VDTP) was originally conceived out of necessity, we discovered that it could be developed into a useful tool in its own right. We have found that a variety of instruments can be downsampled and then resynthesized while maintaining a high quality. Instruments with particularly long attacks or decays must be handled carefully, but so long as the user chooses an appropriate number of partitions (more than 200 may be necessary for long tones) and a sensible windowing strategy (allotting a suitable number of partitions to single frames), nearly any instrument should survive downsampling well. For instruments with large amounts of spectral flux, however, the process of summarizing over groups of frames can lead to a noticeable loss of quality. (See the earlier discussion of the frailty of PCA with respect to decaying strings.)

Our VDTP approach is useful for an additional reason not mentioned earlier: to make musically useful temporal transformations. One of the popular features of additive synthesis is the separation of temporal features from spectral ones, making it possible to resynthesize the tones at durations other than their original without changing pitch or spectral energy distribution. One may transform a 1/2-s tone into a 2-s one or vice versa (see Dolson [37, p. 23]). However, as researchers in timbre know, in real performed notes the attack, sustain, and decay portions of tones do not change according to the same scale. For example, a tone that is lengthened will have a much elongated steady state, but its attack will

change only slightly. The VDTP approach, however, provides the possibility for natural sounding time expansion and contracting by preserving the transients in a natural way and adjusting only the steady portions. If the downsampling partitioning has been chosen carefully for the tone in question, the user would be able to control this quite sensitively.

The total data reduction obtained by the combined results of downsampling and PCA is quite impressive, as high as 82% in the case of the clarinet (see Fig. 12). This suggests that further refinement of the downsampling approach, although outside the scope of the current study, is a worthy topic for further study. An obvious candidate for further improvement would be to use a VDTP rate that would be sensitive to the specific nature of the tone, that is, to increase and decrease in response to the actual variability of the tone over time. The pattern shown in Fig. 6 is one that worked well for most of the instruments we investigated, but it is only one possible pattern. One way to do this would be to employ an error minimization process, as in the work of Serra, Rubine, and Dannenberg [16], [17]. Stapleton and Bass [26] also describe a scheme that is sensitive to variability by following changes in phase in the time-domain waveform. Another possibility, as used by Stautner [24], would be to use time adaptation for high partials, or downsampling time differently for low partials (more coarse) than high partials (more fine). Furthermore, auditory models could be used to limit the number of channels that are analyzed. It might be feasible to average several partials into a single channel if their frequencies are unresolved by the auditory system.

7 SUMMARY

The benefits of using principal components analysis (PCA) for the representation of additive synthesis data sets (partial-based, time-variant representations of musical instrument tones) have been considered. PCA recasts data into a set of basis vectors, or scores, ordered according to the amount of variance accounted for. Thus a likeness of the data can then be reconstructed using a fraction of the data, making it useful as a data-reduction tool. The minimum number that can be used without introducing noticeable differences between the original and a resynthesis was explored in two perception experiments. Nearly identical tones can be attained with a 40–70% data reduction. We outlined the particulars of applying PCA to additive synthesis data, including the advantages of alternative approaches (temporal PCA versus spectral PCA). We also introduced a preprocessing step of variable-duration temporal partitioning (VDTP), which allows for natural-sounding time expansion and contracting of tones, and a method for timbre interpolation.

8 ACKNOWLEDGMENT

Support for G. J. Sandell was provided by the Acoustical Society of America (F. V. Hunt Fellowship), a

research grant in Experimental and Computational Studies of Human Cognition (Science and Engineering Research Council, UK), and research grant 5 P01 DC 00293-11 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. Research was carried out at CNMAT (Center for New Music and Audio Technologies, University of California, Berkeley) and Sussex University (UK). The authors' thanks go to David Wessel, James Beauchamp, Xavier Rodet, John Strawn, Suzanne Winsberg, Chris Plack, Andrew Horner, Brian Link, and two anonymous reviewers for their valuable advice, criticism, and assistance.

9 REFERENCES

- [1] J. C. Risset and D. L. Wessel, "Exploration of Timbre by Analysis and Synthesis," in D. Deutsch, Ed., *The Psychology of Music* (Academic Press, New York, 1982).
- [2] J. A. Moorer, "Signal Processing Aspects of Computer Music: A Survey," in J. Strawn, Ed., *Digital Audio Signal Processing: An Anthology* (W. Kaufmann, Los Altos, CA, 1985).
- [3] J. O. Smith, "Viewpoints on the History of Digital Synthesis," in *Proc. 1991 Int. Computer Music Conf.*, pp. 1–10.
- [4] X. Rodet and P. Depalle, "Spectral Envelopes and Inverse FFT Synthesis," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1046 (1992 Dec.), preprint 3393.
- [5] A. Freed, X. Rodet, and P. Depalle, "Synthesis and Control of Hundreds of Sinusoidal Partial on a Desktop Computer without Custom Hardware," in *Proc. ICSPAT* (San Diego, CA, 1993 Oct.); also, *Proc. 1993 Int. Computer Music Conf.* (Tokyo, Japan), pp. 98–101.
- [6] J. C. Risset and M. V. Mathews, "Analysis of Musical-Instrument Tones," *Phys. Today*, vol. 22, no. 2, pp. 23–30 (1969 Feb.).
- [7] J. M. Grey, "An Exploration of Musical Timbre," Ph.D. thesis, Music Department Rep. Stan-M-2, Stanford University, Stanford, CA (1975).
- [8] K. W. Schindler, "Dynamic Timbre Control for Real-Time Digital Synthesis," *Computer Music J.*, vol. 8, no. 1, pp. 28–42 (1984).
- [9] J. Strawn, "Approximation and Syntactic Analysis of Amplitude and Frequency Functions for Digital Sound Synthesis," *Computer Music J.*, vol. 4, no. 3, pp. 3–24 (1980).
- [10] G. R. Charbonneau, "Timbre and the Perceptual Effects of Three Types of Data Reduction," *Computer Music J.*, vol. 5, no. 2, pp. 10–19 (1981).
- [11] P. Kleczkowski, "Group Additive Synthesis," *Computer Music J.*, vol. 13, no. 1, pp. 12–20 (1989).
- [12] B. Eaglestone and S. Oates, "Analytical Tools for Group Additive Synthesis," in *Proc. 1990 Int. Computer Music Conf.*, pp. 66–68.
- [13] L. H. Sasaki and K. C. Smith, "A Simple Data Reduction Scheme for Additive Synthesis," *Computer*

Music J., vol. 4, no. 1, pp. 22–24 (1980).

[14] I. Bowler, "The Synthesis of Complex Audio Spectra by Cheating Quite a Lot," in *Proc. 1983 Int. Computer Music Conf.*, pp. 79–84.

[15] G. Schwartz, "Sound Synthesis by Hierarchic Sampling," in *Proc. 1985 Int. Computer Music Conf.*, pp. 33–38.

[16] M. H. Serra, D. Rubine, and R. Dannenberg, "A Comprehensive Study of Analysis and Synthesis of Tones by Spectral Interpolation," Tech. Rep. CMU-CS-88-146, Carnegie-Mellon University Computer Science Dept., Pittsburgh, PA (1988).

[17] M. H. Serra, D. Rubine, and R. Dannenberg, "Analysis and Synthesis of Tones by Spectral Interpolation," *J. Audio Eng. Soc.*, vol. 38, pp. 111–128 (1990 Mar.).

[18] A. Horner, J. Beauchamp, and L. Haken, "Methods for Multiple Wavetable Synthesis of Musical Instrument Tones," *J. Audio Eng. Soc.*, vol. 41, pp. 336–356 (1993 May).

[19] R. J. Harris, *A Primer of Multivariate Statistics*, 2nd ed. (Academic Press, New York, 1985).

[20] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *J. Educ. Psychol.*, vol. 24, pp. 417–441, 498–520 (1933).

[21] H. P. Kramer, and M. V. Mathews, "A Linear Coding for Transmitting a Set of Correlated Signals," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 41–46 (1956 Sept.).

[22] S. A. Zahorian and M. Rothenberg, "Principal-Components Analysis for Low-Redundancy Encoding of Speech Spectra," *J. Acoust. Soc. Am.*, vol. 69, pp. 832–845 (1981).

[23] D. Beyerbach and H. Nawab, "Principal Components Analysis of the Short-Time Fourier Transform," in *Proc. 1991 Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, "Digital Signal Processing," pp. 1725–1728.

[24] J. P. Stautner, "Analysis and Synthesis of Music Using the Auditory Transform," Master's thesis, MIT EECS Dept., Cambridge, MA (1983).

[25] S. S. Haykin, *Adaptive Filter Theory* (Prentice-Hall, Englewood Cliffs, NJ, 1986).

[26] J. C. Stapleton and S. Bass, "Synthesis of Musical Tones Based on the Karhunen–Loève Transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, pp. 305–319 (1988).

[27] K. P. Li, G. W. Hughes, and A. S. House, "Correlation Characteristics and Dimensionality of Speech Spectra," *J. Acoust. Soc. Am.*, vol. 46, pp. 1019–1025 (1969).

[28] L. C. W. Pols, L. J. Th. van der Kamp, and R. Plomp, "Perceptual and Physical Space of Vowel Sounds," *J. Acoust. Soc. Am.*, vol. 46, pp. 458–467 (1969).

[29] W. Martens, "Principal Components Analysis and Resynthesis of Spectral Cues to Perceived Directions," in *Proc. 1987 Int. Computer Music Conf.*, pp. 274–281.

[30] D. K. Kistler and F. L. Wightman, "A Model

of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-Phase Reconstruction," *J. Acoust. Soc. Am.*, vol. 91, pp. 1637–1647 (1992).

[31] R. G. Laughlin, "Approximating Harmonic Amplitude Envelopes of Musical Instrument Sounds with Principal Component Analysis," Masters thesis, School of Computing Science, Simon Fraser University, Canada (1989).

[32] R. G. Laughlin, B. D. Truax, and B. V. Funt, "Synthesis of Acoustic Timbres Using Principal Components Analysis," in *Proc. 1990 Int. Computer Music Conf.*, pp. 95–99.

[33] G. J. Sandell and W. M. Martens, "Prototyping and Interpolation of Multiple Musical Timbres Using Principal Component-Based Synthesis," in *Proc. 1992 Int. Computer Music Conf.*, pp. 34–37.

[34] F. Opolko and J. Wapnick, "McGill University Master Samples User's Manual," McGill University, Faculty of Music, Montreal, Que., Canada.

[35] Prosonus, "Prosonus Sound Library," Los Angeles, CA (1988).

[36] M. R. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, pp. 243–248 (1976).

[37] M. Dolson, "The Phase Vocoder: A Tutorial," *Computer Music J.*, vol. 10, no. 4, pp. 14–27 (1986).

[38] J. W. Beauchamp, "Unix Workstation Software for Analysis, Graphics, Modification, and Synthesis of Musical Sounds," presented at the 94th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 41, p. 387 (1993 May), preprint 3479.

[39] X. Serra, "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition," Ph.D. thesis, Department of Music Rep. STAN-M-58, Stanford University, Stanford, CA (1989).

[40] D. L. Wessel, "Low Dimensional Control of Musical Timbre," Tech. Rep. 12, IRCAM, Paris (1978).

[41] W. H. Press, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, London, 1992).

APPENDIX

This appendix describes the computational procedure used for principal components analysis. For specifics of implementation with musical tones, the reader should read this in conjunction with our definitions of spectral and temporal orientation, and spectral and temporal PCA.

Our method is essentially the same as described in Harris [19], but we note that other procedures exist as well for arriving at the same results. PCA routines can also be found in many large statistics software packages such as S, SPSS, and BMDP.

PCA is performed in the following steps:

1) Center the data: for all values in each column, subtract the mean for that column. (Alternatively the

data may be standardized, meaning that the additional step of dividing by the column standard deviations is performed.)

2) Save the column means.

3) Calculate the minor product moment of the input matrix. This is accomplished by postmultiplying the transpose of the input matrix by itself. Dividing this by (nrow-1) gives the covariance matrix for the input.

4) Calculate the eigenvalues and eigenvectors of the input matrix from the covariance matrix. This can be accomplished by using the functions `jacobi ()` and `eigsrt ()` given in Press [41]. When the order of the eigenvectors is sorted according to the magnitude of the eigenvalues, this comprises the principal component weights.

5) Postmultiply the weights by the centered data to yield the principal component scores.

Here is how one does PC resynthesis with N principal components:

1) Extract the first N rows from the scores and the weights.

2) Transpose the weights and postmultiply them by the scores.

3) Add the column means from the input data to each column. (If standardizing rather than centering was performed, then one must also multiply by the column standard deviations here.)

Postmultiplication is performed as follows. Given two matrices, PRE and POST (for prefactor and postfactor matrix, respectively),

$$\text{PROD}_{i,j} = \sum_{k=1}^p \text{PRE}_{i,k} * \text{POST}_{k,j}$$

where

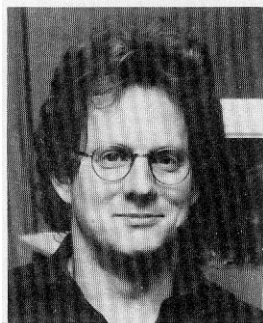
i = columns of PRE

j = rows of POST

p = number of columns of PRE

(The number of columns of PRE must equal the number of rows of POST.)

THE AUTHORS



G. J. Sandell

Gregory J. Sandell was born in Pasadena, CA, in 1957. He received a bachelor of music degree in piano performance from the California State University at Los Angeles in 1980, an M.A. in music theory from the Eastman School of Music in 1983, and a Ph.D. in music theory from Northwestern University in 1991. During 1992 he was at the Center for New Music and Audio Technologies (CNMAT) at the University of California at Berkeley as the F.V. Hunt postdoctoral fellow of the Acoustical Society of America. During 1993-94 he was a research fellow in psychoacoustics in the Department of Experimental Psychology at the University of Sussex (UK). Since 1995 he has been a research associate at the Parmlly Hearing Institute of Loyola University in Chicago.

Dr. Sandell's primary research interests are perception of musical timbre, the perception of simultaneous timbres and its influence on orchestration, and analysis/synthesis of time-variant musical instrument sounds. He maintains a public domain database of musical timbre information called SHARC which is available on the Internet at the World Wide Web address [<http://www.parmly.luc.edu/sharc>].

William L. Martens was born in Dayton, OH, in 1958. He received a bachelor of arts degree in psychology



W. L. Martens

from Miami University (at Oxford) in 1978 and an M.A. and Ph.D. in psychology from Northwestern University in 1982 and 1991. He was also research coordinator at Northwestern Computer Music (1983-1987) and presented lectures on computer sound synthesis and on the psychology of music during that time. From 1987 to 1991 he was vice president of Auris Corporation and from 1991 to 1993, research analyst for the Systems Neurobiology Lab at UCLA. He was responsible for sound spatialization research for the Joint E-Mu/Creative Technology Center from 1993 to 1994.

Dr. Martens specializes in perceptual research (especially spatial hearing) and in audio signal processing research and development. He holds several patents on spatial sound processing technology and has published many papers on both binocular vision and binaural hearing. He joined the Audio Engineering Society in 1984 and has been a member of the Computer Music Association since 1978. He is active in the Interactive Audio Special Interest Group (IASIG) of the MIDI Manufacturers Association (MMA), and is a member of the Virtual Reality Modeling Language (VRML) Architecture Group. He is now an independent contractor in the San Francisco Bay area. Information on his publications, patents, and organizational activities is available on the Internet at the World Wide Web address [<http://www.vrml.org/~wlm>].