Analysis and Synthesis of Musical Transitions Using the Discrete Short-Time Fourier Transform*

JOHN STRAWN**

The Droid Works, San Rafael, CA 94912, USA

To date, the discrete short-time Fourier transform (DSTFT) and similar analysis techniques have generally been used to analyze only individual notes. Problems (and their solution) in using the DSTFT for analyzing transitions *between* performed notes are discussed. Recordings of transitions performed on the trumpet, clarinet, and violin were analyzed with the DSTFT. Based on the responses from 10 musically sophisticated subjects, the DSTFT was shown to be adequate for modeling transitions. To create line-segment approximations for the original DSTFT data, various semiautomatic methods were developed or adapted from the literature on pattern recognition and approximation theory. In a second experiment with the same subjects, line-segment approximations were shown to model musical transitions adequately.

0 INTRODUCTION

Due to the problems involved in recording and analyzing musical instruments, most studies of the physics or the perception of musical sound to date have dealt with isolated notes. With the advent of digital audio and digital signal processing, these constraints need no longer apply. As a first step toward examining musical contexts larger than an individual note, this study concentrates on the spectra of transitions between notes played on musical instruments.

Such transitions include the ending part of the decay of one note, the beginning and possibly all of the attack of the next note, and whatever connects the two notes. To provide a representative sample of performed transitions, a total of 212 two-note pairs on nine orchestral instruments (flute, bass flute, piccolo, clarinet, oboe, bassoon, trumpet, violin, cello) were digitally recorded at the Center for Computer Research in Music and Acoustics (CCRMA); details on the recordings are given in [1], [2]. Each instrument performed four ascending and four descending intervals ranging from major second through minor seventh. Furthermore, the winds and brass played the second note with or without tonguing; and the strings performed the second note with or without bow change. (In this paper, the term tongued will include the transition with bow change on the strings, and untongued will also include transition without bow change.) As many as five separate recordings of each transition were made. Significant differences in both time-varying amplitude and time-varying spectrum were found [1], [2] between tongued and untongued transitions. However, there were no significant differences found for the size of the interval, the direction of the interval, or the size of the instrument (possibly excepting the strings). Having reached these conclusions about the nature of transitions, we performed two experiments to show that a well-known technique for spectral analysis is adequate for analyzing and resynthesizing such transitions.

1 THE DSTFT

The discrete short-time Fourier transform (DSTFT) has proven useful in the analysis of musical signals and speech [3]–[14]. Fig. 1 gives an overview of this process. In effect, a signal is passed through a set of bandpass filters whose center frequencies are equally spaced from dc to one-half the sample rate. In analyzing tones from musical instruments, one usually arranges the filters so that one harmonic falls into each passband. The real and imaginary outputs of the filters shown in Fig. 1 give a time-varying spectral representation of the signal. If the analysis outputs are fed directly to the synthesis part of the technique, the output $y(t)$ is virtually identical to the input $x(t)$. The analysis data may be modified in various ways to produce tones which

* Presented at the 79th Convention of the Audio Engineering Society, New York, 1985 October 12–16; revised 1986 March 27 and August 7.

** Now with S Systems, P.O. Box 623, San Rafael, CA 94915.

are more or less close to the original. Also, the real and imaginary outputs may be converted into time-varying amplitude and frequency terms. The DSTFT and related techniques have been used for several decades to analyze musical instruments, yielding results useful for psychoacoustic researchers [1], [6], [15]–[21] as well as synthesizer manufacturers and the recording industry in general. Thanks to this extensive experience, the DSTFT is well understood.

In order to introduce certain concepts needed later in this paper, the technique will be stated briefly here. For a more intuitive introduction, see [9]. The analysis side of the DSTFT is defined by

$$X(n, k) = \sum_{m=-\infty}^{\infty} x(m)h(n - m)e^{-j2\pi/Nmk}$$

where $x(n)$ is a signal windowed with the low-pass filter $h(n)$, on which there are certain restrictions if an identity system is to be maintained [12]. The dummy variable m allows for the filtering operation with $h(n)$. The spectrum $X(n, k)$ at point n is divided into N frequency bands, or channels, equally spaced from dc to the sample rate f_s and indexed by k. The original signal can be recovered with the inverse DSTFT,

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{m=-\infty}^{\infty} f(n - m)X(n, k)e^{j2\pi/Nmk}$$

where $f(n)$ is also a filter.

In the canonical application, the analysis channels are aligned so that one harmonic of a musical tone falls into a given channel. If $a_k(n)$ and $b_k(n)$ are the real and imaginary outputs of the kth channel, respectively, then the amplitude $A_k(n)$ of the harmonic in the channel may be recovered as follows:

$$A_k(n) = \sqrt{a_k^2(n) + b_k^2(n)} .$$

The instantaneous phase $\phi_k(n)$ is given by

$$\phi_k(n) = \arctan\left[\frac{b_k(n)}{a_k(n)}\right] .$$

The frequency, calculated as the derivative of the phase, is then given by

$$\frac{a(n)\ db(n)/dt - b(n)\ da(n)/dt}{a^2(n) + b^2(n)} .$$

Details of this conversion process are given in [9], [11].

To prepare for the discussion later in this paper, it is necessary to state here that the spectrum $X(n, k)$ is not calculated for every sample of the original signal. Rather, every R point can be skipped; there are restrictions on the relationship between R and N which do not concern us here [11], [12]. Also, it should be pointed out that the selection of $h(n)$ is important. In particular, a longer analysis filter gives better frequency resolution but with a correspondingly coarser resolution in time; and vice versa.

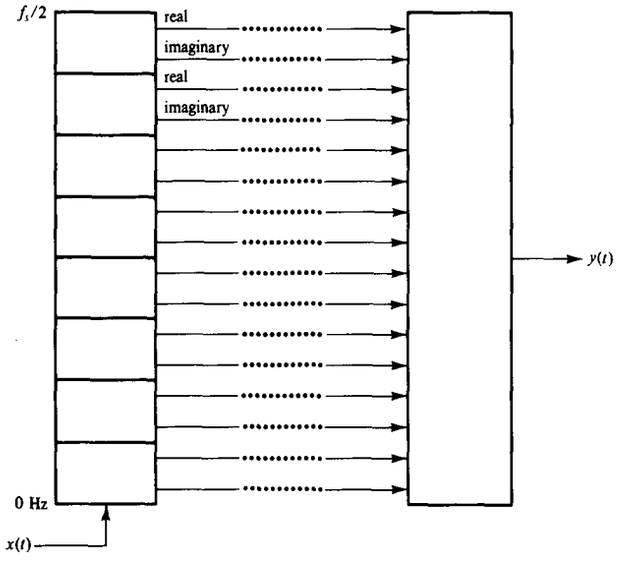


Fig. 1. The DSTFT analysis stage (left) works like a set of band-pass filters spaced equally from 0 Hz to one-half of the sampling rate. If the data analyzed from $x(t)$ are used for resynthesis (right), a signal $y(t)$ results, which is nearly identical to the original. (Reprinted with permission from [9].)

1.1 Problems with the DSTFT

1.1.1 Reliability of Frequency Traces

A perennial difficulty with the DSTFT lies in interpreting its output. Examples of this can be seen in [1], [19], [22]–[23]. Since the amplitude of the recorded signal drops several tens of decibels during many transitions, the frequency traces are especially difficult to interpret because they become unstable at very low amplitudes.

1.1.2 Reliability of Amplitude Estimates

Another problem with using the DSTFT for analyzing transitions is that the center frequencies of the filters remain fixed once set, so that the harmonics of the new note no longer fall onto the analysis channels in a useful way. Furthermore, as the signal leaves one channel and enters the neighboring channel, the amplitude of the signal is subject to distortion [6]. The bandpass filters used to realize the DSTFT have a rolloff of their own, shown in Fig. 2. As long as a spectral component remains in the region shown at A–B in the figure, the magnitude output from the corresponding channel can be trusted. But the spectral component at C has its magnitude modified by the filter's own rolloff. Some possible solutions to these problems, and their pitfalls, are discussed in [1]. It turns out that these problems have no detectable effect on resynthesized tones, as the experiments discussed below will show, so these problems will not be considered further here.

1.2 Examining the Analysis Data

1.2.1 Amplitude Plots

There is also the question of how to display the outputs of the DSTFT. Traditionally, outputs have been presented in the familiar three-dimensional format, as given in [19]. But the current work deals with more than one note. It was certainly possible to run the analysis twice: once for each note, with the filter center frequencies adjusted accordingly. The end of the first note analyzed in this manner looked reasonable; and so did the beginning of the next.

The only way we could find to make useful spectral plots was to splice together these two analyses in a three-dimensional representation. It was necessary to expand our spectral editor [23] to handle these two analyses properly. Fig. 3 (the tongued clarinet ascending major third) shows a sample of the result; more are given in [1], [2]. In this plot, time (shown in seconds) runs from left to right. The fundamental is at the top of the plot; higher order harmonics are plotted along their own axes, which are arranged below the fundamental on the page. One should imagine this spectral plot as "coming out toward" the viewer from the fundamental "at the back." Each harmonic is plotted on a scale of 0 to −60 dB, with 0 dB being the maximum of the strongest harmonic in the entire plot. The first note is at the left; the second, at the right. At the point specified in the legend, the plotting program switches from the analysis for the first note to that of the second; this is approximately the point where the pitch changes. Clearly, there is a spectral rolloff at the end of the first note in the tongued transition of Fig. 3; of the 30 harmonics shown here, perhaps the top 20 drop out. Note that the pattern with which the harmonics drop out and reenter is not entirely regular. However, in general the higher order harmonics leave sooner and reenter later than their lower frequency counterparts. As discussed in [1], [2], this spectral pattern varies significantly for tongued and untongued transitions.

After this work was completed, McAulay and Quatieri [24] introduced the notion of the "birth" and "death" of sinusoidal components. This idea and the automatic method they give for tracking components would be useful for future analyses of the transitions of instruments.

1.2.2 Plots of the Frequencies

For a single note it is possible to create three-dimensional plots of the time-varying frequency traces similar to those for the amplitude traces. However, three-dimensional plots of the frequencies in a transition did not prove to be useful; a few are given in [1].

2 ANALYSIS/RESYNTHESIS WITH THE DSTFT

If the DSTFT were inadequate for analyzing or resynthesizing a transition, one would expect any problems to occur regardless of the instrument being analyzed. Therefore it was sufficient to experiment with just one instrument; of the instruments recorded, the trumpet was arbitrarily chosen here. Also, any inadequacies of the analysis should be more apparent with a larger interval between the notes played. Therefore the largest interval available (minor seventh, commonly abbreviated m7) was chosen. Finally, there should be no differences between ascending and descending intervals in the observed behavior of the analysis nor in the audibility of any distortion produced by the analysis/resynthesis. In short, the ascending minor seventh played on the trumpet formed the basis for this experiment.

At least two possible sources of distortion in the analysis could be identified.

1) As mentioned before, the DSTFT might not be able to track signals adequately in the transition region. One would expect this to be most prominent in the untongued case, since the frequencies are shifting so rapidly (sometimes across just a few periods [1], [2]).

2) The DSTFT models the signal as a group of harmonically related sinusoids. It might not be able to emulate the "puff" of noise at the beginning of a tongued attack.

Preliminary work demonstrated that neither of these produced audible distortion in the transition. Thus this experiment used only the untongued case. The success of the second experiment discussed below shows that using only the untongued case here was reasonable.

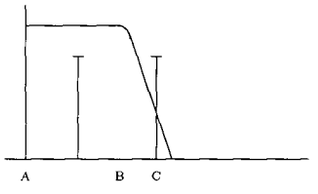


Fig. 2. For a given analysis channel, the amplitude of a spectral component falling in the range A–B is unaffected by the analysis filter's frequency response. This is not the case for a component at, say, C.

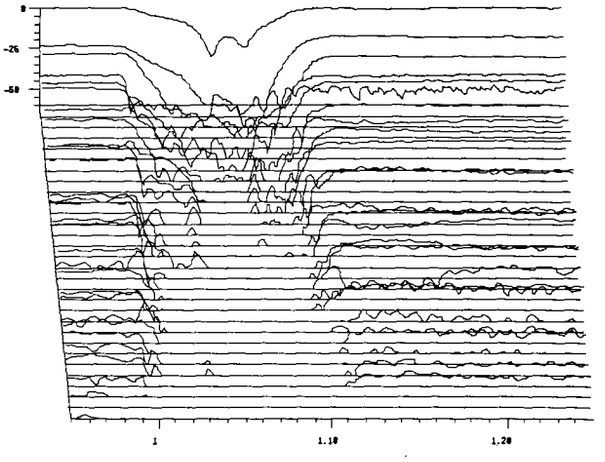


Fig. 3. Time-varying spectral analysis (30 harmonics) of a tongued ascending major third played on the clarinet. The lower note is A220; the splice point is at t = 1.05 s.

2.1 Creating the Stimuli

The recording of the trumpet untongued ascending m7 was resampled [25] to 26 040 Hz to provide a sampling rate into which the fundamentals of both notes would divide easily. Both notes in the recording were analyzed using two sets of analysis parameters. The settings appropriate for the lower note were $N = 100$ and $R = 5$, with $N = 56$ and $R = 8$ for the upper note. The quantity Q mentioned in [9] was set to 1, effectively turning off any further interpolation of the data points as suggested in [11]. (Some modification to the code given in [9] is necessary to make this work.)

The note pair was resynthesized twice, using both analyses. Both resynthesized note pairs sounded completely natural, although slightly low passed when compared with the original. (This problem and some possible solutions are discussed in more detail below.)

The control stimulus was the original recording, shown in Fig. 4(a). To make a test stimulus, it proved impossible to splice the analysis data from the first note directly onto the analysis data from the second note, a procedure implied by Fig. 3. After resynthesis using such a method, the second note still sounded quite natural, but underwent some severe phase distortion, which made it unsuitable for experimental use. The phase distortion apparently occurred because of the abrupt change in analysis parameters.

The final test stimulus was created with a 20-ms cross-fade from the end of the first note, as analyzed with $N = 100$, to the beginning of the second note, as analyzed with $N = 56$. The cross-fade, created with methods given in [1], occurred at the point shown by the arrow in Fig. 4(a). The resulting untongued transition is shown in Fig. 4(b). (This procedure also worked for the tongued case.)

2.2 Experimental Procedure

Experience gathered in creating the stimuli showed that a note resynthesized using all of the analysis data was physically slightly different from the original. It was not necessary here to show that the original and resynthesized tones were physically identical. However, it was necessary to show that the listener could not reliably distinguish between the two. As stated above, the test stimulus sounded slightly low passed when compared with the control stimulus. Thus it proved impractical to conduct an experiment in which the subject decided whether two stimuli were identical, because the notes surrounding the transitions proper were themselves different in the two stimuli. Therefore each subject was asked to state a preference for the transition in one of the two stimuli. If the subject has no clear preference for one of two quite similar stimuli, then we can conclude that the two are perceptually interchangeable. Indeed, they might even be identical for all practical purposes.

In such preference tests it is important to account for any order effects. That is, if A and B are different stimuli, then the preference for A followed by B must be compared with the preference for B followed by A. Also, comparing each stimulus with itself (A:A and B:B) checks for subject bias toward the first or the second stimulus of each pair; such tests are sometimes called "vexierversuche." The subjects in this experiment thus heard four cases, numbered 1–4 in the list below. Each case consisted of one stimulus, followed by a pause of 0.20 s, followed by another stimulus, followed by a pause of 0.50 s. Comparison cases were:

1) Original (control stimulus) versus resynthesized (test stimulus);
2) Resynthesized versus original.

Identical cases were:

3) Original versus original;
4) Resynthesized versus resynthesized.

The comparison cases (1 and 2) were presented three times each; the identical cases (3 and 4) were presented twice each.

The subjects were 10 male volunteers, all trained musicians, all familiar with electronic and computer music. The digital recordings of the cases were resampled to 44.1 kHz and transferred to Sony F1 tape in a randomized order. Three different tapes, each with the cases in a different order, were made. Each subject heard one of the tapes, played back in the fairly dead room where the original recordings were made [1]. The playback level was adjusted to be comfortable and remained constant for all listeners. The subjects completed answer sheets printed on normal 8.5 by 11-in paper. At the end of the experiment, the subjects were encouraged to write their own comments on the answer sheet; some of these are cited below. At the beginning of the experiment, several training examples were presented. These were scored by the test subjects, but were not included in the statistical analysis.

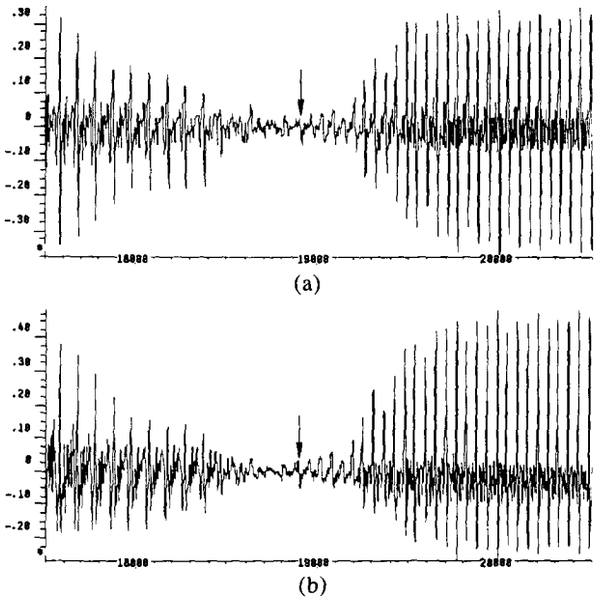


Fig. 4. (a) Untongued trumpet transition. The abscissa shows time in samples, at a sample rate of 25 600. At the point marked by the arrow, a 20-ms cross-fade joins the resynthesized notes. (b) Each note has been analyzed and resynthesized separately.

2.3 Results

Examination of the raw responses showed that it was impossible to find any meaningful overall pattern in the identical cases; that is, there was no significant listener bias toward the first or second stimulus.

For the comparison cases, examination of the raw data showed that subjects 3 and 9 preferred the original over the resynthesis; subjects 1, 2, 4, 5, and 7 preferred the resynthesis over the original (!); and subjects 6, 8, and 10 seemed to prefer neither. Although it might appear surprising that *any* subject would prefer the synthetic stimulus (if this preference were not due to random variation in the responses), this might occur because these particular test subjects are used to working with synthesized sound. At any rate, no consistent pattern appeared in the data for the comparison cases.

This preliminary conclusion was borne out in Table 1, which shows the means of all of the responses for each case averaged across all subjects. It is reasonable to conclude that the subjects could not accurately distinguish between the two stimuli when the mean is close to (case 3) or at (cases 2, 4) the value of 1.5.

The value for case 1 might seem far from the expected mean. It is more meaningful to examine the combined mean of the comparison cases (1 and 2) across all subjects, which is 1.57. This value suggests that the subjects could not accurately distinguish between the original and the resynthesized stimulus; if anything, subjects showed a slight preference for the resynthesized stimulus.

Whether any of the means of Table 1 indicates an actual preference for the resynthesized stimulus or simply chance variation around the "no preference" mean of 1.5 is a question answered by the well-known t test. The results are also shown in Table 1. The t values indicate that none of these means was statistically different from what one might expect from a population of subjects with no preference for one signal over the other.

The question then arose as to whether the means for the four cases were significantly different from each other. If not, then it could be asserted that the seemingly large mean for case 1 was no more significant than the other, smaller means. Analysis of variance implied $p > 20\%$, that is, the probability that the given data would occur due only to chance was greater than 20%. In other words, there was no statistically significant difference among the means of Table 1. Thus, none of the means varied significantly from the value of 1.5, and we concluded that the subjects showed no preference for either the original or the synthetic stimulus.

Table 1. Preferences for first or second stimulus, averaged across all subjects.*

Case	Mean	t
Comparison		
1	1.63	1.49
2	1.50	0.00
Identical		
3	1.55	0.44
4	1.50	0.00

*A value of 1.0 means in cases 1 and 2 that the original was preferred over the synthetic; in cases 3 and 4, that the first of two identical stimuli was preferred.

All of this statistical sophistication may appear to be overkill when one reads the written comments of the subjects, of which these are typical:

"They all sounded rather similar."

"I was not able to hear any differences in any of these pairs (nor between one pair and another)."

2.4 Conclusion

The subjects showed no clear preference for either the original or the resynthesized transition. The transition resynthesized on the basis of full analysis data is therefore perceptually interchangeable with the transition in the original.

3 LINE-SEGMENT APPROXIMATIONS

3.1 Background

The preceding experiment showed that the DSTFT adequately represents the time-varying spectrum in the transition. But spectral analysis of this kind provides too many data for practical work in sound synthesis and for controlled timbral studies. It is commonly accepted that line-segment approximation of the amplitude and frequency traces can produce individual resynthesized tones which sound quite close to the original [17], [19], [21], [22], [26], [27]. The question thus arises as to whether line-segment approximations are adequate for synthesizing musical transitions.

Methods for creating reasonable line-segment approximations were not highly developed [16], [19], [21] when this work started. A search of the literature on approximation theory and pattern recognition showed that several algorithms would be useful [28].

3.2 Creating the Stimuli

For the current work, the split-merge algorithm combined with the adjust procedure, both due to Pavlidis (details and references are given in [28]), were used. The algorithm was applied to amplitude data in the following way.

1) From the analysis, create a spectral average by averaging the data over a specified amount of time. A sample of such a spectral average is given in Table 2. (Other examples are given in [1], [23].) The first column gives the harmonic number. The next two columns show the averaged amplitude, and the third shows average frequency. (The fourth column is explained below.) For amplitudes, this is equivalent to taking the discrete Fourier transform over the time in question, which is selected from the "steady state" of each note.

2) Multiply the averaged amplitude (the second column in Table 2) from each harmonic by some small

constant, say 0.001. This constant varies with instrument, sample rate, N, and R.

3) Use the resulting number as a threshold for the Pavlidis algorithm, with the integral error norm given in [28].

4) The resulting line-segment approximation, typically a dozen segments per harmonic per note, must usually be cleaned up slightly by hand; the editor written for this purpose is discussed in [23].

Note that this process must be done twice to create a single test stimulus for this experiment—once for each of the two notes surrounding the transition.

These two sets of amplitude traces must then be joined by hand on a harmonic-by-harmonic basis. We extended the editor described in [23] to display the analyses for both notes along with a composite function created by splicing the line-segment approximations from the two notes at the point of pitch change. For each harmonic, the user creates a final transition function by hand. Fig. 5 shows the tenth harmonic taken from the two analyses of the tongued ascending third on the trumpet. In each of the three parts of this figure, the end of the first note is shown on the left, and the second note begins on the right. In Fig. 5(a) the analysis parameters were set for the frequency of the first note. The analysis (wavy line) is shown along with the raw output of the Pavlidis split-merge algorithm (straight line). Fig. 5(b) shows a similar analysis, but with the DSTFT set up for the second note (C♯). This part of the figure must be carefully interpreted; the "beating" at the left results when two harmonics of the first note fall into one analysis band. (A good discussion of this phenomenon is given in [6].) Incidentally, the software editor allows the user to view either or both of the original analyses along with either or both of the automatically generated approximations as well as the approximation which the user creates by hand. Fig. 5(c) shows the actual function used in synthesizing the tenth harmonic for the tongued trumpet stimulus. Editing in this manner is not as easy as it might sound. Once the software works, several minutes of console time are needed for each harmonic. For a stimulus with 30 or so harmonics, an hour can be quickly consumed.

The result of this editing is a set of line-segment approximations that more or less accurately capture the amplitude characteristics of the harmonics in the transition. Fig. 6 shows the approximations that were created for the clarinet transition originally shown in

Table 2. Spectral average of steady-state of trumpet tone.*

Channel	Amplitude	dB	Frequency (Hz)	$\text{freq}_n/(\text{freq}_1 \times n)$
1	0.9838	−11.14	274.0896	1.00000
2	1.5478	−7.21	548.1669	0.99998
3	0.4360	−18.21	822.1024	0.99980
4	2.3940	−3.42	1096.2836	0.99993
5	3.5508	0.00	1370.4740	1.00002
6	2.5190	−2.98	1644.5948	1.00003
7	2.0818	−4.63	1918.7072	1.00004
8	1.5529	−7.18	2192.8972	1.00008
9	0.9392	−11.55	2466.9651	1.00006
10	1.1438	−9.83	2741.0100	1.00004
11	0.9841	−11.14	3014.9523	0.99999
12	0.5494	−16.20	3289.2921	1.00007
13	0.6294	−15.02	3563.4460	1.00008
14	0.4979	−17.06	3837.6638	1.00011
15	0.4913	−17.17	4111.7297	1.00009
16	0.4291	−18.35	4385.6819	1.00006
17	0.2915	−21.71	4659.7933	1.00006
18	0.2200	−24.15	4934.0869	1.00010
19	0.2147	−24.36	5208.2567	1.00011
20	0.1637	−26.72	5482.4329	1.00012
21	0.1524	−27.34	5756.4721	1.00010
22	0.1059	−30.50	6030.7901	1.00014
23	0.1006	−30.95	6305.1367	1.00017
24	0.0656	−34.66	6579.3024	1.00018
25	0.0600	−35.43	6853.0558	1.00012
26	0.0403	−38.90	7127.4308	1.00015
27	0.0350	−40.10	7402.1436	1.00023
28	0.0294	−41.63	7675.9214	1.00018
29	0.0264	−42.55	7949.8907	1.00016
30	0.0237	−43.50	8224.8254	1.00026
31	0.0180	−45.87	8520.0140	1.00273
32	0.0131	−48.62	8835.0671	1.00732
33	0.0115	−49.71	9129.9086	1.00939
34	0.0095	−51.40	9464.7775	1.01564
35	0.0075	−53.48	9738.7532	1.01518
36	0.0051	−56.75	10052.0600	1.01873
37	0.0030	−61.23	10337.0000	1.01930
38	0.0016	−66.49	10631.4180	1.02074

* This average spectrum was calculated over 0.1 s. freq_n—frequency of channel n, freq_1—frequency of channel 1 (the fundamental).

Fig. 3.

Risset [21, pp. 36, A-9] was not able to show that including either the "blips" in the trumpet attack nor the slight burst of noise at the beginning of the note had any effect in his resyntheses. However, our experience was that both of these features did in fact make an important difference in how the test stimulus sounded. Much of the time spent in refining the trumpet test stimuli for this experiment was in fine-tuning the blips in the attacks of the first dozen partials or so, and in adding small amounts of amplitude to the higher harmonics right at the attack to simulate the tonguing noise. Without such blips, our experience shows that the resynthesized attacks sound tubby.

For frequency traces, we found that it was adequate to create one line-segment approximation from the fundamental of each note, using the editor just mentioned. The spectral average of Table 2 also contains values (in the right-hand column) for what we term the relative harmonicity of the spectral component—how far it deviates from being an exact multiple of the fundamental. For each harmonic, each point of the hand-made fundamental frequency trace was multiplied by the harmonic number times this relative harmonicity value. This was slightly different from the work by Grey [19], who used a constant-frequency approximation for some experiments, and from Charbonneau's tones [17], where the fundamental frequency trace was multiplied by the (integer) harmonic number. In some instruments, a slightly richer tone results by using the inharmonic case. In particular, the straight-line frequency approximation of Grey is noticeably enriched.

Again, this process must be followed for both notes in the test stimulus. The frequency traces for the two notes are simply spliced, using a vertical transition, at the appropriate point. Fig. 7 shows the frequency function used for the fundamental of the untongued trumpet test stimulus for this experiment. Some activity in the attack of the first note was retained; its aural effect was not as pronounced as the illustration would suggest. We found that as long as the amplitude of the signal was low enough at the point of pitch change, the abrupt transition in frequency between the notes was never audible as such. In their written comments, none of the test subjects in this experiment complained about the quality of the transition synthesized in this manner.

The tongued and untongued ascending thirds from the clarinet and trumpet, and the ascending third from the violin with and without bow change, were used. These intervals had been judged to be representative of the larger set of recordings [1]. For each test stimulus, a two-note pair was created using additive synthesis of the amplitude and frequency line-segment approximations. The control stimuli were the corresponding

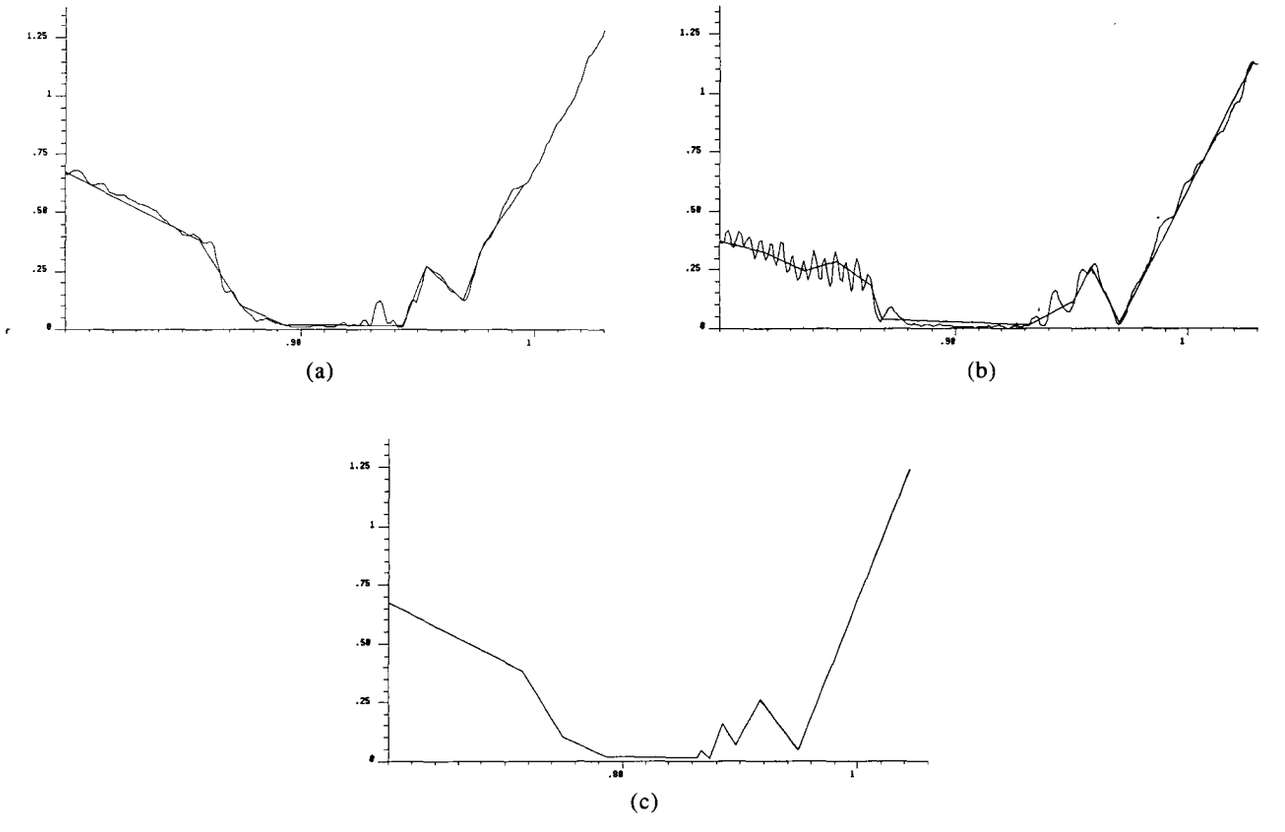


Fig. 5. Editing the amplitude traces in a transition for the tenth harmonic of the ascending third tongued trumpet transition. The abscissa is time in seconds; amplitude on a linear scale is the ordinate. (a) Analysis parameters set for the first note, with line-segment approximation created with the Pavlidis split-merge algorithm. (b) As in (a), but with analysis parameters set for the second note. (c) Composite line-segment approximation created by hand for approximations in (a) and (b).

six original recordings.

The line-segment approximations included harmonics whose amplitudes were above approximately −60 dB from the note's maximum. It was impractical to include harmonics with amplitudes much lower than this, the amplitude and frequency traces being badly degraded by noise. Each note thus contained anywhere from 27 to 40 harmonics.

The transitions in the synthesized stimuli sounded very close to those in the original recordings. However, as in the experiment described above, the notes in the test stimuli were and sounded slightly band-limited. We spent considerable effort trying to solve this pesky problem.

1) We tried splicing from the original data to the line-segment approximation at the very end of the first note, then splicing back to the original data (for the second note) at the very beginning of the second note. Resynthesis using CCRMA's Samson box [29], [30] proved impractical because of the resulting high command rate. Also, even when the notes were resynthesized in software (prohibitively expensive for the amount of computation needed for this experiment), the problem mentioned earlier occurred here as well—there was a nasty phase shift at the splice.

2) Using a very short cross-fade [1], we spliced the resynthesized transition into the original recording, splicing at the end of the first note and again at the beginning of the second. For short cross-fade times (20 ms or so), a perceptible phase shift occurred at each splice. Due to the short duration of the transition, longer splice times, which would probably have made the phase shift inaudible, proved impractical.

3) Following a suggestion by Portnoff [31], we examined the difference signal between the original and the synthetic tones. This difference signal turned out to be a waveform almost identical to the original, except for a "phasing" throughout the duration of the note. Gish [8] included an explicit noise term $n(t)$ in his synthesis model [his Eq. (1)] and claimed: "When the residual error, or noise, $n(t)$, is listened to, it usually sounds just like tape hiss." Ideally, one would like to be able to characterize this noise signal. More work on the time-domain difference signal needs to be conducted.

4) We tried calculating the difference signal by subtracting, on a harmonic-by-harmonic basis, the amplitude and frequency traces of the line-segment approximation from those of the original analysis data. Note that both the original analysis data and the line-segment approximation must be interpolated as needed to bring them to the original sample rate. (Needless to say, the amount of computation is enormous.) These difference signals were used to synthesize a time-domain signal which was then added to the synthesized signal in an attempt to make it sound closer to the original. (Beauchamp [32] also developed a method for approximating the difference signal in this fashion; but he used only the error from the amplitude traces.) The results were inconclusive. This approach needs to be explored further.

5) We tried, without success, to find a way to filter the original to match the quasi-low-passed nature of the synthesized tone. What one really needs here is a time-varying band-reject filter, because it turned out that the spectral differences could not be characterized by a time-invariant low-pass filter alone.

6) We tried to add low-amplitude white or colored noise to the synthesized signal to make it sound closer to the original.

Regarding item 6, Grey [19, p. 37] also found that tape hiss present in the original recording but missing in the line-segment approximation could allow the listener to distinguish the two. Beauchamp likewise reported similar problems, in that in his resynthesized tones, key clicks and "a certain roughness" [32, p. 323] were missing. In the experiment discussed earlier, this was not a problem, as resynthesis with full data captured all of the instrumental noises in the original. For the current work we had mixed success (as did Grey) with trying to add background noise from the original recordings into the synthetic stimuli; so that approach was not followed.

The slightly low-passed nature of the synthetic tones thus made it impossible to design a same/different experiment, or an experiment in which the subjects rated

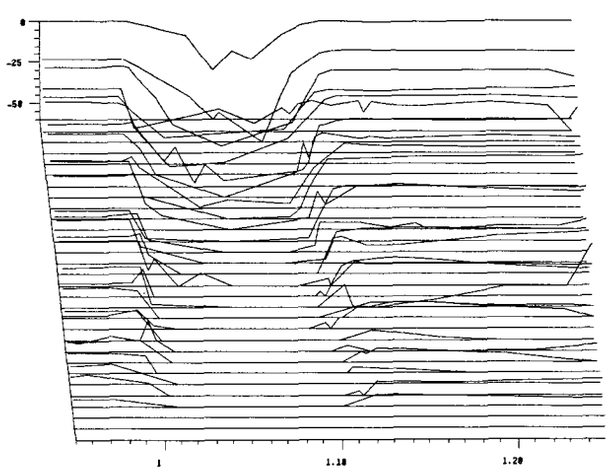


Fig. 6. Line-segment approximations for the clarinet tongued ascending third transition. (See Fig. 3.)

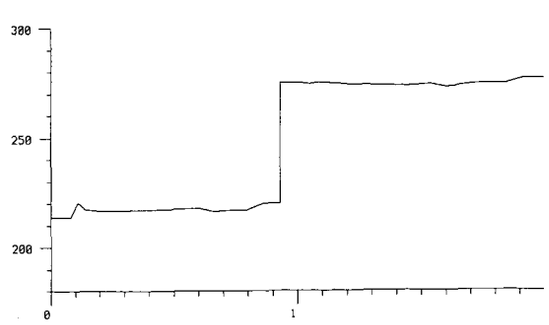


Fig. 7. Fundamental frequency trace for the two-note tongued trumpet test stimulus. The ordinate is frequency in hertz; the abscissa, in seconds.

how different the resynthesized stimuli were from the original stimulus. The notes surrounding the transition in the test stimulus were themselves slightly different from those in the control, which might confuse the test subject.

3.3 Experimental Procedure

Therefore the preference test already discussed was used here as well. The subjects in this experiment heard four cases, numbered 1–4 as before. The comparison cases (1 and 2) were presented three times each; the identical cases (3 and 4) were presented twice each. These four cases were presented for each of the three instruments, using both playing styles (tongued and untongued) for each instrument. It seemed necessary to test more than one instrument, as any failing of the line-segment approximation might well show up for one instrument or playing method but not for another. This experiment followed the previous experiment on the F1 tape described above; both experiments were conducted in one sitting.

3.4 Results

As in the previous experiment, examination of the raw data for the individual subjects' mean responses showed no clear preference for either the original or the resynthesized tones. This conclusion is supported by the subjects' written comments, of which these are typical:

"In a number of cases I heard no difference or at any rate had no preference. . . . "

"Often hard!"

"Impossible!"

Table 3 shows *how often* the subjects preferred the original (cases 1 and 2) or the first stimulus (cases 3 and 4). The right-hand column gives the maximum score possible, derived from the number of subjects (10) times the number of presentations (3 for the comparison cases, 2 for the identical ones). This maximum score would be reached if all subjects preferred the synthetic stimulus in case 1, the original in case 2, or the second of the two identical stimuli in cases 3 and 4. Here again no clear-cut pattern was discernible which might suggest whether the synthesized transition was preferred over the original. For the trumpet and violin in case 2, there was a slight tendency to pick the original over the synthesized tone; recall that in case 2, the original was played second. There seems to be no particular significance to this pattern in the data.

Table 3. Subjects' preferences (line-segment approximations).

Case	Clarinet T	Clarinet U	Trumpet T	Trumpet U	Violin T	Violin U	Maximum possible
Comparison							
1	17	14	15	14	15	15	30
2	17	14	20	19	17	21	30
Identical							
3	7	10	9	12	7	5	20
4	11	6	9	8	8	8	20

T—tongued (with bow change); U—untongued (without bow change).

Table 4 gives the mean, the standard deviation, and the t value for each of the four cases, three instruments, and two playing methods. Only in three instances does the t value imply a probability less than 0.05, which means that for all instruments and playing styles except the violin with no bow change, it is safe to conclude that the observed mean does not vary from the expected mean of 1.5 any more than one would expect from random variation. The p value for case 2 on the tongued trumpet is not considered to be of significance, as case 1 for the tongued trumpet shows no deviation at all from the expected mean of 1.5. Therefore we conclude that in five of the six instrument/playing method combinations the synthetic cases are essentially identical to the originals.

Analysis of variance of the other exception (violin, no bow change) implied that the variation of the means in Table 4 for the violin with no bow change was not just random ($p < 2.5\%$). It is easy to accept the large amount of variation in case 2 of the violin without bow change, as case 1 showed the expected behavior (that is, the mean for case 1 in Table 4 was 1.5), and both cases tested the preference of the original over the synthetic. Examination of the original data for case 3 (not given here) showed that six of the 10 subjects chose the second tone in both trials, which accounts for the large amount of variation seen there. Such a large bias did not occur in any other instance in this experiment. Thus the apparently large variation in the data for this one instrument and playing style is shown to have no real significance.

3.5 Conclusion

The subjects showed no clear preference for either the original or the resynthesized transition. The transition resynthesized using line-segment approximations to the original data, with the frequency traces connected by a straight vertical line, is perceptually interchangeable with the transition in the original.

4 OVERALL CONCLUSION

The model of time-varying spectra based on Fourier methods is adequate for analyzing and resynthesizing transitions between notes, using either the full analysis data or line-segment approximations. The success of these experiments confirms the validity of the conclusions based on plots created with the DSTFT [2].

These experiments suggest a method for creating transitions between two notes. (Others are given in [1].) After the notes have been analyzed with the DSTFT or some other suitable technique, one can create a transition by extending, say, the lowest 10 or 20 harmonics of the first note; their summed amplitude should be 10 to 40 dB down [1], [2] from the peaks of the notes. At

the point where the second note is to start, splice these harmonics onto the amplitude traces of the attack of the second note. The frequencies can jump vertically from the first note to the next note, right at the point of pitch change.

5 ACKNOWLEDGMENT

The work described here was performed at the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, and was supported in part by the System Development Foundation. The author gratefully acknowledges the contributions of a) John Chowning, Earl Schubert, James A. Moorer, John Pierce, and John Grey for their help in designing the experiments; b) David Wessel, Stephen McAdams, and Al Bregman for participating in lengthy conversations on the ideas presented here; c) Mark Dolson and John Gordon for their assistance with the DSTFT; d) Emily Bernstein, David Burkhardt, Angela Sohn, Dexter Morrill, and David Jaffe for recording the test tones; e) Phil Gossett for his modification of the Sony F1; and f) Jim Aiken, Thom Blum, Chris Chafe, Doug Fulton, David Jaffe, Kyle Kashima, Douglas Keisler, Brian Schober, Xavier Serra, and Amnon Wolman for voluntarily submitting themselves to the experiments. The support of the Stanford Artificial Intelligence Laboratory, Lucasfilm, and The Droid Works is also gratefully acknowledged.

6 REFERENCES

[1] J. Strawn, "Modeling Musical Transitions." Ph.D. dissertation, CCRMA/Dept. of Music, Stanford University, Stanford, CA, Dept. of Music Rep. STAN-M-26, 1985.

[2] J. Strawn, "Orchestral Instruments: Analysis of Performed Transitions." *J. Audio Eng. Soc.*, vol. 34, pp. 867–880 (1986 Nov.).

[3] J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. ASSP-25, pp. 235–238 (1977).

[4] J. B. Allen and L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," *Proc. IEEE*, vol. 65, pp. 1558–1564 (1977 Nov.).

[5] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[6] M. B. Dolson, "A Tracking Phase Vocoder and Its Use in the Analysis of Ensemble Sounds," Ph.D. dissertation, California Institute of Technology, Pasadena, 1983.

[7] J. L. Flanagan and R. M. Golden, "Phase Vocoder," *Bell Sys. Tech. J.*, vol. 45, pp. 1493–1509 (1966).

[8] W. C. Gish, "Analysis and Synthesis of Musical Instrument Tones," presented at the 61st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 26, p. 998 (1978 Dec.), preprint 1410.

[9] J. W. Gordon and J. Strawn, "An Introduction to the Phase Vocoder," in *Digital Audio Signal Processing: An Anthology*, J. Strawn, Ed. (William Kaufmann, Los Altos, CA, 1985), pp. 221–270.

[10] S. Holtzman, "Non-Uniform Time-Scale Modification of Speech," M.Sc. and E.E. dissertation, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, 1980.

[11] J. A. Moorer, 1978. "The Use of the Phase Vocoder in Computer Music Applications." *J. Audio Eng. Soc. (Engineering Reports)*, vol. 26, pp. 42–45 (1978 Jan./Feb.).

[12] M. R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, 1978.

[13] M. R. Portnoff, "Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. ASSP-28, pp. 55–102 (1980).

[14] R. W. Schafer and L. R. Rabiner, "Design and Simulation of a Speech Analysis–Synthesis System Based on Short-Time Fourier Analysis," *IEEE Trans.*

Table 4. Statistical analysis (line-segment approximations).*

Case	Mean	Standard deviation	t	p	Mean	Standard deviation	t	p
Clarinet		Tongued				Untongued		
1	1.43	0.50	−0.72		1.53	0.50	0.36	
2	1.60	0.49	1.10		1.47	0.50	−0.36	
3	1.65	0.48	1.37		1.50	0.50	0.00	
4	1.45	0.50	−0.44		1.70	0.46	1.90	
Trumpet		Tongued				Untongued		
1	1.50	0.50	0.00		1.53	0.50	0.36	
2	1.67	0.47	1.90	<0.10	1.63	0.48	1.49	
3	1.55	0.50	0.44		1.40	0.49	−0.89	
4	1.55	0.50	0.44		1.60	0.49	0.89	
Violin		Bow change				No bow change		
1	1.50	0.50	0.00		1.50	0.50	0.00	
2	1.60	0.49	1.10		1.70	0.46	2.35	<0.05
3	1.65	0.48	1.37		1.75	0.43	2.52	<0.02
4	1.60	0.49	0.89		1.60	0.49	0.89	

* A mean value of 1.0 in case 1 means that the original was preferred. The same value in case 2 means that the synthetic was preferred. In cases 3 and 4, this value means that the first of two identical stimuli is preferred. Mean values may range from 1.0 to 2.0. All t values imply $p > 0.10$ unless shown otherwise.

Audio Electroacoust., vol. AU-21, pp. 165–174 (1973).

[15] H. Backhaus, "Über die Bedeutung der Ausgleichsvorgänge in der Akustik," *Z. tech. Physik*, vol. 13, no. 1, pp. 31–46 (1932).

[16] J. W. Beauchamp, "A Computer System for Time-Variant Harmonic Analysis and Synthesis of Musical Tones," in *Music by Computers*, H. von Foerster and J. W. Beauchamp, Eds. (Wiley, New York, 1969), pp. 19–62.

[17] G. Charbonneau, "Timbre and the Perceptual Effects of Three Types of Data Reduction," *Computer Music J.*, vol. 5, no. 2, pp. 10–19 (1981).

[18] M. D. Freedman, "A Method for Analyzing Musical Tones," *J. Audio Eng. Soc.*, vol. 16, no. 4, pp. 419–425 (1968 Oct.).

[19] J. M. Grey, "An Exploration of Musical Timbre." Ph.D. dissertation, Dept. of Psychology, Stanford University, Stanford, CA, Dept. of Music Rep. STAN-M-2, 1975.

[20] D. A. Luce, "Physical Correlates of Nonpercussive Musical Instrument Tones," Ph.D. dissertation, Dept. of Physics, Massachusetts Institute of Technology, Cambridge, 1963.

[21] J.-C. Risset, "Computer Study of Trumpet Tones," Bell Laboratories, Murray Hill, NJ, typewritten mss.; *J. Acoust. Soc. Am. (Abstracts)*, vol. 38, p. 912 (1965).

[22] J. A. Moorer, J. M. Grey, J. Snell, and J. Strawn, "Lexicon of Analyzed Tones. Part 1: A Violin Tone," *Computer Music J.*, vol. 1, no. 2, pp. 39–45 (1977); "Lexicon of Analyzed Tones. Part 2: Clarinet and Oboe Tones," *Computer Music J.*, vol. 1, no. 3, pp. 12–29 (1977); "Lexicon of Analyzed Tones. Part 3: The Trumpet," *Computer Music J.*, vol. 2, no. 2, pp. 23–31 (1978).

[23] J. Strawn, "Editing Time-Varying Spectra," *J. Audio Eng. Soc.*, accepted for publication.

[24] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. 34, pp. 744–754 (1986 Aug.).

[25] J. O. Smith and P. Gossett, "A Flexible Sampling-Rate Conversion Method," *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing* (San Diego, CA, 1984 Mar.), vol. 2, pp. 19.4.1–19.4.2.

[26] J. A. Moorer, "Signal Processing Aspects of Computer Music—A Survey," *Proc. IEEE*, vol. 65, pp. 1108–1137 (1977 Aug.); reprinted in *Digital Audio Signal Processing: An Anthology*, J. Strawn, Ed. (William Kaufmann, Los Altos, CA, 1985), pp. 149–220.

[27] J.-C. Risset and M. V. Mathews, "Analysis of Musical Instrument Tones," *Physics Today*, vol. 22, no. 2, pp. 23–40 (1969).

[28] 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).

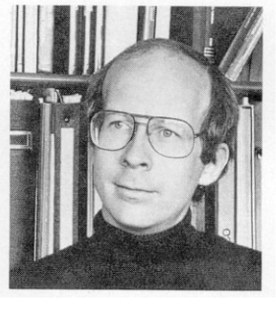
[29] P. R. Samson, "A General-Purpose Digital Synthesizer," *J. Audio Eng. Soc.*, vol. 28, pp. 106–113 (1980 Mar.).

[30] P. R. Samson, "Architectural Issues in the Design of the Systems Concepts Digital Synthesizer," in *Digital Audio Signal Engineering: An Anthology*, J. Strawn, Ed. (William Kaufmann, Los Altos, CA, 1985), pp. 61–93.

[31] M. R. Portnoff, Lecture, Stanford University, Stanford, CA, 1985 May 25.

[32] J. W. Beauchamp, "Data Reduction and Resynthesis of Connected Solo Passages Using Frequency, Amplitude, and 'Brightness' Detection and the Nonlinear Synthesis Technique," in *Proc. 1981 International Computer Music Conf.*, L. Austin and T. Clark, Eds. (North Texas State University, Denton, TX), pp. 316–323.

THE AUTHOR



John Strawn was born in 1950 in Ohio. He received a Bachelor of Music degree (double major in organ and music theory) from Oberlin Conservatory in 1973. From 1973 through 1975 he studied music history and theory on a Fulbright in Berlin. In 1976 he traveled through the Middle East and Asia to Japan, where he conducted independent research on the electronic music scene with a Thomas Watson Fellowship.

Strawn studied with John Chowning at the Center for Computer Research in Music and Acoustics (CCRMA), Department of Music, Stanford University, graduating with a Ph.D. in 1985. During his years at Stanford he was an editor of The *Computer Music Journal* (1977–1982) and served as a consultant for a number of digital audio synthesizer manufacturers such as Mattel Electronics, Music Technology (GDS and Synergy synthesizers), and Kurzweil. In 1984 he established the Computer Music and Digital Audio Series (published by William Kaufmann, Inc.), for which he is series editor. Together with Curt Roads he has published *Foundations of Computer Music* (MIT Press, 1985) and is the author of numerous papers and translations.

From 1985 January through 1986 September, Dr.

Strawn was a member of the Digital Audio Research and Development Group of The Droid Works, an affiliate of Lucasfilm. There he was primarily involved in digital audio signal processing research (including microcode implementation) for the Droid Works' digital audio products such as the SoundDroid. He is now active as a consultant, working in digital signal processing and computer programming. A member of the AES, he is Chairman of the 1987 International Conference on Music and Digital Technology.