ANALYSIS AND SYNTHESIS OF MUSIC USING THE AUDITORY TRANSFORM

by

John Paul Stautner

S.E., Massachusetts Institute of Technology
('79)

Submitted to the Department of
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
in Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY, 1983

© Massachusetts Institute of Technology, 1983

Signature of Author

Department of Electrical Engineering
and Computer Science, May 18, 1983

Certified by

Campbell L. Searle
Thesis Supervisor

Accepted by

Chairman, Department Committee

Archives

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEP 1 1983

LIBRARIES
ANALYSIS AND SYNTHESIS OF MUSIC USING THE AUDITORY TRANSFORM

by

John Paul Stautner

Submitted to the Department of Electrical Engineering and Computer Science on May 19, 1983, in partial fulfillment of the requirements for the degree of Master of Science

ABSTRACT

A simplified model of auditory perception, the auditory transform, is developed using a generalized short-time Fourier transform log-magnitude analysis of acoustic signals. The Short-time Fourier transform log-magnitudes are computed using a filterbank which was designed to incorporate psychoacoustic and physiological observations of auditory sensitivity. Examples of the auditory transform analysis of sounds are given.

An analysis/synthesis algorithm using both the magnitude and phase of the complex filterbank outputs is presented. An iterative extension of this algorithm leads to sound synthesis from short-time Fourier transform magnitude only. An example of sound synthesis from the auditory transform domain requiring signal reconstruction from short-time Fourier transform magnitude only is given.

Further analysis of the auditory transform is used to detect musical features. Pitch information is available in both the frequency and time dimension, and an algorithm is developed for the detection of residue pitch. The spectral and temporal features of the smoothed auditory transform are analyzed using principal components analysis. The analysis results in characteristic patterns and associated weighting functions which may be used for the the modification of sounds, the detection of musical performance gestures and the control of computer sound synthesis.

Thesis Supervisor: Professor Campbell L. Searle

Title: Professor of Electrical Engineering
ACKNOWLEDGEMENTS

Many thanks go to Barry Vercoe for the freedom and support he offered me to pursue graduate studies in music and technology at the Experimental Music Studio. His enthusiasm and confidence inspired me and provided the impetus for much of this research.

Campbell Searle contributed immensely to my education and to many of the ideas in this thesis. I am deeply grateful for our hours and hours of discussion and probing, and for the insights he shared with me.

Much of this thesis would not have possible without the valuable contributions of Bruce Musicus in the Digital Signal Processing Group, and Jim Anderson in the Communications Biophysics Group. I owe much to the many discussions I had with them.

My Research Assistantship was partly supported by Teledyne Acoustic Research. I am grateful for their support. Special thanks go to Bob Berkovitz and Gray Abbott.

I am also grateful to the System Development Foundation, through which much of this research was made possible.

I appreciate the kindness of Norman Farkas and David Horowitz for providing the musical examples.

My parents have always been an inspiration, and have encouraged me throughout my life. I would like to express my love for them, and thank them and the rest of the family for their endless encouragement and concern.

Finally, I would like to acknowledge my wife Kim for giving me strength and persistence. She kept together my mind and body, and supported me throughout the various stages of this research. Her patience and love were especially valuable.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>2</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>3</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>4</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>5</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Overview</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2 AUDITORY MODELING</td>
<td>8</td>
</tr>
<tr>
<td>2.1 Psychoacoustics</td>
<td>8</td>
</tr>
<tr>
<td>2.2 Physiology</td>
<td>10</td>
</tr>
<tr>
<td>2.3 Basic Model Description</td>
<td>11</td>
</tr>
<tr>
<td>2.4 Envelope Compression</td>
<td>15</td>
</tr>
<tr>
<td>2.5 Some Comments on the Model</td>
<td>16</td>
</tr>
<tr>
<td>CHAPTER 3 IMPLEMENTATION AND ANALYSIS Response</td>
<td>18</td>
</tr>
<tr>
<td>3.1 Infinite Impulse Response Implementation</td>
<td>18</td>
</tr>
<tr>
<td>3.2 Examples of Sound Analysis</td>
<td>24</td>
</tr>
<tr>
<td>3.3 Sampling Rates</td>
<td>34</td>
</tr>
<tr>
<td>CHAPTER 4 SYNTHESIS FROM THE AUDITORY TRANSFORM</td>
<td>36</td>
</tr>
<tr>
<td>4.1 Complex Analysis and Synthesis</td>
<td>36</td>
</tr>
<tr>
<td>4.2 Synthesis from the Log-Magnitude Samples</td>
<td>41</td>
</tr>
<tr>
<td>CHAPTER 5 ANALYSIS AND SYNTHESIS OF MUSICAL FEATURES</td>
<td>46</td>
</tr>
<tr>
<td>5.1 Residue Pitch Detection</td>
<td>46</td>
</tr>
<tr>
<td>5.2 Temporal and Spectral Patterns</td>
<td>56</td>
</tr>
<tr>
<td>5.3 Acoustic Modification using Principal Components</td>
<td>71</td>
</tr>
<tr>
<td>5.4 Onset Detection</td>
<td>78</td>
</tr>
<tr>
<td>CHAPTER 6 CONCLUSION AND SUGGESTIONS FOR FURTHER RESEARCH</td>
<td>91</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>94</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>104</td>
</tr>
</tbody>
</table>
1. **INTRODUCTION**

1.1. Motivation

The purpose of this thesis is to implement a simplified computational model of human peripheral auditory processing for use in computer music. We wish to investigate the applications of the model for analysis, synthesis and control of musical sounds.

1.2. Overview

It seems natural that algorithms for sound analysis and synthesis in computer music could incorporate the findings of psychoacoustic and physiological studies on human auditory perception. Acoustic signal analysis algorithms based on models of human auditory perception would exhibit sensitivities similar to the sensitivities of the human auditory system. Sound synthesis performed from the standpoint of the desired auditory representation would define the components of an acoustic signal in a perceptually significant way.

In the human auditory system, the stages of processing an acoustic signal to derive perceptually relevant quantities may be termed the **auditory transform**. The transform is a complicated function, not well understood despite much observation and inquiry. It does, however, represent all that is critical to the understanding of timbre and to the synthesis and control of musical nuance. By creating a simplified model of auditory analysis, we may find a representation of sound which can serve as a template or specification for the perceptually vital characteristics of the acoustic waveform, and which may be used in computer music analysis and synthesis applications.
A computer model of the auditory transform may be thought of as a psychoacoustic preprocessor for acoustic analysis. The outputs of the preprocessor may be used for the detection of auditory features, such as pitch, timbre, loudness, and other characteristics. Such a model would be useful for analysis and transcription of musical performance gestures. Since the model is based on the sensitivities of human hearing, the analysis would characterize only the audible characteristics of the analyzed sounds.

Several schemes for signal analysis have been developed based on models of the auditory analysis (Searle et al., 1979; Zwicker, et al., 1979; Petersen, 1980; Lyon 1982). Our approach for computer music will be to make a very simplified model in terms of familiar signal processing theory. Since the nature of auditory perception is an ongoing research endeavor, it is important to design the model in a flexible manner which allows a variety of parameter specifications. The analysis parameters used in this thesis are approximations to the data which can be easily refined later.

A further possibility of the auditory model would allow signal synthesis from the auditory domain. Once features of interest have been detected in analysis, they could be modified, and the corresponding acoustic waveform synthesized. Therefore, the analysis model must be cast in such a way that it can also be used for signal reconstruction. Again, we desire generality, so that the parameters of the model may be widely varied without destroying the conditions necessary for synthesis.

Conditions necessary for complete signal reconstruction from the auditory domain imply that no information is lost in the analysis process.
Our approach, however, will be to attempt to characterize the analysis result in a manner which allows a summarized or data-reduced description of the patterns present in the auditory domain. These patterns may then be expanded back to the original data rate prior to signal reconstruction. We expect that since the acoustic signal is characterized only in terms of auditory sensitivities as represented by the auditory transform model, a parsimonious specification of the features present in the auditory analysis domain may be developed which captures the audible details of the sound.
2. AUDITORY MODELING

The science of modeling the auditory system involves integration of the many complicated details of auditory behavior. In this thesis, we focus attention on some of the more commonly known properties of the auditory system which have been observed by researchers in psychoacoustics and auditory physiology.

2.1. Psychoacoustics

Since the time of Ohm and Helmholtz, psychoacoustic investigators have attempted to determine the frequency analyzing properties of the auditory system. Beginning with Helmholtz's notion that the ear acts as a bank of resonators (Helmholtz, 1885) systematic research has shown that the resolving capabilities of hearing have non-uniform characteristics, and that the shapes of the assumed auditory filters depend on the frequency region under observation.

Curves obtained from pure tone masking (Wegel and Lane, 1924) and noise masking (Egan and Hake, 1950) studies give some indication of the frequency characteristics of the auditory filters. The resulting masking curves are a measure of the level at which a tone at a particular frequency must be presented in order to be just perceptible in the presence of a masking tone or noise. These experiments suggested that the ear behaves like a set of tuned filters or series of resonant systems, each of which responds selectively to some frequencies while attenuating those frequencies outside of its acceptance band.

Experiments by Fletcher (1940) led to the concept of analysis bandwidths or critical bands to which the ear is sensitive. Fletcher,
using noise of constant spectral power density and varying bandwidths to mask a sine tone, found that the threshold to detect the tone was influenced by the bandwidth of the noise only if the bandwidth was less than a certain critical value. For noise bandwidths greater than the critical bandwidth, he found that the power of the signal at threshold was approximately equal to the product of the critical bandwidth and the noise power density. For noise bandwidths less than the critical bandwidth, the power of the signal at threshold was simply equal to the product of the noise bandwidth and the noise power density. These studies suggested that in the process of signal detection, the auditory system is sensitive to power over limited bandwidths of frequencies, which were later named the critical bandwidths.

Numerous other studies have been undertaken since Fletcher's experiments to determine the nature and characteristics of critical band phenomena. These studies included experiments in loudness summation, masking, phase sensitivity and many other experiments (Zwicker el al., 1957; Scharf, 1970). Most of these experiments result in estimates of the critical bandwidth which are about 2.5 times wider than the original results of Fletcher, but which have the same frequency dependence (Scharf, 1970).

Despite some objections (Nordmark, 1978) that critical band analysis fails to explain various phenomena such as the beats in mistuned consonances, and more recent observations in phase sensitivity over broad frequency intervals, the predominating psychoacoustic evidence suggests that acoustic analysis in terms of critical bands is an appropriate first stage for a music analysis system.
2.2. Physiology

Physiological studies have shown that the relationship of the pressure at the eardrum to the mechanical motion at the basilar membrane may be approximated as a linear system (Weiss, 1966). It is believed that the middle ear behaves essentially as a linear dynamic system (Sisbert, 1973a). Research on the function of the inner ear has shown that the frequency selectivity of the approximately linear response of the basilar membrane varies exponentially with position along the cochlear partition (von Bekesy, 1960). This variation is thought to describe the function of the cochlea as a mechanical spectrum analyzer.

Studies of the physiological data of basilar membrane motion (Flanagan, 1965) and observation of neural response data for clicks and tones (Searle, 1982) have led to estimates of the effective time window for the mechanical short-time analysis made by the basilar membrane. An approximation to the temporal envelope $w_k(t)$ of the displacement response of the basilar membrane, at a point maximally responsive to radian frequency $\Omega_k$, has the form

$$w_k(t) = \begin{cases} \frac{1}{2} \alpha_k^3 t^2 e^{-\alpha_k t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

(1)

where $t$ is time. The parameter $\alpha_k$, which depends on the analysis center-frequency $\Omega_k$, determines the effective duration of the analysis window.

Evidence has shown that the mechanical to neural transduction process involves a non-linear rectification stage, followed by a low-pass smooth-
ing filter and subsequent probabilistic device (Siebert, 1973a and 1973b). The transduction process is thought to function as an envelope detector at high frequencies, but tends to follow the details of the rectified waveform at low frequencies. The detection non-linearity may be modeled as a saturating half-wave rectifier with square-law characteristic.

The auditory nerve fibers rarely fire at sustained rates greater than about 300 firings per second (Green, 1976). Recent evidence suggests, however, that the auditory system utilizes temporal information contained in the distribution of intervals between nerve firings (Siebert, 1982). In our model, then, we should allow a temporal analysis bandwidth greater than 300 samples per second for each channel. Each analysis channel of the model could be thought to represent the information contained in the firing interval patterns of a neuron or group of neurons with similar auditory sensitivity.

2.3. Basic Model Description

A simplified model of the functioning of the inner ear resulting from psychoacoustic and physiological considerations (Siebert, 1973b) is summarized in Figure 1a. At each center-frequency or place along the basilar membrane, the frequency analysis of the ear is modeled by a linear band-pass filter, followed by a non-linear half-wave rectifier, a smoothing filter for envelope detection, and subsequent probabilistic device for neural firings.

A functionally similar implementation may be obtained using a generalized version of short-time Fourier transform (STFT) magnitude analysis for several analysis channels with index k spanning the frequency range.
Figure 1. Auditory analysis models. A) Simplified model from Siebert (1973a and 1973b). B) Short-time Fourier transform log-magnitude analysis.
The generalized STFT $\mathbf{X}(\Omega_k, t)$ at analysis frequency $\Omega_k$ and at time $t$ of an input signal $x(t)$ is defined as

$$\mathbf{X}(\Omega_k, t) = \int_{\tau = -\infty}^{\infty} x(\tau) w_k(t-\tau) e^{-j\Omega_k \tau} d\tau$$  \hspace{1cm} (2)$$

where $w_k(t)$ is an analysis window which depends on the analysis frequency $\Omega_k$, and $j = \sqrt{-1}$.

An equivalent form of Equation 2 obtains the linear filtering interpretation of the STFT:

$$\mathbf{X}(\Omega_k, t) = e^{-j\Omega_k t} \int_{\tau = -\infty}^{\infty} x(\tau) w_k(t-\tau) e^{j\Omega_k (t-\tau)} d\tau$$  \hspace{1cm} (3)$$

or

$$\mathbf{X}(\Omega_k, t) = e^{-j\Omega_k t} \left[ x(t) \ast (w_k(t) e^{j\Omega_k t}) \right]$$  \hspace{1cm} (4)$$

where $\ast$ indicates convolution. We recognize the term inside the brackets in Equation 4 as the convolution of the input signal $x(t)$ with a complex bandpass filter having impulse response $h_k(t)$ where

$$h_k(t) = w_k(t) e^{j\Omega_k t}$$  \hspace{1cm} (5)$$

Therefore, the STFT magnitude may be found directly by computing the magnitude of the bandpass filter output:
\[ |\mathcal{X}(\Omega_k(t))| = |\chi(t) \otimes h_k(t)| \] (6)

The STFT model of auditory analysis is shown in Figure 1b. Notice that the smoothing filter of Figure 1a has disappeared. Here the analysis windows \( w_k(t) \) are chosen to correspond with the analysis windows given in Equation 1. The complex bandpass filter impulse responses of Equation 5 are then given by

\[ h_k(t) = \frac{1}{2} \alpha_k^3 t^2 e^{-\alpha_k t} \sin \Omega_k t \] (7)

The frequency dependent parameter \( \alpha_k \) is defined as

\[ \alpha_k = \begin{cases} 75 \times 2\pi & \frac{\Omega_k}{2\pi} < 500 \text{ Hz} \\ 0.15 \times \Omega_k & \frac{\Omega_k}{2\pi} \geq 500 \text{ Hz} \end{cases} \] (8)

to model the critical bandwidths. The frequency response bandwidth between half-power points for the filters is plotted along with the critical bandwidth data in Figure 2.
2.4. Envelope Compression

Once the signal envelopes have been detected at the various center frequencies, we wish to compress and quantize the envelope levels in a manner consistent with psychoacoustic observations in intensity perception. In this way, we may avoid excessive precision in representing the levels, while maintaining the perceptually relevant differences in intensity.

It is well known in psychoacoustics that the just noticeable difference in loudness of a tone corresponds roughly with the logarithm of the intensity level, although the sensitivity and range of the variation
depends on frequency (Roederer, 1975). The analysis channels of our model should therefore measure the logarithm of the intensity of the envelope fluctuations.

Experiments by Riesz (1928) were performed to find the just detectable change in intensity for tones at various frequencies. He found that the most sensitive region is in the neighborhood of 1000 Hz, and that the just noticeable difference in intensity depends on the level of the tone presented. We shall take the smallest noticeable difference he found, about .2 decibels, as the quantization unit for all the envelope intensity levels.

2.5. Some Comments on the Model

The advantage of the auditory analysis model presented here is that it performs a rather simplified frequency analysis of sound while maintaining a correspondence with the observed behavior of the auditory system. Furthermore, the formulation in terms of the generalized short time Fourier transform allows us to understand the model in terms of familiar signal processing theory.

Among the limitations is the fact that the model does not account for the effects of adaptation and saturation of the nerve firings. Physiological data also suggest that the frequency response curves up to the neural level are actually asymmetric, with the high frequency slope falling off much more rapidly than the low frequency slope (Kiang, 1974). No attempt has been made to match these characteristics.

It can be shown that for an appropriately chosen smoothing filter, the auditory analysis model of Figure 1a (before the probabilistic device)
approximates the measurement of STFT magnitude (Flanagan, 1965; Anderson and Searle, 1983). Computation of STFT magnitude therefore implicitly implements an envelope detector and smoothing device. The STFT log-magnitude analysis of Figure 1b serves to function as a simplified simulation of auditory analysis model of Figure 1a.

Current understanding of auditory physiology indicates that the neural firings tend to be synchronized with the phase of the stimulus waveform for up to 3 or 4 kHz waveform frequency (Johnson, 1980). The analysis configuration of the auditory model in Figure 1a, using a half-wave rectifier, will tend to follow the details of the rectified waveform for all analysis bandwidths provided the smoothing filter has broad enough bandwidth. Envelope detection using the magnitude-squared of the STFT will not follow the details of the filtered waveform. A phase-locked temporal response will be absent in the STFT model for certain input signals such as a sine wave. The structure of the temporal response will nonetheless reflect the interaction of frequency components within the particular analysis bandwidth.

For our present purposes the STFT model will adequately imitate auditory analysis. As we shall see, the analysis and synthesis formulation is not restricted to any particular window characteristics, so these may be easily changed to obtain a more accurate representation. Other parameters such as the frequency dependence of the filter bandwidths, and the degree of frequency response overlap for adjacent filters may also be chosen with little or no restriction.
3. **IMPLEMENTATION AND ANALYSIS RESPONSE**

A digital implementation of the STFT log-magnitude model of auditory analysis presented in Chapter 2 is developed. The infinite impulse response implementation which is presented is especially suited for use on computers with limited memory space. A computer program was written allowing the specification of some of the auditory filterbank parameters, such as the frequency range and number of filters. Examples of the auditory transforms of a variety of input signals are presented.

3.1. Infinite Impulse Response Implementation

An infinite impulse response (IIR) implementation of the bandpass filters given in Eq. 7 may be obtained using the impulse invariant design procedure (Oppenheim and Schafer, 1975). The unit sample impulse response \( h_{D_k}(n) \) of the complex bandpass filter in the \( k \)th analysis channel, corresponding to analog center-frequency \( \omega_k \), is given by

\[
 h_{D_k}(n) = T h_k(nT) \tag{9}
\]

where \( T \) is the sampling period, \( h_k(t) \) is the continuous time impulse response from Eq. 7, and \( n \) is the sample number. Carrying out the substitution, we find

\[
 h_{D_k}(n) = \frac{1}{2} \alpha_k^3 T^3 n^2 e^{-\left(\alpha_k T + j\omega_k\right)n} \tag{10}
\]

where \( \omega_k \) is the digital center-frequency \( \omega_k T \). From here on we will refer to the impulse response \( h_{D_k}(n) \) of the digital implementation simply as \( h_k(n) \).
The Fourier transform \( H_k(e^{j\omega}) \) of \( h_k(n) \) is given by

\[
H_k(e^{j\omega}) = \frac{\gamma_k e^{-j\omega} (1 + \gamma_k e^{-j\omega})}{(1 - \gamma_k e^{-j\omega})^3}
\]  \hspace{1cm} (11)

where \( \omega \) is the digital frequency. \( \gamma_k \) is the complex coincident pole location,

\[
\gamma_k = e^{-\alpha_k T} + j\omega_k
\]  \hspace{1cm} (12)

and \( C_k \) is a scaling factor giving unity gain at the filter response center frequency:

\[
C_k = \frac{1}{2} \alpha_k^3 T^3
\]  \hspace{1cm} (13)

A convenient implementation of the filter \( h_k(n) \) may be obtained by noticing that \( H_k(e^{j\omega}) \) is a cascade of first order filters with filter coefficient \( \gamma_k \):

\[
H_k(e^{j\omega}) = \left( \frac{1}{1 - \gamma_k e^{-j\omega}} \right) \left( \frac{\gamma_k e^{-j\omega}}{1 - \gamma_k e^{-j\omega}} \right) \left( \frac{1 + \gamma_k e^{-j\omega}}{1 - \gamma_k e^{-j\omega}} \right) C_k
\]  \hspace{1cm} (14)

The digital flow graph implementing the filter is shown in Figure 3. Each branch of the network diagram is labeled with the corresponding frequency domain operation. Arrows converging on a node imply summation. The short-time Fourier transform log-magnitude may be calculated directly by first finding the magnitude of the filter output, and then computing the logarithm of the result.

The bandpass filter implementation of Figure 3 requires three complex
multiplication operations, three and one half adds, and two real multiplies per output point. Three additional floating point operations are required whenever the log-magnitude is computed. Since the output signal may be sampled at a rate much lower than the input sampling rate, the final complex addition and scaling by \( C_k \) may be avoided except when computation of the log-magnitude is actually required. This results in filter operation using three complex multiplies and two and one-half complex adds for all intermediate points.

The IIR implementation is quite useful when working with limited computer memory space. Finite impulse response (FIR) filter implementations, however, are expected to be less computationally intensive (Schwede, 1983). An FIR implementation would have the advantage that intermediate output points need not be computed as in the IIR case. Therefore, in computing environments where a large amount of memory space is available for storing the complex bandpass filter impulse responses, a "hopped" FIR implementation would appear to be advantageous.

A filterbank design program was implemented, allowing specification of the frequency range and the amount by which adjacent filter frequency responses are overlapped. The low frequency and high frequency endpoints were specified as the half-power point for the lower skirt of the lowest frequency, and the half-power point for the upper skirt of the highest frequency filter, respectively. The point of overlap was defined as the attenuation in decibels where adjacent filter responses would coincide.

The parameter \( \alpha_k \), which determines the bandwidths of the filters at the various frequencies, was defined to have the values given in Eq. 8. The frequency response for a filter with center-frequency at 988 Hz is
Figure 3. IIR implementation of complex bandpass filter for auditory analysis.
Figure 4. Frequency response of the one of the bandpass filters used for auditory analysis. Center-frequency is at 988 Hz.
Figure 5. Algorithm for computation of auditory transform.
shown in Figure 4.

Computation of the auditory transform is shown in Figure 5. The auditory transform samples corresponding to an input signal are found by computing the log-magnitude of the sampled output channels of the auditory filterbank. These values are then scaled and quantized to integer values where each unit corresponds to .2 decibels. Values of the magnitudes below a predetermined threshold (such as zero) are set to the value of the threshold to avoid excessively negative logarithms.

3.2. Examples of Sound Analysis

In order to better understand the auditory transform analysis, we first view examples of the analysis in response to ideal, synthetic input waveforms. Our plots will show the magnitude dependence as a function of time and frequency.

The example of Figure 6 is the auditory transform of a unit sample impulse occurring at .01 seconds. The filterbank consists of 24 filters with center-frequencies spanning the range from 99 Hz to 3806 Hz (see Table 1). The filters have impulse responses given in Eq. 10, where the parameter \( \alpha \) is determined by each center-frequency according to Eq. 8. Each analysis channel output is plotted separately in Figure 6, with the low frequency channels at the top of the figure. This viewpoint is particularly useful for observing the temporal details of the analysis result. In this case, we see clearly the individual shapes of the analysis windows. The maximum level of 360 units (corresponding to 72 decibels) is attained by the highest frequency channel (number 23).

In order to present the analysis in a manner which allows us to
better interpret the information visually, we plot the auditory transform magnitude contours as a surface in three-dimensional space. The three-dimensional representation of the auditory response to the unit sample impulse is shown in Figure 7. The horizontal axis shows the analysis channel center frequency, the vertical axis is the log-magnitude, and the time axis increases into the page. For high frequency channels, we notice that the response is immediate and dies away very quickly, thereby accurately localizing the occurrence of the unit sample in time. The low frequency channels have a slower, more smeared out response corresponding to the longer analysis windows.

Figure 8 shows the auditory response to a sine tone burst at 988 Hz. Note that the energy is initially widely smeared in frequency, with the high frequency channels responding almost immediately after the onset as before. After a few milliseconds, however, a peak appears at the frequency of the tone as the transient effects die away.

Figure 9 shows the auditory analysis of a complex tone with fundamental frequency 200 Hz and 24 equal amplitude harmonics. Notice that the lowest five harmonics are clearly resolved, showing up as long narrow ridges extending in temporal direction. The upper harmonics, however, become more and more difficult to pick apart as adjacent harmonics fall within the same critical bandwidth. Instead we see a temporal beating pattern which results from the interaction of adjacent harmonics within a critical bandwidth. The periodic shape of the channel outputs is particularly apparent in the corresponding temporal plots of Figure 10. The frequency of the periodicity corresponds to the fundamental frequency of 200 hz. This is an example of the temporal variations in STFT magnitude caused
Table 1.

Filter center-frequencies for auditory transform analyses in Figures 6 through 10.

<table>
<thead>
<tr>
<th>Analysis channel</th>
<th>Center frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>1</td>
<td>161</td>
</tr>
<tr>
<td>2</td>
<td>223</td>
</tr>
<tr>
<td>3</td>
<td>285</td>
</tr>
<tr>
<td>4</td>
<td>347</td>
</tr>
<tr>
<td>5</td>
<td>409</td>
</tr>
<tr>
<td>6</td>
<td>471</td>
</tr>
<tr>
<td>7</td>
<td>535</td>
</tr>
<tr>
<td>8</td>
<td>605</td>
</tr>
<tr>
<td>9</td>
<td>684</td>
</tr>
<tr>
<td>10</td>
<td>774</td>
</tr>
<tr>
<td>11</td>
<td>874</td>
</tr>
<tr>
<td>12</td>
<td>988</td>
</tr>
<tr>
<td>13</td>
<td>1117</td>
</tr>
<tr>
<td>14</td>
<td>1263</td>
</tr>
<tr>
<td>15</td>
<td>1428</td>
</tr>
<tr>
<td>16</td>
<td>1614</td>
</tr>
<tr>
<td>17</td>
<td>1824</td>
</tr>
<tr>
<td>18</td>
<td>2062</td>
</tr>
<tr>
<td>19</td>
<td>2331</td>
</tr>
<tr>
<td>20</td>
<td>2635</td>
</tr>
<tr>
<td>21</td>
<td>2979</td>
</tr>
<tr>
<td>22</td>
<td>3367</td>
</tr>
<tr>
<td>23</td>
<td>3806</td>
</tr>
</tbody>
</table>
Figure 6. Auditory transform of a unit sample impulse.
Figure 7. Auditory transform of a unit sample impulse, plotted as a surface in three dimensions.

FREQUENCY
(Channel Number)
Figure 8. Auditory transform of a sine tone burst at 988 Hz.
Figure 9. Auditory transform of a complex tone with fundamental frequency 200 Hz. Notice the resolved harmonics as well as the temporal periodicity.
Figure 10. Auditory transform of complex tone (same as Figure 9) plotted to show temporal periodicity more clearly.
Table 2.

Filter center-frequencies for auditory transform analysis of clarinet tone in Figure 11. Also used for clarinet melody in Section 5.1.

<table>
<thead>
<tr>
<th>Analysis channel</th>
<th>Center frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>1</td>
<td>102</td>
</tr>
<tr>
<td>2</td>
<td>144</td>
</tr>
<tr>
<td>3</td>
<td>186</td>
</tr>
<tr>
<td>4</td>
<td>228</td>
</tr>
<tr>
<td>5</td>
<td>270</td>
</tr>
<tr>
<td>6</td>
<td>312</td>
</tr>
<tr>
<td>7</td>
<td>354</td>
</tr>
<tr>
<td>8</td>
<td>396</td>
</tr>
<tr>
<td>9</td>
<td>438</td>
</tr>
<tr>
<td>10</td>
<td>480</td>
</tr>
<tr>
<td>11</td>
<td>523</td>
</tr>
<tr>
<td>12</td>
<td>569</td>
</tr>
<tr>
<td>13</td>
<td>619</td>
</tr>
<tr>
<td>14</td>
<td>673</td>
</tr>
<tr>
<td>15</td>
<td>731</td>
</tr>
<tr>
<td>16</td>
<td>794</td>
</tr>
<tr>
<td>17</td>
<td>863</td>
</tr>
<tr>
<td>18</td>
<td>938</td>
</tr>
<tr>
<td>19</td>
<td>1018</td>
</tr>
<tr>
<td>20</td>
<td>1106</td>
</tr>
<tr>
<td>21</td>
<td>1202</td>
</tr>
<tr>
<td>22</td>
<td>1306</td>
</tr>
<tr>
<td>23</td>
<td>1419</td>
</tr>
<tr>
<td>24</td>
<td>1542</td>
</tr>
<tr>
<td>25</td>
<td>1675</td>
</tr>
<tr>
<td>26</td>
<td>1820</td>
</tr>
<tr>
<td>27</td>
<td>1977</td>
</tr>
<tr>
<td>28</td>
<td>2147</td>
</tr>
<tr>
<td>29</td>
<td>2332</td>
</tr>
<tr>
<td>30</td>
<td>2532</td>
</tr>
<tr>
<td>31</td>
<td>2750</td>
</tr>
<tr>
<td>32</td>
<td>2987</td>
</tr>
</tbody>
</table>
Figure 11. Auditory transform of a clarinet note.
by the interaction of frequency components within the same critical bandwidth.

Figure 11 shows the auditory transform analysis of a single clarinet note. The filterbank used for the analysis consisted of 33 filters spanning the range from 60 Hz to 2987 Hz (see Table 2). The first and third harmonics are particularly apparent in the plot, showing up as narrow ridges extending upward near channel 5 and channel 16 respectively (corresponding to a fundamental frequency of about 262 Hz). Slight variations in the frequency of the harmonics are clearly seen as shifts in the horizontal position of the ridges going up the page. The fifth and seventh harmonics can also be identified individually, as well as the very weak presence of the second harmonic. Again, the temporal beating resulting from the interaction of upper harmonics within the same analysis bandwidth is a prominent feature of the auditory representation.

3.3. Sampling Rates

The plots in the previous section were made by sampling all of the output channels at the same rate. A more appropriate method of sampling the channel outputs would be to determine a sampling rate consistent with the bandwidth of the analysis bandpass filters. In order to make the book-keeping in the algorithms easier, we choose the sampling rates of the channels to be related to the signal input sampling rate (ISR) by factors of \( \left( \frac{1}{2} \right)^m \), for integer m.

Let us first consider the sampling rate required to represent the output of an ideal bandpass filter (Crochiere and Rabiner, 1981). The sampling frequency necessary is equal to the bandwidth of the
filter. Since the auditory bandpass filters are not ideal bandpass filters, there will always be some signal degradation except when the filter output is sampled at the same rate as the input. The amount of degradation can be controlled, however, by defining an effective bandwidth during the filterbank design procedure.

Since the bandpass filter responses have smooth response skirts, the effective bandwidth must be defined relative to the response attenuation characteristics. The effective signal bandwidth for each bandpass filter output may be defined as the frequency interval separating similar response attenuations on each side of the filter center-frequency, where the attenuation value is selected during the filterbank design procedure. The actual sampling rates are then set to the lowest value of \((\frac{1}{2})\text{ISR}\) which is also greater than the effective bandwidth. Since the magnitude-squared signal has approximately twice the bandwidth of the bandpass signal, care should be taken to choose high enough sampling rates to represent the temporal channel signals adequately.

By sampling the analysis channel outputs at a rate consistent with each of the bandpass filter bandwidths, we expect to preserve the information in each analysis channel in a uniform way. This strategy can also be thought of as a method for sampling the auditory transform surface in terms of frames which have finite duration and which span the frequency range. The minimum possible duration of a frame is equal to the maximum sampling period of among the analysis channels.
4. **SYNTHESIS USING THE AUDITORY TRANSFORM**

In applications where we wish to make modifications on an acoustic signal in the auditory transform domain, an inverse transform is required in order to synthesize sound from the modified representation. Siebert (1960) has shown that signal recovery is theoretically possible from short-time Fourier transform magnitude only. Recently, algorithms have been developed for signal recovery from sampled STFT magnitude in practical applications (Nawab, 1982; Anderson and Searle, 1983; Griffin and Lim, 1983).

An analysis/synthesis method using the magnitude and phase resulting from the complex bandpass filter analysis is first presented. An iterative extension of this algorithm is then used for synthesis from STFT magnitude only.

4.1. Complex Analysis and Synthesis

We consider first the case where we preserve both magnitude and phase data resulting from a complex filterbank, with no reduction in sampling rate (Musicus, 1982). The complex bandpass filter outputs \( y_k(n) \) in the \( k^{th} \) analysis channel resulting from input signal \( x(n) \) is given by:

\[
y_k(n) = x(n) \otimes h_k(n) \tag{15}
\]

where \( h_k(n) \) is the bandpass filter impulse response for the \( k^{th} \) analysis channel. The Fourier transform of Equation 15 is
\[ Y_k(e^{i\omega}) = X(e^{i\omega}) H_k(e^{i\omega}) \quad (16) \]

Since there is one such equation for each of the \( k \) analysis channels, the Fourier transform \( X(e^{i\omega}) \) of \( x(n) \) is overdetermined by a set of \( N \) equations, where \( N \) is the total number of analysis filters. The least-square solution \( \hat{X}(e^{i\omega}) \) for \( X(e^{i\omega}) \) is given by

\[
\hat{X}(e^{i\omega}) = \frac{1}{\sum_{k=0}^{N-1} |H_k(e^{i\omega})|^2} \cdot \sum_{k=0}^{N-1} H_k^*(e^{i\omega}) Y_k(e^{i\omega}) \quad (17)
\]

from which we may find \( \hat{X}(e^{i\omega}) \) provided the summation in the denominator of Equation 17 does not evaluate to zero for any value of the frequency \( \omega \).

We now interpret this least-square solution. Each complex channel output resulting from the complex bandpass filters is passed through a bandpass filter having the conjugate symmetric impulse response. These signals are then summed and passed through an equalizing filter with frequency response \( H_{EQ}(e^{i\omega}) \) given by

\[
H_{EQ}(e^{i\omega}) = \frac{1}{\sum_{k=0}^{N-1} |H_k(e^{i\omega})|^2} \quad (18)
\]

Since the input signal \( x(n) \) is a real signal for acoustic data, the Fourier transform \( X(e^{i\omega}) \) has conjugate symmetry on the frequency axis:
\[ \mathcal{X}^*(e^{j\omega}) = \mathcal{X}(e^{j\omega}) \]  

(19)

It is therefore possible to solve Eq. 17 for values \( 0 < \omega < \pi \), and use the conjugate symmetry property to fill in \( \hat{\mathcal{X}}(e^{j\omega}) \) for values of \( \omega \) less than 0. This result may be achieved by redefining the equalizing filter as

\[
H'_E(e^{j\omega}) = \begin{cases} 
H_{EQ}(e^{j\omega}) & \omega = 0 \text{ and } \pi \\
2H_{EQ}(e^{j\omega}) & 0 < \omega < \pi \\
0 & -\pi < \omega < 0
\end{cases}
\]  

(20)

Taking the real part of the equalizer output then restores the negative frequency components of \( \hat{\mathcal{X}}(e^{j\omega}) \).

The synthesis filters \( H_k(e^{j\omega}) \) and equalizing filter \( H'_E(e^{j\omega}) \) used for signal reconstruction are in general non-causal. In practice, they can be implemented with FIR filters by introducing a delay to make them realizable.

For IIR analysis filters with frequency response \( H_k(e^{j\omega}) \), synthesis involves a more complicated procedure. To see how the corresponding non-causal IIR synthesis filter with frequency response \( H^*_k(e^{j\omega}) \) may be implemented, consider a signal \( \mathcal{Z}(\eta) \) to be filtered. We observe first that the corresponding time reversed and complex-conjugated signal, \( \mathcal{Z}^*(-\eta) \), passed through the filter with frequency response \( H_k(e^{j\omega}) \), results in a signal \( \mathcal{Y}_k(\eta) \) with frequency response \( \mathcal{Y}_k(e^{j\omega}) \) given by
\[
Y_k(e^{j\omega}) = H_k(e^{j\omega}) Z^*(e^{j\omega}) \tag{21}
\]

Time reversal and complex conjugation of \(Y_k(n)\) gives the signal \(V_k(n)\),

\[
V_k(n) = Y_k^*(-n) \tag{22}
\]

After taking the Fourier transform of Eq. 22 and using Eq. 21 to substitute for \(Y_k(e^{j\omega})\), we find

\[
V_k(e^{j\omega}) = H_k^*(e^{j\omega}) Z(e^{j\omega}) \tag{23}
\]

Therefore, an IIR implementation of a filter with response \(H_k^*(e^{j\omega})\), where \(H_k(e^{j\omega})\) has a causal and stable impulse response, may be achieved by time reversal and complex conjugation before and after filtering by \(H_k(e^{j\omega})\).

Since the IIR analysis/synthesis implementation requires storage of intermediate data files, it is generally desirable to decimate the complex channel outputs. In the present implementation, this is achieved first by choosing the effective bandwidth of the bandpass filter output as described in Section 3.3. The complex channel outputs are then sampled at a sampling rate equal to the effective bandwidth in Hz.

Restoration of the original sampling rate prior to the synthesis filters is achieved by first upsampling the channel signals and then using a linear interpolator for each center frequency. If the center frequency of the channel is \(\omega_k\) and the intermediate results have been decimated by a factor of \(l_k\), then the impulse response \(h_{\text{ori}}(n)\) of the interpolator is
Figure 12. Analysis and least-squares synthesis using the complex auditory filterbank. Here both magnitude and phase information are preserved.
given by

\[ h_{I_k}(n) = \begin{cases} 
(1 - \frac{|n|}{l_k})e^{j\omega_k n} & -l_k < n < l_k \\
0 & \text{otherwise}
\end{cases} \] (24)

This interpolator is by no means ideal for restoring the sampling rate of the decimated complex bandpass signals. The implementation of Eq. 24 is convenient, however, requiring only two input samples at a time to generate the intermediate points.

Research currently in progress will determine the optimum procedure for signal reconstruction given the sampled STFT representation for arbitrary and non-uniformly spaced analysis windows (Musicus et al., 1983). These results will lead to efficient synthesis algorithms which include implicitly the rate at which the STFT of a signal is sampled.

The complex analysis and and least-squares synthesis procedure discussed in this section is shown in Figure 12. In this figure, the box labeled "interpolation" implies both zero-padding by factor \( l_k \) and filtering with \( h_{I_k}(n) \).

4.2. Synthesis from the Log-Magnitude Samples

Our interest in magnitude-only reconstruction is from the standpoint of auditory analysis, modifications, and subsequent synthesis. We have seen that a simplified model of auditory analysis detects the signal intensity within the critical band frequency bands spanning the audible
range. Signal modification should occur, then, in terms of these magnitude quantities. Furthermore, after synthesis is complete, subsequent auditory transform analysis of the synthesized signal should reflect the desired modifications.

We adopt here a simple iterative extension of the complex analysis and synthesis approach, whereby a signal is repeatedly analyzed and synthesized subject to the desired magnitude constraints (Nawab, 1982; Griffin and Lim, 1983). The iteration is started by making an arbitrary estimate of the signal to be reconstructed. Random noise or even a zero signal is adequate for this purpose. This initializing signal is then analyzed using the auditory filterbank, and both the magnitudes and phases in the analysis result are preserved. The magnitudes are then modified by replacing them with the known magnitudes of the signal to be synthesized. A signal is then synthesized using the synthesis filters and equalizing filter described in Section 4.1. This signal is analyzed again, and the process is repeated. During this process, the phases of the complex auditory filterbank analysis outputs may change as the signal converges to the desired magnitude characteristics.

Synthesis is complete when the analysis magnitudes of the synthesized signal are sufficiently close to the desired magnitudes. The error may be monitored by observing the sum of the absolute values of the differences between the desired log-magnitude samples and the log-magnitude samples resulting from the current version of the synthesized signal. A flow graph for this reconstruction procedure is shown in Figure 13.

The iterative magnitude-only approach to signal reconstruction can be used to synthesize a signal given only its STFT magnitude samples. Figure
14a shows the acoustic waveform plot of a single tuba note. An auditory transform analysis was made using 42 filters spanning the frequency range from 79 to 3177 Hz and overlapped at the -.6 db point. The channels were sampled and quantized so the resulting data rate was eight times greater than the original data rate. Signal reconstruction from the auditory-transform log-magnitude samples was then initiated with random noise of the same duration as the original signal. After 65 iterations, the reconstructed time domain signal appears as in figure 14b and sounds nearly identical to the original waveform.

Although signal reconstruction from STFT log-magnitude is still inefficient from the standpoint of data rate and computation, the general procedure may be applied successfully with rapid convergence when a good initial guess exists for starting the reconstruction iteration. In applications where a signal is to be analyzed, modified in the auditory domain, and then synthesized, the unmodified signal may serve as the initializing signal in the reconstruction iteration. Convergence in this case can be quite rapid. An example of signal modification and reconstruction using a variation of the iterative algorithm presented here is given in chapter 5.
Figure 13. Signal reconstruction using only the auditory transform log-magnitude samples. The auditory transform of a signal is iteratively modified until it is close enough to the desired result. (The auditory filterbank analysis and synthesis stages are equivalent to the complex analysis and synthesis stages of Figure 12).
Figure 14. A) Original acoustic waveform of a tuba note. B) Acoustic waveform of tuba note reconstructed from the auditory transform representation (log-magnitudes only), using the iterative procedure of Figure 13. Time axis is labeled in samples, for sampling period of .0001 sec.
5. ANALYSIS AND SYNTHESIS OF MUSICAL FEATURES

The time, frequency, and magnitude representation of sound presented in the preceding chapters may be used for the detection of features in musical signals, such as the pitch, spectral evolution, and onset times of acoustic events. As we have seen, the magnitude samples of the short-time Fourier transform analysis can be used to fully reconstruct an acoustic signal. Therefore, the features characterizing the acoustic waveform exist on the three dimensional auditory transform representation. In order to analyze, modify, and synthesize musical sounds, we attempt to separate the features and obtain alternative presentations of the auditory transform data which highlight the desired perceptible quantities.

5.1. Residue Pitch Detection

There are many auditory mechanisms which are believed responsible for the detection of pitch. Most of these mechanisms can be grouped into one of two types of theories. The place theory postulates that the auditory system determines pitch from the spectral analysis pattern. Residue or temporal theories presuppose that the cues for pitch are found in the temporal firing patterns of the neurons. Both place and timing are probably important to the pitch percept in some combination (Siebert, 1973a).

We have seen in Section 3.2 that the lowest few harmonics of a complex tone are individually resolved in the auditory transform representation. Psychoacoustic pitch detection algorithms based on spectral analysis have already been developed (Goldstein, 1973; Terhardt et al., 1982), and could potentially be applied using these resolved components. We present here a method for the estimation of the fundamental frequency
of complex tones based on the temporal fluctuations of the auditory transform envelope signals.

Our approach is to implement a second stage of STFT magnitude analysis at the output of each initial analysis channel. By analyzing the temporal variations of each channel using another stage of running spectral analysis, we may identify the various rates of modulation occurring in the channels. A similar approach has already been psychoacoustically justified as a possible model for higher level processing in the auditory system (Martens, 1982). The second analysis stage gives a measure of the periodicities present in the envelope signals.

As Figure 9 and Figure 11 in Section 3.2 demonstrate, the auditory transform of a complex harmonic tone resolves the first few harmonics in the spectral dimension. For higher harmonics, two or more adjacent partials fall within the bandwidth of the analysis channel, resulting in temporal beating. Let us consider a signal \( x(n) \) comprised of two harmonic overtones of the fundamental frequency \( \omega_0 \),

\[
x(n) = \cos[m\omega_0 n] + \cos[(m+1)\omega_0 n]
\]

where \( m \) is an integer. Figure 15a shows the frequency response of the signal components and a neighboring complex bandpass analysis filter. Next, let us imagine the various stages of the auditory transform analysis channel detecting \( x(n) \).

Equation 25 can also be expressed as
\[ x(n) = \frac{1}{2} \left\{ (e^{j\omega_0 n} + e^{-j\omega_0 n}) + (e^{j(m+1)\omega_0 n} + e^{-j(m+1)\omega_0 n}) \right\} \] (26)

The output \( x_2(n) \) of the complex bandpass filter (see Figure 15b) contains only the positive frequency components,

\[ x_2(n) = \frac{1}{2} \left[ a e^{j\omega_0 n} + b e^{j(m+1)\omega_0 n} \right] \] (27)

\[ = \frac{1}{2} e^{j\omega_0 n} \left[ a + b e^{j\omega_0 n} \right] \]

Here the complex values \( a \) and \( b \) are the filter frequency response at the location of each partial. In the next stage of analysis, we find \( x_3(n) \) corresponding to the magnitude-squared of \( x_2(n) \).

\[ x_3(n) = \left| x_2(n) \right|^2 \] (28)

\[ = \frac{1}{4} \left\{ |a|^2 + |b|^2 + 2|ab^*| \cos(\omega_0 n - \phi) \right\} \]

where \( \phi \) is the phase of \( ab^* \). The output of the analysis channel therefore has a component at frequency \( \omega_0 \) in addition to the constant value \( \frac{1}{4}(|a|^2 + |b|^2) \).

In general, the magnitude-squared channel output resulting from input signals consisting of several adjacent and harmonically related frequency components with fundamental frequency \( \omega_0 \) will be periodic with frequency \( \omega_0 \). The effect of the filter response shape is to enhance the beating
of adjacent partials near the response peak or center frequency. While subsequent logarithmic compression may introduce some distortion of the temporal fluctuations, a frequency analysis of the output signal $\gamma(n)$ of the auditory transform channel should show a peak at a frequency corresponding with the predominant frequency spacing of partials within the response bandwidth of that channel.

If a spectral analysis is performed on all of the channels, the resulting magnitude spectra may be averaged to determine the predominant and frequently occurring frequency intervals between partials. The resulting histogram would then show peaks at these values, presumably estimates of the various fundamental frequencies of complex tones present in the acoustic input. A flow graph for the computation of a running histogram or "pitch periodogram" is shown in Figure 16. After spectral analysis has been performed on each of the auditory transform channels, the magnitude spectra are superimposed to form the histogram. Therefore, any periodicities which predominate across the auditory analysis channels will show up as peaks in the histogram. The histogram is recomputed at regular intervals in time, limited only by the sampling rates of the auditory transform channels.

A clarinet performance of the melody in Figure 17a was analyzed (see Table 2), and the pitch periodogram was computed and is shown in figure 17b. The second stage of spectral analysis was performed using a uniform filterbank with rectangular frequency windows of 40 Hz bandwidth, with center-frequencies spanning the range from 210 Hz to 890 Hz. The fundamental frequency is clearly visible as the largest peak, including the details of the trills. Notice that the pitch contour becomes compressed
Figure 15. A) Frequency response of two harmonically related cosine waves (solid lines), compared with the frequency response of a complex bandpass filter (dotted line). B) Processing stages for the corresponding auditory transform analysis channel.
Figure 16. Computation of pitch periodogram based on temporal periodicities in the auditory transform channels. A second stage of spectral analysis is performed on each channel, and the results are superimposed to display the predominating periodicities.
Figure 17. A) Clarinet melody from Debussy, Premiere Rhapsodie (Durand, 1910)
B) Pitch periodogram of clarinet performance of melody in A.
(continued on next page)
Figure 18. A) Synthesis of a signal comprised of two complex harmonic tones. B) Pitch periodogram of signal synthesized in A.
in frequency for the low notes. A possible improvement would be to choose the frequency selectivity of the second spectral analysis stage to correspond with human pitch sensitivity or the musical scale. Also apparent in Figure 17b is a peak at twice the fundamental frequency, presumably caused by the strong, but more widely spaced odd harmonics of the clarinet falling within the same response curve for some of the auditory filters. (The missing note at 2.1 seconds was missing in the original recording).

Since the pitch periodogram displays the predominating frequency intervals present in the auditory filters spanning the frequency range, it may be applied to the detection of fundamental frequencies of two or more simultaneous harmonic sources. A signal was synthesized as shown in Figure 18a, comprised of two complex harmonic tones with fundamental frequency 250 Hz and 600 Hz. An auditory transform analysis was then performed, and the pitch periodogram was computed. The second stage of spectral analysis was performed using a uniform filterbank with rectangular frequency windows of 40 Hz bandwidth, spanning the frequency range from 50 Hz to 890 Hz. The result, shown in figure 18b, demonstrates that both the fundamental frequencies appear prominently as peaks in the pitch periodogram (The presence of the 600 Hz fundamental shows up in 610 Hz analysis channel). The large peak at low frequency comes from the slowly varying and constant signal components in the auditory analysis channels.

These results indicate that pitch detection based on the temporal variation of the auditory transform surface has potential application to the separation of simultaneous pitches of musical sounds. A more powerful method may eventually be achieved by combining these temporal results with
pitch detection algorithms which account for the spectral resolution of individual partial frequencies available on the auditory transform surface as well.

5.2. Temporal and Spectral Patterns

We wish now to observe the broader variations in shape of the auditory transform magnitude contours in order to gain insight into the characteristic spectral and temporal patterns for musical sounds. By inserting a smoothing filter into each analysis channel just after the magnitude-squared operation, we may effectively filter out temporal fluctuations resulting from the pitch, and reduce the bandwidth of the channel. In doing so, however, the temporal onsets of the envelope signals will be somewhat smeared out as well. The resulting analysis yields a temporally smoothed version of the auditory transform analysis surface.

No fundamental changes are required in the original analysis procedure in order to obtain the smoothed version of the auditory transform. Since the logarithms of the magnitude-squared signals are available on the auditory transform surface, the corresponding magnitude-squared signal may be recovered by taking the inverse logarithm. These magnitude-squared channel signals are then smoothed using the smoothing filters, and the logarithms of the resulting signals are computed to obtain the smoothed auditory transform surface.

The impulse response of the smoothing filters used for filtering the magnitude-squared signals were Hamming windows. Each channel was filtered using filters of identical length. Because of the multiple sampling rates, the effective temporal extent of the filters varied depending on the chan-
nel signal being smoothed. In general, the high frequency channels with high sampling rates and broad bandwidth have smoothing filters with shorter duration than the low frequency channels which have narrow bandwidth and low sampling rates.

Figure 19 shows the auditory transform analysis of a short excerpt from a sentence of speech. Sixteen analysis channels were used, spanning the frequency range from 99 Hz to 3374 Hz (see Table 3). Figure 20 shows the entire sentence after the magnitude-squared signals of each channel have been smoothed using a thirteen point hamming window, and decimated by a factor of 4 (see Table 4). Comparison of the smoothed version from frame 60 to 70 with the corresponding unsmoothed segment shown in Figure 19 suggests that although the spectral shape and the onset characteristics of the words are smeared, they stand out quite clearly in the smoothed version.

The question we ask now is: are there some underlying patterns or combination of patterns which could be used to characterize the fluctuations of the smoothed auditory transform surface? If such patterns can be identified which could be used to closely represent the various fluctuations of the surface resulting from a particular kind of sound, then these patterns can be used for analysis and identification of sounds as well as for synthesis of sounds.

First we must choose a unit of analysis for interpreting the local variations in the auditory transform surface. We require a unit of analysis which allows us to observe the variations both in frequency and in time. One possibility is to use the frame concept introduced in Section 3.3. A single analysis frame spans the entire frequency range and has
Table 3.

Filter center-frequencies and sampling rates for auditory transform analysis of speech (Figures 19, 24, and 25).

<table>
<thead>
<tr>
<th>Analysis channel</th>
<th>Center frequency (Hz)</th>
<th>Sampling Rate (fraction of 10000 Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99</td>
<td>1/32</td>
</tr>
<tr>
<td>1</td>
<td>189</td>
<td>1/32</td>
</tr>
<tr>
<td>2</td>
<td>279</td>
<td>1/32</td>
</tr>
<tr>
<td>3</td>
<td>369</td>
<td>1/32</td>
</tr>
<tr>
<td>4</td>
<td>459</td>
<td>1/32</td>
</tr>
<tr>
<td>5</td>
<td>554</td>
<td>1/32</td>
</tr>
<tr>
<td>6</td>
<td>664</td>
<td>1/32</td>
</tr>
<tr>
<td>7</td>
<td>796</td>
<td>1/32</td>
</tr>
<tr>
<td>8</td>
<td>954</td>
<td>1/16</td>
</tr>
<tr>
<td>9</td>
<td>1143</td>
<td>1/16</td>
</tr>
<tr>
<td>10</td>
<td>1369</td>
<td>1/16</td>
</tr>
<tr>
<td>11</td>
<td>1639</td>
<td>1/16</td>
</tr>
<tr>
<td>12</td>
<td>1963</td>
<td>1/8</td>
</tr>
<tr>
<td>13</td>
<td>2352</td>
<td>1/8</td>
</tr>
<tr>
<td>14</td>
<td>2817</td>
<td>1/8</td>
</tr>
<tr>
<td>15</td>
<td>3374</td>
<td>1/8</td>
</tr>
</tbody>
</table>

Table 4.

Filter center-frequencies and sampling rates for smoothed auditory transform analysis of speech (Figure 20).

<table>
<thead>
<tr>
<th>Analysis channel</th>
<th>Center frequency (Hz)</th>
<th>Sampling Rate (fraction of 10000 Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99</td>
<td>1/128</td>
</tr>
<tr>
<td>1</td>
<td>189</td>
<td>1/128</td>
</tr>
<tr>
<td>2</td>
<td>279</td>
<td>1/128</td>
</tr>
<tr>
<td>3</td>
<td>369</td>
<td>1/128</td>
</tr>
<tr>
<td>4</td>
<td>459</td>
<td>1/128</td>
</tr>
<tr>
<td>5</td>
<td>554</td>
<td>1/128</td>
</tr>
<tr>
<td>6</td>
<td>664</td>
<td>1/128</td>
</tr>
<tr>
<td>7</td>
<td>796</td>
<td>1/128</td>
</tr>
<tr>
<td>8</td>
<td>954</td>
<td>1/64</td>
</tr>
<tr>
<td>9</td>
<td>1143</td>
<td>1/64</td>
</tr>
<tr>
<td>10</td>
<td>1369</td>
<td>1/64</td>
</tr>
<tr>
<td>11</td>
<td>1639</td>
<td>1/64</td>
</tr>
<tr>
<td>12</td>
<td>1963</td>
<td>1/32</td>
</tr>
<tr>
<td>13</td>
<td>2352</td>
<td>1/32</td>
</tr>
<tr>
<td>14</td>
<td>2817</td>
<td>1/32</td>
</tr>
<tr>
<td>15</td>
<td>3374</td>
<td>1/32</td>
</tr>
</tbody>
</table>
Figure 19. A segment of the auditory transform of a sentence of speech. The underlined portion of "Their hot proteins were paced on their breakdowns" is shown in the plot.
Figure 20. Smoothed auditory transform of the sentence, "Their hot proteins were paced on their breakdowns."
Figure 20 (continued).
finite duration. The duration for the analysis frames was chosen to be 25.6 milliseconds, long enough to include most perceptible pitch periods, and shorter than the minimum time span to hear echoes (Blauert, 1974).

We now resort to the method of principal components analysis to find the recurring patterns and variations in the data frames. In principal components analysis, relationships between the elements in a set of data waveforms are found by calculating the covariance matrix for the waveforms. This is used to determine a set of orthogonal basic waveforms (principal components) and weighting coefficients which can be used to represent the original data waveforms.

The specification of each frame of the auditory analysis by a weighted sum of principal components can be thought of as a sort of Fourier analysis using the data dependent principal components instead of sine waves or complex exponentials. Each frame, consisting of M samples, is fully specified by the M principal components and the M associated weighting coefficients (or weighting function) for the frame.

The advantage of principal component analysis is in its potential for data reduction. The analysis summarizes the underlying relationships between the samples in the frames optimally so that the fewest possible orthogonal components account for the maximum possible variance in the data. It is therefore possible to use a limited number of the lowest components to characterize the frames. The principal components are optimal in the sense of yielding a least-mean-square fit for the particular set of data waveforms from which they have been derived.

Let us designate the successive smoothed auditory transform analysis
frames as an $N \times N$ matrix $A$, where the column vectors of dimension $N$ are the log-magnitude samples in an auditory transform frame, and the row dimension $N$ is the total number of frames representing the duration of the analyzed sound. We seek now to characterize the frames $A$ using a matrix $P$ and $W$, where $P$ is a $M \times M$ matrix with columns containing the principal components, and $W$ is an $M \times N$ matrix of principal component weighting coefficients. Each column of $W$ specifies a set of weights required to reconstruct the frame in the corresponding column of $A$ using the principal components:

$$A = PW$$

(29)

The determination of the principal components and the weighting coefficients to reconstruct the individual data waveforms is essentially implemented by means of a least-mean-square error fitting process (Glaser and Ruchkin, 1976). The first principal component is simply the least-mean-square error fit of a single waveform to the entire data set. The second principal component, orthogonal to the first, is just the least-mean-square error fit to the residual from the fit of the first waveform. The process is continued until the data waveforms are fully specified by the set of components.

Principal components may be found using the Karhunen-Loeve expansion by first computing the $M \times M$ matrix of covariance coefficients $R$ associated with the data frames $A$:

$$R = \frac{1}{N} AA^T$$

(30)
The principal component matrix may then be determined with the relation
\[ P = \sqrt{\mathbf{U}} \]  \hspace{1cm} (31)

where the columns of \( \mathbf{U} \) contain the eigenvectors associated with \( \mathbf{R} \). Further details of principal component analysis may be found in (Glaser and Ruchkin, 1976).

The principal component matrix \( \mathbf{P} \) was determined using a standard statistical analysis package (Nie et al., 1975). Each frame of the smoothed and downsampled auditory transform surface of the speech was submitted to the principal components analysis. A frame consists of 64 samples, arranged as shown in Figure 21.

The first nine principal components, accounting for 97.9 percent of the variance in the smoothed auditory transform frames from the sentence of speech are shown in Figure 22. The plots are padded out in the temporal direction in order to better display the shapes, and have also been normalized to unit amplitude. There are two samples in the temporal direction at the low frequency side, and eight samples in the temporal direction at the high frequency side of the plot, as shown in Figure 21. The analysis channel center frequencies are given in Table 4.

The first component is basically flat, acting more or less as a shape determining the average strength of the samples in the frame. The second component shows frequency dependence only, and acts as a sort of "brightness function," measuring spectral tilt. The third and fourth principal components have temporal and spectral dependence, and appear to complement each other. Both have a spectral characteristic which is U-shaped,
Figure 21. Configuration of frames used for principal component analysis of smoothed auditory transform of speech shown in Figure 20.
Figure 22. First nine principal components of the smoothed auditory transform of speech in Figure 20. Percentage of variance accounted for by each component is also indicated.
Figure 22 (continued)
Figure 22 (continued)
Figure 22 (continued)
measuring signal strength at the extremes of the frequency range versus the signal strength near the center of the frequency range. Superimposed on this frequency dependence is a temporal decay characteristic for the third component, and a temporal growth characteristic for the fourth component. The remaining components show further dependencies on frequency and time, each acting as a template or pattern which may be used to characterize the modulations of the smoothed auditory transform surface for the speech.

5.3. Acoustic Modification using Principal Components

In order to verify that a limited set of principal components can indeed be used to characterize the time varying timbre of the speech waveform, synthesis of a waveform using the components is required. Consider the complete principal component representation of a single frame in Equation 29. A single auditory transform frame, represented by vector \( \hat{a} \) of length M, is specified fully by the principal component matrix P and the vector \( \hat{w} \) of weights for that frame:

\[
\hat{a} = P \hat{w} \tag{32}
\]

Alternatively, we could rearrange this expression to determine the weighting function once we know the principal components. Since the principal component matrix is orthonormal, the inverse is equal to the transpose. Multiplying each side of Eq. 32 by P-transpose then yields

\[
P^T \hat{a} = \hat{w} \tag{33}
\]
Equations 32 and 33 may be viewed as an analysis/synthesis pair which can be used to analyze and synthesize the shapes of the auditory transform frames in terms of a set of principal components. Equation 33 is the analysis equation, and yields the weighting function or strengths of the individual components in the frame. Equation 32 is the synthesis equation, which uses the weighting function to reconstruct the auditory transform frame.

We may use the analysis and synthesis equations to analyze a sound and impose the desired patterns onto the auditory transform surface of another sound. First, the principal components and associated weighting functions for the auditory transform frames of a waveform such as speech are computed. Of these, some of the components and functions (usually the lowest few, accounting for the greatest variance on the original auditory transform surface) may be retained for later use in modifying the auditory transform frames of another sound. We call these the "desired principal components" and the "desired weighting functions."

Our approach for signal modification will be to first find the auditory transform frames of a sound to be modified. Each frame is then analyzed using Equation 33 with the desired principal components. Next, the weighting functions determined in the analysis are set equal to the desired weighting functions. The auditory transform frames are then reconstructed using the synthesis equation (Eq. 32), and the acoustic waveform is reconstructed from the modified frames.

The sound modification algorithm is shown in Figure 23. Notice the similarity to the iterative algorithm presented in Section 4.2. The desired principal components and the desired weighting functions are
indicated in the middle of the diagram. Usually a few iterations are required before the modified sound takes on the desired magnitude characteristics.

Since only a few of the total principal component weighting functions will actually be modified in the analysis/synthesis procedure, it is desirable to avoid computations associated with the principal components which will not be adjusted. Let us rewrite Equation 32 as

$$a_i = \sum_{j=0}^{M-1} P_{ij} w_j$$  \hspace{1cm} (34)

and let us modify the weights of only the first $L$ components, where $L$ is less than $M$. Equation 34 can then be rewritten as

$$a_i = \sum_{j=0}^{L-1} P_{ij} w_j + q_i$$  \hspace{1cm} (35)

where

$$q_i = \sum_{j=L}^{M-1} P_{ij} w_j$$  \hspace{1cm} (36)

After modification of the first $L$ weights, the expression for reconstructing the modified frame $\hat{a}'$ is given by

$$\hat{a}'_i = \sum_{j=0}^{L-1} P_{ij} w'_j + q_i$$  \hspace{1cm} (37)

where the weights $w'_j$ for $j$ less than $L$ are the new modified weights. The computation of $q_i$ can be avoided by recognizing from Eq. 35 that $q_i$ is given by the relation
\[ q_i = a_i - \sum_{j=0}^{L-1} p_{ij} \cdot w_j \]  \hspace{1cm} (38)

Combining Equations 37 and 38, we see that the modified frame is given by

\[ a'_i = a_i + \sum_{j=0}^{L-1} p_{ij} (w'_j - w_j) \]  \hspace{1cm} (39)

Therefore, the computations involve only the principal components actually being modified on the auditory transform surface.

A bandlimited pulse train with fundamental frequency 90 Hz was modified using the analysis/synthesis procedure of Figure 23. A signal corresponding to the entire sentence of speech analyzed earlier in Figure 20 was synthesized. A section of the auditory transform surface of the pulse train is shown in Figure 24. After five iterations, a short segment the auditory transform of the reconstructed signal appears as in Figure 25. This segment may be compared with the similar segment from the original speech shown in Figure 19. The entire sentence which was reconstructed has a new pitch (corresponding to the fundamental frequency of 90 Hz), is quite intelligible, and retains the timbral characteristics of the original sentence.

In resynthesizing the speech, a data rate considerably lower than the original waveform data rate was attained. Every 256 samples of the input acoustic waveform was characterized using 9 integer weighting values corresponding to the first nine principal components of the smoothed speech. This suggests that the spectral evolution of sounds may be controlled at a data rate well below the original data rate, using a language
Figure 23. Modification of selected principal components in a waveform using an iterative procedure. The auditory transform of the waveform is first analyzed in terms of the desired principal components. The weighting functions are then set equal to the desired weighting functions, and a waveform is reconstructed. The procedure is repeated until the weighting functions are sufficiently close to the desired values.
Figure 24. Segment of the auditory transform of a bandlimited pulse train with fundamental frequency 90 Hz.
Figure 25. Segment of auditory transform of bandlimited pulse train after modification in terms of the nine speech principal components in Figure 22 (and their associated weighting functions). This segment corresponds to the segment of the original speech shown in Figure 19.
of patterns which are intimately related to both a model of human auditory perception, and to the characteristic shapes of sounds as represented by the model.

5.4. Onset Detection

The principal component analysis of the auditory transform surface leads to useful applications for analyzing musical performance. A 15.6 second segment of tabla (Indian classical drums) performance was digitized and analyzed. Tabla sounds are produced by two tuned drums played by the left and right hands. The left hand drum (bayan) has a low pitch which may be dynamically varied to create glissandos by varying the pressure of the hand on the drum head. The right hand drum (tabla) normally has a single, fixed pitch, although a lower pitch mode may be attained using an open handed drum stroke. A complex language of pitched sounds and articulate slapping sounds comprises the tabla performance (further details can be found in Wade, 1979). Of particular interest in analyzing the tabla performance are the onset times of the drum strokes, the pitches played, and identification of the articulated sounds.

A segment of the smoothed auditory transform analysis is shown in Figure 26. Table 5 gives the analysis center-frequencies and sampling rates. Each onset is rather clearly identifiable as a sharp rise in the surface, occurring across most of the frequency range. The onsets have been numbered on the figure for convenience (A plot of the entire 15.6 second example is given in the appendix for reference). The first few partials associated with the tabla are also clearly identified as long narrow ridges extending in the time direction.
Onset number 1 consists of both the left and right hand drum, as evi-
denced by the very low frequency growth of energy of the bayan, and the
presence of high frequency growth of energy and following partials for the																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																											

The first six principal components account for 93.8 percent of the
variance in the frames of the smoothed auditory transform surface and are
shown in Figure 27. The weighting functions for the full 15.6 seconds of
the performance are shown in Figure 28. Each principal component must
account for some aspect of the characteristic shapes of the tabla sounds.
We may attempt to assign some meaning to the components in order to inter-
pret the weighting functions.

The first component, as in the first principal component for the
speech in Section 3.2, functions as a sort of average level for the sam-
ples in the analysis frame. Observation of the weighting function for
the first component shows a peak corresponding to the growth of energy at
the onsets.

While the weighting function of the first component may be used for
localizing the onset times, the weighting function for the third principal
component has strong negative peaks at every onset which are more sharply
defined. Observation of the shape of the third principal component sug-
gests that it functions as a decay detector, since the samples decrease in
Table 5.

Filter center-frequencies and sampling rates for smoothed auditory transform analysis of tabla performance (Figure 26 and Appendix)

<table>
<thead>
<tr>
<th>Analysis channel</th>
<th>Center frequency (Hz)</th>
<th>Sampling Rate (fraction of 10000 Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45</td>
<td>1/128</td>
</tr>
<tr>
<td>1</td>
<td>124</td>
<td>1/128</td>
</tr>
<tr>
<td>2</td>
<td>203</td>
<td>1/128</td>
</tr>
<tr>
<td>3</td>
<td>282</td>
<td>1/128</td>
</tr>
<tr>
<td>4</td>
<td>361</td>
<td>1/128</td>
</tr>
<tr>
<td>5</td>
<td>440</td>
<td>1/128</td>
</tr>
<tr>
<td>6</td>
<td>520</td>
<td>1/128</td>
</tr>
<tr>
<td>7</td>
<td>609</td>
<td>1/128</td>
</tr>
<tr>
<td>8</td>
<td>713</td>
<td>1/128</td>
</tr>
<tr>
<td>9</td>
<td>835</td>
<td>1/128</td>
</tr>
<tr>
<td>10</td>
<td>978</td>
<td>1/64</td>
</tr>
<tr>
<td>11</td>
<td>1145</td>
<td>1/64</td>
</tr>
<tr>
<td>12</td>
<td>1341</td>
<td>1/64</td>
</tr>
<tr>
<td>13</td>
<td>1570</td>
<td>1/64</td>
</tr>
<tr>
<td>14</td>
<td>1838</td>
<td>1/64</td>
</tr>
<tr>
<td>15</td>
<td>2152</td>
<td>1/32</td>
</tr>
<tr>
<td>16</td>
<td>2519</td>
<td>1/32</td>
</tr>
<tr>
<td>17</td>
<td>2948</td>
<td>1/32</td>
</tr>
</tbody>
</table>
Figure 26. Beginning segment of the smoothed auditory transform of tabla performance. Each onset is numbered to the left of the plot. The full 15.6 second performance is shown in appendix. (Tabla performance by Norman Farkas)
Figure 26 (continued)
Figure 27. Configuration of frames used for principal component analysis of the smoothed auditory transform of tabla performance.
Figure 28. First six principal components of smoothed auditory transform of tabla performance. Percentage of variance accounted for by each component is also indicated.
Figure 28 (continued)
Figure 28 (continued).
Figure 29.
Weighting functions of first six principal components for 15.6 sec. tabla performance
value in the temporal direction, combined with a U-shaped frequency dependence. For most of the duration of the performance, the resonance of the drum strokes are decaying slowly and the weighting function for the third component takes on small positive values to indicate this fact. At an onset, however, just the opposite happens and the weighting function responds by going strongly negative.

The second principal component has mostly frequency dependence and has large values at low frequencies. It would appear that this component functions to detect the presence of low frequencies associated with the bayan and possibly the lower resonant mode of the tabla. This is especially apparent for drum strokes involving only the tabla, as in onsets 5 and 6. The weighting function takes a sharp dip downward, indicating the absence of energy growth at low frequencies. These downward dips occur consistently for drum strokes involving the tabla alone at its high resonant mode. Onsets associated with the low resonant mode of the tabla, such as onsets 7, 15, 20, 31, 39 and 47, are indicated with positive slopes of the weighting function. The low frequency bayan sounds, particularly those employing wide frequency glides as in onsets 51 and 54, have particularly strong positive slopes for this weighting function.

The fourth principal component has primarily frequency dependence, showing a strong peak at the predominating partial frequency of the tabla. It serves to indicate the presence of resonance associated with that frequency. At most onsets, the weighting function for the fourth component goes near zero or slightly negative. After the onsets of resonant tabla strokes, the function takes on increasing strength, indicating the presence of ringing at the resonant mode after other transients associated
with the onset have died away. After slapping sounds on the tabla, however, as in onsets 10, 12, 17, 28, 42 and 44, the weighting function stays at a value near zero, indicating the absence of ringing.

The sixth principal component appears to account for the open-handed drum stroke, associated with the lowest partial resonant mode of the tabla. Each occurrence of the open-handed tabla drum stroke, as in onsets 7, 15, 20, 31, 39, and 47, displays a prominent negative shape just after the onset in the weighting function for that component. A negative weight for the sixth component would add energy at the frequency associated with the alternative mode, and subtract energy at the frequency normally corresponding to the lowest partial frequency of the tabla sound.

The principal component weighting functions may be used to transcribe the gestures of the tabla performance. Although it is possible to visually associate the behavior of the principal component weighting functions with the various classes of sounds articulated in the tabla performance, the problem remains to create an automated procedure for transcription. Furthermore, it is not clear at this point whether the principal components themselves should be used for the pattern detection. Other patterns based on the principal components can be found to characterize the data by using the method of varimax rotation (Glaser and Ruchkin, 1976) or other available rotations.

Automated detection of onset times using the third principal component is relatively easy, however, since the peaks are well defined. A computer program was implemented to identify the peaks and find the corresponding onset times. The onset times were used to define the onset times for the synthesis of computer generated sounds. The rhythm of the
computer performance matched the rhythm of the original tabla performance.
6. CONCLUSION AND SUGGESTIONS FOR FURTHER RESEARCH

Auditory transform analysis provides a framework for analyzing and understanding acoustic signals in the context of human auditory sensitivities. Beginning from basic considerations of some of the more commonly known psychoacoustic and physiological observations of auditory behavior, a simplified model of auditory analysis was developed and applied to the analysis of musical sounds. The model, based on a generalized short-time Fourier transform log-magnitude analysis of acoustic signals, serves as a basis for the identification of musical features such as the pitch, timbre and onset characteristics. Furthermore, the capability for acoustic signal synthesis from the log-magnitude samples allows specification or modification of sounds directly in the auditory domain.

Since the auditory transform analysis attempts to characterize sounds in terms of human auditory sensitivities, the analysis result should contain only information relevant to the audible characteristics of the sounds. Therefore, we expect that the auditory transform samples represent the perceptible differences in sounds, and may serve to potently specify the audible details of musical sounds. This expectation was born out by the summarized representations of sound available in the auditory domain using the method of principal components analysis. The method leads to compact descriptions of the auditory transform samples which may be effectively used for both analysis and synthesis of musical sounds.

A question remains as to the exact tuning of the auditory transform filterbank parameters. The degree of filter response overlap, quantization, and channel sampling rates should be determined such that the time, frequency and intensity resolution of human hearing is matched. Solving
this problem first requires consideration of the optimal way to extract the time, frequency, and intensity information from the auditory transform surface. For example, the presence of a sine tone shows up at various levels in all of the analysis channels. All of these channels may be used to optimally determine the differences in frequency and strength of the tone. The combination of channel quantization, filter overlap and sampling rates should be such that a determination may be made to just within the limits of human auditory capabilities. Similar limits should be attained for signals which have temporal fine-structure and which test the temporal limits of auditory perception.

Various other improvements can be made to the algorithms discussed. While the infinite impulse response implementation of the auditory filter-bank is useful when memory space is at a premium, a finite impulse response implementation should result in more efficient operation. The generalized short-time Fourier transform log-magnitude analysis and synthesis allows considerable flexibility in the choice of filterbank parameters, such as the window shapes, center-frequencies, bandwidths, and degree of overlap. These parameters should be refined to correspond more closely with critical band studies and physiological models of auditory filtering. These aspects may rendered without significantly affecting the efficiency of the analysis and synthesis algorithms.

One approach to sound synthesis using the auditory transform discussed in this thesis involved synthesis directly from the log-magnitude samples. Another intriguing possibility is the use of the auditory transform and feature detection algorithms for the control of synthesis. This leads to computer music applications where the gestures of acoustic
instrument performances may be tapped and applied to the control of computer generated sounds. This possibility was briefly explored using the tabla onset information in Section 5.4.

The algorithms discussed in this thesis based on the auditory transform analysis and synthesis may be used in computer music applications. The residue pitch detection algorithm can be combined with analysis of the individually resolved partial frequencies for the extraction of pitch information. Principal components can be used to impose the timbral characteristics of a sound onto another acoustic signal. The principal component weighting functions can be used to monitor musical performance and control computer synthesis. The auditory transform provides a framework for analyzing and understanding musical signals in a manner which reflects the human perception of the signals. Using this model, we may gain access to the perceptually relevant components of acoustic signals for the analysis, synthesis, and control of sounds in computer music applications.
APPENDIX

The smoothed auditory transform of the 15.6 second tabla performance discussed in Section 5.4 is shown here for reference.Tabla performance is by Norman Farkas.
REFERENCES


Siebert, W.M., 1982. Lecture given during auditory modeling seminar series, Research Lab of Electronics, MIT, Cambridge, MA


Wegel, R.L. and Lane, C.E., 1924. "The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear," Physical Review, 23(2)


