The Computer Analysis of Polyphonic Music

by

Charles Richard Watson

A thesis submitted in fulfilment of

the requirements for the degree of

Doctor of Philosophy

Basser Department of Computer Science

University of Sydney

1985

ABSTRACT

The Computer Analysis of Polyphonic Music

This thesis describes research on the computer analysis of polyphonic music, concentrating on the physical problem of identifying and tracking simultaneous tones in an acoustical signal. Except for J.A. Moorer's work and the results in this thesis, little progress has been made. Identifying the harmonics of each note is a difficult problem.

The analysis procedure involves analog to digital conversion of recorded music, pitch estimation, and grouping pitch estimates into notes. Known signal processing algorithms are applied to the problem, and a new spectral extraction procedure improves the results. Error measures are defined to determine the accuracy and sensitivity of the algorithm. Several musical examples are tested. The pitches of notes in a synthesized Trio are determined with 99% accuracy, (90% accuracy for the notes of a woodwind Trio).

The response of the human cochlea to polyphonic tones is simulated. The observed amplitude modulation of the response between harmonics is enough to distinguish pairs of superimposed tones.

Keywords:

pitch estimation, music analysis, artificial intelligence, automatic music transcription, musical acoustics, auditory modelling, signal processing.

ACKNOWLEDGEMENTS

I would like to express sincere gratitude to my supervisors. Dr. Michael Kassler's guidance was invaluable in directing me to this research problem and to the relevant literature; his critical acuity has been a great stimulus. Dr. Don Herbison-Evans assisted with algorithmic and signal processing problems. Dr. Sylvan Elhay and Dr. Charles Pearce were helpful and encouraging with chapter eight and the final revision of the thesis.

My thanks go to the computing support staff at the University of Sydney; especially John Holden for the hours spent diagnosing hardware faults in the A/D conversion equipment.

The Post-Graduate Scholarship assistance, provided by the Australian Commonwealth Department of Education, was gratefully received.

Finally, I thank my wife, Pauline, for her patient support.

DEDICATION

The human brain is one of the most complex signal processors known to man. This thesis is dedicated to its Creator.

CONTENTS

CHAPTER SEVEN

Evaluation of Analysed Music

CHAPTER EIGHT

Error Measures for Comparing Music Analyses

CHAPTER ONE

The Computer Analysis of Music

1.1 Introduction : Statement of the Problem

Human beings are adept at distinguishing simultaneous sounds. A person can understand one of several people speaking at the same time (the so called cocktail party effect). An orchestral conductor can tell if one instrument out of a hundred is playing out of tune. How this discrimination is made from the acoustical signal is not well understood.

The central problem considered in this thesis is the automatic separation of simultaneous tones from an acoustical signal. This is a difficult problem, because the harmonic frequencies of two superimposed tones are interspersed. For consonant chords, some harmonics of different tones coincide. As an example, when two tones sound an octave apart, the harmonics of the higher tone coincide with the even harmonics of the lower tone.

For the work described in this thesis, recorded music is converted to a digital signal, which is analysed at successive times to determine the pitches of the constituent notes. These pitch estimates are then grouped together in time, to determine the pitches, starting times and durations of the notes.

This work develops many computing techniques, but also contributes to the fields of musical acoustics, signal processing, and auditory perception. The identification of simultaneous musical tones is an intelligent human activity requiring many years of aural training. In this sense, the automated analysis of

musical sounds is an Artificial Intelligence problem.

1.2 The Automated Analysis of Musical Pitch : A Review

In 1843 G.S. Ohm proposed the theory that the human ear acts as a Fourier analyser, separating the harmonic components of a periodic sound wave. Twenty years later Hermann von Helmholtz was first to observe the spectra of sustained musical tones using acoustical resonators.

M. Metfessel, et al. (1926) developed a stroboscopic instrument to display the fundamental frequency of tones. This system was limited by the inability to report dynamics or frequency fluctuations.

Electronic technology overcame these limitations. A.W. Hull (1933) designed and built a frequency meter using thermionic valves. F.V. Hunt (1935) improved the design by low-pass filtering the signal to enhance the fundamental. Juichi Obata et al. (1937) used low-pass filtering and automatic gain control to enable a larger dynamic range of signals to be used.

Similar devices have since been built by C. Seeger (1951) and P.A. Tove et al. (1966). After low-pass filtering and rectifying the incoming signal Tove's device charges two capacitors to determine the period of the signal. This is mapped into a one octave frequency range. The Seeger melograph model C (Moore 1974) has overlapping one-third octave band-pass filters. Output from the filter of lowest frequency with significant signal level drives a frequency meter. Spectral analysis of the musical tones can also be done.

Bengtsson (1972), and Piszscalski (1979) have suggested using such

methods to analyse polyphonic music, but the only successful attempt before this thesis is by J.A. Moorer (1975). Moorer considers only two part music, with the intervals between parts ranging from a minor third to a minor seventh. This avoids the problem, mentioned earlier, of simultaneous tones at an octave apart.

All these methods depend on the presence of a significant component at the fundamental frequency, which is not always present, especially for low pitched sounds. Tones with weak fundamentals can be detected using the methods developed here.

The analysis of a synthesized Trio by J.S. Bach, described in chapter 8, has an error rate similar to that of Moorer's guitar example. However, the Trio has three musical parts spanning nearly four octaves, in comparison to Moorer's two parts spanning two octaves. All the notes of the trills in the synthesized Trio, some with note durations as small as 30 milliseconds, are correctly identified.

1.3 An Overview of the Thesis

The first three chapters introduce the subject. The second chapter discusses musical sound in terms of physical acoustics, perceptual psychology and music theory. Signal processing algorithms used in such areas as automatic speech recognition and the analysis of seismic waves are presented in chapter 3.

The central chapters (4 to 8) of the thesis present the author's contribution to the analysis of polyphonic music. Chapter 4 describes the hardware and software systems developed for the analysis of musical signals. The fifth chapter gives the algorithms used for estimating musical pitch. Chapter 6 presents

the software to group pitch estimates into musical notes and display them in standard music notation. In chapter 7 the algorithms given in chapters 4, 5 and 6 are evaluated. Chapter 8 describes the work to determine the accuracy and sensitivity of the analyses.

The mechanisms of hearing in the human ear and auditory cortex, and a computer simulation of the cochlear response to musical sounds are described in chapter 9. A neural pitch-determining mechanism is proposed to account for the human ability to distinguish polyphonic sounds.

Finally, chapter 10 summarizes the results of the research described in this thesis and suggests some possible areas for future research.

CHAPTER TWO

The Nature of Musical Sound

2.1 Introduction

Although the emphasis of this research is on the signal processing and algorithmic aspects of the analysis of music, it is useful to explore the relationship with the disciplines of Physics, Psychology and Music.

Here, sound refers to the sensation resulting from vibration within the ear, or any vibration that can cause such a sensation. A tone is a sound that is perceived to have pitch or musical height. A pure tone by definition has only one sinusoidal component. Tones can be considered as the superposition of pure tones, called the partials of the tone. If the frequencies of the partials are multiples of the fundamental (lowest) frequency, they are called harmonics. A note is the perceived musical entity (or the symbol in music notation) associated with a tone. Music is a sequence of sounds, some of which may overlap in time.

Polyphony refers to the simultaneous sounding of two or more melodic parts, whereas monophonic music has at most one instrument sounding at any time.

Little is known about how the human brain processes musical sounds. Roederer (1977) suggests that the desire to listen to music arises from the redundant speech processing areas of the minor half of the human cortex. There are many cases of people with brain-damage, where lesions in the left (or dominant) hemisphere have made them unable to understand or generate speech, yet they

can recall melodies, sing the words and apparently appreciate music with their unimpaired minor hemisphere.

2.2 Physical Correlates of Musical Attributes

There are several perceptual attributes of musical sounds; the most important are pitch, loudness, timbre and, for multiple tones, consonance and dissonance. This section considers their relationship to the corresponding physical attributes.

2.2.1 The Relationship Between Pitch and Frequency

Pitch is usually defined as perceived frequency and is measured in mels. The pitch of 1,000 mels is assigned to a pure tone with frequency 1,000 cycles per second (1 kHz). Pitch can differ from frequency by about 1% as frequency and loudness change. The pitch of a tone usually corresponds to the fundamental frequency; the harmonics are generally not perceived as separate components. There are some tones (such as those of bells and cymbals) with non-periodic wave-shape, that have ambiguous pitches. Other tones, such as piano tones, have upper partials with slightly higher frequencies than the corresponding harmonic frequencies. The pitch then corresponds to the average frequency difference of the strongest partials, and not that of the fundamental frequency (Houtsma et al. 1972). Many tones with the fundamental completely absent are still perceived as having the pitch of the missing fundamental (Schouten et al. 1962). In the remainder of this thesis a logarithmic pitch scale is used, so musical intervals have a constant

pitch difference, independent of pitch. The interval of an octave corresponds to the frequency ratio of 2:1 or a pitch difference of 12 semitones.

Since 1939, absolute musical pitch has been generally accepted as 440 Hz for *A* above middle *C*. This will be assumed here. The pure tone frequencies perceivable to the human ear range from about 20 Hz (cycles per second) to about 20 kHz (20,000 Hz). The fundamental frequencies of orchestral instruments range from 30 Hz on the contra-bassoon to 2 kHz on the piccolo flute (Apel 1970).

Signal processing techniques can estimate pitch. The following example illustrates the difficulty in using time-based methods. Figure 2.1 is a plot of a 40 millisecond segment of the steady state of a bassoon tone. The note is *E* below middle *C*. The period is 6 milliseconds. However the zero-crossings suggest a period of one third of this value. This is caused by a strong third and sixth harmonic (see figure 2.2 - the spectrum of the tone). There is no superimposed note at three times the fundamental frequency.

2.2.2 The Relationship Between Loudness and Intensity

Intensity is the power of a sound signal per unit area measured in watts per square metre (W/m^2). A pure tone at 1 kHz, with intensity 1 W/m^2 is near the threshold of pain, while the same tone with an intensity of 10^{-12} W/m^2 is near the threshold of hearing. To encompass this wide range of intensities, the logarithmic decibel (dB) scale is used. The dB level is defined as $10\log_{10}(I_1/I_2)$, where I_1, I_2 are the intensities of the compared signals. The reference intensity,

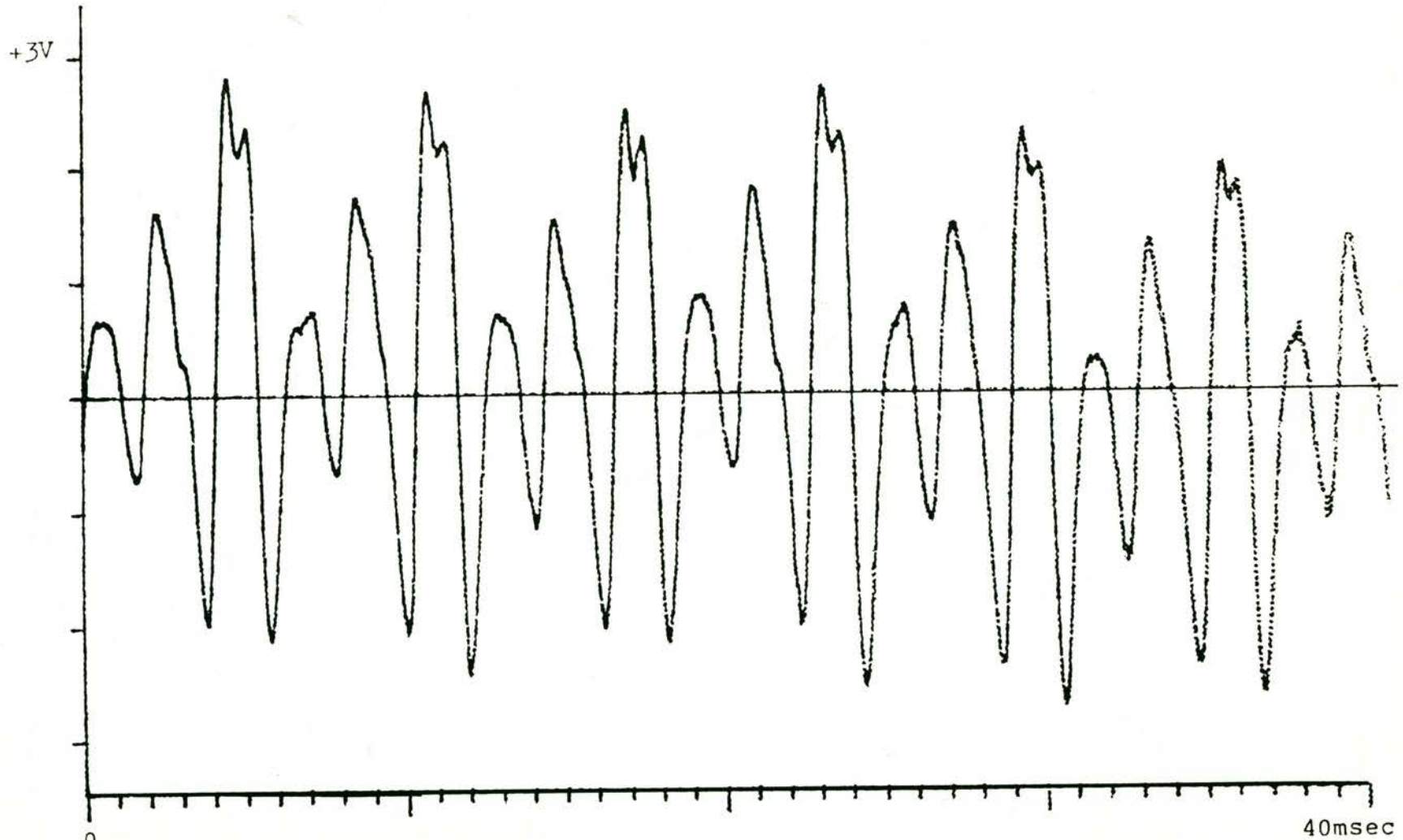


Figure 2.1 is a plot of a 40 millisecond segment of the steady state of a bassoon tone. The note is E below middle C.

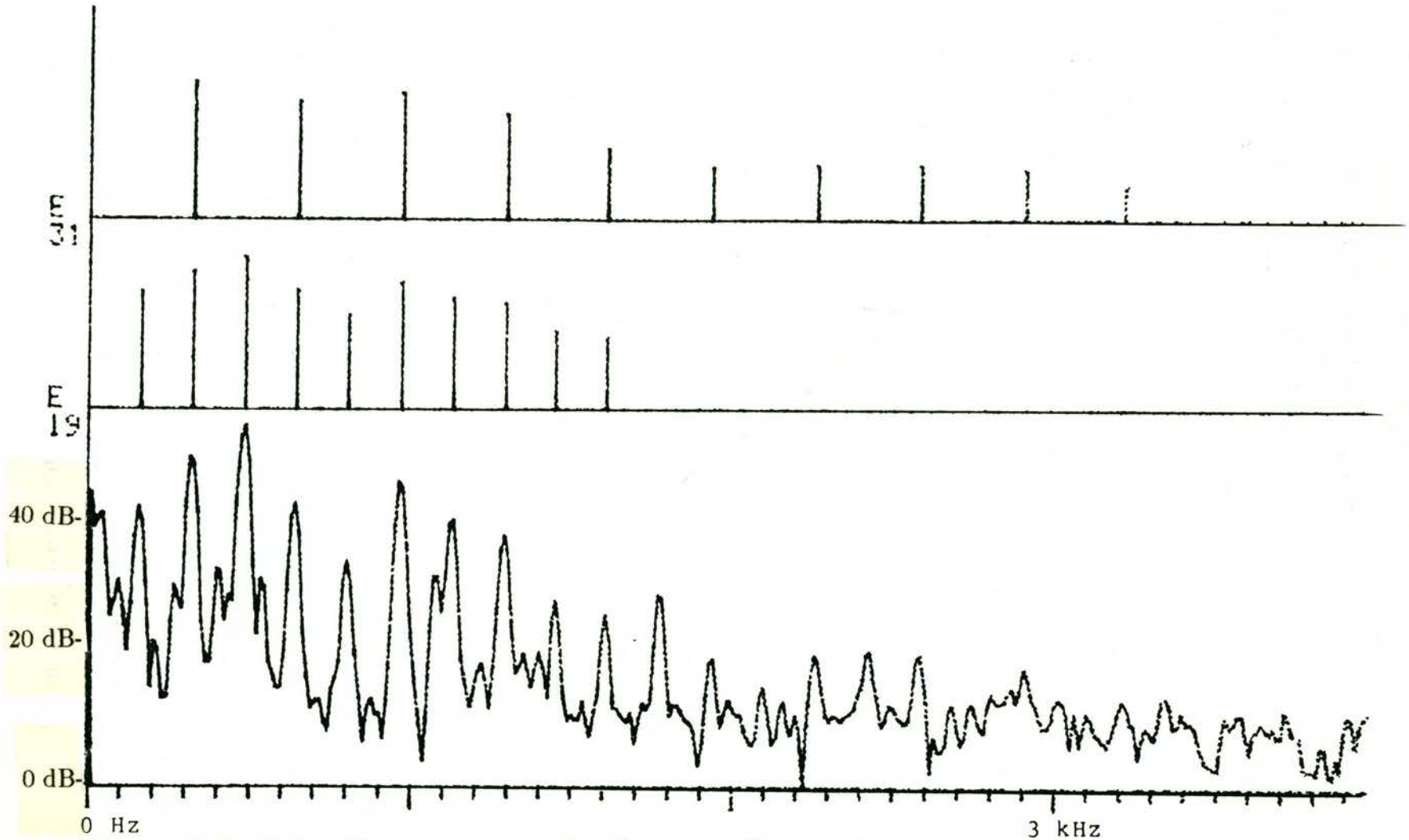


Figure 2.2 plots the spectrum of the previous time segment with two proposed estimates of the note that is present.

I_2, is generally taken as $10^{-12}\ W/m^2$. With this reference the threshold of pain at 1 kHz is about 120 dB; the threshold of hearing is 0 dB; the musical dynamic level of fortissimo is 80dB and pianissimo is 40dB. SometimesdB refers to the relative intensity of two signals, for example signal to noise ratio.

Loudness, measured in phons, is the perceived dB level of a signal. A pure tone of frequency 1 kHz and intensity 0 dB has a loudness of 0 phons. For a pure tone of fixed intensity, loudness is maximal at about 1 kHz, and decreases as the frequency tends toward the extremes of the audio range (20 Hz to 20,000 Hz). A pure tone must be increased in intensity by a factor of about 20 to sound as loud at 100 Hz or at 10 kHz as it does at 1 kHz (Fletcher 1934). Loudness can also be altered by spectral characteristics. For example, a tone of high intensity, with reverberation added, can be perceived as quieter than one of lower intensity, but without reverberation. It appears to be coming from further away (Chowning et al. 1974).

2.2.3 Correlates of Timbre

The physical correlates of timbre are more elusive than those of pitch and loudness. John M. Grey (1975) differentiated musical instrument tones using three measurable attributes: the spectral extent, the inharmonicity of the onset, and the synchrony of the upper harmonics. This is insufficient to apply to the automatic discrimination of simultaneous tones; human listeners are still much better than computer programs at identifying musical instruments.

2.2.4 Dissonance and Consonance

Consonance refers in this thesis to chords (simultaneous tones) with only consonant intervals (pitch differences of 3 , 4, 5, 7, 8, 9, or 12 semitones) between the tones. Dissonance refers to chords containing at least one dissonant interval, i.e. a pitch difference of 1, 2, 6, 10, or 11 semitones. Dissonance increases with the amount of audible beating between the harmonics of the constituent tones.

Consider two superimposed pure tones $\sin(at)$ and $\sin(bt)$, where a, b are different constants and t is time. If a and b are close enough, beating (amplitude modulation) occurs. The beating frequency equals the difference between the input frequencies because

$$\sin(at) + \sin(bt) = 2\sin(\frac{(a+b)t}{2})\cos(\frac{(a-b)t}{2}).$$

If this beating frequency is increased beyond about 20 Hz, it is perceived as roughness or dissonance. This roughness continues as the beat frequency increases through the lower audible range, until two distinct tones are heard with no roughness. Two pure tones of nearly equal frequency are said to lie within the critical bandwidth for that frequency, if they cannot be identified by a human listener as two distinct tones. The critical bandwidth increases slightly with frequency, but here dissonance is considered to be caused by a beating frequency between 20 and 100 Hz, and consonance to be the absence of this beating.

To take some examples: two notes an octave apart are consonant, because the harmonics of the upper note coincide with the even harmonics of the

lower note. For real musical tones the fundamental frequencies could vary by 1 percent, but most of the signal energy is in the lower harmonics below 2 kHz, so beating would still be within the critical bandwidth. For the major third with frequencies 400 Hz and 500 Hz, the difference between adjacent harmonics is 100 Hz. This is on the border of dissonance. If the notes are played an octave lower (200 Hz and 250 Hz), the difference frequency of 50 Hz is within the dissonant region of the critical bandwidth. It is interesting that although a major third is considered to be consonant by most musicians, only octaves and fifths are allowed in the bass of consonant cadences. A thousand years ago when Gregorian chant began, thirds were rarely used. Thirds may have been introduced into polyphonic music over the following centuries, as dissonance became more acceptable.

2.3 The Perception of Musical Events

Standard Music Notation (SMN), evolved with extensive use by musicians, reveals much about the perception of music. A musician perceives music as a series of discrete events, interrelated by rhythm and key, and not as continuously changing superimposed signals. The perception of speech is similar; the phonetic sounds of speech are grouped in the context of natural language grammar to determine the underlying semantics.

Although there is a close correspondence between perceived musical events and the notes of SMN, the acoustical signal is different. In performance, the pitch and duration of notes can differ considerably from the values designated

by SMN. There are many ambiguities in producing SMN from the acoustical signal. For example, a piece of music could be written in 4/4 time or 2/4 time for the same performance.

2.3.1 The Perception of Pitch

There is more to an acoustical signal than meets the auditory cortex. Music is usually performed in a reverberant environment, which can sustain sounds by more than one second. This means that while a note is being played, the reverberation of the preceding notes may also be present in the acoustical signal (see figure 2.7). Most musical instruments have natural resonances (formants). Signals at these resonant frequencies are reinforced (see figure 2.2). This can cause narrow spectral peaks that have no relation to the perceived pitch. For stringed instruments, such as piano, guitar or violin, undamped strings vibrate in sympathy with the string being played. Mechanical moving parts of instruments produce extraneous noises. Examples are the click of the keys of a woodwind or brass instrument, or the rasp of fingers sliding along guitar strings. Yet the listener can mask these extraneous signals and perceive only the tone being played.

Steady-state frequency differences of tones as little as 0.3% (a twentieth of a semitone) can be distinguished by musicians. However, harmonic frequencies can vary as much as a semitone during a performed note, yet the note can still be perceived as having a single pitch.

Many musical intervals are exaggerated in performance by as much as

two percent, i.e. up to one third of a semitone sharp (Ward 1970). Therefore, as well as masking out extraneous sounds, the human auditory system tolerates deviations in pitch, and perceives pitch within the musical context.

A tune or sequence of notes can be memorized or identified independently of the key in which it is played, although some musicians, young children, and non-western peoples are able to recall absolute pitch with high accuracy.

2.3.2 The Perception of Time and Duration

Although we can differentiate time differences of as little as 10 milliseconds, we tolerate variations of more than 100 milliseconds in the times and the durations of performed musical tones. Durations vary with articulation from legato to staccato. Even when a performer tries to produce exact durations, these can vary by 100 milliseconds. Vocal, woodwind and brass performers must take a breath from time to time, thereby shortening a tone. The transition from one string to another in bowed violin music differs from the transition from note to note on the same string. Also reverberation extends the effective duration of tones.

Because of these phenomena, the time difference between the start of successive tones (inter-onset time) is more useful in determining durations than the time difference between the start and finish of the tone (Tucker 1977). Even the inter-onset times vary. An onset of a tone is usually accompanied by an increase in intensity, and a non-periodic or noisy signal (inharmonicity), followed by a change in fundamental frequency. These cues are not always present. The

fundamental frequency can glide from one pitch to another without inharmonicity and with little change in intensity.

In human perception of music, the durations appear to be determined from the inter-onset times and are interpreted in terms of the current tempo, which constantly adapts to the incoming tones. This behaviour is modelled by the music plotting procedures described in chapter 6.

2.3.3 Musical Examples

The following examples illustrate the difficulties in determining the pitches and start and finish times of musical events.

Figure 2.3 plots amplitude versus time for the first 150 milliseconds of a piano tone. The amplitude decays after the kinetic energy of the hammer transfers to the strings, so there is no steady state.

Figure 2.4 shows the time varying spectra of the piano tone of figure 2.3, and the transition into the following tone, spanning a total of 600 milliseconds. The spectra are 10 milliseconds apart, and the frequency ranges from 0 to 3 kHz. The decay of the first tone is masked by the broad-band onset of the second tone. Amplitude modulation occurs in the harmonics of the tones. Beating of the treble strings can account for this.

Figure 2.5 shows the time varying spectrum (0 to 800 Hz) of the lower harmonics of a bassoon playing the three notes; *A* (110 Hz), *G* (98 Hz) and *C* (131 Hz) in succession, spanning 1.2 seconds. Observe the low frequency distortion between 0 and 70 Hz, and the prolongation of harmonics caused by

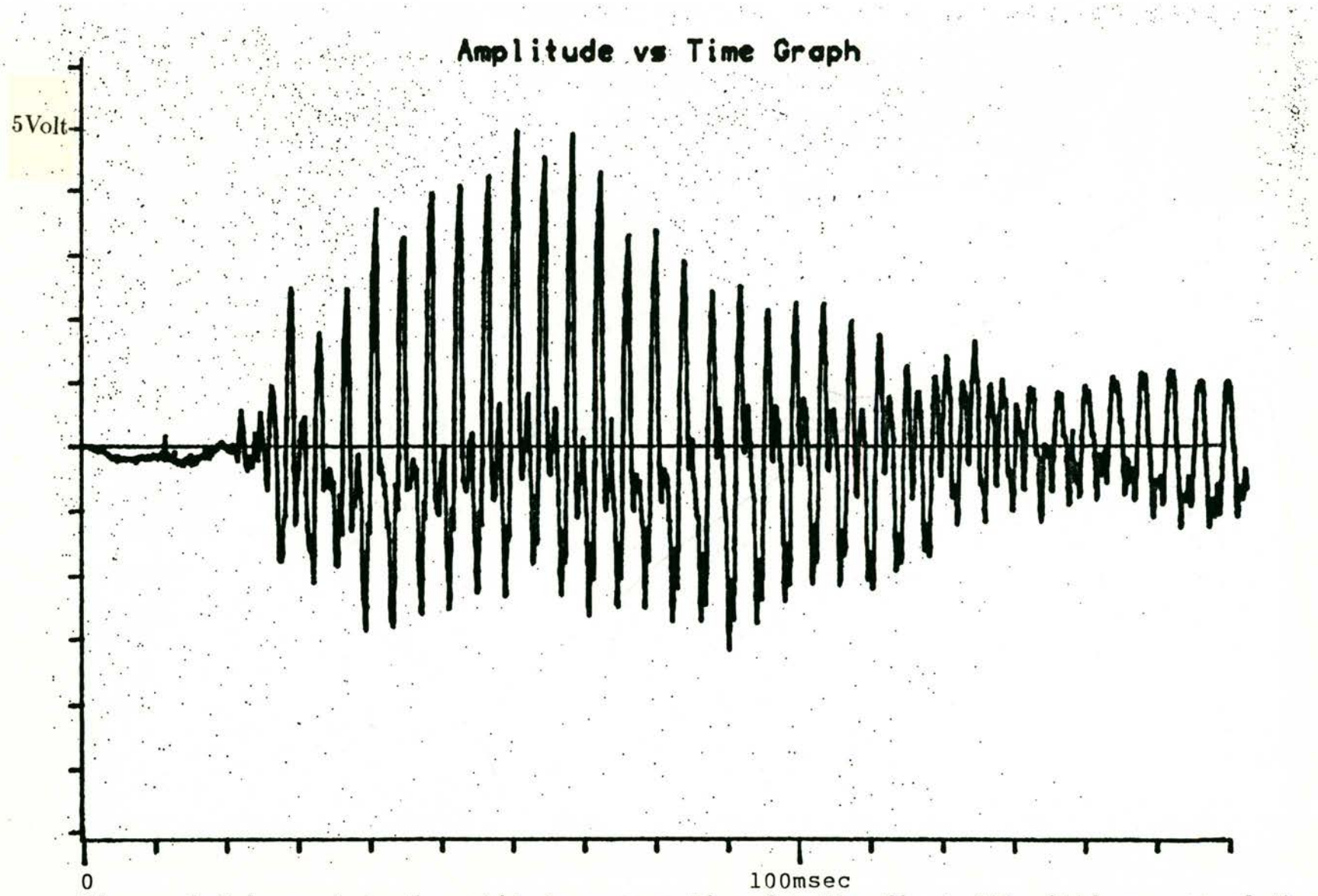


Figure 2.3 is a plot of amplitude versus time for the first 150 milliseconds of the attack of a piano tone.

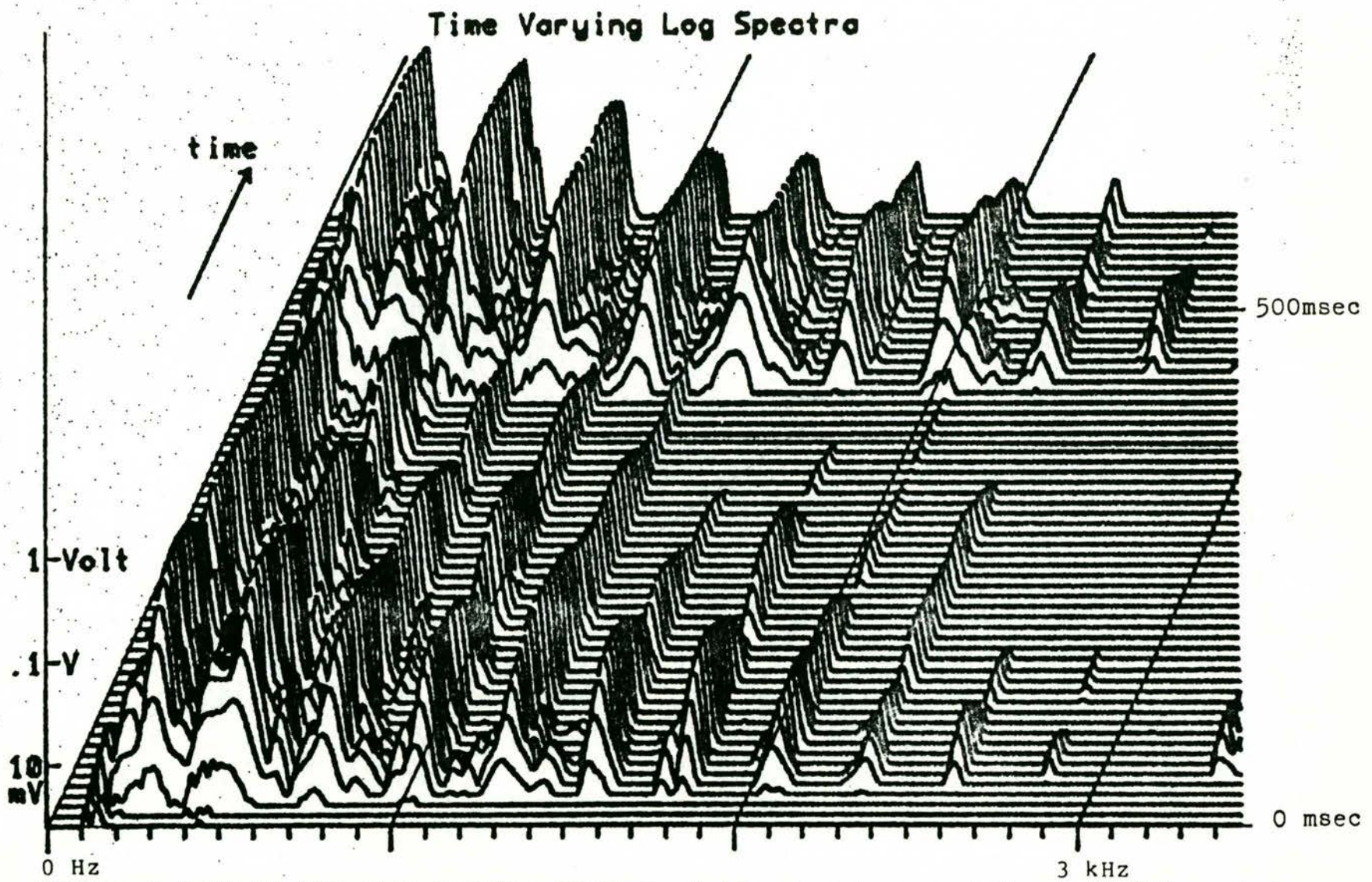


Figure 2.4 plots the time varying spectra of the same tone and the transition into the next tone, spanning a total of 600 milliseconds. Each spectrum is advanced by 10 milliseconds from the previous one, and the frequency range is from 0 to 3 kHz.

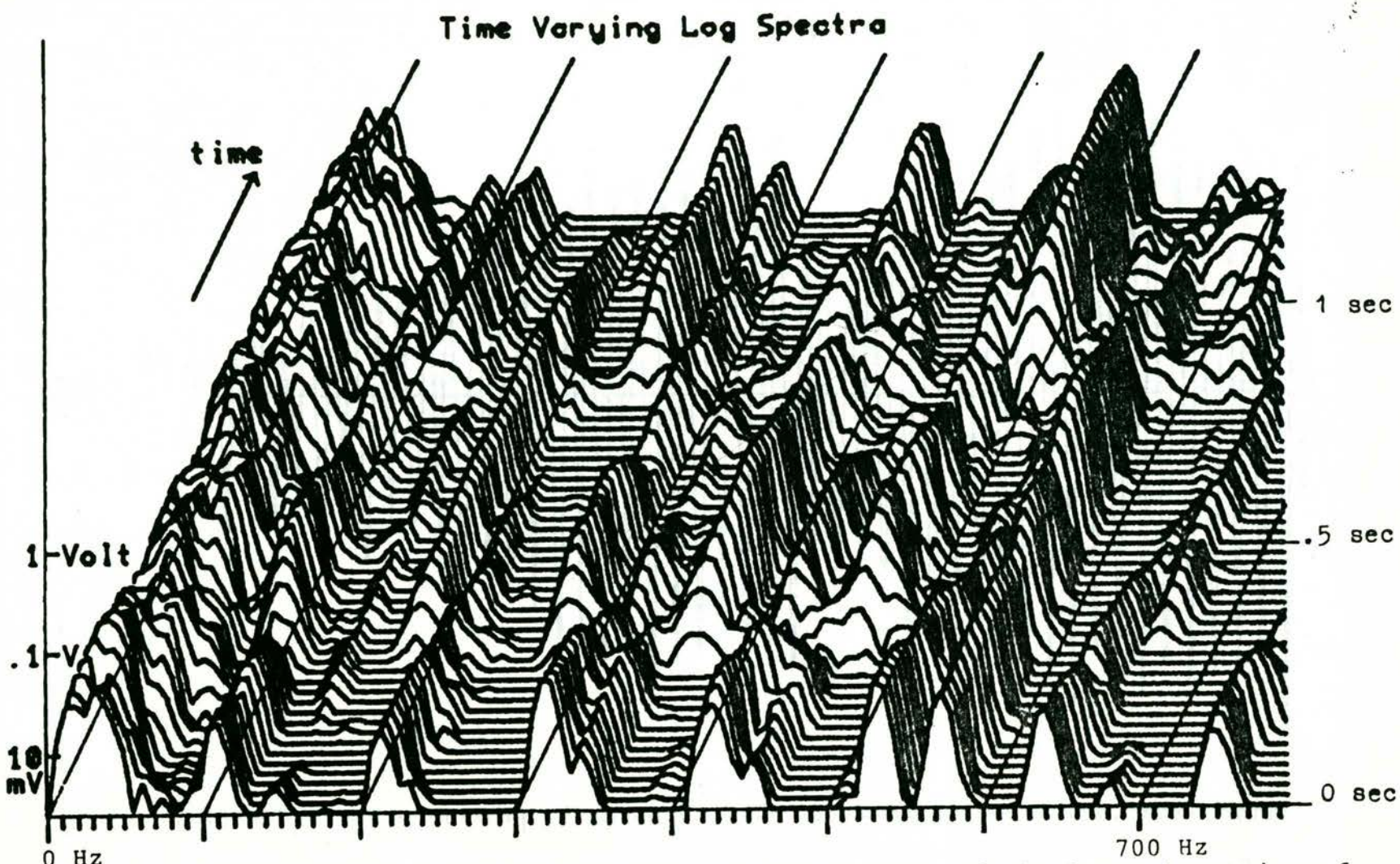


Figure 2.5 shows the time varying spectrum of 1.2 seconds of the lower harmonics of a bassoon playing the three notes; A (110 Hz), G (98 Hz) and C (131 Hz) in succession.

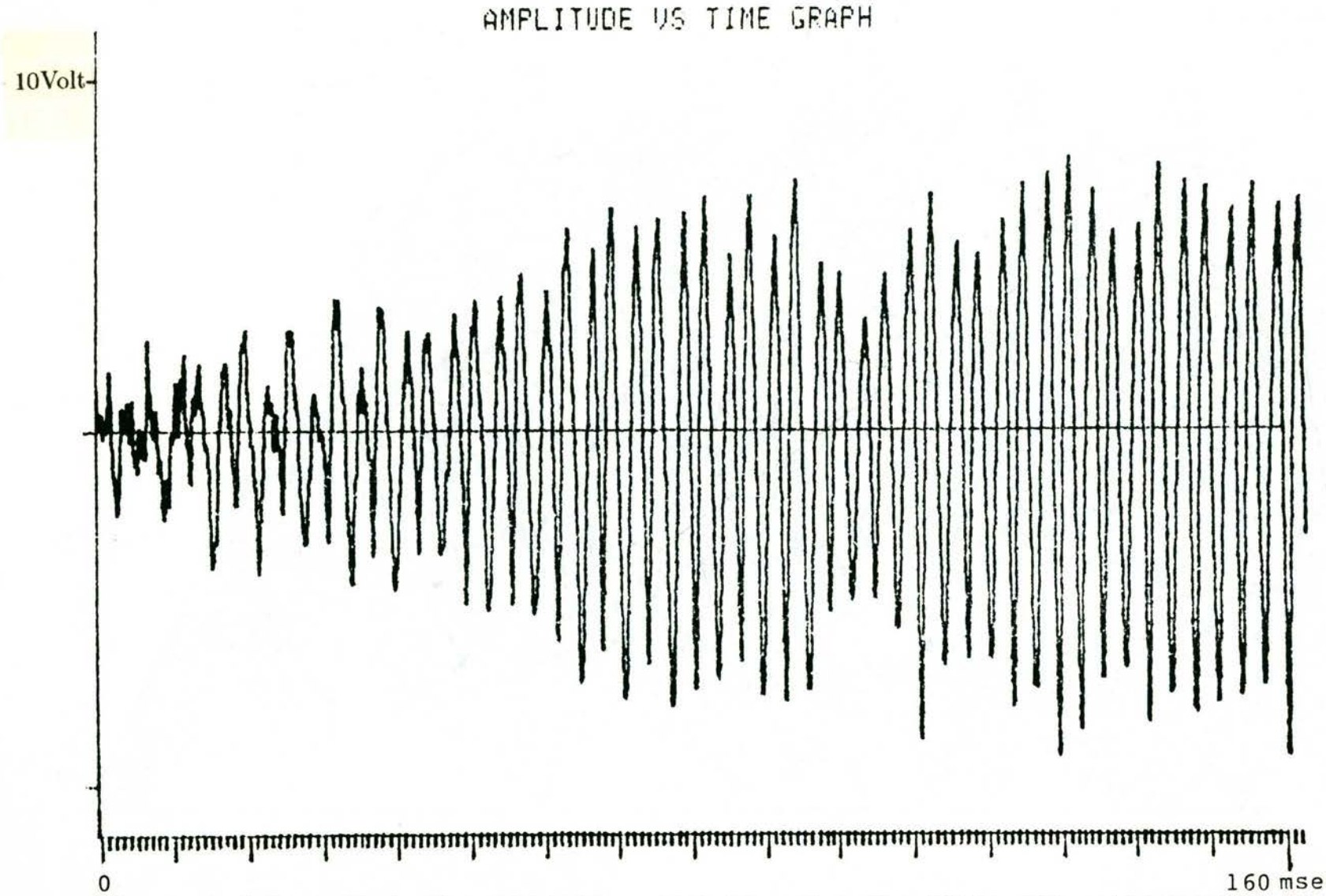


Figure 2.6 is a plot of amplitude versus time for the first 160 milliseconds of the attack and steady-state of a cello tone.

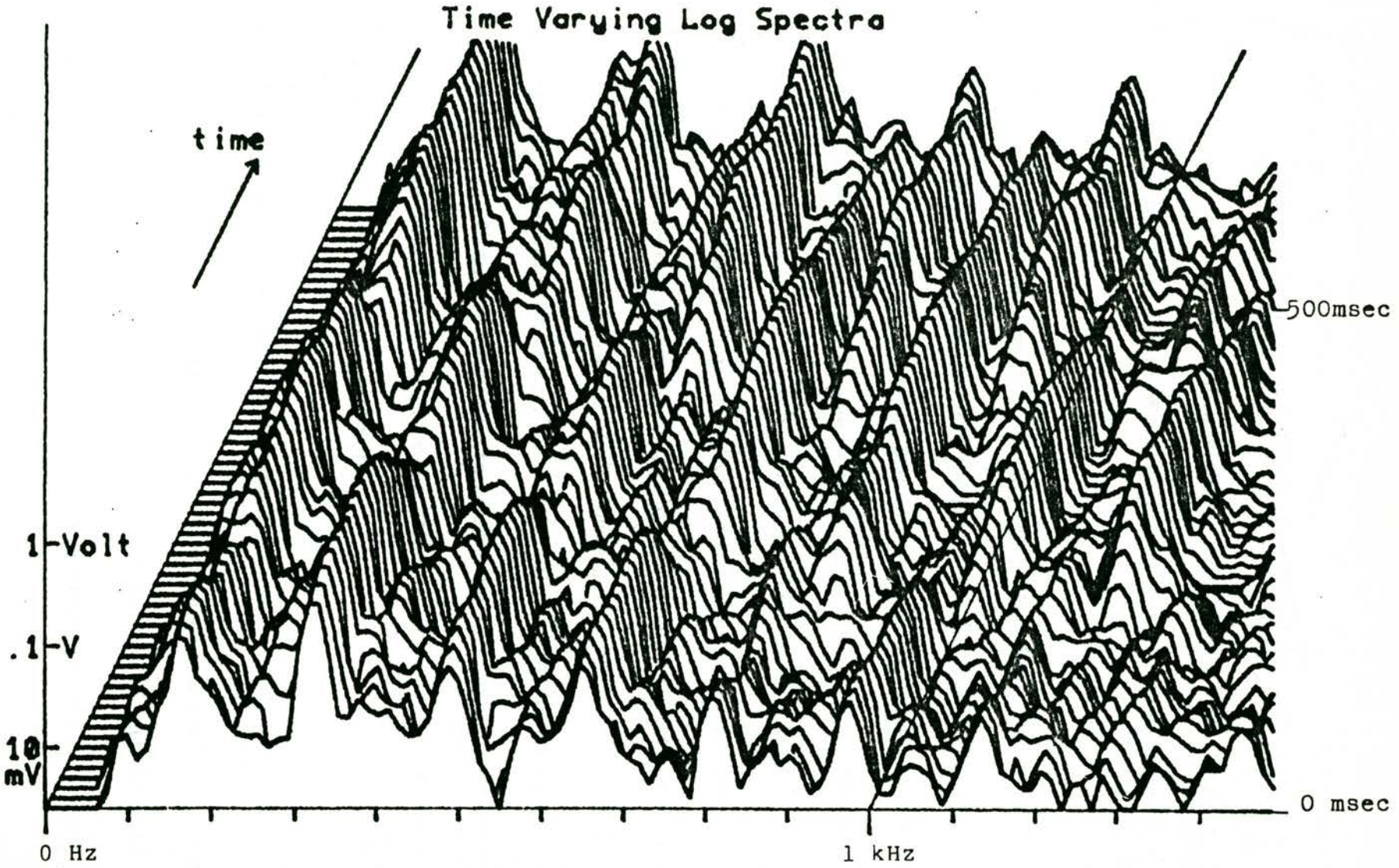


Figure 2.7 plots the time varying spectra of the same tone (E below middle C) and the transition into the next tone (F below middle C), and the next (G below middle C), spanning 600 milliseconds. Each spectrum is advanced by 10 milliseconds from the previous one, and the frequency range is from 0 to 1.5 kHz.

reverberation. Also the 5th and 6th harmonics have more energy than the lower harmonics. This corresponds to Lehman's (1964) first formant of the bassoon.

Figure 2.6 plots amplitude versus time for the first 160 milliseconds of the onset and steady-state of a cello tone. The first 20 millisecond is non-periodic.

Figure 2.7 displays the time varying spectra of the tone in figure 2.6 (*E* below middle *C*), the transition into the next tone (*F* below middle *C*), and the next (*G* below middle *C*), spanning 600 milliseconds. The time between spectra is 10 milliseconds, and the frequency ranges from 0 to 1.5 kHz. The signals of the different notes appear to overlap in time. This overlap may be caused by reverberation in the instrument and in the recording room, because the notes are played on the one string. The harmonics of the note *F* continue throughout the note *G*. The reverberation of the lower harmonics of the note *E* merge with the harmonics of the note *F*. It is physically impossible to separate these. Increasing the frequency resolution to resolve the harmonics would require the effective time window to be increased to half a second. The spectra would then contain components from several notes.

2.4 Musical Scales and Temperament

Musical scales probably evolved from a need to limit the range of possible pitches, as an aid to the perception and memory of melodies. Although musicians can differentiate up to 200 pitches within an octave, the major and minor scales have only seven notes. This corresponds to our short term memory

capacity.

The perfect fourth and fifth with frequency ratios of 4:3 and 3:2 respectively are obvious candidates for notes in the scale, and occur in the major, minor and many non-European scales, such as the Asian pentatonic scales. There is a pitch equivalence between octaves. Two notes an octave apart give a similar sensation of pitch (chroma). Therefore tones outside the range of an octave can also be mapped into the notes of the scale.

The modern major and minor scales have evolved from the ancient Grecian modes, which derived from the theoretical scale attributed to Pythagoras (Apel 1970). This scale appeared at about the same time in China. It includes the perfect fifth and the octave. The Pythagorean tone of frequency ratio 9:8 is two fifths minus one octave (3:2 times 3:2 divided by 2:1). Therefore the Pythagorean major scale is made up of two tones, a Pythagorean semitone (ratio 256:243), three tones and another semitone.

In the mid 16th Century, Zarlino proposed the "just" scale using the ratios 5:4 and 6:5 for the major and minor thirds. This created harmonious major chords.

Both these systems, though theoretically elegant, restricted the musician to the one key, because many of the pitches differed too much from their enharmonic equivalents in related keys.

In the late 16th Century, "mean-tone" temperament came into prominence. Most of the early works of J.S. Bach and Handel were written in this temperament. The perfect third (ratio 5:4) was derived by flattening the perfect

fifth by 5 cents (5% of a semitone) and generating the scale in the same way as Pythagorean temperament, but using this flattened fifth. This enabled harmonious music to be played in six related major keys and three related minor keys, but transposition to more distant keys produced "wolf" tones up to half a semitone from their correct pitch.

The modern system of equal temperament divides the octave into twelve equal semitones. This allows complete freedom in modulation but sacrifices the harmonicity of the earlier scales. The tempered semitone has a frequency ratio of 1.05946 ($\sqrt[12]{2}$), and is divided into 100 cents.

Table 2.1 gives a comparison (in cents above the tonic C) of the scales described here.

Table 2.1

Comparison of Scales

note	Pythagorean	Just	Mean-tone	Equal
C	0	0	0	0
D	204	204	193	200
E	408	386	386	400
F	498	498	503	500
G	702	702	697	700
A	906	884	890	900
B	1110	1088	1083	1100
C	1200	1200	1200	1200

The frequency discrimination of the ear is about 5 cents, therefore the equal tempered and mean-tone fifths are indistinguishable from the perfect fifth. The equal tempered third is 14 cents above the perfect (just intonation) third and differs noticeably.

Table 2.2 gives the fundamental frequencies of the major triad based on *G* at 400 Hz for the various intonations.

Table 2.2

Comparison of Scales

note	Pythagorean	Just	Mean-tone	Equal
G	400.0	400.0	400.0	400.0
B	506.2	500.0	500.0	504.0
D	600.0	600.0	598.1	599.3

Considering the consonance of the various temperaments, from table 2.2 it can be seen that the 8th harmonic of the note *B* beats with the 10th harmonic of *G*, at 32 Hz for equal temperament and at 50 Hz for Pythagorean temperament. These are therefore sources of dissonance. But the 8th harmonic of *D* and the 12th harmonic of *G* in mean tone temperament beat at 15 Hz, which is not perceived as dissonance.

In practice the tuning of musical instruments can differ by a few percent from these ideal tunings. Well tuned instruments can fall out of tune with temperature and humidity variations, and even during performance. Pitch variation in most performed music is similar to the differences in the modes of intonation. Equal temperament will be assumed for the remainder of this thesis. It is easy to calculate, because pitch is directly proportional to the logarithm of the frequency.

2.5 Problems Relating to Polyphonic Perception

This section considers what is known about the perception of polyphony

and looks at some of the problems which must be overcome in the automated analysis of polyphonic music.

Consider an inverted chord, for example, G seventh in the second inversion with frequencies 300 Hz, 400 Hz, 500 Hz, and 700 Hz. The frequencies of the harmonics of these notes are divisible by 100 Hz. The chord has a periodicity of 10 milliseconds. Most musicians would be able to identify the four individual notes, when performed on musical instruments. Chowning (1981) showed that a such a chord synthesized with exact frequencies, is perceived as a single tone with a pitch of 100 Hz. But if the frequency of each note varies independently by 1%, the harmonics fuse together and four natural sounding notes are perceived.

A similar phenomenon occurs with stopped organ pipes where several tones of different frequency are perceived as a single sound. The resultant bass saves the cost of manufacturing of large organ pipes (Apel 1970). Independent vibrato of musical parts is also an important cue in the perception of unison or quasi-consonance, because of the varying beating between parts.

Rasch (1978) has shown that a difference of as little as 10 milliseconds in onset time is enough to distinguish polyphonic sounds. His examples use a rapid onset of less than 1 millisecond followed by an exponentially decaying amplitude envelope. The onset transients of musical instrument tones can last as long as 200 milliseconds, and differ widely even when the note is repeated by the same performer.

Another important cue in the discrimination of simultaneous tones is the stereo effect. Sound is located by the phase and amplitude difference of

the signals arriving at each ear. The diffraction pattern of the outer ear is also important in sound localization (Schroeder 1975).

Melodic expectation, within the musical context of key signature and tempo, helps to determine the musical events in an acoustical signal. Timbre of instruments also assists in the tracking of individual parts of music.

The following musical examples illustrate some of the difficulties in determining the tones in polyphonic music.

Figure 2.8 shows the time varying spectra of the final 200 milliseconds of a bassoon tone (*E* of frequency 165 Hz), followed by a chord of three tones; a bassoon (*A* at 220 Hz), and two oboes (*C*♯ at 554 Hz and *E* at 660 Hz). Each spectrum advances by 10 milliseconds from the previous one, and the frequency ranges from 0 to 3 kHz. The onsets of the three superimposed tones are difficult to distinguish.

Figure 2.9 shows all the harmonics of a half second transition between two chords in the woodwind Trio. The notes of the first chord are; *B*♭ p13 (pitch 13), *B*♭ p25, and *G* p34. The notes of the second chord are; *C* p15, *G* p22, and *E*♭ p30. The exact time of transition of the harmonics differs by as much as 100 milliseconds.

Figure 2.10 is a plot of amplitude versus time for two oboes and a bassoon playing simultaneously. It is difficult to determine from the plot which notes are played. A prominent periodic feature is the series of marked peaks with a time difference of 3.46 milliseconds, but this periodicity does not correspond to any of the three fundamental frequencies.

Figure 2.11 is the spectrum of the superimposed tones and the three correct estimates of the notes being played.

Figure 2.12 displays the result of the harmonic summing algorithm, described in section 5.5. The *D* and *B♭* estimates an octave below the correct estimates are eliminated, because the even harmonics are significantly stronger than the odd harmonics. The *G* estimate an octave above the correct value cannot be eliminated by the criteria of the earlier monophonic example (figure 2.2). This estimate is stronger than would be expected from the bassoon alone, because the 2nd, 4th and 6th harmonics of the *D* also contribute to this.

The example of figures 2.10 to 2.12 is used throughout the thesis to illustrate the spectral extraction procedure (section 5.6), to compare pitch estimating techniques (section 7.2), and to simulate the response of the human cochlea to simultaneous tones (section 9.4).

2.6 Conclusion

This chapter has investigated the nature of musical sound. The physical and perceptual attributes of musical tones were discussed and compared. Examples of musical tones were presented to demonstrate the difficulties in predicting perceptual attributes, especially pitch, from acoustical signals.

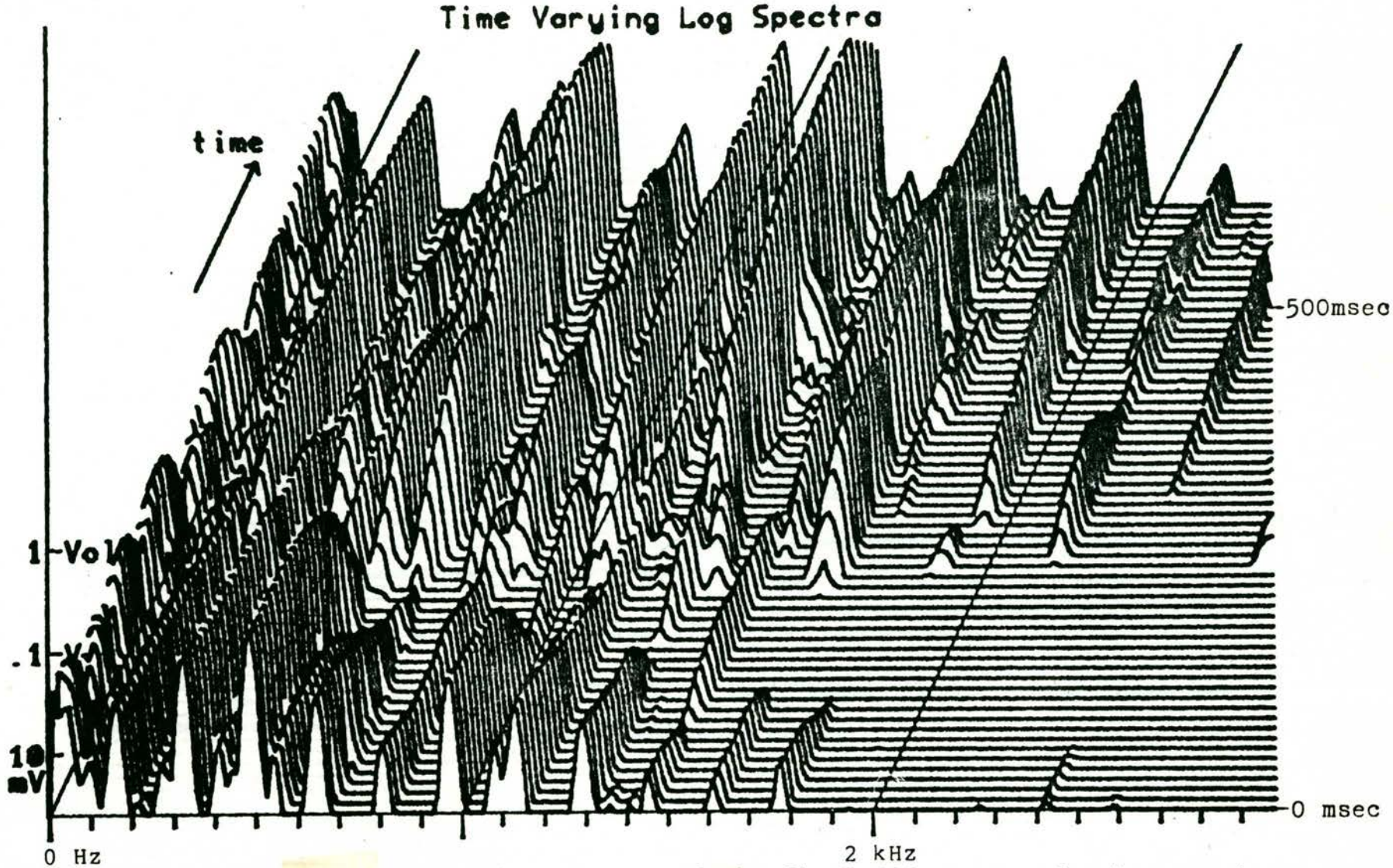


Figure 2.8 plots the time varying spectra of the final 200 msecs. of a bassoon tone (E of frequency 165 Hz), followed by a chord of three tones; a bassoon (A at 220 Hz), and two oboes (C# at 554 Hz and E at 660 Hz). Each spectrum is advanced by 10 milliseconds from the previous one.

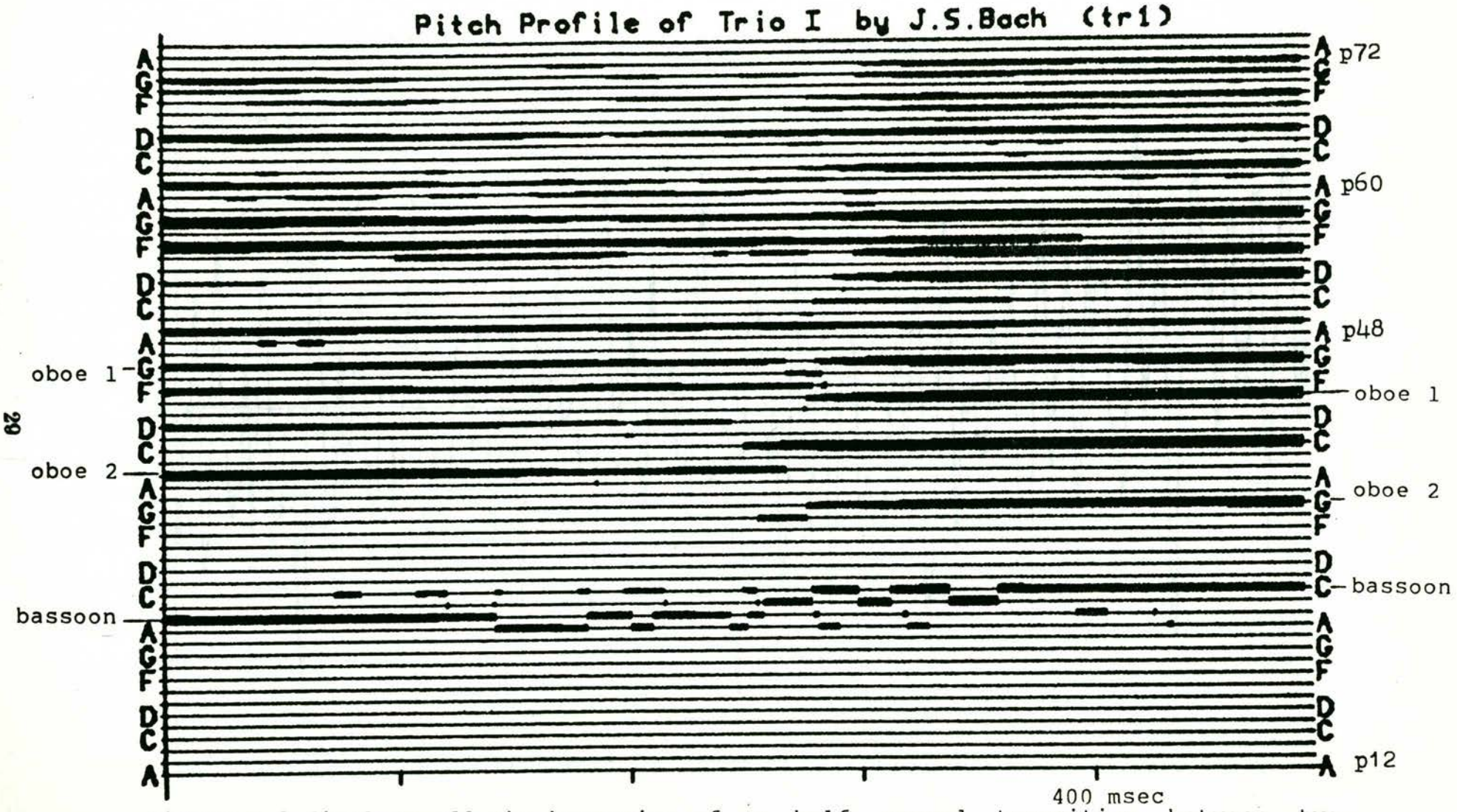


Figure 2.9 displays all the harmonics of a half second transition between two chords from the Trio by J.S.Bach.

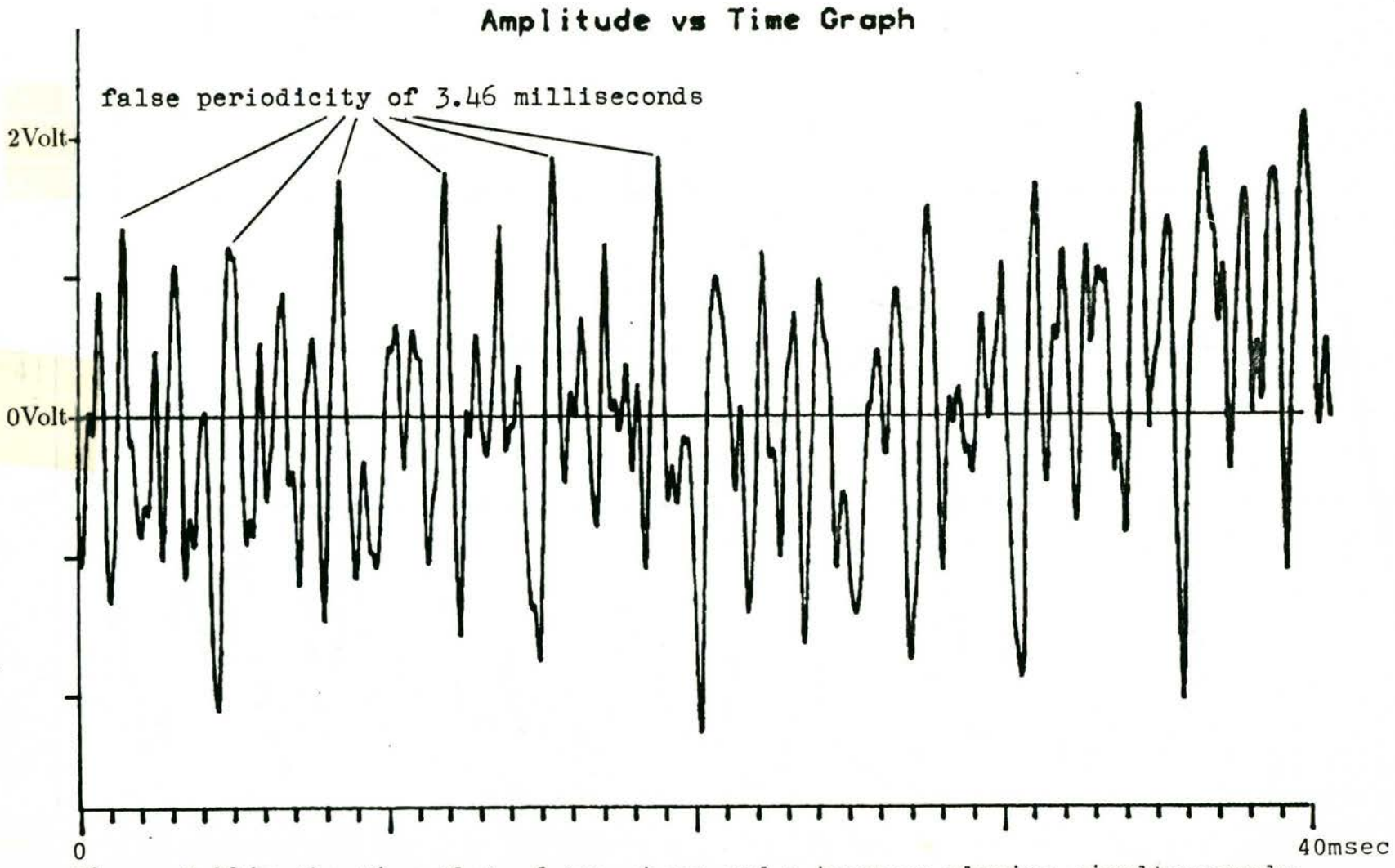


Figure 2.10 is the time plot of two oboes and a bassoon playing simultaneously.

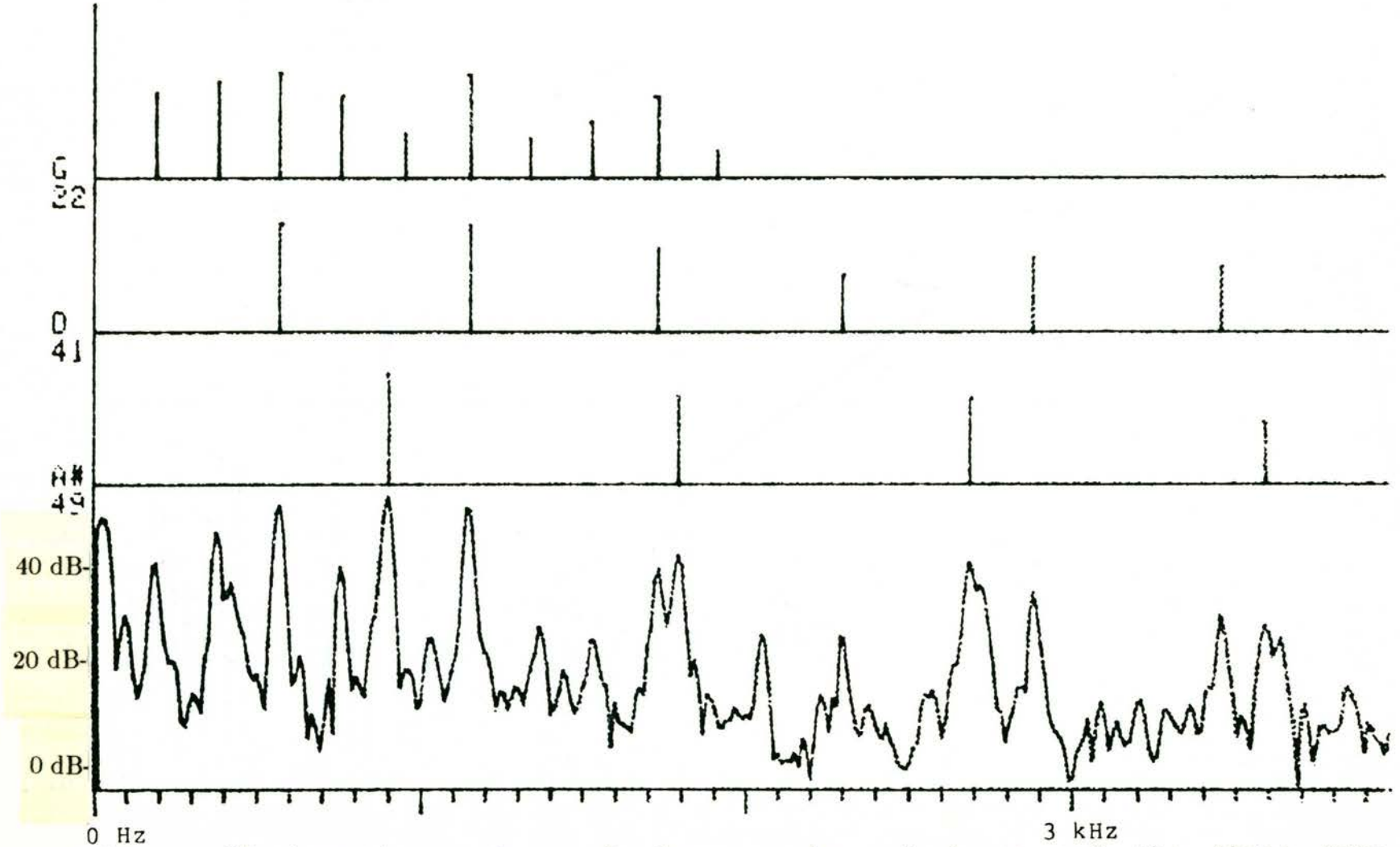


Figure 2.11 gives the spectrum of the superimposed tones and the three true estimates of the notes being played.

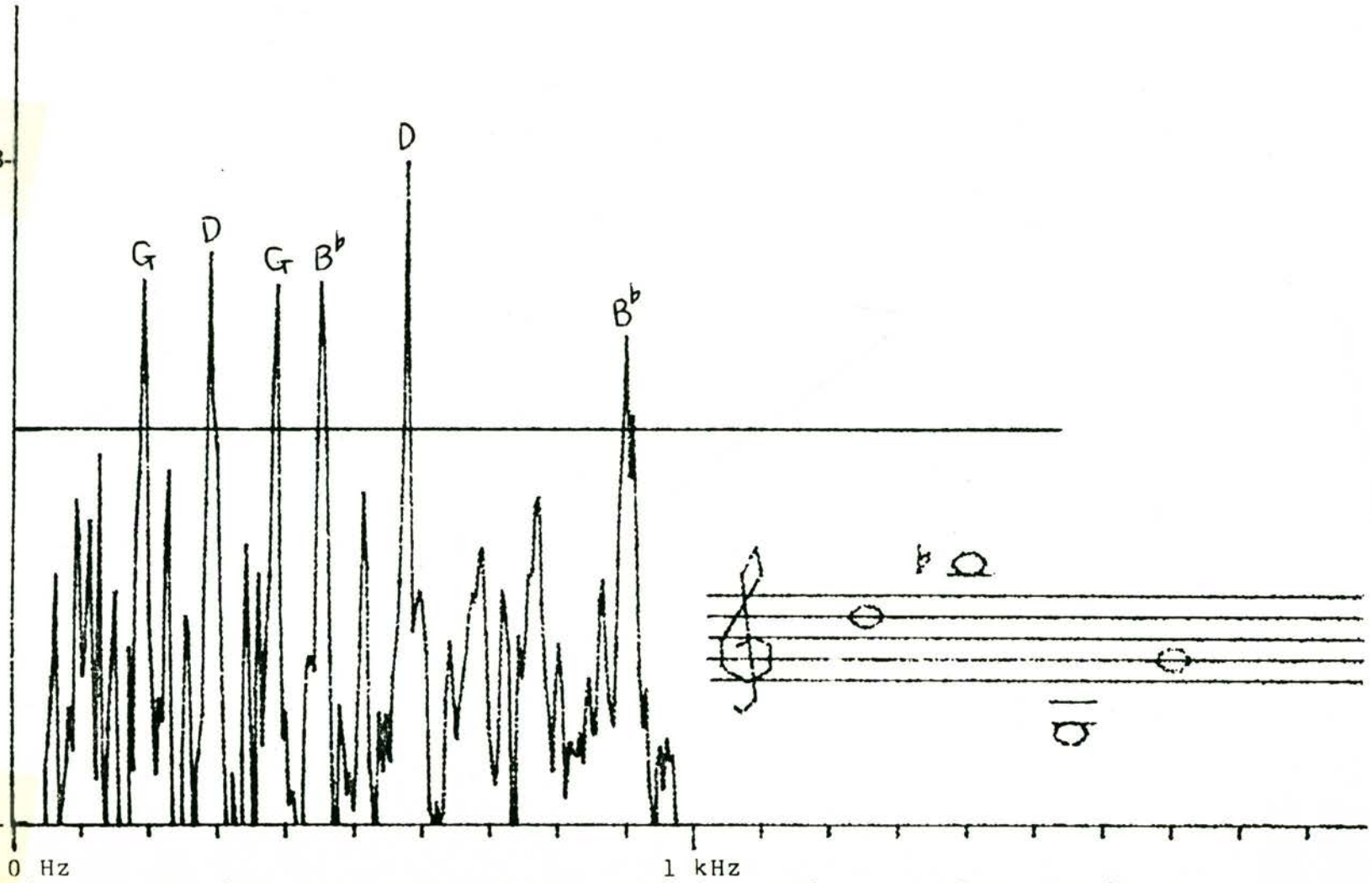


Figure 2.12 gives the result of the harmonic summing algorithm applied to figure 2.10

CHAPTER THREE

Pitch Estimating Algorithms: The State of the Art

3.1 Introduction

This chapter introduces standard signal processing techniques used for pitch estimation in speech and music. Oppenheim and Schafer (1975) give a detailed discussion. The techniques are applied in either the time domain (section 3.2) or the frequency domain (section 3.4). Section 3.3 describes the method of spectral analysis used in this thesis. All the techniques of this chapter are applied to a finite set of data (sampling window) sampled at equal time intervals. Time is assumed to take integral values, the integer N denotes the width of the sampling window, $j = \sqrt{-1}$ and $\text{cis}\ z = \cos z + j \sin z$.

3.2 Time Domain Methods

A function, $x(t)$, is said to be almost periodic if there exists a period T, and a function $e(t)$, such that; $x(t) = x(t+T) + e(t)$, and $e(t)$ is small with respect to $x(t)$. For the steady state of most musical tones, the maximum value of $e(t)$ is only a few percent of that of $x(t)$. If $e(t)$ is zero for all t then $x(t)$ is said to be periodic, but for signals sampled from the real-world this rarely occurs.

3.2.1 Autocorrelation

The discrete autocorrelation $A(t)$, of a function $x(t)$, is:

$$A(t) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n+t),$$

where t is an integer.

The autocorrelation function is equivalent to the inverse Fourier transform of the power spectrum of $x(t)$. The evaluation of $A(t)$ (for $t = 0, N-1$) requires $O(N^2)$ multiplications whereas the fast Fourier transform (see section 3.3) requires only $O(N \log N)$. Therefore for large N (eg. $N > 100$), $A(t)$ is more economically derived using the fast Fourier transform.

If $x(t)$ has period T, then $A(nT)$ is locally maximal for integral n.

3.2.2 Average Squared Difference Function

The average squared difference function (ASDF) $S(t)$, of $x(t)$ is:

$$S(t) = \frac{1}{N} \sum_{n=0}^{N-1} [x(n) - x(n+t)]^2,$$

where t is an integer.

If $x(t)$ has period T, then $S(nT)$ is locally minimal for integral n.

3.2.3 Average Magnitude Difference Function

The average magnitude difference function (AMDF) $M(t)$, of $x(t)$ is:

$$M(t) = \frac{1}{N} \sum_{n=0}^{N-1} |x(n) - x(n+t)|,$$

where t is an integer. This function is also called the optimum comb filter (Moorer 1974).

If $x(t)$ has period T, then $M(nT)$ is locally minimal for integral n.

From Schwarz's inequality, $M(t)^2 = b(t)S(t)$, where $b(t) \leq 1.0$. Ross et al. (1974) showed empirically, from a wide range of speech samples, that $b(t)$ is not strongly dependent on t, and lies in the range [0.4 , 1.0]. The AMDF is more

economical to compute than the ASDF, where multiplication takes significantly longer than addition.

3.2.4 Linear Predictive Coding

Linear Predictive Coding (LPC) was developed by several workers (Schroeder 1970, Markel 1972, Makhoul 1973, Maksym 1973) for the automatic recognition of speech. The speech signal, $x(n)$, is modelled as a signal generated by the vocal cords (glottis) which is filtered by the vocal tract:

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + e_n,$$

where $A = \{a_k\}(k = 1, \ldots, p)$ is the set of filter coefficients modelling the vocal tract and $E = \{e_n\}(n = 1, N)$ is the set of glottal excitation samples. LPC finds A to minimize $\sum_{n=1}^{N} e_n^2$, where N is some large integer. This technique is not applicable to pitch recognition in music, because it assumes white noise or a periodic pulse train as the signal source. Moorer (1974) found it was not useful for distinguishing simultaneous tones.

3.2.5 Spectral Flattening

Spectral flattening is a pre-processing technique that improves pitch estimation. Methods such as cubing the signal values (Tucker 1977) can improve pitch estimation by reducing the amplitude of the strongest spectral peaks of the signal. Inverse filtering (Markel 1972) flattens the spectral envelope of a signal, by finding the LPC filter coefficients, determining the inverse filter, and applying

it to the signal. Tucker used this technique for enhancing the pitch estimation of musical instruments such as the bassoon. The author applied spectral flattening techniques to simultaneously sounding musical tones, but found no improvement in pitch detection for polyphonic music.

3.3 Spectral Analysis

Let $x(n)$, (for $n = 0, \ldots, N-1$) be sampled data spanning one period of a signal. The discrete Fourier transform (DFT) of $x(n)$ is:

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \exp\left(\frac{-2\pi jkn}{N}\right),$$

for $k = 0, \ldots, N-1$.

To evaluate this transform the summing term must be calculated N^2 times. (Cooley, Tukey 1965) developed the fast Fourier transform (FFT) to evaluate the DFT in $O(Nlog(N))$ multiplications, which is a significant computational saving for large N. Detailed accounts of the FFT can be found in the literature (Cooley, Tukey 1965), Brigham (1975), Winograd (1978)). Section 4.3.2 describes the implementation used in this thesis.

3.3.1 Spectral Leakage

Spectral peaks are broadened, if the width of the sampling window is not a multiple of the period of the signal. This is called spectral leakage, and results from the incorrect assumption that the signal is infinitely repeated outside the sampling window. To avoid this, the sampled points may be either scaled in time to an exact multiple of the period, or multiplied by a bell shaped

weighting function (denoted W_n, for $n = 0, \ldots, N-1$). The effect of this weighting function is to reduce the discontinuity at the boundary of the periodic extension. Multiplication in the time domain is equivalent to convolution in the frequency domain, therefore this weighting function broadens the spectral peaks to the shape of the weighting function's frequency response.

3.3.2 Weighting Functions

The Dirichlet Kernel is introduced to explain spectral leakage. Consider a digital signal that is unity in the range $(0, N-1)$, and zero elsewhere. The frequency response of this signal is the Dirichlet kernel:

$$D(n) = \frac{1}{N} \frac{\mathrm{cis}\left(\frac{\pi n}{N}\right) \sin \pi n}{\sin\left(\frac{\pi n}{N}\right)}$$

Therefore,

$$\lim_{N \to \infty} D(n) = \frac{\sin(\pi n)}{(\pi n)}.$$

This is a good approximation to $D(n)$, if N is large.

The Dirichlet kernel, $D(n)$ is zero for all integral, non-zero n. It has a maximum amplitude of 1 at $n = 0$ (main lobe) and locally maximal amplitudes (sidelobes) between the points of zero amplitude. If N is large the sidelobes occur at $\pm 1.5, \pm 2.5, \pm 3.5$, etc.

A delta function, $\delta(n)$, is defined to be unity for $n = 0$, and zero for all non-zero real n. The convolution of a Dirichlet kernel, $D(n)$, with a delta function, $\delta(n)$, is also a Dirichlet kernel. More precisely, the convolution $D(n) * \delta(n-k) = D(n-k)$, for all real k. If k is integral then the convolution

is a delta function, $\delta(n-k)$, otherwise $D(n-k)$ takes non-zero values for all integral n. In the time domain, this is equivalent to saying that if the width of the window is a multiple of the period of the sampled signal, no spectral leakage occurs, and the DFT has a single non-zero value corresponding to this period.

Consider the following weighting function:

$$W_n = \sum_{i=0}^{4} a_i \cos\left(\frac{2\pi i n}{N}\right).$$

The frequency response of such a function is the weighted sum of shifted Dirichlet kernels. When this is applied to a periodic signal before calculating the DFT, the spectral peaks of the harmonics are broadened to the shape of the frequency response of this weighting function. This shape can be tailored to the application by altering the weightings.

As an example of this weighting function, consider the Hann window, with frequency response,

$$\frac{D(n)}{2} + \frac{D(n-1)}{4} + \frac{D(n+1)}{4}$$

The first sidelobe has a maximum amplitude of, $D(2.5)/2 + D(1.5)/4 + D(3.5)/4$ or .0243, which is 32 dB below the main lobe with maximum amplitude 1.

The Hamming window has frequency response:

$$.54D(n) + .23D(n-1) + .23D(n+1)$$

Here, the weighting of the central kernel, $D(n)$, is increased and the weighting of the shifted kernels $\{D(n+1), D(n-1)\}$ decreased. A value of zero is produced

in the first sidelobe, making the second sidelobe at $n = 3.5$ the strongest at 43 dB below the main lobe.

Harris (1978) found that by adding further Dirichlet kernels at two and three DFT points from the centre, he could reduce the maximum sidelobe amplitude to 67 dB and 92 dB respectively below the main lobes, at the cost of widening the main lobes. The weighting functions used in this thesis can be found in Table 3.1.

Table 3.1

Spectral Weighting Functions

name	sidelobe attenuation	6 dB bandwidth	weightings				
			a_0	a_1	a_2	a_3	a_4
Rectangle	13 dB	1.21	1.0	0	0	0	0
Hann	32 dB	2.00	0.5	0.5	0	0	0
$\cos^3 x$	39 dB	2.32	0.375	0.5	0.125	0	0
$\cos^4 x$	47 dB	2.59	0.3125	0.4687	0.1875	0.03125	0
Hamming	43 dB	1.81	0.54	0.46	0	0	0
3 term minimal Blackman-Harris	67 dB	1.81	0.4232	0.4975	0.079	0	0
4 term minimal Blackman-Harris	92 dB	2.72	0.3587	0.4882	0.1412	0.0117	0
Gaussian	69 dB	2.52	not applicable				

The sidelobe attenuation is the difference in decibels between the maximum amplitude of the main lobe and that of the highest sidelobe. The 6dB bandwidth is the range of frequencies with response within 6dB of the main lobe, and is used as a measure of the width of spectral peaks.

The 3 term minimal Blackman-Harris weighting function is used in this

thesis, because it provides the narrowest peak consistent with sufficient sidelobe attenuation (67 dB) to isolate harmonics from noise. The signal to noise ratio of recorded music is about 60 dB, corresponding to an error of 0.1%

The Gaussian weighting function is defined by:

$$W_n = \exp -\alpha\left(\frac{n}{N}\right)^2,$$

where α is a constant.

The Gaussian window has sidelobe attenuation similar to, and spectral peaks half as wide again as the 3 term minimal Blackmann-Harris window.

The heterodyne filter, used by Moorer (1975), is equivalent to a DFT applied to a sampling window containing one period of the signal. If the signal is dissonant or has no readily identifiable period, spectral leakage occurs. By comparison, the weighting functions have the advantage that the leakage of spectral peaks has a predictable shape and can be confined to effect only closely neighbouring peaks.

Figure 3.1 compares a spectral analysis using the Gaussian weighting function (upper spectrum) with the heterodyne filter (lower graph). The signal is a steady state bassoon tone with fundamental frequency 128 Hz. The first, 4th, 5th, 8th, and 9th harmonics of the heterodyne filter produce narrow peaks, because their frequencies are close to the exact harmonic frequencies determined by the average magnitude difference function. The 2nd, 3rd, 6th and 7th harmonics, however, deviate far enough from the exact harmonic frequencies to cause severe spectral leakage over a wide portion of the spectrum. Any har-

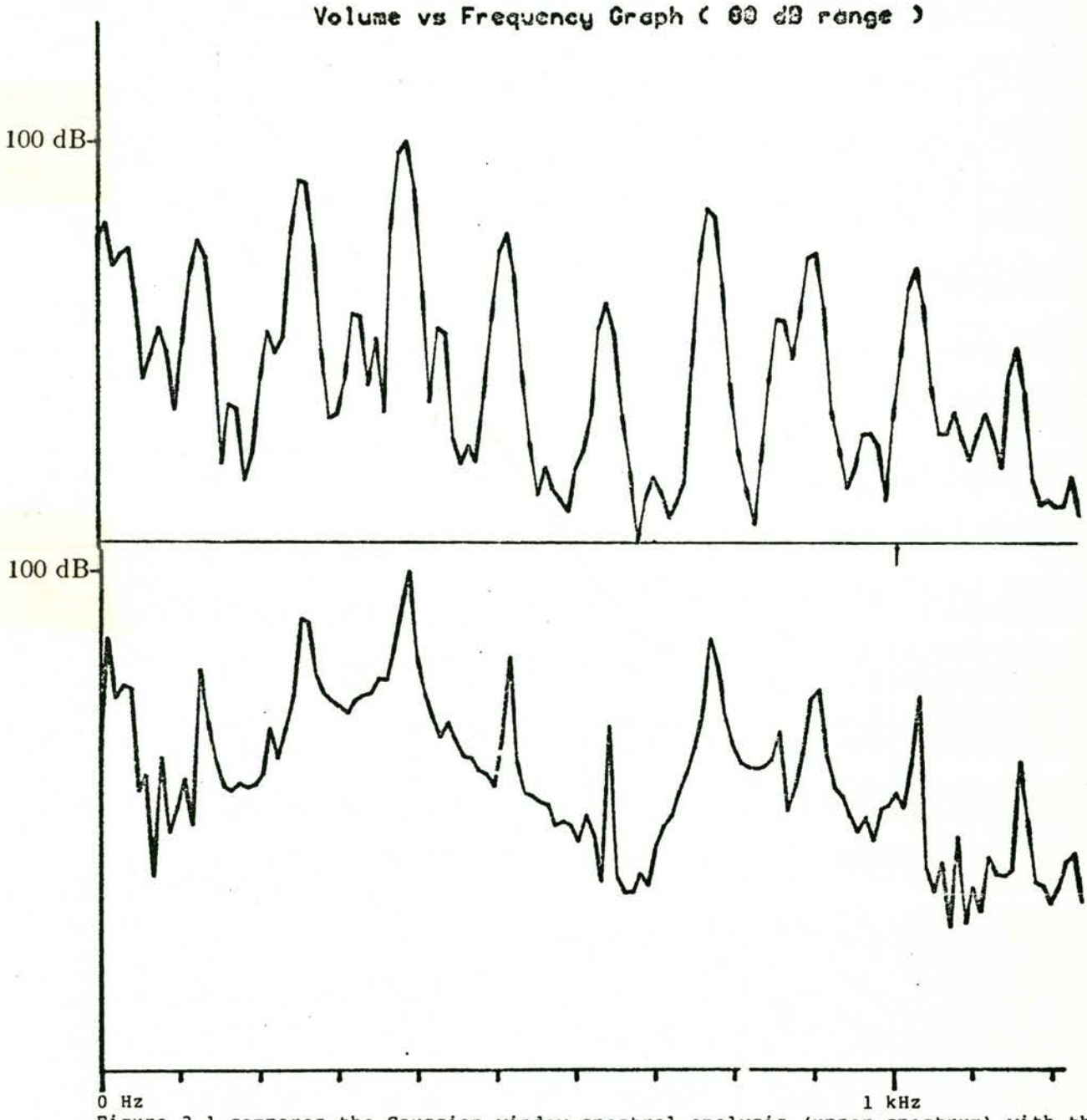


Figure 3.1 compares the Gaussian window spectral analysis (upper spectrum) with the Heterodyne filter output (lower graph). The tone is the steady state of a bassoon.

monic signal with frequency between these spectral peaks, and intensity 25 to 50 dB below the peaks would be masked by the heterodyne filter, but not by the Gaussian spectrum.

3.4 Frequency Domain Methods

3.4.1 Period Histogram

The harmonic frequencies of a periodic signal are multiples of the fundamental frequency. Schroeder (1970) suggested exploiting this in pitch estimation of speech, although no results were presented. His method was to make a histogram of the submultiples of the harmonic frequencies, and take the most frequent submultiple as the estimate of the fundamental. For example, given the harmonic frequencies 300, 400, 500 Hz, the submultiples are (150,100,75,60,...), (200,133,100,80,...) and (250,167,125,100,83,...) respectively. The most frequent submultiple, 100 Hz, is taken as the fundamental frequency.

Piszczalski (1979) applied this method to musical tones, using a histogram of the highest common factor of pairs of harmonic frequencies. A similar method was implemented by the author to determine the pitch of simultaneously sounding tones (see section 5.4).

3.4.2 The Cepstrum and Deconvolution

The cepstrum is the inverse Fourier transform of the log power spectrum of the signal. The cepstrum of two convolved signals is the sum of their cepstra. The name cepstrum is coined by reversing the "spec" in the word

spectrum. Time is termed quefrency (coined from frequency), and frequency differences are termed repiodicity (coined from periodicity).

The cepstrum was introduced in 1963 by Bogert et al. as a heuristic technique for separating seismic signals from their echoes. The echo impulse response is convolved with the seismic signal. In the frequency domain the frequency responses of the signal and its echo are multiplied. By taking the logarithm of the amplitude the signal spectra are additive and can therefore be easily separated.

Stockham et al. (1975) used the same method for improving old sound recordings, notably those of Enrico Caruso. The frequency response of early recording equipment is constant throughout the recording, therefore it is possible to determine, and then filter out the characteristic metallic sound of the old recordings.

Similar work has been done on the removal of blurs from video images (Oppenheim et al. 1968), and the removal of room reverberation from audio signals (Schafer 1969).

With speech signals, the cepstrum can be used to deconvolve the vocal tract response from the glottal source (see 3.2.4).

3.4.3 Walsh Transform

The Walsh Transform is based on a set of orthonormal binary functions, in the same way that the Fourier transform uses orthonormal sinusoidal functions (K.G. Beauchamp 1975).

Walsh transforms are generally faster to calculate than Fourier transforms, because logical operations are used instead of multiplications and additions. Tucker (1977) concludes that there is no advantage in using Walsh transforms to analyse music, because of the strongly sinusoidal nature of musical signals.

3.5 Conclusion

This chapter presents signal processing techniques currently used for the estimation of pitch in speech and music. The usefulness of these techniques is discussed for analysing polyphonic music. The method of spectral estimation used in this thesis is described in more detail.

CHAPTER FOUR

Computer Hardware and Software for the Analysis of Music

4.1 Introduction

This chapter describes the software and hardware used in this research for the automated analysis and transcription of recorded music. This includes the analog-to-digital conversion of music (section 4.2), the interactive software for plotting signals, several signal processing procedures, and original procedures for automatically determining pitches of notes (section 4.3). The system for transcribing these notes into standard music notation is described in chapter 6. The author wrote all the programs mentioned in this chapter, except the UNIX system software.

Figure 4.1 shows the interface between the computer and the musician. Music can be entered as a sound recording, or it can be played in on an electronic keyboard. Output can be heard via an organ or digital-to-analog converter, or plotted in standard music notation.

4.2 Hardware for Digitizing Music

Conventional audio recording and amplification devices represent sound as a continuously varying voltage signal. Before this signal can be processed by a general purpose digital computer, the analog voltage has to be measured and recorded at successive points in time. This process, termed analog-to-digital (A/D) conversion, first low-pass filters the analog input, then samples and holds

COMPUTER - MUSIC INTERFACE

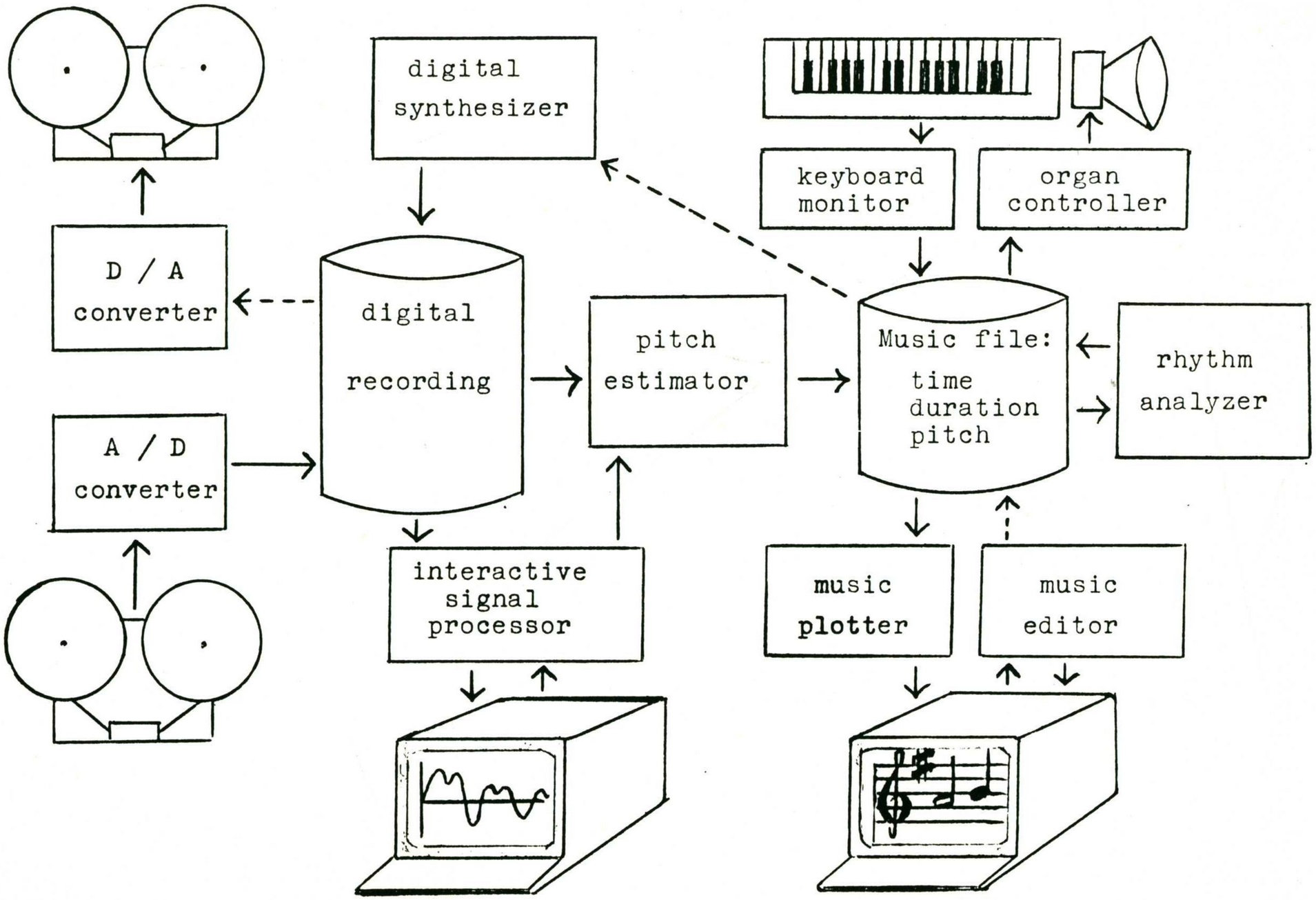


Figure 4.1 is the interface between the computer and the musician. Music can be input as a live performance or as a sound recording via the A/D converter, or it can be played in on an electronic keyboard or via a music notation editor. Conversely, music can be output as sound via the organ or D/A converter, or plotted on a graphics terminal.

the instantaneous voltage until the conversion is complete, and finally, stores the digital value. A/D conversion is controlled by a clock that samples the signal at a constant rate.

4.2.1 Digitization Errors

Two major types of error produced by A/D conversion are referred to as aliasing and quantization error.

First, aliasing is caused by the ambiguity of digitizing signals with frequencies greater than half the sampling frequency. Indeed, there is no way to distinguish between a signal of frequency f and its alias of frequency $(f_s - f)$, where f_s is the sampling frequency. This problem is countered by sampling at twice the highest frequency of interest, and applying a low-pass filter to the input signal with cut-off frequency less than $f_s/2$ (the so called Nyquist frequency).

The second type of error results from the quantization of the analog signal. For example, the A/D converter used for this research maps the voltage range of plus or minus 10 volts into a 12 bit register. Therefore a single digital step corresponds to an analog difference of 10/2048 volts (4.88 millivolts), which represents a maximum signal to noise ratio of 74 dB. This compares favourably with the signal to noise ratio of analog recording techniques, which range from 55 dB for portable cassette recorders to 80 dB for professional studio recorders. Blesser (1978) gives a comprehensive examination of digitization.

The sampling rate of 25.6 kHz was used because the significant frequency range of musical signals lies below 10 kHz. 25 kHz was also used by J.A.

Moorer (1975). The value of 25.6 kHz makes the UNIX block size of 256 words correspond to 10 milliseconds of signal.

In the final stage of A/D conversion, storage of data was critical. The data had to be transferred to disk storage on a second computer, because there was no mass storage device on the computer performing the A/D conversion. The speed of the data link (9600 baud) between the computers, caused the A/D conversion to be interrupted when all the memory buffers were filled with data. This meant that the maximum duration of continuously digitized data was only one second. To record longer segments it was necessary to replay the analog recording repeatedly. The A/D conversion was triggered by the first sampled data that exceeded the recorded noise threshold. Successive segments of data (1 second in duration) were recorded after each replay. The digitized data contained discontinuities between the segments, caused by minor variations in response to the analog trigger, and differences in the analog signal during successive replays. When a Fourier transform is applied to a time segment containing a discontinuity, spectral leakage occurs (see 3.3.1).

A high-speed link was later installed between the computers to enable large segments of data to be transferred continuously. Double buffering was used on both machines to provide a continuous flow of data, uninterrupted by system overheads such as initiating direct memory access (DMA) transfers.

4.2.2 Detection and Correction of Discontinuities

The digitized signal is validated to check for discontinuities, and to de-

termine if the correct recording amplitude was used. If the recording amplitude is low, the quantization noise can be too great. However, when the voltage of the signal being converted is too large, signals are peak-clipped at the maximum recordable values. For example, if the input voltage is 11 volts, one volt above the maximum, a value corresponding to 10 volts would be recorded. The sampled points at this maximum value are counted, and if they are too numerous, the piece of music must be re-recorded at a lower amplitude level.

Discontinuities occur when the A/D converter is driven near the maximum conversion speed. The data register occasionally gives incorrect readings, which cause spikes in the sampled data. These errors occurred less than one in 10,000 samples. The recording is smoothed by polynomial interpolation between the points adjacent to the erroneous samples.

Spikes and discontinuities are detected by calculating the first and second order differences between sample values, and determining when the absolute second order difference exceeds a constant (generally one hundred points or about 500 mV). Second order differences are necessary because first-order differences do not detect cusps; that is, points where the signal is continuous, but its first derivative is discontinuous.

4.3 Interactive Software for Audio Signal Analysis

An interactive system is used for studying musical signals, and for evaluation of signal processing algorithms applied to these data.

4.3.1 A Description of the Interactive Program

This program interactively computes signal processing functions and represents them in graphic form on the terminal. It is invoked with a data file as an argument. After opening files and initializing data, the user is prompted for commands. The help command gives information on the commands that are available. There are about 30 commands, and further procedures are easily compiled and linked to the program.

The program acts on complex floating point data stored in a buffer. In the time domain only the real part is used, while in the frequency domain both real and imaginary parts are used. The contents of this buffer can be saved and recalled if different data segments need to be compared.

Three ways of introducing data to the buffer are: reading from the data file at a given time, recalling a saved buffer, or using a synthesis routine to superimpose sine waves, sawtooth waves or pulse waves of given phase, frequency and amplitude. This synthesis routine is used for calibration and testing.

The size of the buffer is determined by the user. The maximum buffer size of 4096 corresponds to 160 milliseconds of sound, and is too large to isolate some musical events. A buffer size of 1024 or 2048, (40 or 80 milliseconds), is usually enough to separate rapidly changing musical events, while maintaining adequate frequency resolution for isolating harmonics in the spectra.

Data is plotted on a graphics terminal. This is an economical representation, and is a convenient way for the user to identify the relevant characteristics of the data. Long listings of data may also be printed for detailed examination.

Voltage can be plotted against time. In the frequency domain, spectra can be graphed either as log amplitude, amplitude, complex amplitude, or phase, versus frequency. The signal processing techniques of chapter 3 can be invoked interactively.

Finally, the harmonic grouping algorithms described in chapter 5 can be applied and the results plotted. Spectral peaks can be isolated and the frequencies compared for near integral ratios, to give the most likely fundamental frequency estimates (Schroeder (1970), Noll (1964), Piszczalski (1979)). Cepstra, the harmonic summing algorithm, and the spectral extraction method can also be applied.

4.3.2 Fast Fourier Transform Implementation

Spectral analysis is central to this research. Considering the amount of processing time spent evaluating the fast Fourier transform (FFT) and its inverse, some care has been taken to improve this procedure.

The author improved the FFT with pointers to arrays instead of indices. The sine and cosine functions are evaluated in the order required by the FFT when the program begins. This enables pointers to the sine and cosine tables to be merely incremented during FFT evaluation. The ordering of these tables is independent of the value of N, so the tables do not have to be re-evaluated if N is changed. See Appendix IV for the C source code of the FFT program used.

4.4 Software for Automated Pitch Recognition

The interactive system was used to test algorithms, and determine optimal parameters for segments of musical data. These algorithms were then applied repeatedly to generate pitch estimates for complete musical pieces. The time between pitch estimates must be small enough to detect the briefest of musical events. This is typically 10 to 50 milliseconds.

This system can also be used for simulating the response of the human cochlea to musical sounds. This is dealt with in chapter 8.

APPENDIX IV

An In-Place Fast Fourier Transform

```
# define begin {
# define end }
float co[N],si[N];
float xr[N],xi[N];

/*
**      compute FFT of xr,xi
*/
FFT(n,nu)
int n,nu;
begin
     int i;
     for (i=0;i<n;i++)
          xi[i] = -xi[i] ;
     IFFT(n,nu);
     for (i=0;i<n;i++)
     begin
          xr[i] = xr[i]/n ;
          xi[i] = -xi[i]/n;
     end
end

/*
**      find the nu bit reverse of k
*/
int BitReverse(k,nu)
int k,nu;
begin
     int i,k1,k2,kk;
     k1 = k;
     kk = 0;
     for (i=1;i<=nu;i++)
     begin
          k2 = k1>>1;
          kk = (kk<<1) - (k2<<1) + k1;
          k1 = k2;
     end
     return(kk);
end
/* end BitReverse */
```

```
/*
**      Set up sine, cosine tables.
**      Entries are sorted in the order required for IFFT().
**      i.e. 0, pi/2, pi/4, 3pi/4, pi/8, etc.
*/
SinCosTable(n,nu)
int n,nu;
begin
     int i;
     float arg,p;
     p = 3.1415926 / n ;
     for (i=0;i<n;i+= 2)
     begin
          arg = p * BitReverse(i,nu) ;
          co[i+1] = -(si[i] = sin(arg)) ;
          arg = p * BitReverse((i+1),nu) ;
          co[i] = si[i+1] = sin(arg) ;
     end
end

IFFT(n,nu)
/* This procedure computes the Inverse Fast Fourier
** Transform of n points, n is 2 to the power of nu;
** xr, xi are the real and imaginary parts
** respectively to be transformed
*/
int n,nu;
begin
     float tr,ti,*c,*s,*r1,*r2,*i1,*i2;
     int i,j,k,m,n2;
     n2 = n>>1;
     for (m=0;m<nu;m++)
     begin
          c = co;
          s = si;
          r2 = n2 + (r1 = xr);
          i2 = n2 + (i1 = xi);
          for (i=0;i<n2;i++)
          begin
               tr = *r2;
               ti = *i2;
               *r2++ = *r1 - tr;
               *i2++ = *i1 - ti;
               *r1++ += tr;
               *i1++ += ti;
          end
```

```
        k = 1<<m;
        for (j=1;j<k;j++)
        begin
            r1 += n2;
            i1 += n2;
            r2 += n2;
            i2 += n2;
            c++;
            s++;
            for (i=0;i<n2;i++)
            begin
                tr = (*r2)*(*c)+(*i2)*(*s);
                ti = (*i2)*(*c)-(*r2)*(*s);
                *r2++ = (*r1) - tr;
                *i2++ = (*i1) - ti;
                *r1++ += tr;
                *i1++ += ti;
            end
        end
        n2 >>= 1;
    end
    r1 = xr;
    i1 = xi;
    for (i=0;i<n;i++)
    begin
        if ((j=BitReverse(i,nu)) > i)
        begin
            r2 = &xr[j];
            i2 = &xi[j];
            tr = *r2 ; ti = *i2;
            *r2 = *r1; *i2 = *i1;
            *r1 = tr ; *i1 = ti;
        end
        r1++;
        i1++;
    end
end     /* end IFFT */
```

CHAPTER FIVE

Algorithms for the Estimation of Pitch in Polyphonic Music

5.1 Introduction

This chapter describes the algorithms implemented by the author for the separation of simultaneously sounding musical tones. These algorithms estimate instantaneous pitch. Chapter 6 describes the grouping of these estimates in time to determine the musical notes. Chapter 7 compares these algorithms by applying them to musical examples.

Four basic algorithms are considered for the estimation of pitches of simultaneous tones. These are:

(1) Moorer's comb and heterodyne filtering method,

(2) cepstral analysis,

(3) the harmonic frequency ratio algorithm, and

(4) the harmonic summing algorithm.

All these algorithms act in the frequency domain to group the harmonics of the tones. Moorer's method traces the time varying frequency and amplitude of the harmonics, which are then grouped to identify tones. The cepstrum (developed for seismic signal processing), and the harmonic ratio algorithm have limited success at distinguishing polyphonic tones. The fourth algorithm, developed by the author, will be treated here in greater detail. Section 5.6 introduces some original heuristics and a spectral extraction procedure to improve the accuracy of estimation of the above four algorithms.

All these algorithms work well for monophonic music, with most of the incorrect estimates occurring an octave above or below the correct ones. For polyphonic music, the number of harmonically related errors increases dramatically with the number of parts.

Efficiency was not a consideration in the design of these algorithms; the main objective was to discover techniques that worked. The double precision arithmetic (56 bit mantissa) and the spectral resolution (12 Hz per DFT point) were sufficient to prevent round-off errors.

The estimates determined by the algorithms have three attributes: a pitch, a likelihood, and a strength. Pitch is $12\log_2(f/55)$ where f is the frequency; that is, the number of semitones above ***A*** of frequency 55 Hz. The likelihood is a measure of the probability that the pitch of the estimate corresponds to the pitch of a tone in the signal, and one estimate is said to be more likely than another if its likelihood is greater. Strength is a measure of the dB level of the constituent harmonics, and one estimate is said to be stronger than another if its strength is greater.

5.2 James A. Moorer's Method

The method used by J.A. Moorer (1975) finds the period of the signal using an average magnitude difference function (see section 3.2.3). A heterodyne filter is used to extract the harmonic components for this fundamental period. The harmonic components are band-pass filtered and then comb-filtered to give the exact harmonic frequencies. These harmonic components are traced over

time, and a function based on amplitude envelope and frequency variation is used to select the best traces. These are then grouped to determine the estimates.

Spectral leakage of the heterodyne filter (see section 3.3.1) caused many of the spurious traces that Moorer had to remove later. This also masked some low amplitude harmonics. Moorer states that many chords (especially discords) have ambiguous periodicity. That is, there is no frequency that is a common factor of all the harmonic frequencies of the simultaneous tones. Therefore for such chords, the heterodyne filter tuned to any period will fail to detect some harmonics. For this reason the spectral analysis method of section 3.3 is used for the other algorithms described in this chapter.

Moorer applies the following constraints on the music he analyzes:

(a) Only two part music is considered.

(b) The intervals between parts range from a minor third to a minor seventh. This avoids the problem, mentioned earlier, of discriminating simultaneous tones an octave or more apart.

(c) The steady state portion of the tones must be sustained for at least 80 milliseconds.

(d) No trills or vibrato are allowed.

(e) The lowest three partials must be present. Harmonics are rejected as candidates for a fundamental if there is another spectral peak at one half or one third of its frequency. This restricts the range of intervals to less than one octave.

The music analyzed in this thesis has none of these constraints. Note durations as small as 30 milliseconds can be detected, and up to 5 simultaneous tones have been correctly identified (see section 5.5).

5.3 Cepstral Analysis

The spectrum (or log power spectrum) of a harmonically dense tone is a series of equidistant peaks. Applying a Fourier transform to this spectrum should therefore produce a peak corresponding to the frequency difference between peaks. In this way the cepstrum (being the inverse Fourier transform of the log power spectrum) can be used to determine pitch. For simultaneous tones the equidistant spectral peaks of the tones are superimposed, therefore the cepstrum will produce peaks for all the fundamentals present. Many other pitch estimates are also produced which are harmonically related to the tones (especially at octaves). The autocorrelation function can be used to discriminate polyphonic tones, for the same reasons, as it is equivalent to the inverse Fourier transform of the power spectrum. Section 7.2 shows examples of the autocorrelation and cepstral analysis of music signals.

5.4 Frequency Ratios of Harmonics

The frequencies of the partials of a tone are close to multiples of the fundamental frequency, even for tones with vibrato or tremolo (frequency or amplitude variation). When several tones are played simultaneously, the frequency ratios are closer to small integer ratios for harmonics within a tone, than for

harmonics from different tones.

Piszczalski (1979) exploited this to determine the pitch content of monophonic music (see section 3.4.1). They also suggested the feasibility of applying it to polyphonic music, though no successful results were reported. Here follows a description of the method of Piszczalski as implemented by the author.

To begin, the most significant spectral peaks are determined from the spectrum. Rather than use a fixed amplitude threshold to select spectral peaks, the local average for 10 DFT points on either side of peaks is subtracted from the spectrum. This allows low intensity isolated peaks to be detected while rejecting the sidelobes of high intensity peaks.

Then for every combination of two such peaks, common factor frequencies (CFFs) are stored, ordered, and clustered to determine the fundamental frequency estimates. For every pair of spectral peaks with frequencies f_1, f_2, and integers i, j, the tolerance function, defined as $|if_1 - jf_2|/(f_1 + f_2)$ is used to determine whether the peaks belong to the same tone or to different tones. If this tolerance function has a magnitude less than a fixed tolerance, then a CFF is produced with frequency: $f = (f_1/2j + f_2/2i)$.

The frequencies $f/2, f/3$ etc. are also used as CFFs. The initial strength of a CFF is the sum of the decibel levels of the harmonic peaks at f_1 and f_2. This biases the CFFs toward the higher intensity harmonics. The CFFs with frequencies f/n, where n is a positive integer, are weighted by $(0.5)^n$, because the sequence $\{f/n\}$ clusters together as n increases. This decreases the

strength of the lower frequency CFFs to compensate for the larger number of CFFs in the low frequency clustered groups.

After sorting all the CFFs in ascending order, they are grouped in the following way:

- consecutive frequencies differing by more than a fixed tolerance are assigned to different groups;
- the frequency of a group is taken as the average (weighted by CFF strengths) of all the frequencies for that group, and the likelihood of a group is the sum of all the estimate strengths for that group;
- the frequencies of the groups with highest likelihood estimates are rounded to a twelve-tone-per-octave logarithmic pitch and stored.

As an example, the harmonic ratio algorithm is here applied to the same woodwind Trio chord given in figure 2.10 (G minor chord with fundamental pitches of 22, 41, and 49). The frequencies and log amplitudes of the 13 strongest spectral peaks in the range, 50 to 5,000 Hz are:

	Frequency(Hz)	Log Amplitude (dB)
1:	901	73
2:	1152	70
3:	1370	40
4:	1531	37
5:	1730	55
6:	1795	58
7:	2053	38
8:	2300	38
10:	2882	49
11:	3457	43
12:	3592	41
13:	3842	25

The second and fifth spectral peaks are due to the second and third harmonics of the tone *D* with a fundamental frequency of 576 Hz. If $f_1 = 1152, f_2 = 1730, i = 3$ and $j = 2$, then the tolerance function, $|if_1 - jf_2|/(f_1 + f_2)$ equals 0.0014, (compared to 1) which is small enough to accept f_1 and f_2 as the frequencies of the second and third harmonics of a tone. This tolerance function typically ranges from 0 to 10. The strength of this CFF is 125 (the sum of the dB levels: 70 and 55) and the frequency is 576 Hz (the average of f_1/j and f_2/i).

This frequency comparison is applied to every pair of spectral peaks and the resulting CFFs are clustered to obtain the following list of fundamental estimates:

tone	pitch	likelihood
C	15	194
G	22	434
B♭	25	113
C	27	189
D	29	353
D♯	30	150
G	34	655
B♭	37	307
D	41	801
G	46	203
B♭	49	639

Three of the four most likely estimates of the Trio chord are correct (pitches 22, 41, 49), but the incorrect estimate (pitch 34), an octave above the correct estimate (pitch 22), has a higher likelihood. The use of heuristic weighting and spectral extraction overcome this (see 5.6).

5.5 Harmonic Summing Algorithm

The harmonic summing algorithm evaluates the function:

$$H_h(f) = \sum_{i=0}^{h} 10 \log_{10} S(if).$$

where $S(f)$ is the instantaneous power spectrum, f the frequency and h the number of harmonics being considered (h is assumed constant). Most tones are harmonically dense, some having strong formants, or low amplitude fundamentals. This function enhances the detection of the fundamental frequency of such tones.

The highest peaks are quadratically interpolated to determine the frequencies and hence the pitches of the estimates. The strength of an estimate is the average decibel level of the harmonic peaks. That is,

$$H_h(f)/h,$$

where f is the fundamental frequency of the estimate. Simultaneous tones differing in intensity by as much as 30 dB can still be distinguished from each other.

This function H can correctly identify simultaneous tones even when only a few harmonics are present. The most spectacular result is the correct identification of a 5 tone pianoforte chord consisting of $B\flat$, $B\flat$, F, $B\flat$ and D in ascending order, which is the last chord of figure 7.32. Taking the number of harmonics h as 8, the 5 estimates of greatest strength corresponded exactly to the tones that are present. It is, however, not obvious from the pitch estimates

whether 4, 5, or 6 notes are being played (see figure 5.1). The number of simultaneous tones varies even for music with a fixed number of parts, because some parts can rest while others sound. Therefore the number of simultaneous tones cannot be assumed as a fixed parameter. A more advanced method such as that in 5.6 needs to be applied.

Most erroneous pitch estimates are harmonically related to the tones being played. For example, if 2 tones with fundamental frequencies 200 Hz and 300 Hz are played simultaneously with harmonic frequencies 200, 400, 600, ... and 300, 600, 900, ..., then the combined spectrum appears to contain a tone with fundamental frequency 100 Hz, but missing the fundamental, 5th, 7th, and 11th harmonics etc. Also the incorrect estimate with harmonic frequencies 600, 1200, 1800 .. is reinforced by both tones. When three or more tones are playing, the number of harmonically related errors increases dramatically. Incorrect estimates at an octave above and below the played tones are the major problem.

Moorer avoided the problem of harmonically related errors by requiring that the fundamental of each tone be present. This criteria is not adopted here because of possible interference with low frequency peaks such as those caused by wow or flutter. Besides, the human ear can detect tones with lower partials absent (Schouten et al. 1962). Furthermore, many of the low pitched tones encountered in this research have strongly attenuated lower partials.

These harmonically related erroneous pitch estimates are generally not as strong as the estimates for the correct tones. A notable exception is the instance of an inverted major or dominant seventh chord. Consider the chord

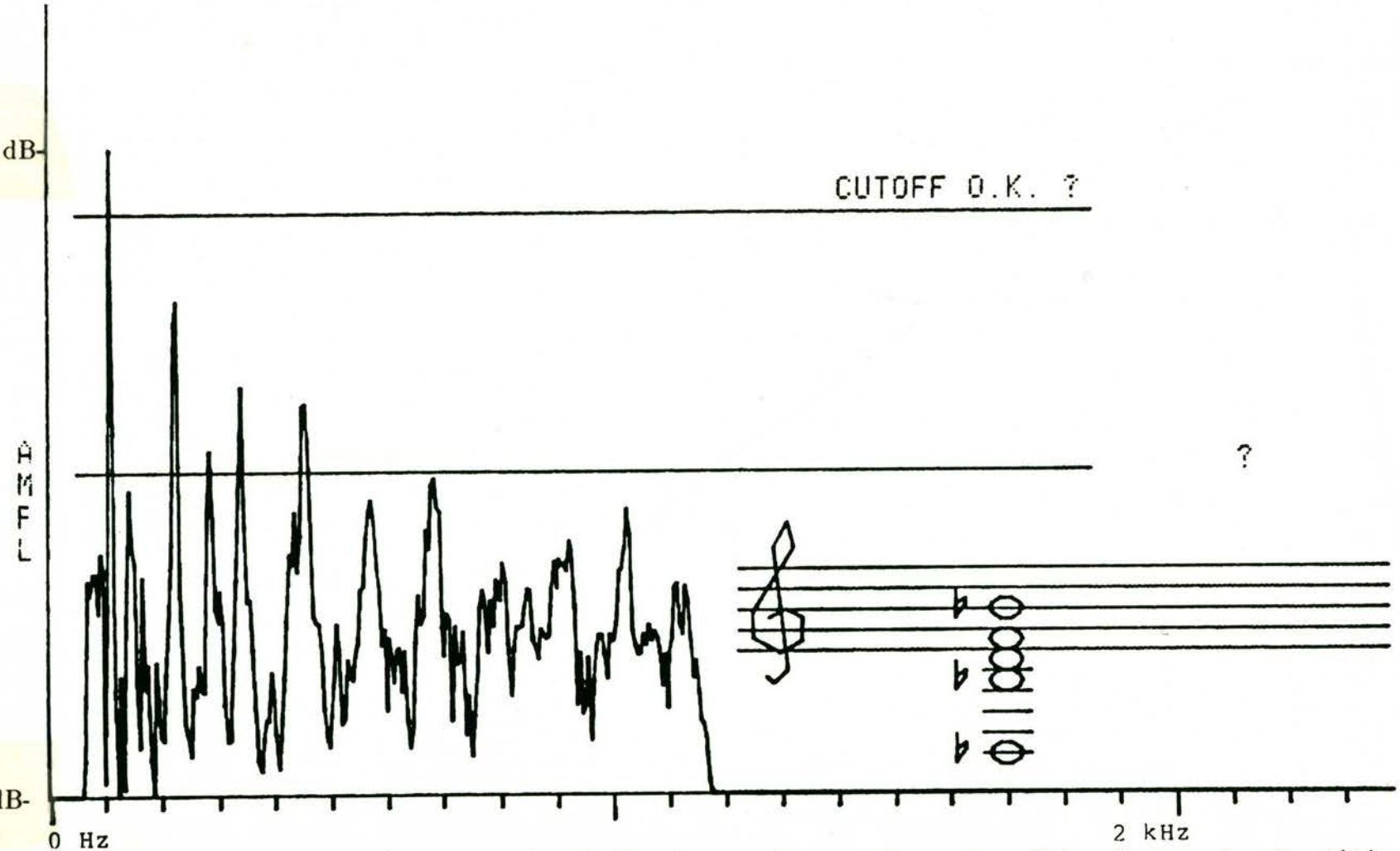


Figure 5.1 presents the out-put of the harmonic summing algorithm for a chord with five notes. The strongest pitch estimates corresponded exactly to the notes that were present. (B-flat, B-flat, F, B-flat & D)

with fundamentals 300, 400, 700, and 1000 Hz.

An estimate with fundamental frequency 100 Hz is missing the 1st, 2nd, 5th and 11th harmonics, and an estimate at 200 Hz is only missing the fundamental, 11th, 13th etc., and has harmonics 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000, 2400 i.e. all harmonics from 2nd to 10th. Therefore the estimate at 200 Hz appears a likely candidate when in fact it is absent.

Figure 5.2 compares the spectra of this second inversion chord transposed up a minor tenth (15 semitones). The chord is from the Menuet by J.S. Bach, and played on pianoforte. The four correct estimates (*D* - pitch 41, *G* - pitch 35, *B♭* - pitch 25, and *F* - pitch 20) and the incorrect estimate (*B♭* - pitch 13) are shown above the spectrum. The dots on the estimates give the point where the exact harmonic should be, while the vertical lines show the nearest spectral peak, corresponding to the respective harmonic. Notice that the 2nd to 10th partials of the incorrect estimate (pitch 13) all correspond to spectral peaks.

5.6 Spectral Extraction and Pitch Determining Heuristics

The algorithms described in sections 5.2 to 5.5 can be improved by two methods. One method is to apply heuristics to discriminate between correct and harmonically related incorrect estimates. The other is to iteratively calculate the likelihoods of estimates and attenuate the harmonics of the most likely estimates. This second method can determine the number of tones being played, because after all the harmonics of the correct estimates have been extracted, the

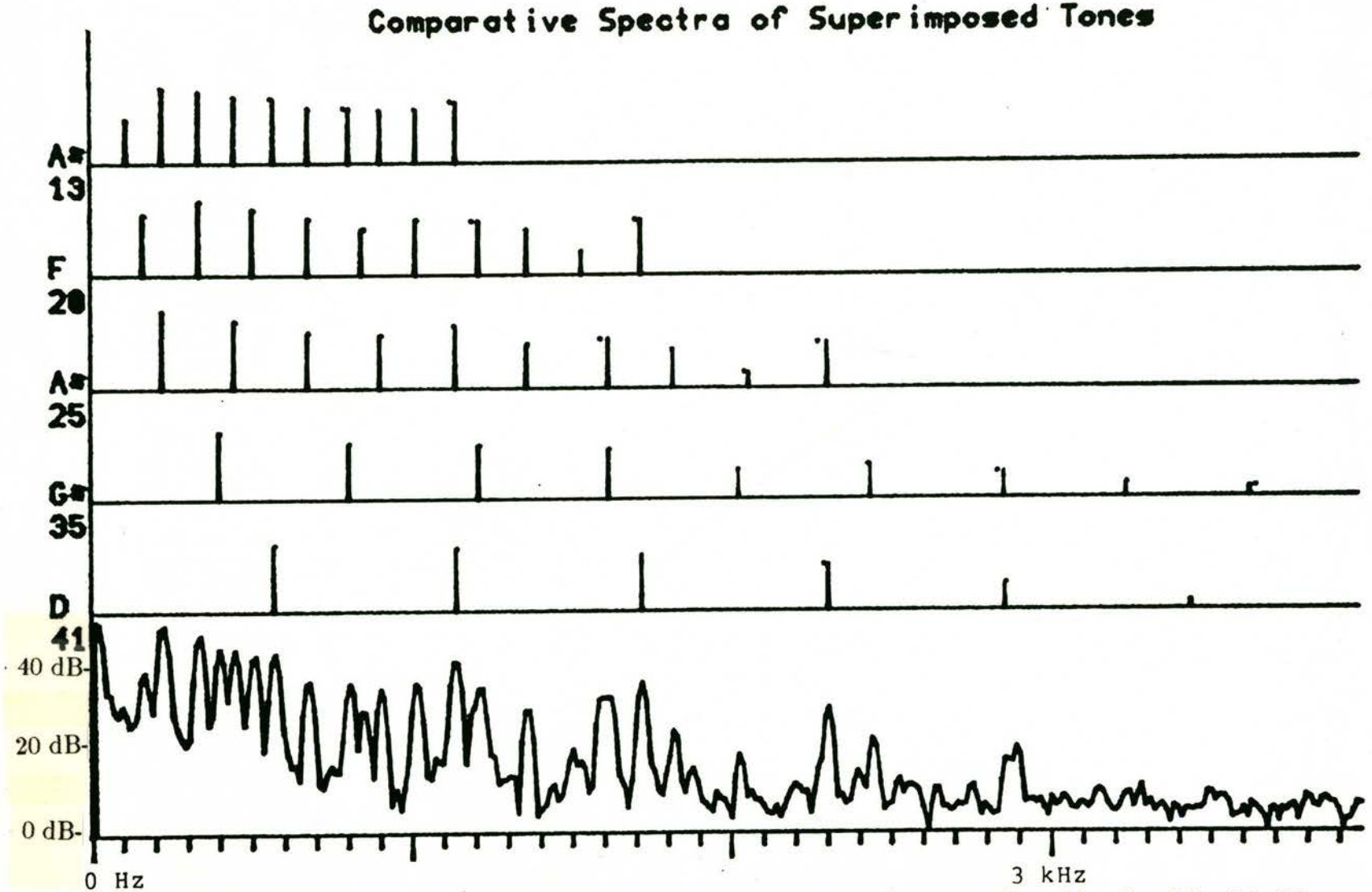


Figure 5.2 compares the spectra of four correct estimates (D p41, G p35, A#p25 and F p20) and an incorrect estimate (A#p13) above the original spectrum from the Menuet.

residual spectrum provides no further estimates of significant strength. The best results reported in this thesis use both these methods applied with the harmonic summing algorithm of section 5.5.

A set of weightings for the heuristic functions was found empirically to work well on several polyphonic pieces. These weightings were chosen and are held constant throughout the analysis of a piece of music. The weightings and error rates are considered in chapters 7 and 8.

5.6.1 Heuristics For Finding the Best Estimate

These heuristics help to distinguish between correct estimates and harmonically related erroneous ones.

The five heuristic functions are:

H1:

odds minus evens :- If the sum of the log amplitudes of the even harmonics exceeds that of the odd harmonics, a fraction of this difference is subtracted from the likelihood of the estimate.

The reason for this is, if the even harmonics are significantly stronger than the odd harmonics, then it is likely that there is a tone an octave above the current estimate but not at the current estimate.

H2:

octave, twelfth, 2 octaves below :- If there is an estimate with non-zero likelihood an octave, a twelfth or two octaves below another estimate, a fraction

of that likelihood is subtracted from the likelihood of the higher pitch estimate.

This compensates for the harmonic summing algorithm producing peaks at multiples of the fundamental frequency of the correct estimates. For a single tone, the sum of the even harmonics of a tone will produce a strong estimate as well as the sum of all the harmonics. Two tones at an octave apart can only be detected if the estimate of higher pitch is significantly stronger than the estimate of lower pitch. It was not necessary to consider estimates on the harmonic series more than two octaves away, because their contribution to the higher pitched estimates was insignificant.

H3:

centre of gravity :- A fraction of the difference between the centre of gravity (amplitude weighted mean frequency) of the harmonic peaks and half the number of harmonics considered (median frequency) is added to the likelihood of the estimate.

This biases the estimation in favour of the tones of higher pitch, where there is less interference from the harmonics of the other tones that are present. This also helps to distinguish between correct and incorrect estimates. The latter rarely have monotonically decreasing harmonic amplitudes.

H4:

harmonic amplitude increments :- A fraction of the sum of all positive steps in log amplitude from one harmonic to the next is subtracted from the

likelihood.

This heuristic function is zero if the harmonic amplitudes are monotonically decreasing, and positive otherwise. An incorrect estimate would result from the harmonics of harmonically related tones, but with some harmonics missing, and should therefore have a series of amplitudes that is not monotonically decreasing. This therefore biases the estimation in favour of the correct estimate, and also the higher pitch estimates.

For most tones, the amplitude decreases as the harmonic number increases. Three exceptions are:

(a) clarinet, or stopped organ pipes where the even harmonics are absent,

(b) low pitched sounds (piano, bassoon, etc.) where fundamental and lower harmonics are attenuated and,

(c) stringed or woodwind instruments where formants or resonances cause variations in spectra. The music considered in this thesis includes piano and bassoon tones, testing the generality of the algorithms.

H5:

variance of harmonic frequencies :- A fraction of the variance (mean squared difference) of the frequencies of partials from the corresponding ideal harmonic frequencies is subtracted from the likelihood. The ideal harmonic frequencies refer to exact multiples of the fundamental frequency.

The frequencies of partials of a tone have near integral ratios. An incorrect estimate contains several harmonics from harmonically related tones,

which would lack the synchrony of the harmonics of a single tone, and should therefore have a larger variance. Most pipe or stringed instruments exhibit phase locked partials, and the relative phases of the harmonics in the steady-state is stable even in the presence of amplitude or frequency variation (Beauchamp 1974, Fletcher 1978). This is not true of multiple stringed instruments like the piano.

Table 5.1 shows the differences between partial frequencies and the corresponding harmonic frequencies, for some estimates from the woodwind Trio chord shown in figure 2.10. The partial frequencies are found by climbing the spectral peak from the harmonic frequency to the maximum point. The right hand column (sigma) gives the standard deviation of the differences. The three correct estimates; *G* of pitch 22, *D* of pitch 41, and *B♭* of pitch 49, have the smallest standard deviations.

Table 5.1

Difference Between Harmonic and Partial Frequencies

pitch	1	2	3	4	5	6	7	8	sigma
G 10	4.50	0.55	0.00	0.64	-9.32	0.62	0.00	0.40	4.65
G 22	-1.50	-1.33	-1.35	-1.81	0.15	0.51	4.21	-0.14	1.97
D 29	17.14	-4.23	10.97	-1.43	-0.97	-1.14	3.83	-1.94	7.99
G 34	-3.27	-4.23	0.41	-0.82	5.66	-0.27	0.42	2.20	3.08
B♭ 37	-23.85	1.14	7.19	-0.73	-2.79	-1.44	-1.52	-0.53	9.46
D 41	-5.06	0.53	1.11	-0.48	0.77	0.53	-0.22	0.45	2.01
G 46	-10.10	-3.27	-2.17	2.76	-0.60	-0.93	1.74	4.41	4.59
B♭ 49	3.67	-0.07	-1.49	0.34	1.59	1.47	-2.27	0.83	1.93

Taking a weighted sum of these heuristic functions is a simplistic method

for deriving the likelihoods of estimates. The five heuristics used here are by no means the only heuristics that could be applied. However, the success of this method justifies the simplification. These heuristics could also be applied to other estimation techniques, such as autocorrelation.

The heuristics in this section will henceforth be referred to as H1, H2, H3, H4, and H5.

5.6.2 Iterative Extraction of Tones From Spectra

Each estimate has a measure of likelihood, based on the strength of the estimate, and a linear combination of 5 weighting functions. These weighting functions help to differentiate between correct and incorrect estimates.

The algorithm is as follows:

```
for every new time
begin
        determine the log spectrum ;
        compute the harmonic sum ;
        while there remains a significant estimate
        begin
                calculate the heuristics and likelihoods ;
                with the most likely estimate
                do
                        record strength and pitch of the estimate ;
                        attenuate the harmonics in the log spectrum;
                end ;
                compute the harmonic sum of the residual spectrum;
        end ;
end ;
```

An estimate with strength greater than 40 dB is considered significant. This was found to work well for the music considered here, because of the small dynamic range. To generalize this for music with a wide dynamic range, a better criteria would be to consider only estimates within 30 dB of the strongest estimate at that time. This is comparable to the human auditory masking of simultaneous tones. All points in a spectral peak are attenuated equally to give

an unaltered frequency when quadratically interpolated on the next iteration. No peak is attenuated below the ambient noise level.

Choosing the amount of attenuation to be applied to the harmonics of the most likely estimate requires a compromise. If the attenuation is too small, peaks that are shared by more than one tone are not changed to the extent of distorting the spectra of the other tones, but more iterations are required before a new pitch is found. Processing time is reduced if a higher level of attenuation is used, but incorrect estimates are selected more often.

The example of figure 2.10 is used here to illustrate the extraction procedure. This figure gives the result of the harmonic summing algorithm applied to a woodwind Trio chord. The bassoon is playing *G* 22 (the tone *G* with pitch 22), and the oboes are playing *D* 41 and *B♭* 49. The incorrect estimates *D* 29 and *B♭* 37 (an octave below the correct tones) can be eliminated by applying H1 (see 5.6.1), because their even harmonics have greater intensity than their odd harmonics. These incorrect estimates are due to the presence of the tones an octave above. The estimate *G* 34, an octave above the correct tone *G* 22, cannot be eliminated by comparison of likelihoods. This estimate is stronger than would be expected from the bassoon alone, because the 2nd, 4th and 6th harmonics of the *D* 41 also contribute.

Tables 5.2a to 5.2e give pitch estimates for the example of figure 2.10 for each iteration of the extraction procedure.

The pitch, fundamental frequency, dB level of the first 8 harmonics, strength and likelihood of each estimate are given. The strength is the result

Table 5.2
Harmonic levels before extraction of most likely estimate (denoted *)
Table 5.2a - first iteration

Pitch	Freq(Hz)	1	2	3	4	5	6	7	8	strength	Likelihood
G 22	191	56	64	71	51	30	70	30	37	51	35
D 29	289	21	71	43	69	27	55	21	27	41	0
G 34	383	64	51	70	37	15	37	57	7	42	21
B♭ 37	449	41	73	30	58	20	54	13	40	41	0
D 41	575	71	70	52	37	49	43	25	25	46	* 93
G 46	769	51	37	37	18	25	26	20	8	27	20
B♭ 49	897	73	58	54	41	25	20	19	19	38	74

Table 5.2b - second iteration

Pitch	Freq(Hz)	1	2	3	4	5	6	7	8	strength	Likelihood
G 22	191	56	64	59	51	30	58	30	37	48	36
D 29	289	21	59	43	58	27	45	21	22	37	0
G 34	383	64	51	58	37	15	31	57	7	40	21
B♭ 37	449	41	73	30	58	20	54	13	40	41	0
C 39	513	25	38	37	38	24	21	40	17	30	16
D 41	575	59	58	45	31	41	36	21	21	39	53
G 46	769	51	37	31	18	25	21	20	8	26	13
B♭ 49	897	73	58	54	41	25	20	19	19	38	* 78

Table 5.2c - third iteration

Pitch	Freq(Hz)	1	2	3	4	5	6	7	8	strength	Likelihood
G 22	191	56	64	59	51	30	58	30	37	48	41
D 29	289	21	59	41	58	29	45	21	22	37	0
G 34	383	64	51	58	37	15	31	47	18	40	22
B♭ 37	450	41	60	33	46	14	45	13	34	35	0
C 39	514	25	38	37	36	24	21	34	17	29	21
D 41	575	59	58	45	31	41	36	21	21	40	* 54
G 46	769	51	37	31	18	25	21	16	8	25	7
B♭ 49	897	60	48	45	34	21	16	16	16	32	40

Table 5.2d - fourth iteration

Pitch	Freq(Hz)	1	2	3	4	5	6	7	8	strength	Likelihood
G 22	191	56	64	44	51	30	42	30	37	44	* 39
D 29	288	21	44	41	42	29	38	21	21	32	0
G 34	383	64	51	42	37	15	31	47	18	38	21
B♭ 37	449	41	50	30	40	20	38	13	29	32	0
C 39	514	25	38	37	36	24	21	29	17	28	22
D 41	570	44	42	38	21	24	25	21	20	29	33
G 46	769	42	37	21	18	21	20	14	8	22	6
B♭ 49	898	50	40	38	29	21	14	15	14	27	18

Table 5.2e - fifth iteration

Pitch	Freq(Hz)	1	2	3	4	5	6	7	8	strength	Likelihood
G 22	190	47	53	41	46	30	40	26	31	39	20
C 27	256	29	25	42	38	23	32	40	38	33	4
D 29	288	21	41	41	40	29	38	21	21	31	0
F 32	341	32	15	38	34	36	38	18	41	31	6
G 34	380	53	46	40	31	18	20	35	17	32	15
B♭ 37	449	41	50	30	40	20	38	13	29	32	0
C 39	514	25	38	32	36	24	21	29	17	27	24
D 41	570	41	40	36	20	24	25	21	20	28	* 33
B♭ 49	898	50	40	38	29	21	14	15	14	27	18
B♭ 49	908	50	24	42	31	22	21	18	20	28	21

of the harmonic summing algorithm, and the likelihood is the strength plus the weighted sum of the heuristics described in section 5.6.1. The heuristic weightings are 0.600, 0.500, -1.000, 1.000, 0.100 for H1, H2, H3, H4, H5 respectively.

At each iteration the harmonics of the most likely estimate (indicated by * in Table 5.2) are attenuated by 15 dB. The likelihoods are then re-evaluated. When an estimate is extracted it is selected as one of the tones present. The iterative extraction is repeated until another estimate is found (*C* 39), but its strength (26 dB) is too small to support the hypothesis of a fourth tone with this pitch.

Although the incorrect estimate, *D* 34, is significant in the first iteration (Table 5.2a), it is not selected by the iterative extraction procedure. The extraction procedure is one of the major discoveries of this work, providing a marked improvement in the determination of simultaneous tones.

5.6.3 Coincidence of Harmonics

A major problem is the coincidence of the harmonics of different tones. For example, for two tones at a musical fifth apart, the second harmonic of the upper tone is coincident with the third harmonic of the lower tone. Spectral peaks are broadened because of the finite sampling time. The effective width (6dB attenuation bandwidth) of peaks is typically 30 Hz, and is independent of frequency. If the frequencies or amplitudes vary within the time window, the peak width is even greater.

If two overlapping peaks differ in phase by 180 degrees, they cancel.

This is, however, a rare event. Assume that the phase of the harmonics, and their log amplitudes, are random variables with constant probability density functions. In that case, for two independent overlapping peaks to cancel each other to the extent of leaving a hole at least 20 dB below the highest peak, their phases must differ by between 174 to 186 degrees, and their amplitudes must differ by less than 1 dB. The likelihood of these peaks completely cancelling each other is less than 0.001, assuming a 30 dB dynamic range.

Overlapping harmonics that do not cancel still cause problems. The larger peaks distort the spectrum of a tone with lesser peaks. Superimposed peaks can combine to form a single peak that is removed as much as 30 Hz from the ideal harmonic frequencies. This makes the variance heuristic too large and can therefore reject the correct estimate. When attenuating peaks, if one side of the peak has a phase different from the other side, then it is assumed that two nearby peaks are present, and only half of the peak is attenuated on the side nearest the ideal harmonic frequency. This method was applied to the spectral attenuation, but the resulting improvement was insignificant.

5.7 Deconvolution of Reverberation

Reverberation can be a major problem in the analysis of musical tones. Note durations are prolonged by reverberation, causing overlap. The effect is more serious with short tones. The resonant characteristics of the instruments and the recording room also distort the spectra. Stockham et al. (1975), and Schafer (1969) showed that recorded signals can be modelled as the convolution of the

source signal and the impulse response of the room or instrument reverberation. This means that in the frequency domain the reverberation frequency response is multiplied with the spectrum of the source signal. In the log spectrum the signal and reverberation are additive, and if enough is known about them they can be separated (deconvolved). The room and recording reverberation is nearly constant throughout a piece of music, but the resonances of different instruments are only present when those instruments are playing.

In this work an adaptive deconvolution method is applied to spectra before pitch extraction. A decaying average of previous spectra is used to attenuate the current spectrum. The parameter controlling this is the half life of the decaying average of previous spectra used to attenuate the current spectrum. Unlike Stockham's work, where the recording reverberation is assumed to be constant, this method can adapt to the changing reverberation of the instruments that are present. An interesting bonus of this method is that for the bassoon, formants are attenuated over a series of tones, enhancing pitch detection. For the piano, tones sustained by reverberation are attenuated, which helps to differentiate tones of shorter duration.

5.8 Conclusion

Several low-level techniques applied to spectra are considered for discriminating simultaneous tones. Some heuristic functions are presented to improve the discrimination, and a spectral extraction procedure is described which iteratively attenuates the spectral harmonics of the most likely estimates, thereby avoiding many harmonically related errors.

CHAPTER SIX

Algorithms for the Analysis and Plotting of Music

6.1 Introduction

This chapter describes the algorithms for grouping pitch estimates (resulting from the algorithms of chapter 5), determining the notes, and plotting the music. Music can be displayed on the system either as a pitch profile (section 6.2), or as standard music notation.

The automatic plotting of music notation is a non-trivial problem. Details of other systems for music printing can be found in Kassler (1977), Smith (1973), Tucker (1977), Boker-Heil (1972), and Byrd (1974). All these systems are aids to musical typesetting and require substantial human intervention to determine the musical symbols and to position them. The music plotting software described here was developed by the author to provide a rapidly produced, and easily read representation of musical data. Music notation is plotted automatically from the pitch estimate data. It was not intended to produce high quality output suitable for printing. The work in this thesis considers the issues of automatic determination of key and tempo, and the horizontal and vertical positioning of notes, bar-lines, and accidentals, but further research must be done before fully automated music printing can be realized.

The music shown in the figures of this thesis is produced automatically from analysed sound and is not post-edited. Plotting is controlled by a set of run-time parameters.

6.2 The Pitch Profile

The pitch profile consists of notes plotted on horizontal lines. These lines represent equally tempered semitones spanning five octaves, and centred at $E\flat$ above middle C. This corresponds to a fundamental frequency range of 60Hz to 1.7kHz. The vertical width of the line is a measure of the strength or loudness of the note, and the horizontal position and length give the time of occurrence and duration respectively of the note. The pitch profile is the most direct representation of pitch estimates.

6.3 An Overview of the Music Notation Plotting Programs

Plotting standard music notation is more involved. Pitch estimates must first be grouped into notes as described in section 6.5. The times and durations of these notes must then be fitted to the correct musical durations (section 6.6). The key (tonality) must also be determined for plotting accidentals and key signatures.

Two programs are used. One is tailored to music in which a fixed number of instruments are either playing or resting, e.g. the *Brandenburg* woodwind Trio. This is called the fixed part analysis. The other more general program is for music where the number of parts varies or is not known in advance, (e.g. for piano). Here the plotting of too many rests would be a hindrance, not a help, in reading the music. This is called the general part analysis.

Optional run-time parameters include:

- the starting time and finishing time of the music to be plotted,

- a scale factor for the plotting dimensions,
- the minimum duration of notes to be plotted,
- the minimum strength of pitch estimates that are considered,
- the tempo,
- the maximum number of parts,
- and a flag for producing detailed information about the analysis.

Music can be entered into these plotting programs via an electronic organ keyboard, and data can be automatically played on the organ. Input from the organ already has the information on the starting and finishing times of the tones. The tone starts when a key on the keyboard is depressed and ends when it is released. Music entered via the organ does not require the grouping procedures described in section 6.5.

The tempo (or number of crotchets per minute) is determined (see section 6.6) and the times and durations of all notes (in milliseconds) are scaled to the equivalent musical durations. Finally, the notes are allocated in order of pitch to the different parts (bass, tenor, alto, or soprano), and the music is plotted.

Musical data are represented internally as a structure called an Event, which has three attributes: the Time at which the Event begins, the Duration of the Event and the Pitch (being the number of semitones above the note ***A*** with fundamental frequency 55 Hz). An Event with a negative Duration represents the end of an Event stream, while a negative Pitch denotes a rest with an associated Time and Duration.

6.4 Deleted

6.5 Grouping the Pitch Estimates

Pitch estimates have three attributes: a pitch, a time of occurrence and a strength (or measure of loudness).

Pitch estimates are grouped together in time by leaky-bucket integrators; one for each pitch. For each point in time and for every estimate at that time, the level in the bucket is increased, if the strength of the estimate is large enough, otherwise it is decreased. When a bucket is full it overflows and can be filled no more. The onset or beginning of a tone is the time at which an empty bucket begins to refill. The finish time of a tone is the time at which a bucket is completely emptied, minus the time taken to empty a full bucket. The duration of a note is the finish time minus the onset time.

Sporadic pitch estimates will only partly fill the bucket, which will soon empty, so the corresponding duration will be insignificant. A strong tone may be masked for a short time by the onset or vibrato of another tone, or by noise. During this time the bucket will partially empty but will be replenished when the masking ceases. This helps to correctly identify the single note, instead of two consecutive notes.

The time for the bucket to freely empty, is typically 40 to 100 milliseconds. The minimum strength is set at 40 dB below the maximum recorded amplitude to reject any estimates caused by noise or signal distortion.

An alternative approach is to set the onset time to the time when the bucket is first full, and the finish time to when it completely empties. This means that the note duration is smaller, if the bucket level rises slowly. This approach

is adopted when the number of musical parts is known (fixed part analysis). On filling a bucket, a new note is opened (i.e. created) for that part, and the previous note closed and recorded, provided its duration is large enough. If the bucket of the current note for any part empties, then that note is closed and a rest is opened. The part to which a note belongs is determined from the pitch ordering of all the notes that are currently sounding. The note with the highest pitch is assigned to the highest part. If some of the other parts are resting, then the note is assigned to the part which previously had the pitch nearest to the pitch of the current note. Rests are treated in the same way as notes. They require the same minimum duration, and are sorted and aligned with the other notes.

This process of grouping the pitch estimates into notes is similar to low-pass filtering the output for each pitch.

6.5.1 Detection of Rapidly Changing Note Sequences

For the music considered in this thesis, most passages of rapidly changing notes proceed by small musical intervals; generally a single step of one or two semitones. To incorporate this, and to reject sporadic estimates far removed from the expected path of the various parts, the following heuristics are included. Where the number of parts is known (fixed part analysis), the minimum duration required to accept the previous note is correspondingly increased, if the current candidate for a note is more than 2 semitones from the previous pitch for that part. For example, in the Trio, if the oboes moved by 1,2,3,4,5 or

more semitones, then the preceding note had to be at least 60,60,100,140 or 180 milliseconds in duration respectively to be accepted. This allows the detailed explication of trills and other ornaments without causing numerous sporadic errors at other times in the music.

Where the number of parts is not known (general part analysis), notes that are 1 or 2 semitones removed from the strongest estimate at any time are forcibly closed at the time of filling of the bucket of the strongest estimate. This masks any weaker estimates within two semitones of this strongest estimate. For example, in the Piano Partita (Figure 7.24), a common mordent is *B♭*, *C*, *B♭*, *A*, then *B♭*. The three *B♭* notes appear to merge into one, because each *B♭* is sustained until the hammer re-hits the string. Therefore without this masking, only a single *B♭* tone would be detected instead of three with the *C* and *A* interleaved.

6.6 Determination and Scaling of Tempo

Tempo (the number of crotchets per minute) can be given as a parameter to the plotting program, or determined automatically by one of two methods.

The first method used here is to sample the local amplitude at regular time intervals (typically 10 msec to 100 msec), and compute the spectrum of this set of values. Local amplitude refers to the average amplitude over a small interval (typically 1 millisecond). The largest peak in the range 0.3 to 2 Hz is taken as the tempo. Then by looking at the spectrum at one half, one third, and

one quarter of this frequency, the beat (2/4, 3/4 or 4/4 time) can be determined. This method gave a tempo of 119 crotchets per minute, and a beat of 3/4 time for the Trio example in section 7.4.4.

The second method, developed by Harris (1982), finds the most frequent inter-onset time. The histogram of the difference in starting time of the notes is maximal for the tempo (see table 7.1).

Once the tempo is determined, the times and durations of the notes are scaled to match the internal representation of duration. The beat is used for placement of bar lines and time signatures. This scaling works well for short segments of music without tempo variation or pauses.

Internally the duration of a crotchet is represented by 24, so that compound time or triplets can be accommodated, and notes as brief as a demi-semiquaver (an eighth of a crotchet) can be represented.

6.7 Music Analysis Procedures

These analysis procedures determine the key signature for the plotting and harmonic analysis of the music. They assume major or harmonic minor tonality.

6.7.1 Determination of Key

Key (tonality) is determined by minimizing the number of chromatic notes for all the harmonic minor and major keys. This is done at the beginning of a piece of music to determine the initial key signature, and whenever a note

is encountered that is not in the current key, to determine if a modulation (key change) has occurred, or just a transitory chromatic. The key is determined locally; that is, only the next 20 notes are used to determine the key.

6.7.2 Harmonic Analysis

These programs attempt to determine the basic harmony (or the chords) of the piece of music. The chord at a given time is determined by considering the intervals between all the notes sounding at that time that are in the current key. If there exists an interval of a third within the chord, then a third below the lower note is found repeatedly until the root of the chord is determined. The root of a chord has no third below it. Intervals are determined by the number of steps in the scale of the current key. If there is a note at an interval of a seventh or ninth above this root, a flag is set to show that the chord is a seventh or ninth chord respectively.

Applying this method to the *C* 6th chord with notes *C, E, G, A,* would be give *A* minor 7th (*Am7*). To resolve this musical context must be taken into account. Also the chord with notes *G, B, F, A* could be a *G* ninth or an *F* eleventh chord, depending on which third is found first. The analysis is relative to the current key. Unless a modulation (change of key) is detected, thirds are ignored if either note is not in the current key. Only chords with the root and third in the key are detected.

6.8 Music Plotting

Music is plotted as a series of straight lines on a graphics terminal or a plotter. Figure 6.1 is an example of an expanded plot of some music showing how the treble clef, notes, and accidentals are constructed.

6.8.1 Horizontal Positioning of Notes in a Bar

The simplest form of horizontal positioning of notes is to separate them in direct proportion to their duration. The problem with this is that the short notes become too crowded.

The method adopted here is to position each note at a horizontal distance proportional to a constant plus the duration of the previous note. This constant is equivalent to a crotchet, so that a crotchet takes 2 units of space while a quaver takes 1.5 units of space. Notes starting at the same time are plotted in the same vertical line. Notes with differing onset times are plotted in the order of their onset times. Consequently, if a crotchet is played simultaneously with two quavers, the crotchet is spaced with the two quavers and take up the same room of 3 units. If the crotchet is played on its own without simultaneous notes of shorter duration, it takes only 2 units. If an accidental is required, all the notes at that time are moved right to make room for the accidental. Therefore synchrony of simultaneous notes is maintained while preventing crowding of passages with accidentals and notes of short duration.

Bar lines are automatically inserted. If the starting time of the note to be plotted is greater than the product of meter and the number of bars already plotted, a new bar is placed before plotting the note.

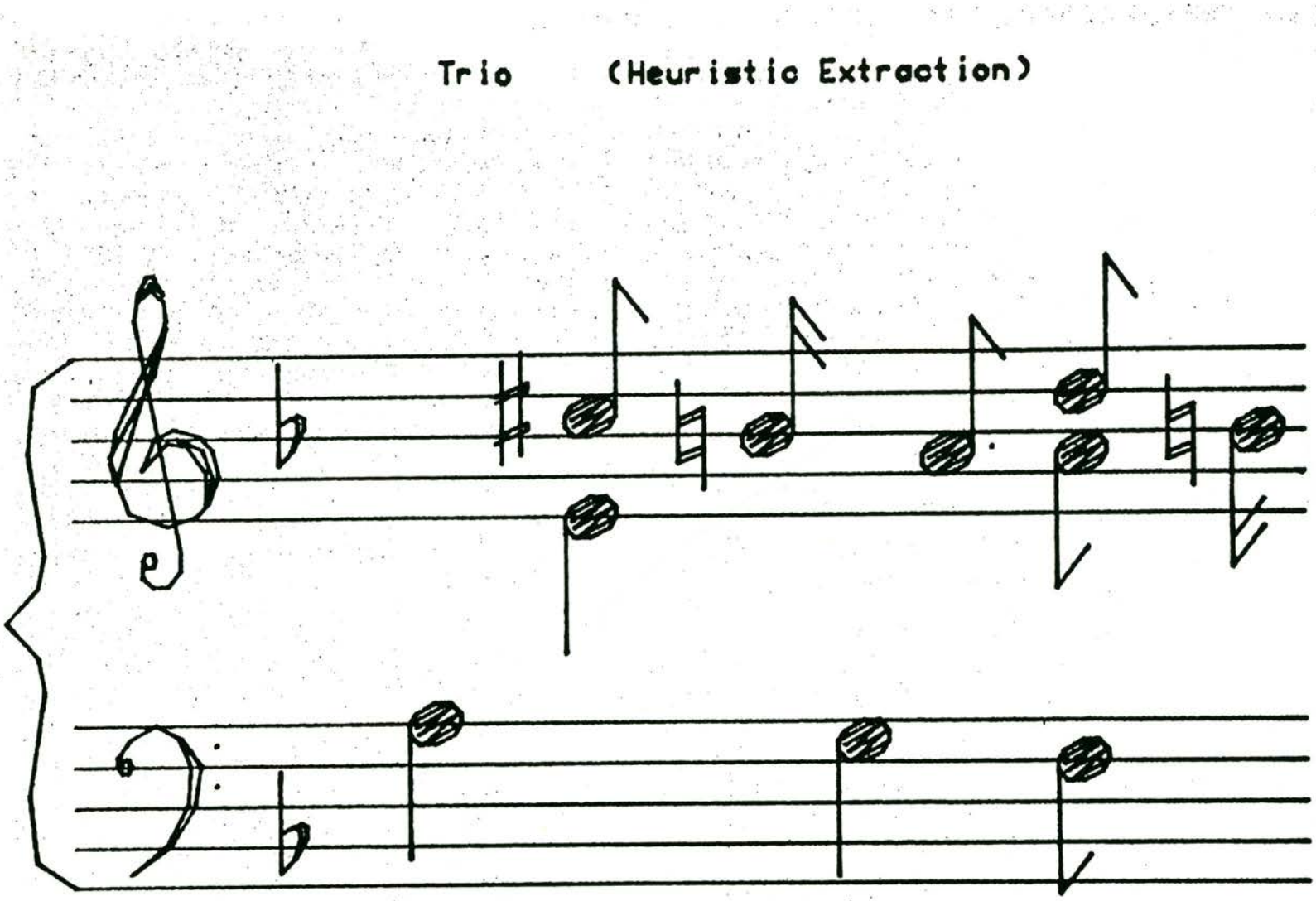


Figure 6.1 is an example of an expanded plot of some music showing how the treble clef, notes and accidentals are constructed.

One problem with allocating musical times by dividing by a constant, is notes that are nearly simultaneous may have different onset times, and therefore may be plotted separately. For example with a crotchet duration of 500 milliseconds and 2/4 time, two notes with onset times 980, and 1010 milliseconds would be plotted on opposite sides of the first bar-line. Whether two notes should be considered simultaneous is determined by the variable, *overlap*, which is the amount of time both notes are playing minus twice the difference in onset times.

Table 6.1 gives three examples of the *overlap* of pairs of notes. The times are given in milliseconds.

Table 6.1

Example	note A			note B			*overlap*	
	onset	finish	duration	onset	finish	duration		
1	0	100	100	20	120	100	40	yes
2	0	100	100	60	120	60	-80	no
3	0	500	500	100	500	400	200	yes

If *overlap* is sufficiently large, as in Example 1 and Example 3 (table 6.1), and their onset times do not differ by more than 160 msec., then the onset times are given the same value and the notes are then plotted on the same vertical line. This gives a significant improvement to the vertical alignment of simultaneous notes.

6.8.2 Determination of Accidentals

Rests are represented as negative pitch values. For notes with positive pitch, the vertical plotting position is determined from the pitch and current

key. There is an ambiguity here; for example $G\sharp$ in the key of A would be $A\flat$ in the key of $E\flat$, and would therefore require a different vertical position depending on the key. The following procedure determines which accidental is required if Pitch is not in the current key.

(a) If Pitch of the note is not in the current relative major key and the note is the seventh or leading note, then the key is minor and the note is plotted as a sharp or natural, not a flat. For example, to plot $B\natural$ in the key of C minor (relative minor of $E\flat$), $B\natural$ is the seventh note of the scale, and $E\flat$ is not in the key of F major; therefore a natural is plotted. If the key had been D minor and the leading note $C\sharp$, the leading note would be written as a sharp and not as a $D\flat$, because the note F is in the key of F major.

(b) If the note is not the leading note of a minor key, and Pitch is in the key of C, a natural is written, otherwise a sharp or a flat depending on whether the tonic of the current key is in the key of G. This minimizes the harmonic distance of the note from the current key; that is, the number of steps of a fifth from the tonic of the key to the note. For example, in the key of D major, a $B\flat$ would be written as a flat and not as an $A\sharp$. $B\flat$ is 4 steps from D in the cycle of fifths, while $A\sharp$ is eight steps from D.

6.8.3 Vertical Positioning of Notes

The current key and the pitch of the note to be plotted are used to determine the vertical positioning. A note with pitch 40, for example, could be plotted as a $C\sharp$ or at a higher vertical position as a $D\flat$.

If the vertical position is outside the range of the bass or treble five lines, then ledger lines are drawn to the note.

To plot a note, a semi-breve is first drawn. If the duration of the note is less than a minim, the plotted semi-breve is filled and the required tail is added. If the duration is divisible by the duration of a dotted demi-semi-quaver, then a dot is placed after the note to show that the note is prolonged by half as much again.

6.8.4 Allocation of Musical Parts

In the fixed part analysis of the woodwind Trio, the following assumptions are made:

-at most, only a fixed number of instruments are playing at any one time,

- each part is required to lie in a range of pitches, typical for the instrument.

-the parts never cross each other.

The higher pitched first oboe is denoted by upward stems and the second oboe by downward stems, while the bassoon is placed on its own on the bass clef. If there are insufficient estimates in the oboe range, a rest is plotted in the treble clef.

In the general part analysis the number of parts can vary, so the plotting of rests is suppressed to avoid cluttering the output. This is useful for plotting piano music where the number of simultaneous notes can vary greatly. The parts (numbered 0 to 3) represent the soprano, alto, tenor and bass parts respectively. The parts are determined by counting the number of coincident

notes with a higher pitch. Two notes are coincident if the first note to finish does so at least a demi-semi-quaver (50 to 100 milliseconds) after the other note begins. If a note is allocated to soprano and has a pitch less than ***E*** above middle ***C***, it is reallocated to the alto part. If the alto is below middle ***C***, it is moved to the tenor part. The lowest note is finally assigned to the bass. The soprano and alto are plotted in the treble clef, while the tenor and bass are in the bass clef. The soprano and tenor have upward stems while the alto and bass have downward stems. If the number of overlapping notes exceeds four, then some of the notes will be assigned to the same part depending on the pitch of the notes.

6.9 Conclusion

This chapter describes programs to output the results of analysed music. This includes the grouping of pitch estimates in time to determine the onset and finish times of notes, the determination of key, tempo, and beat, and the plotting of notes in the required positions with the appropriate accidentals.

The programs assume major or minor tonality, and equal temperament.

There are several ways the music plotting could be improved. Notes of duration less than a crotchet are plotted as independent short notes, and could be beamed together in groups of a crotchet duration. Notes of long duration are not split and tied across bar lines. The placement of bar-lines could be made more accurate by using adaptive beat tracking (see Harris 1982). Staves could be scaled horizontally to make them line up at the right-hand side of the page. And finally, the part allocation or voicing algorithm could be improved to track

parts as independently moving melodies, by minimizing note steps from one note to the next in each part.

Despite these limitations, the plotting system does provide an easily read output for musical data, a vast improvement on deciphering listings with thousands of numbers and letters. Instead of requiring human intervention for the determination and positioning of musical symbols, the plotting is fully automated.

CHAPTER SEVEN

Evaluation of Analysed Music

7.1 Introduction

This chapter presents the results of applying the algorithms of chapters 5 and 6 to music examples. Section 7.2 compares the low level techniques described in chapter 5, namely, autocorrelation, the cepstrum, the harmonic summing algorithm, and the frequency ratio algorithm. The iterative extraction procedure is considered in section 7.3, and section 7.4 evaluates the musical analysis and plotting described in chapter 6. The benchmark test is treated in section 7.5. Section 7.6 applies the error measures defined in chapter 8 to the analyses in this chapter.

The analysed pieces of music are:

(a) Fugue number 11 from the 48 *Preludes and Fugues,*

(b) Prelude from *Partita* number 1,

(c) Menuet II from *Partita* number 1,

(d) Trio I for two oboes and bassoon from the *Brandenburg Concerto* 1 performed by L.A. Philharmonic on Deutsche Grammophon recording 2707 098.

All the music is by J.S. Bach, and (a), (b), and (c) are played on piano, performed by Dinu Lipati on EMI recording HQM 1210.

Much of the data analysed is polyphonic piano music. A piano tone is a difficult musical tone to analyse. It lacks a steady state, and has a complicated decay function due to beating of multiple strings and inharmonicity of partials, caused by string stiffness (the higher partials tend to be sharp). Cues such as quasi-consonant beating (chorus effect) do not apply. These problems are

compounded when more than one tone sound simultaneously. Yet the algorithms developed in this thesis give good results even for piano music.

7.2 Comparison of Low Level Techniques

Figure 7.1 shows the cepstrum of the chord with two oboes and a bassoon. This is the same chord as that shown in figure 2.10. The oboes are playing a $B\flat$ (frequency 930 Hz) and D (590 Hz), and the bassoon a G (196 Hz). In this example all the notes present are correctly detected, but some harmonically related incorrect estimates also occur. The errors mainly occur at an octave, twelfth, etc. below the correct estimates. These errors occur at multiples of the fundamental period and could be corrected using a procedure similar to heuristic H2 (see 5.6.1), but applied in the time domain.

Figure 7.2 plots the autocorrelation of the same chord as figure 7.1. Again the harmonically related errors occur, and the second oboe estimate (560 Hz) is a semitone flat.

Figure 7.3 shows the cepstrum of the steady state of a bassoon tone; the same tone shown in figure 2.1. In this monophonic case the cepstrum gives a single unambiguous peak corresponding to the tone.

7.2.1 Fugue Example

The Fugue is used for comparison of low level techniques, because the first four bars are monophonic and the remainder is polyphonic. Bar lines and bar numbers have been inserted by hand on all the pitch profiles to simplify

comparison with the written music.

Figure 7.6 shows all the harmonics of the first 10 seconds of the Fugue. Many of the notes can be determined by observing just the fundamentals. Such a technique, however, would not apply to music with attenuated fundamentals.

Figure 7.7 gives the cepstrum of the first 10 seconds of the Fugue. The monophonic section of figure 7.7 (bars 1 to 4) is clearly tracked (with another false estimate at an octave below), but the polyphonic section (bars 5 to 8) contains many harmonically related errors. Similarly, the autocorrelation of the same music (figure 7.8) fails to discriminate the polyphonic part.

Figure 7.9 displays the output of the harmonic ratio algorithm and figure 7.10 displays the output of the harmonic summing algorithm for the Fugue. The harmonic summing algorithm appears to be the best of the low level techniques, and the harmonic ratio algorithm the worst.

Figure 7.13 gives the original written music for comparison.

7.2.2 Woodwind Trio Example

Figure 7.14 shows the cepstrum of the first 12.5 seconds of the Trio. The oboes are clearly tracked, but the bassoon is almost undetectable. Similarly the autocorrelation (figure 7.15) provides a clear trace for the oboes, but not the bassoon part.

Figure 7.16 plots the output from the harmonic ratio algorithm for the first 9 seconds of the Trio. This algorithm is more successful at detecting woodwind tones than piano tones (figure 7.9), because the frequencies of the

harmonics are close to integral ratios of the fundamental.

Figure 7.17 displays the output from the harmonic summing algorithm applied to the first 12.5 seconds of the Trio.

Figure 7.18 plots the harmonic summing algorithm applied to the first 12.5 seconds applied to the Trio with heuristic H1 used. Here, estimates with even harmonics stronger than odd harmonics are suppressed (see 5.6.1).

Figure 7.19 plots the harmonic summing algorithm applied to the first 12.5 seconds of the Trio with heuristics H1, H2, H3, H4 and H5 used (respective weightings are: 1,.3,-1,1,.1), but with no spectral extraction. The results improve as more heuristics are introduced.

7.3 Heuristic Extraction

Figure 7.11 plots the harmonic summing algorithm applied to the first 10 seconds applied to the Fugue with heuristics H1 to H5 applied, and with the best estimate iteratively extracted from the spectrum (see section 5.6).

Figure 7.13 shows the original written music for comparison.

Figures 7.24 and 7.25 show the the results of the harmonic summing algorithm applied to the first 20 seconds of the Partita Prelude, with heuristics H1 to H5 applied, and with iterative extraction of estimates.

Figure 7.27 shows the original written music for comparison.

Figures 7.28 and 7.30 display the harmonic summing algorithm applied to the first 20 seconds of the Menuet, with heuristics H1 to H5 applied and iterative extraction.

Figure 7.32 shows the original written music for comparison.

Figure 7.20 plots the output to the harmonic summing algorithm applied to the first 12.5 seconds of the Trio with heuristics H1 to H5 applied and with the best estimate iteratively attenuated.

Figure 7.21 plots the output of the harmonic summing algorithm applied to the first 12.5 seconds of the Trio with heuristics H1 to H5 applied and the best estimate iteratively attenuated, but without the recorded reverberation deconvolved. All the previous examples have reverberation automatically removed. Here, reverberation can sustain signals up to half a second after the finish of some notes.

In all these examples the results are improved by applying the the iterative extraction procedure.

7.4 Music Analysis and Plotting

All the music plotted in this section uses pitch data from the heuristic extraction procedure (section 5.6), and deconvolution of reverberation. Errors have been circled by hand.

7.4.1 Fugue Example

Figure 7.12 shows the music output for the Fugue, derived from figure 7.11. The note ***B*** preceding the ***A*** in the trill at time 8 seconds and in bar 7 is missing from the analysis. The pitches of the other notes are correct. There are, however, several errors in the music plotting. The crossing of the parts (bar 6)

cannot be determined by the algorithm used, because the note with the lowest pitch, at any time, is assigned to the bass part. The *D* (p29) in bar 7 and the *D* (p29), *C* (p27), and *D* (p29) in bar 8 are all incorrectly assigned to the bass, because there is no detectable note below them. The *B*♮ (p26) at the beginning of bar 8 is played as a passing note to *C* (p27).

Here, the minimum duration of notes included in the analysis is 100 milliseconds, and the pitch bucket filling time (see 6.5) is 100 milliseconds. The tempo (178 crotchets per minute) and the starting time are set as parameters to fit the bar lines correctly.

7.4.2 Partita Example

Figure 7.26 shows the music output for the Partita Prelude. The minimum duration is 60 milliseconds, the bucket filling time is 120 milliseconds, and the tempo is 51.3 crotchets per minute. Several notes of long duration are split, due to masking by other notes (eg. the *B*♭ p25 in the first bar). The *C* (p39) in the middle of bar 2 (figure 7.26) is incorrect. The *F* (p32) at the end of bar 2 is incorrect. The *D* (p41) at the beginning of bar 3 is not strong enough to cause the note *C* (p39) in the previous bar to finish (see section 6.5.1). The *C* (p39) is therefore assumed to continue throughout the duration of the *D* (p41). The same phenomenon occurs at the beginning of the next bar. In bar 4 the note *G* (p46) is missing and the note *C* (p39) is incorrect.

7.4.3 Menuet Example

Figure 7.29 shows the music output for the first 10 seconds of the Menuet. The *B♭* (p37) in bar 3 (figure 7.29) is missing. The *E* (p19) in bar 6 is incorrect. The grace note (*D* p41) in the last bar is not separated from the rest of the chord, because its overlap is too great (see 6.8.2).

Figure 7.31 shows the music output for the second 10 seconds of the Menuet. The *F* (p32) in bar 1 is incorrect. The *B♭* (p13) in bar 2 is incorrect. The *D* (p29) in bar 4 is incorrect. The *E♭* (p30) in bar 5 is incorrect. The incorrect note, *B♭* in the bass at the end of the fifth bar in figure 7.31 (at 5.5 seconds on figure 7.30) is the same problem as the second inversion chord presented in section 5.5 and displayed in figure 5.2. The *G* (p46) in bar 6 is missing. The *D* (p29) in bar 6 is missing. The minimum duration is 80 milliseconds, and the bucket filling time is 100 milliseconds. The masking of adjacent notes is suppressed. The tempo is 152 crotchets per minute.

7.4.4 Trio Example

Figure 7.22 shows the music output for the Trio. Here, the general part analysis is used. The bassoon is not assumed to lie below a pitch of 30 (as in the fixed part analysis), but is taken as the note with lowest pitch. The minimum duration is 80 milliseconds, and the bucket filling time is 120 milliseconds. Apart from the exact transcription of the trills, the pitches of all the notes are correct. The grace notes of the top oboe (*F* and *D*) in bars 5 an 6 respectively, are played on the recording but are not shown in figure 7.23. The first bassoon tone of bar 8 fails to line up with the oboes. This is because the early part of the bassoon

tone is masked by the onset of the oboes. The note ***A*** p36, in bar 6, at 8.5 seconds in figure 7.20 is really two notes, the first played by the top oboe and the second played by the bottom oboe. Figure 7.22 plots this as one note, but in figure 6.1 of the last chapter, the fixed part algorithm of section 6.8.5 is used, and both notes are shown.

The tempo of the Trio was determined automatically from the inter-onset times. Table 7.1 shows the inter-onset time histogram for the Trio. The tempo suggested by the histogram is 120 crotchets per minute. The tempo used to generate figure 7.22 is 121.5 crotchets per minute.

Table 7.1

Inter-Onset Time Histogram of Trio (first section)

```
250    ****************
275    *************
300    ********
325    ******
350    *******
375    **********
400    ********
425    ***********
450    ************
475    *************************
500    ************************************
525    ***************************
550    ********************
575    ******
600    *********
625    **************
650    ************
675    ********
700    ******
725    *************
750    ************
```

7.5 Trio Benchmark

The second half of the Trio was reserved as a benchmark test. Parameters were chosen on the basis of earlier analyses, but the results reported here

are the first, unaltered automated analysis of the music.

This section of the Trio starts at the twentieth bar (twelfth bar of the second section) of figure 7.23. At the end of the page of music the second section is repeated. The notes of this piece are specified by either the onset time or the number of bars from the twentieth bar.

Figures 7.33 to 7.36 show the pitch profile (spanning a total of 50 seconds) for the Trio benchmark. Full heuristic extraction is used. Figures 7.37 to 7.40 give the music output for the benchmark. Here, the fixed part analysis is used (see section 6.5).

The errors in the pitch, onset and durations of notes are treated in chapter 8. Other errors include:

- split notes (eg. the note ***A*** p36 in bar 2),
- enharmonic errors (bar 1, bassoon plotted as $D\flat$ instead of $C\sharp$)
- the exact transcription of the trills,
- the tying of notes across bar lines (eg the note ***A*** p12 in bar 3 should continue into bar 4)
- premature bar lines. (eg. bars 23, 24, and 29)

7.6 Error Analysis of the Music Examples

This section tabulates the error measures defined in chapter 8 for the music examples of this chapter.

Except for the Benchmark (see 8.3), the exact onset times of the notes were not known. Therefore E1 and E2 measures are not considered here. Match-

ing of events is based on the author's comparisons of pitch profiles with the written music. Matching of events in the Trio Benchmark is treated formally in chapter 8. The definitions of E3, E4, and E6 are given in section 8.2.4.

Table 7.2

Heuristic Weightings

name		PARAMETER WEIGHTINGS H1	H2	H3	H4	H5
Benchmark	Fig 7.36	.6	.5	-1	1.2	.1
Trio	Fig 7.21	.6	.3	0	1	.1
Prelude	Fig 7.11	.4	.15	1.5	.6	.1
Partita	Fig 7.24	.4	.15	1.5	.6	.1
Menuet	Fig 7.28	.4	.15	1.5	.6	.1

Table 7.3

Error Measures

name		Inclusion Errors Total Events	E3	E4	E6	Exclusion Errors Total Events	E3	E4	E6
Benchmark	Fig 7.36	318	3.8	3.1	0.6	350	15.1	9.3	11.7
Trio	Fig 7.21	80	0	0	0	80	6.2	0	3.7
Prelude	Fig 7.11	69	0	0	0	70	1.4	0	1.4
Partita	Fig 7.24	123	2.8	0	0	116	2.9	2.9	2.9
Menuet	Fig 7.28	128	5.2	0.8	0.8	122	2.8	0	0

Increasing the bucket filling time can overcome the problem of analyses splitting notes into a sequence of notes of shorter duration. This can only be done at the expense of introducing other errors, such as failing to detect notes of short duration. The erroneous pitches in these examples are introduced at the stage of estimating the pitches (section 5.5) and cannot be improved by altering the plotting parameters (eg. the bucket fill time and the minimum strength).

7.7 Conclusion

A range of musical examples is used to test the algorithms of chapters 5 and 6. The low level techniques described in sections 5.2 to 5.5 are compared. The harmonic summing algorithm appears to give the best results. The pitch estimation is further improved by the application of heuristics and a procedure to iteratively extract the most likely estimates from the spectra. The additional errors in the automated transcription to music notation are considered.

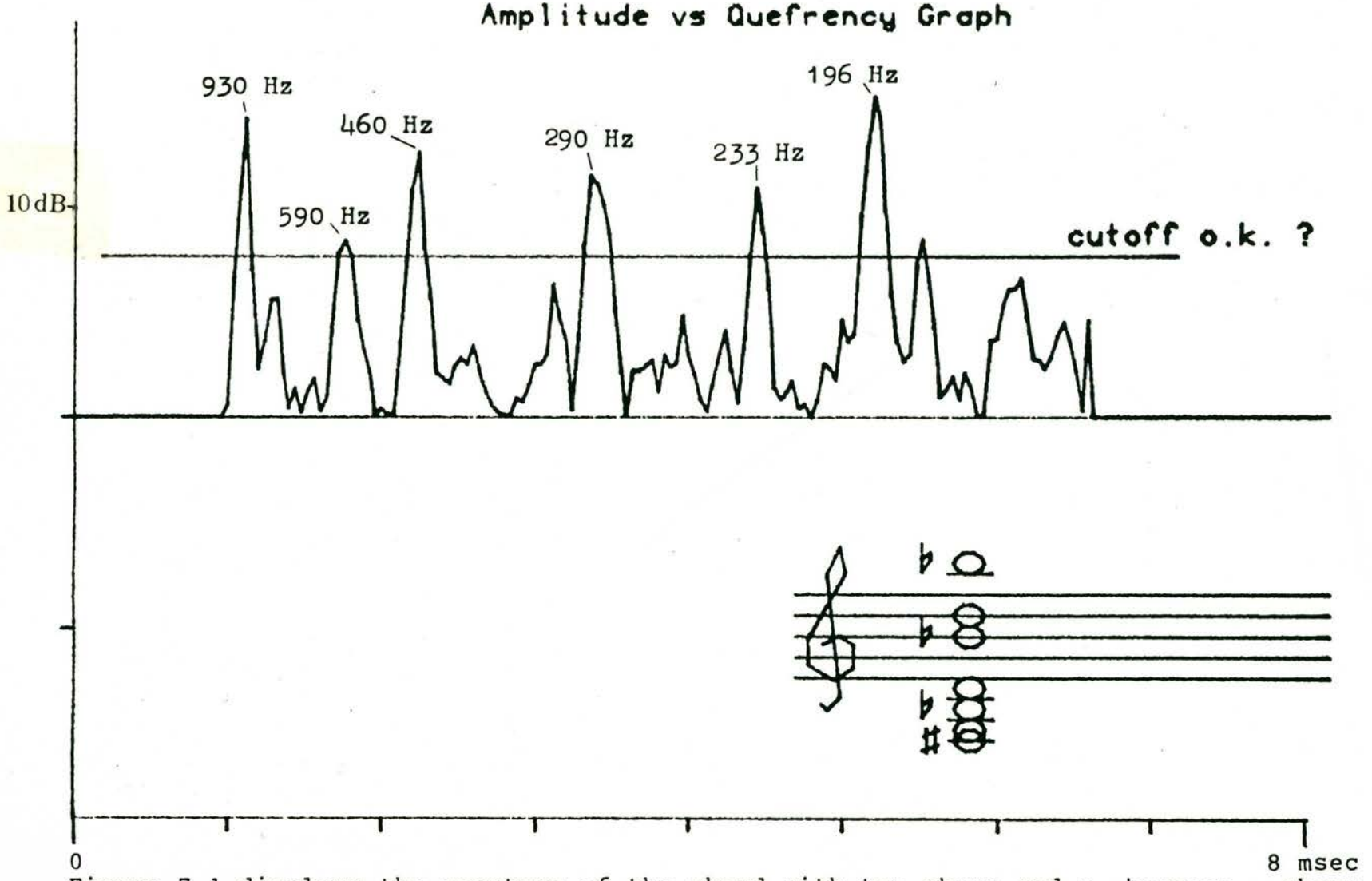


Figure 7.1 displays the cepstrum of the chord with two oboes and a bassoon, shown in figure 2.10.

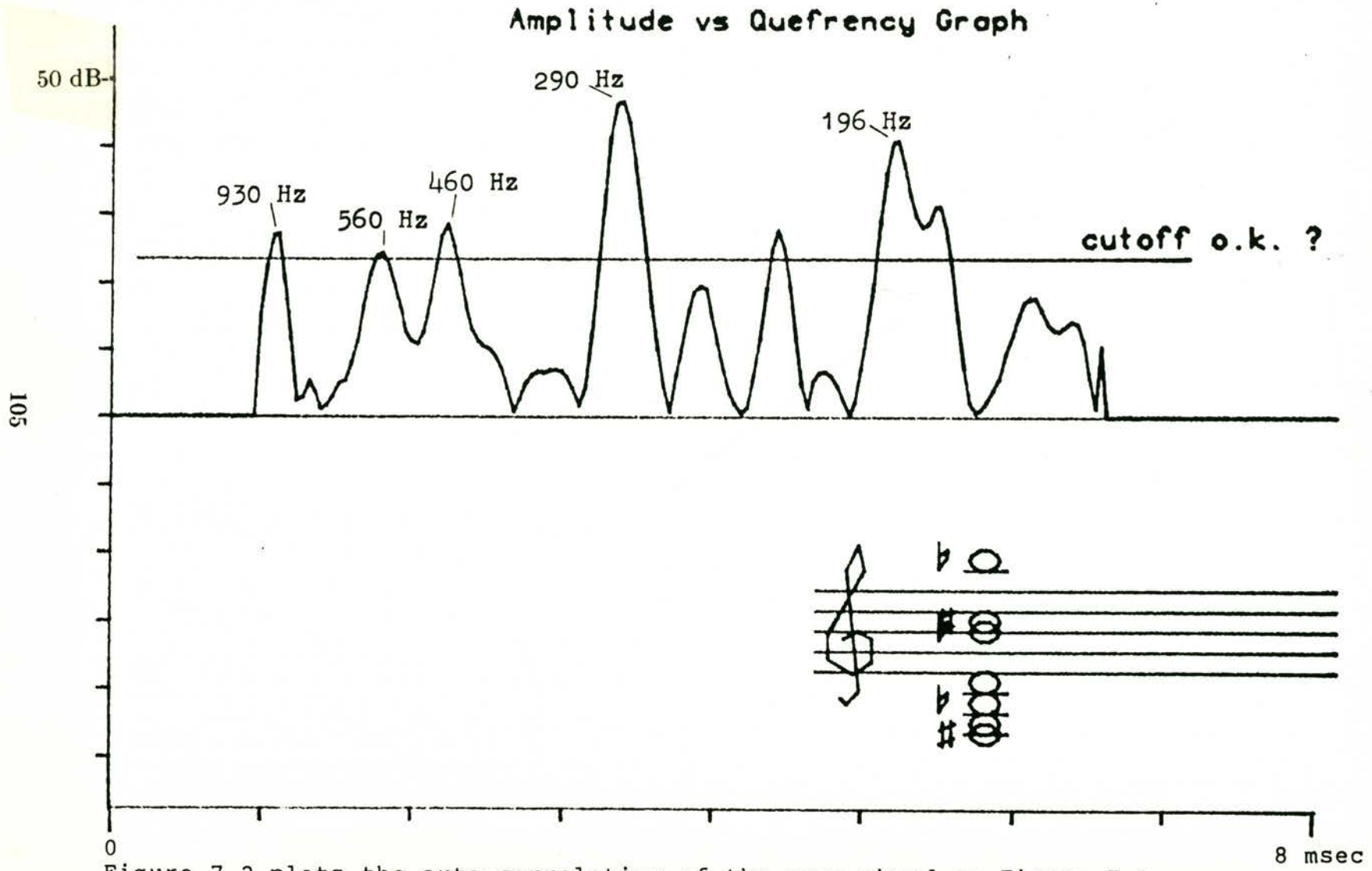


Figure 7.2 plots the auto-correlation of the same chord as Figure 7.1.

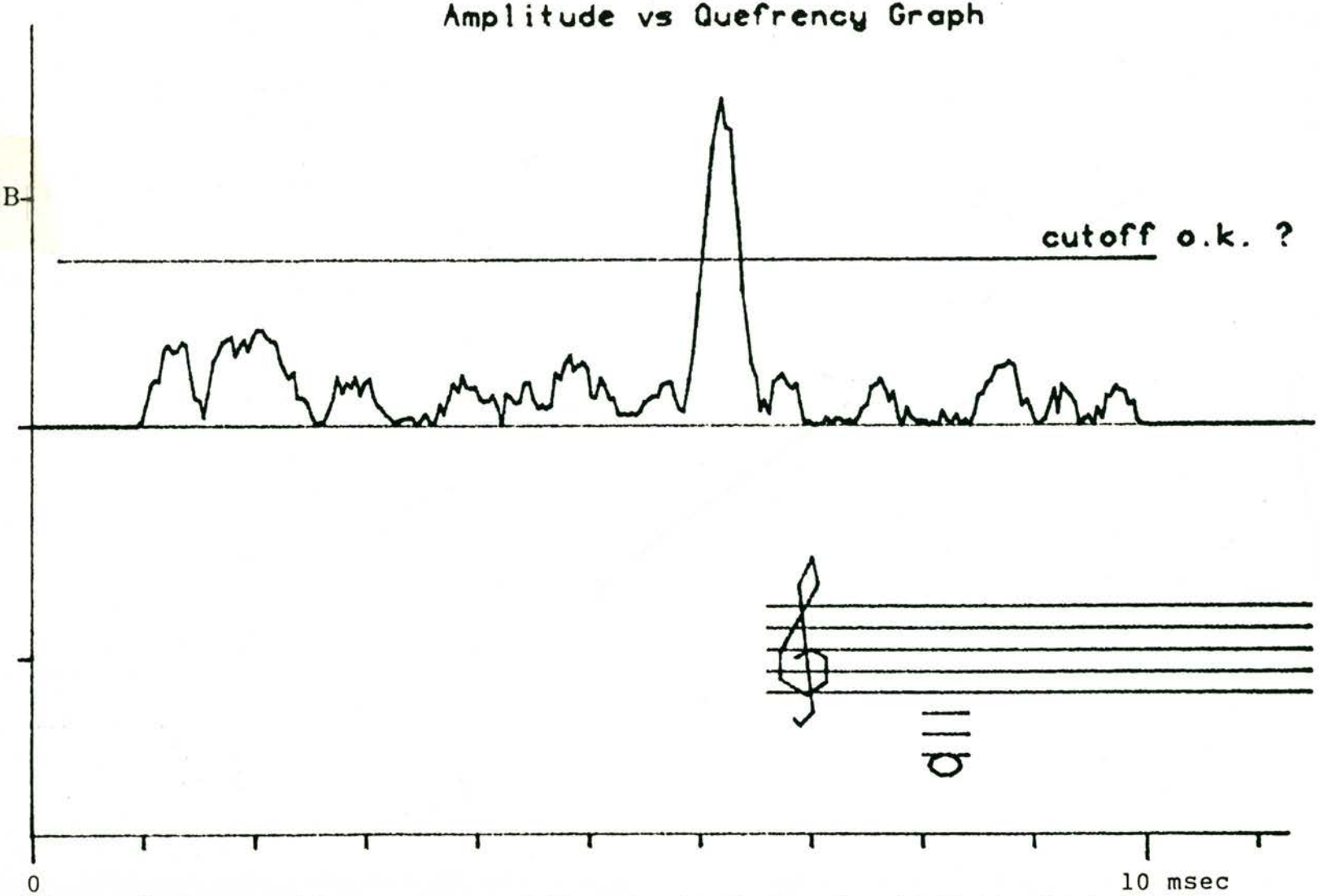


Figure 7.3 shows the cepstrum of the steady state of a bassoon given in figure 2.1.

Figures 7.4, 7.5 are deleted.

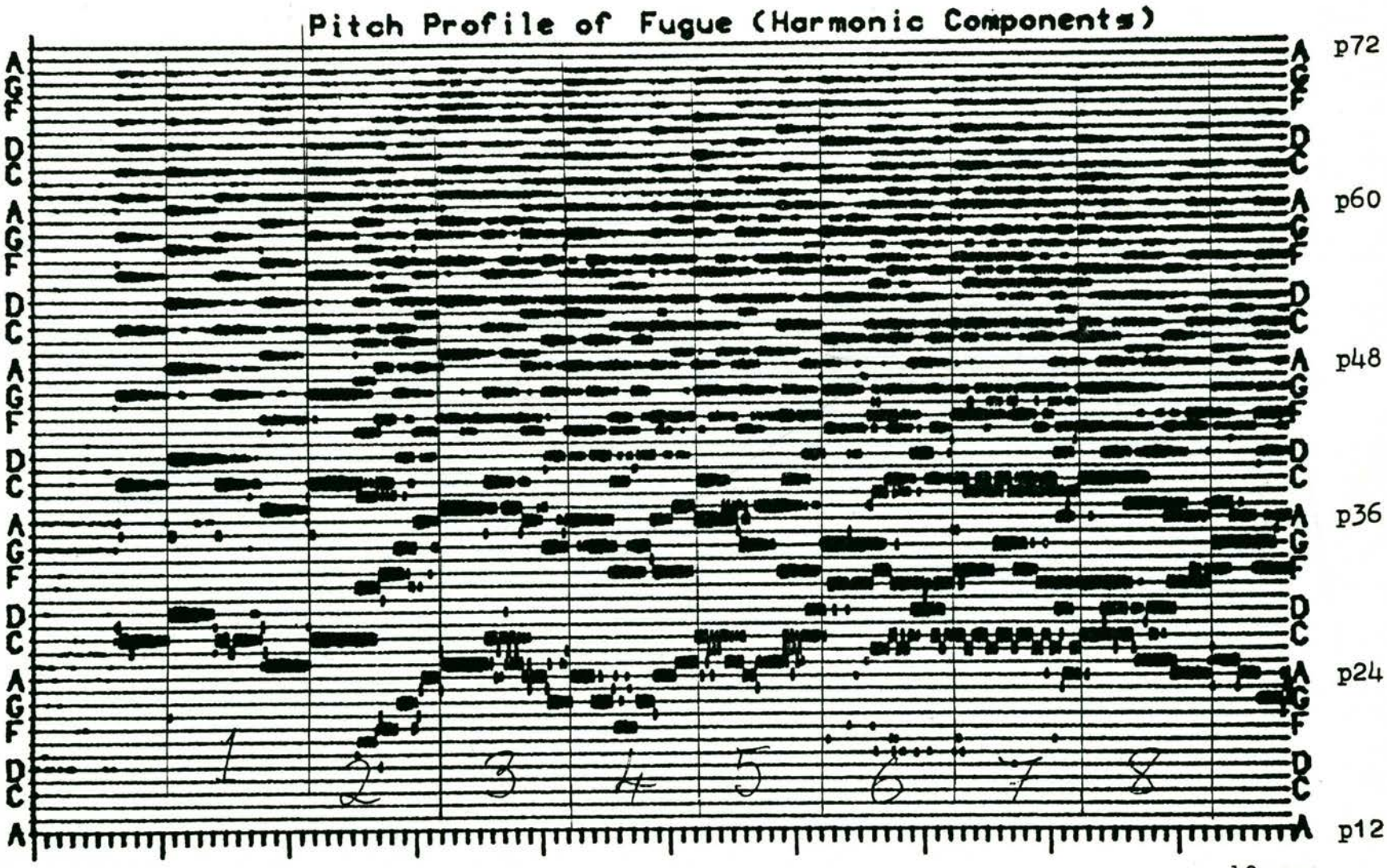


Figure 7.6 displays all the harmonics of the first 10 seconds of the Fugue number 11 for piano by J.S.Bach.

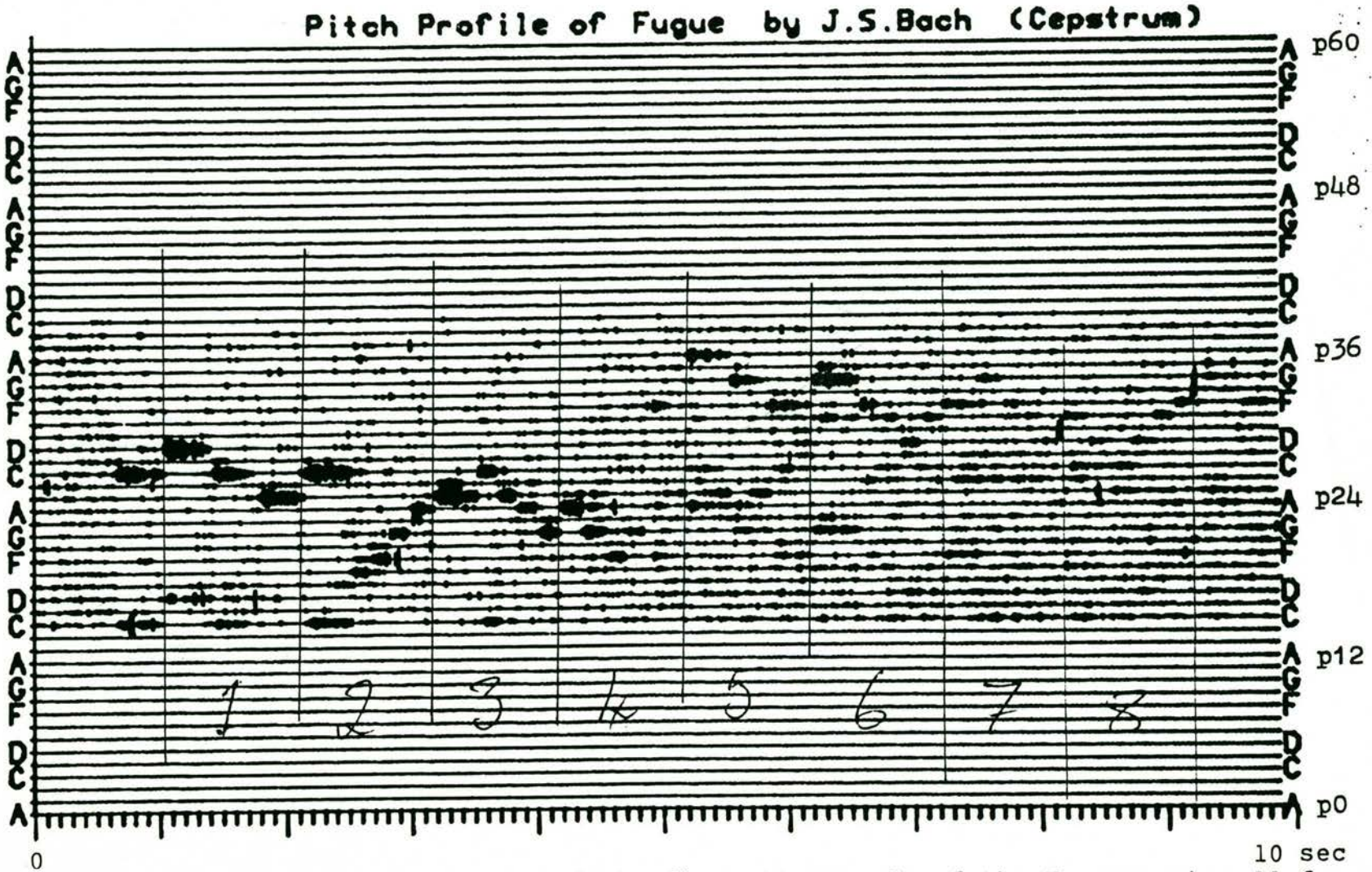


Figure 7.7 displays the cepstrum of the first 10 seconds of the Fugue number 11 for piano by J.S.Bach.

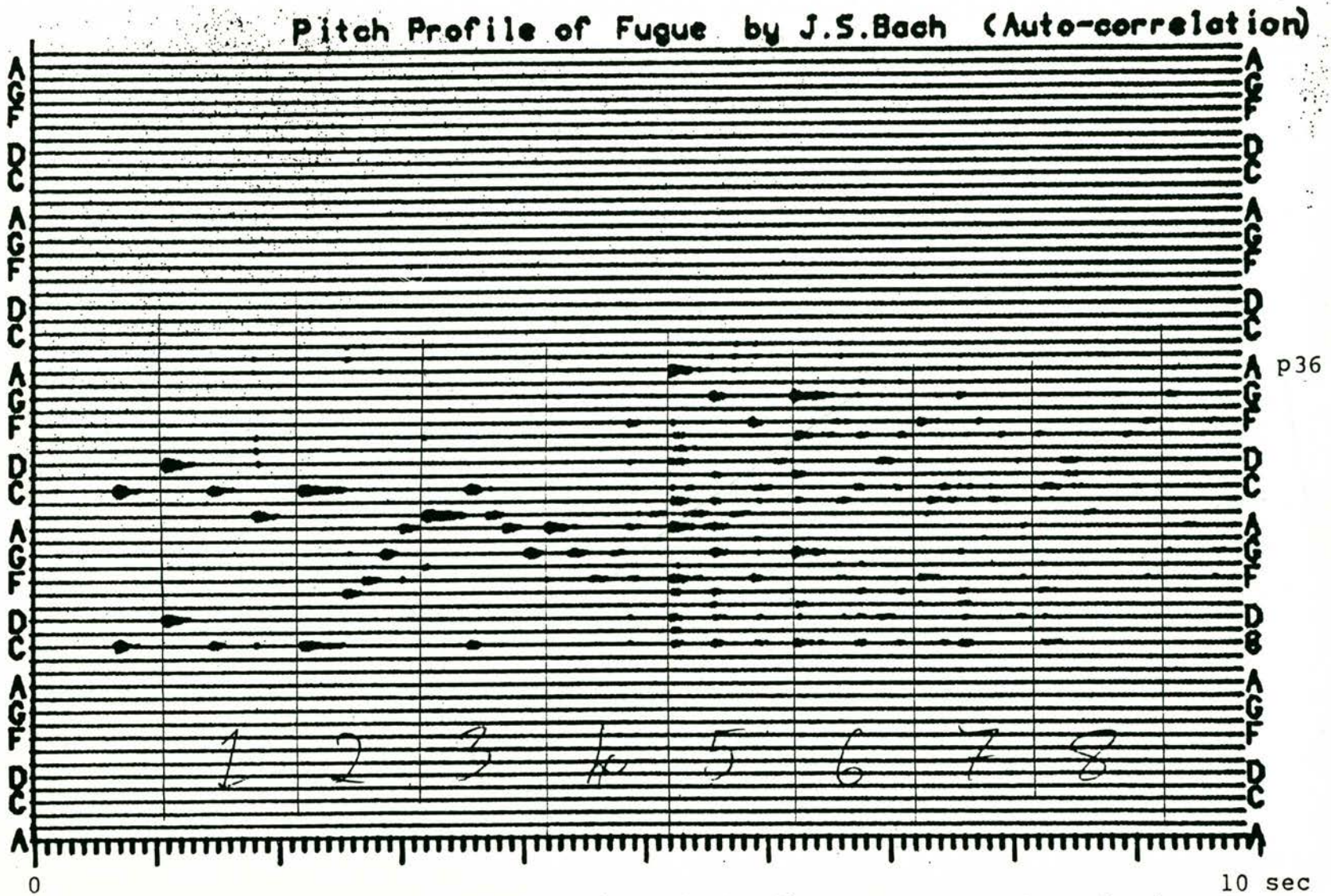


Figure 7.8 displays the auto-correlation of the first 10 seconds of the Fugue number 11 for piano by J.S.Bach.

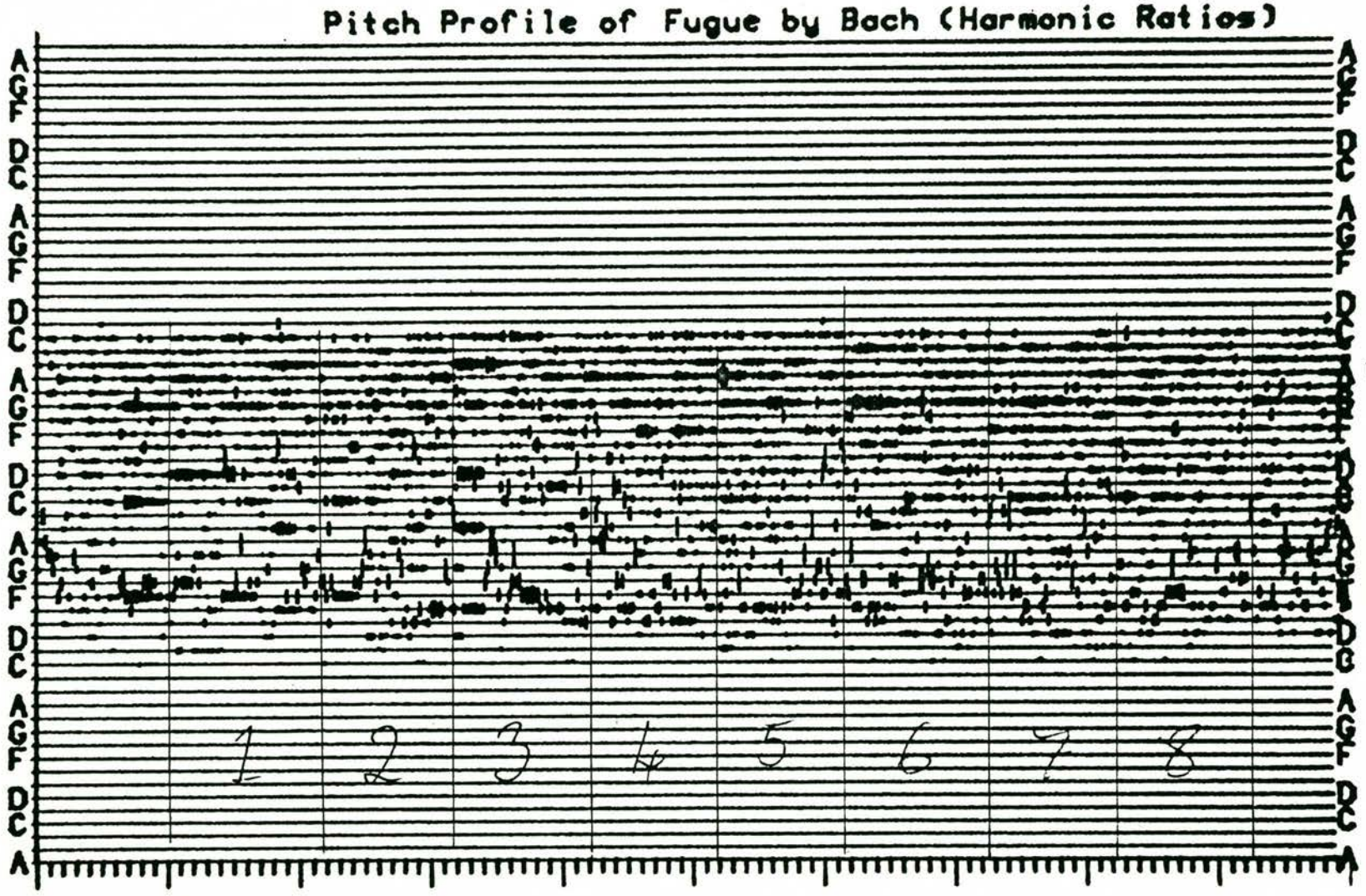
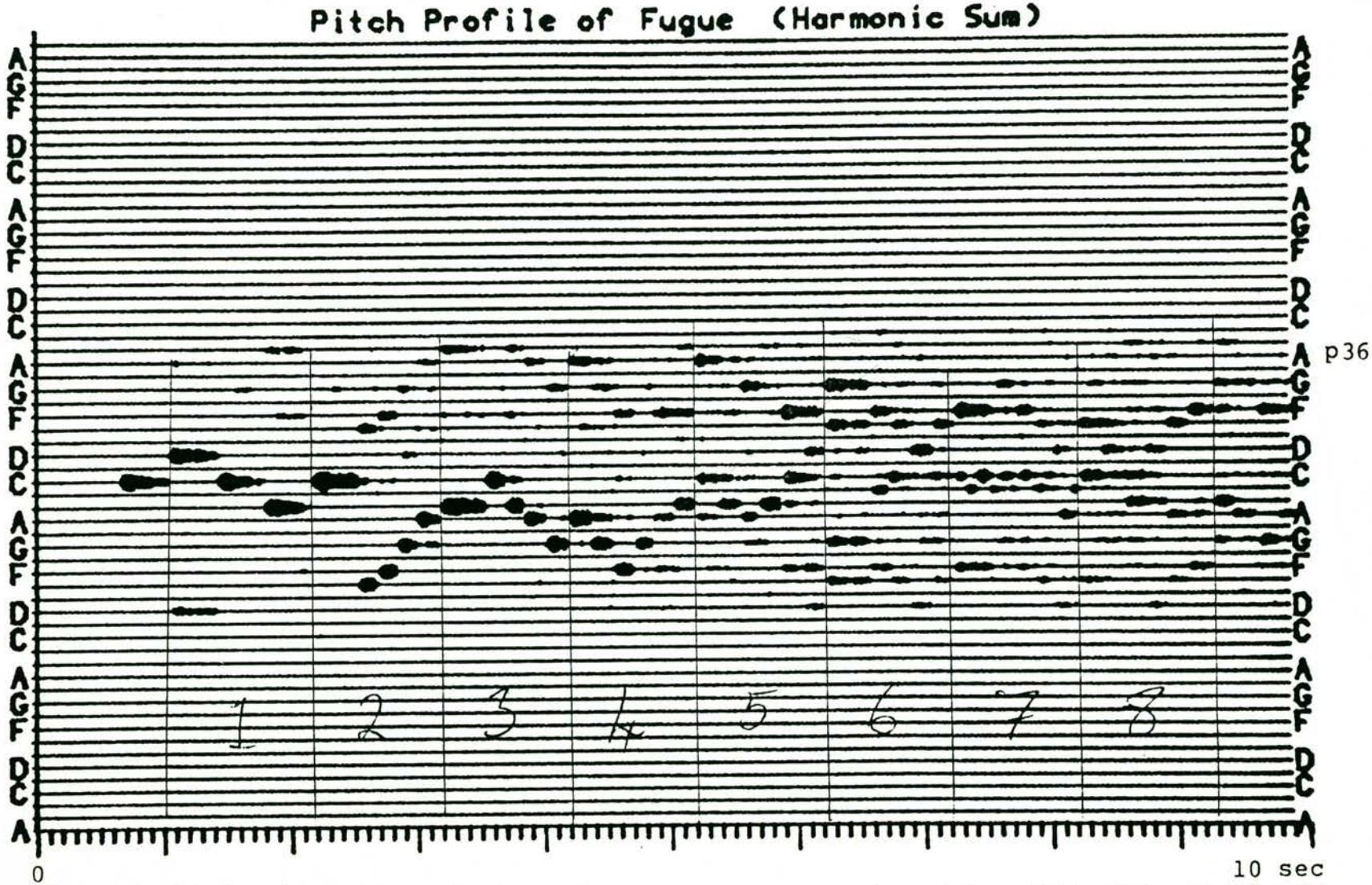
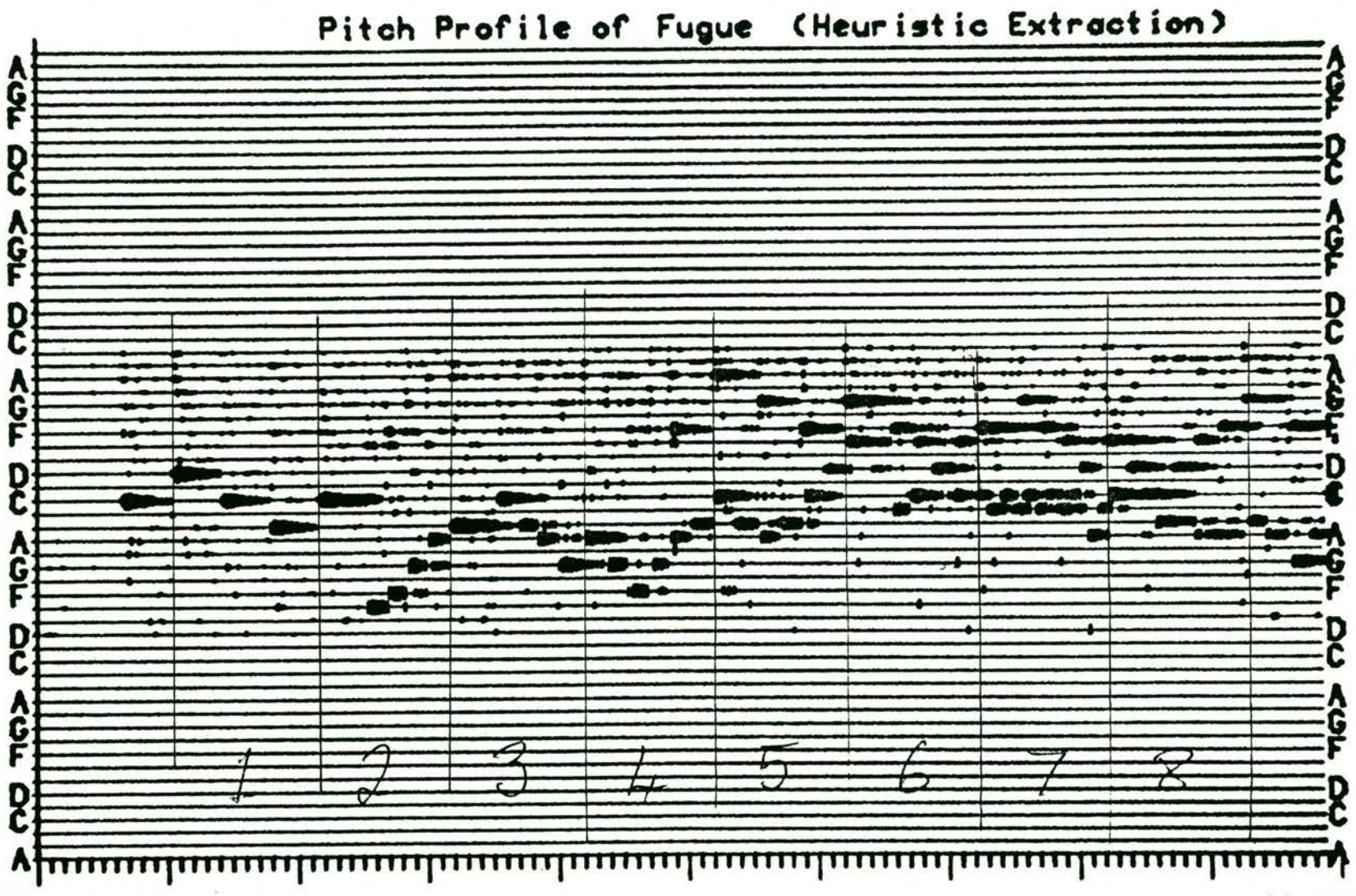


Figure 7.9 displays the output of the harmonic ratio algorithm for the first 10 seconds of the Fugue.



p36

Figure 7.10 displays the output of the harmonic sum algorithm for the first 10 seconds of the Fugue.



p36

Figure 7.11 plots the harmonic summing algorithm of the first 10 seconds of the Fugue with heuristics H1 to H5 applied.

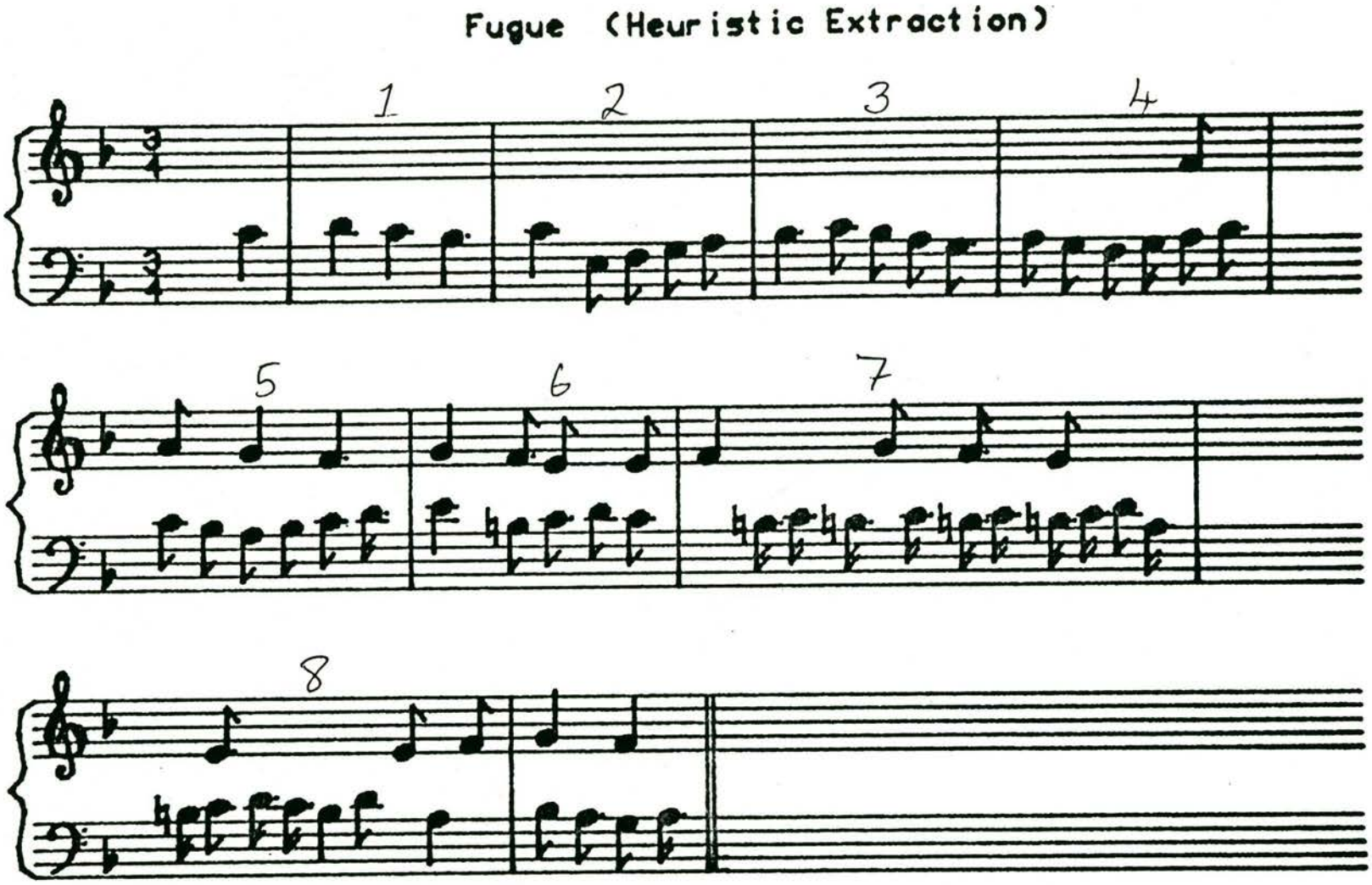


Figure 7.12 shows the music output for the Fugue, derived from figure 7.11. The pitches of all the notes are correct.

FUGUE XI

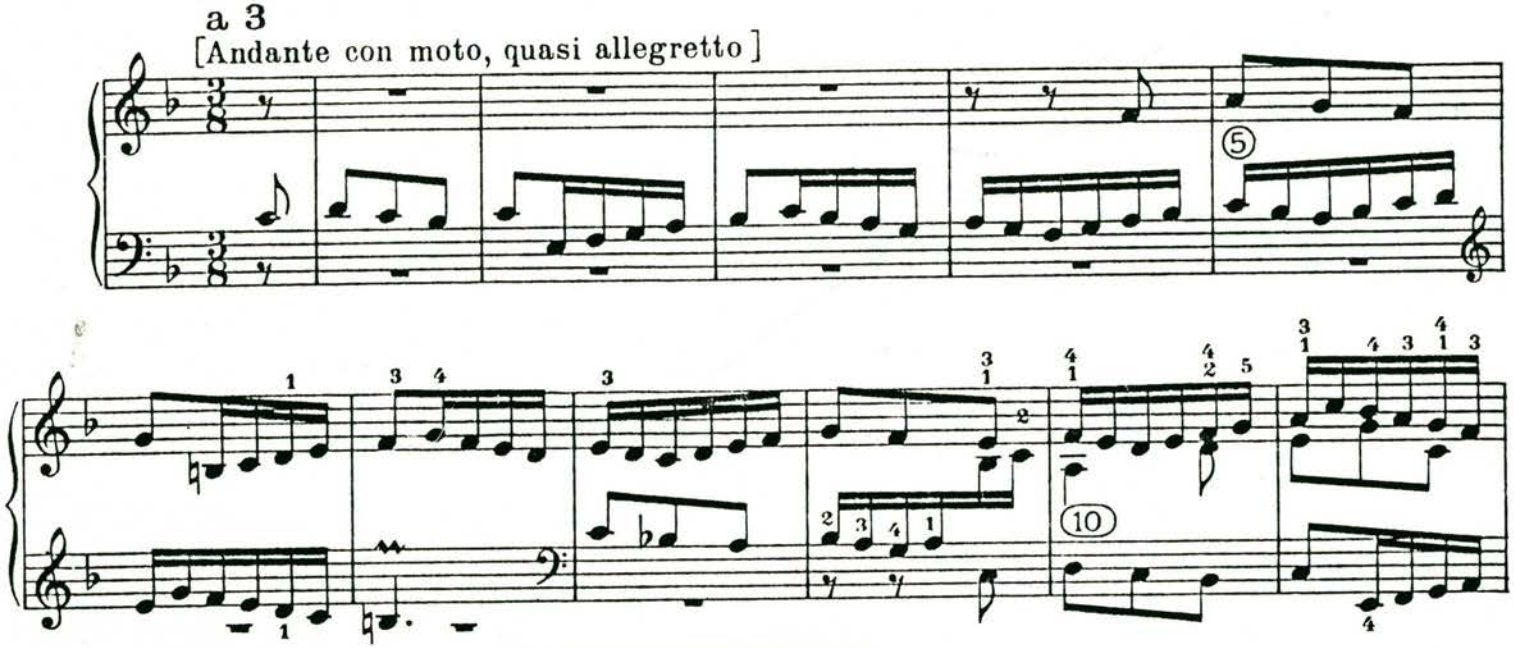
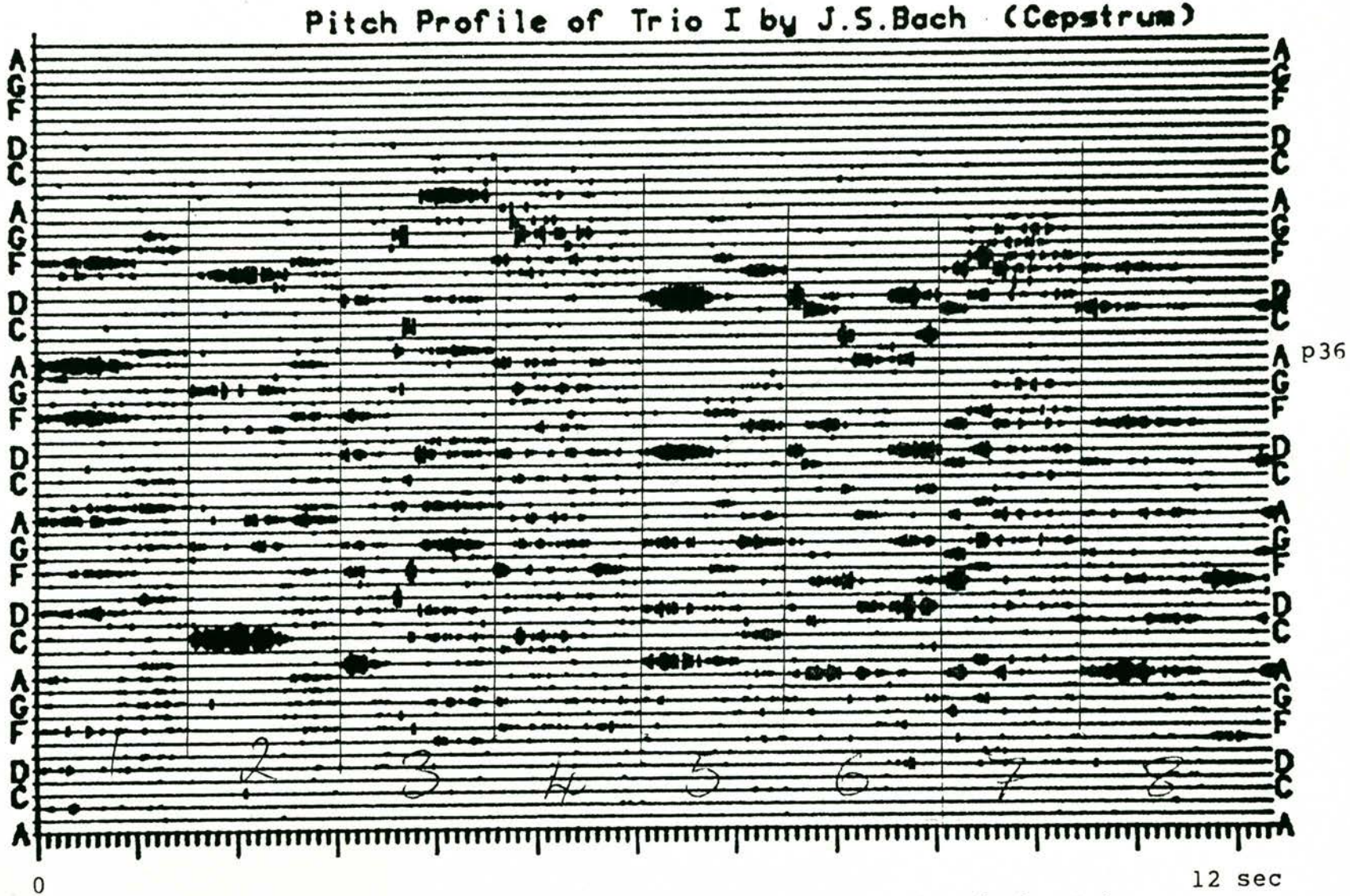


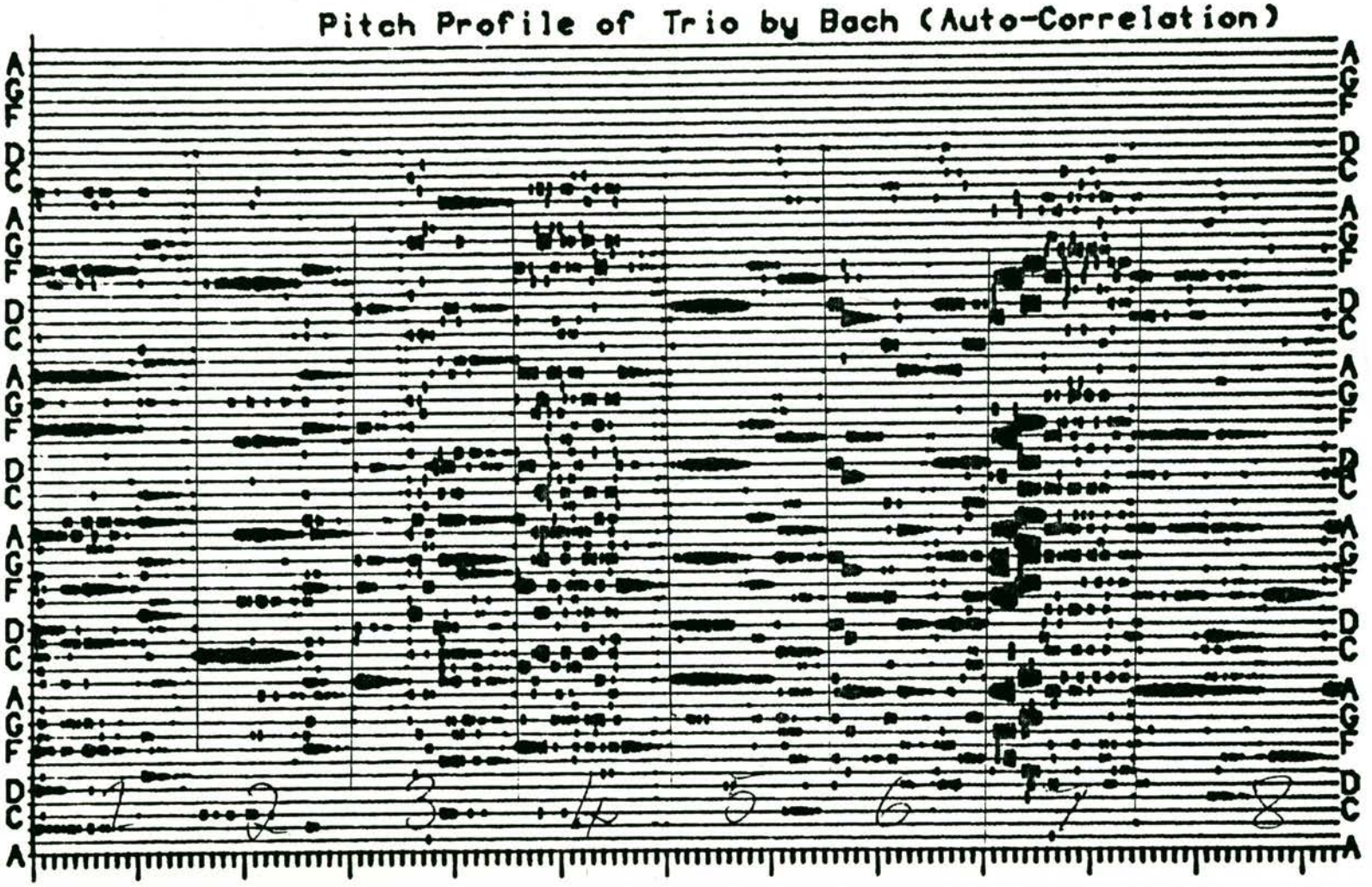
Figure 7.13 is the original written music of the Fugue.

(from J.S. Bach, Partiten 1-3, Urtext Edition).



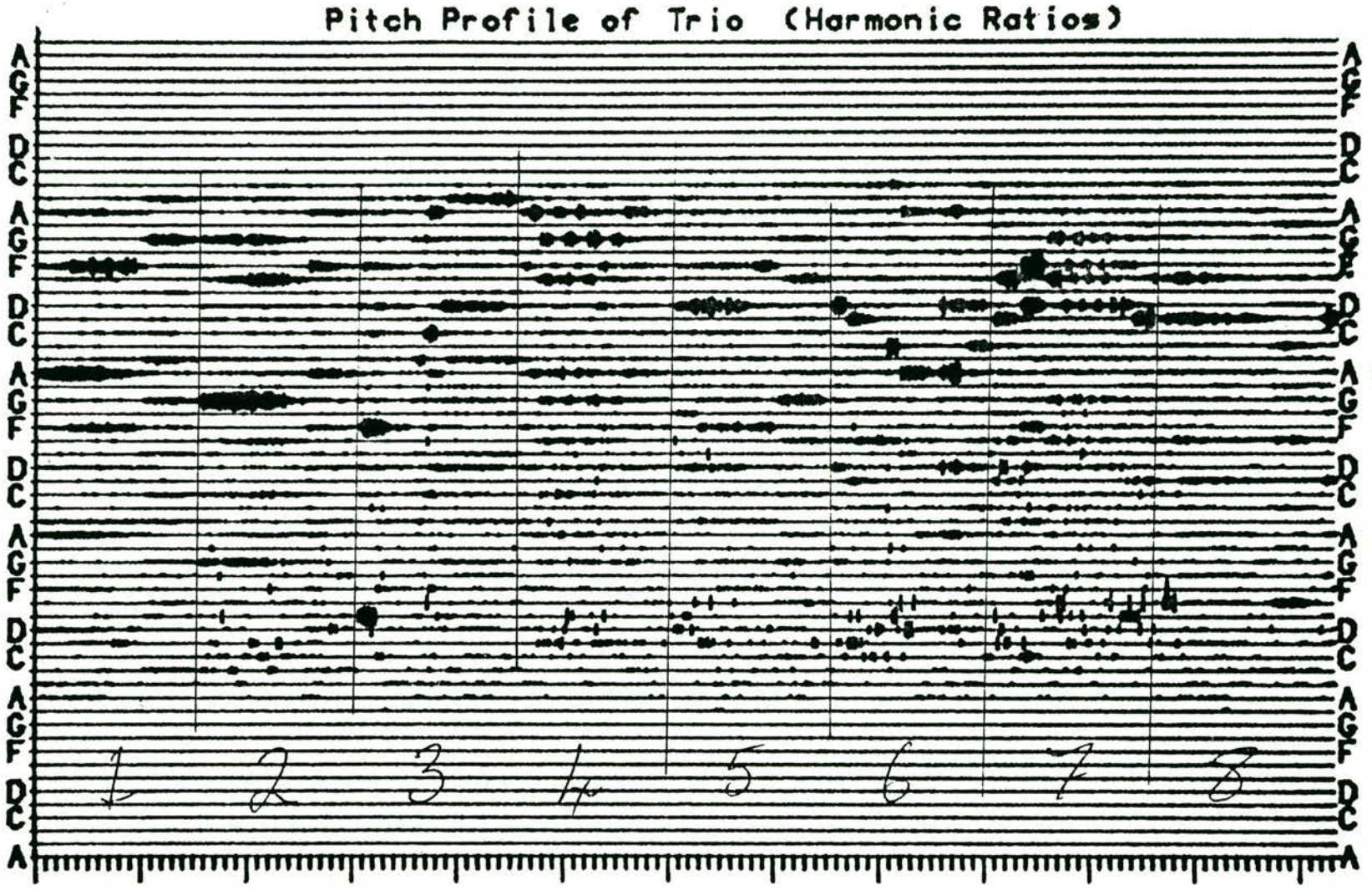
p36

Figure 7.14 displays the cepstrum of the first 12.5 seconds of the Trio.



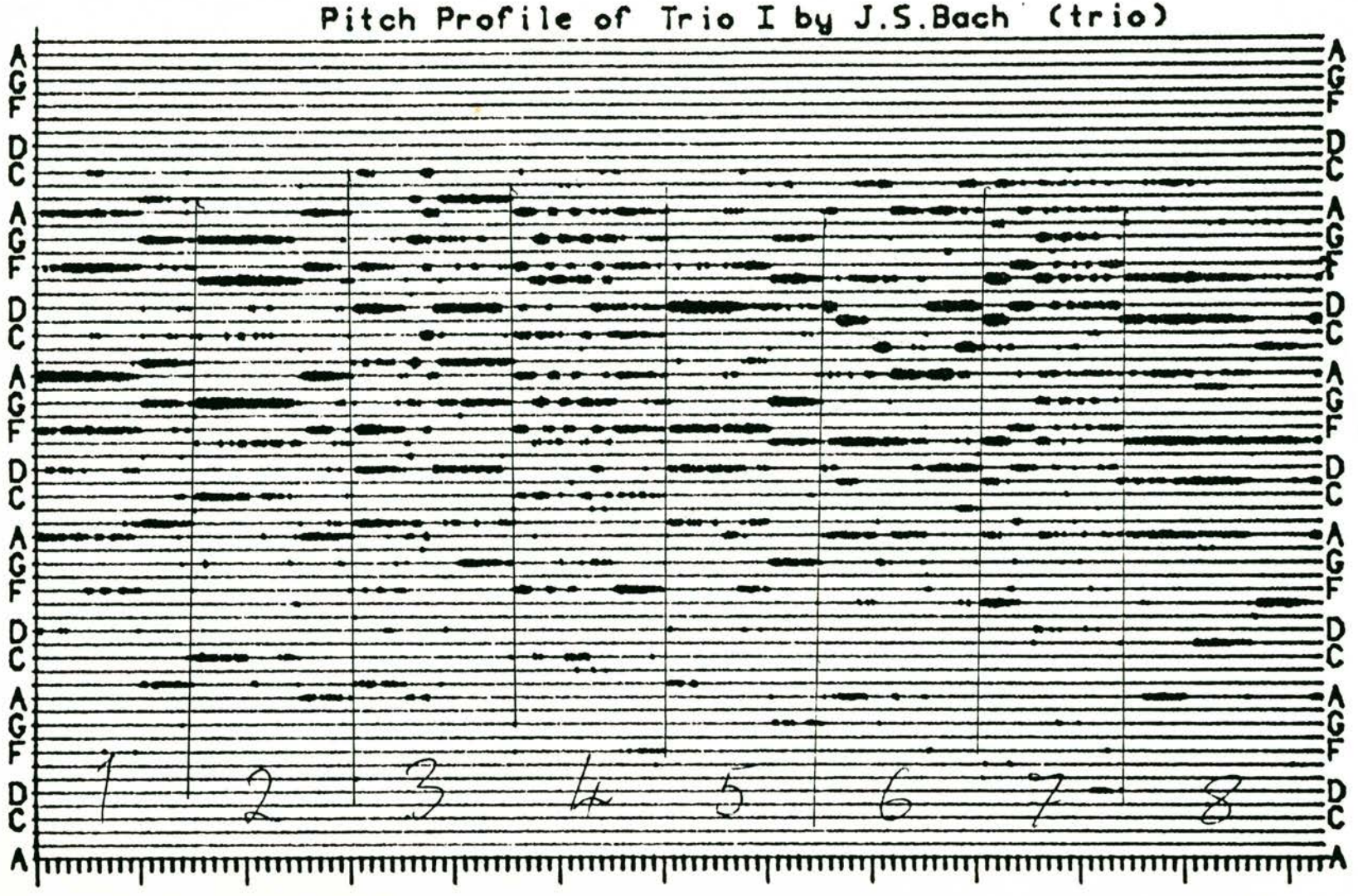
p36

Figure 7.15 displays the auto-correlation of the first 12.5 seconds of the Trio.



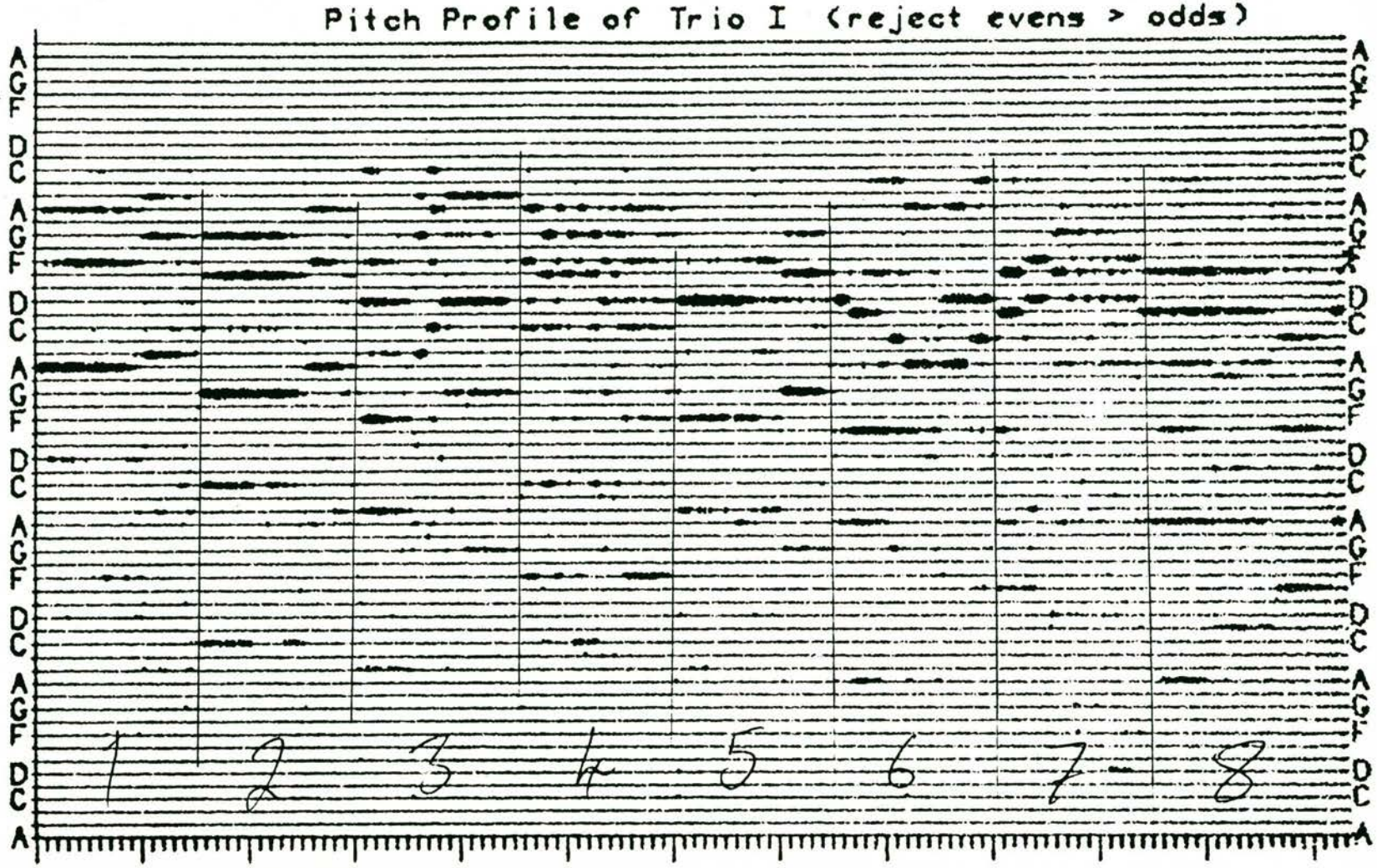
p36

Figure 7.16 plots the output from the harmonic ratio algorithm for the first 9 seconds of the Trio.



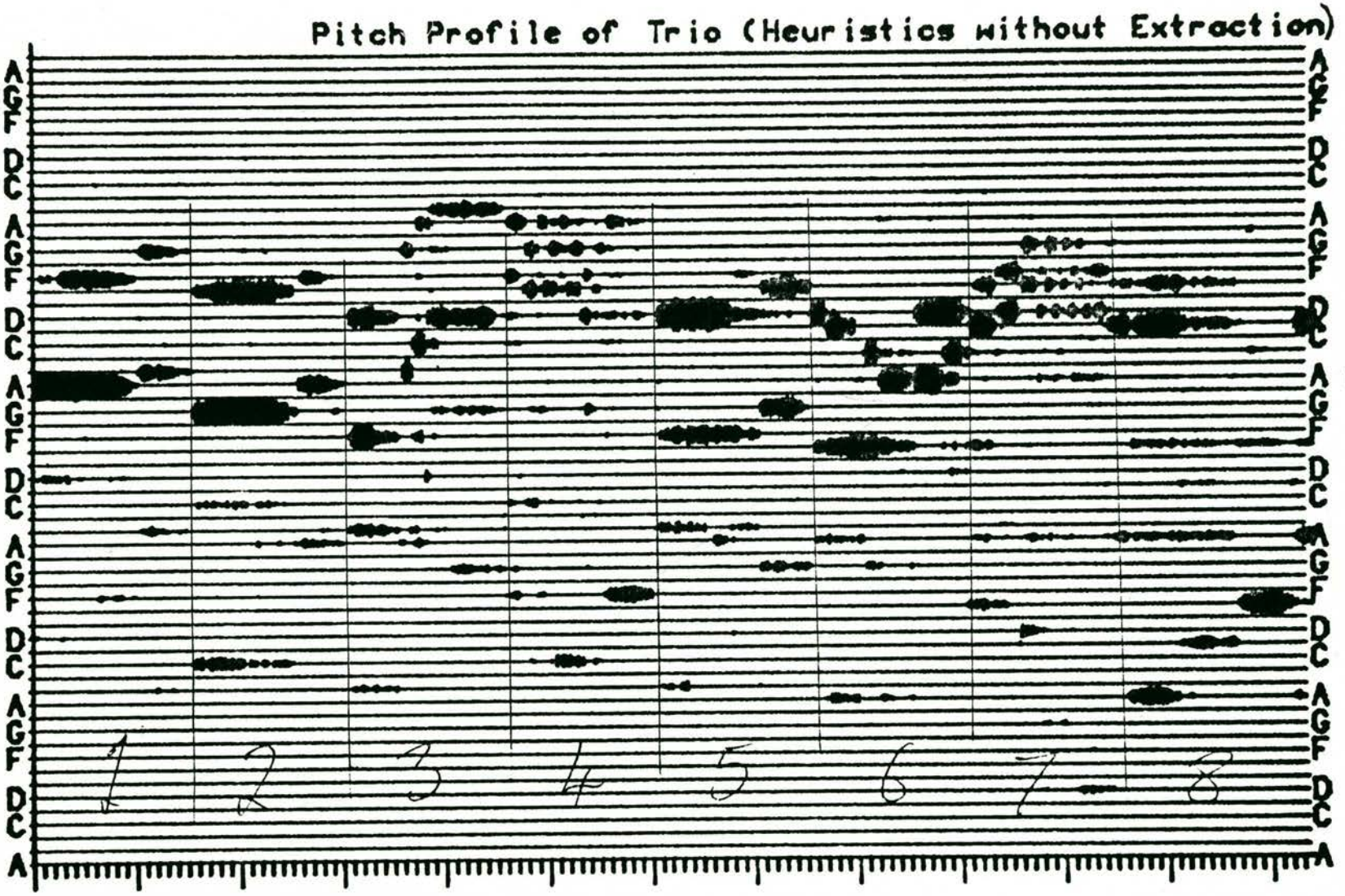
p36

Figure 7.17 displays the harmonic summing algorithm of the first 12.5 seconds of the Trio.



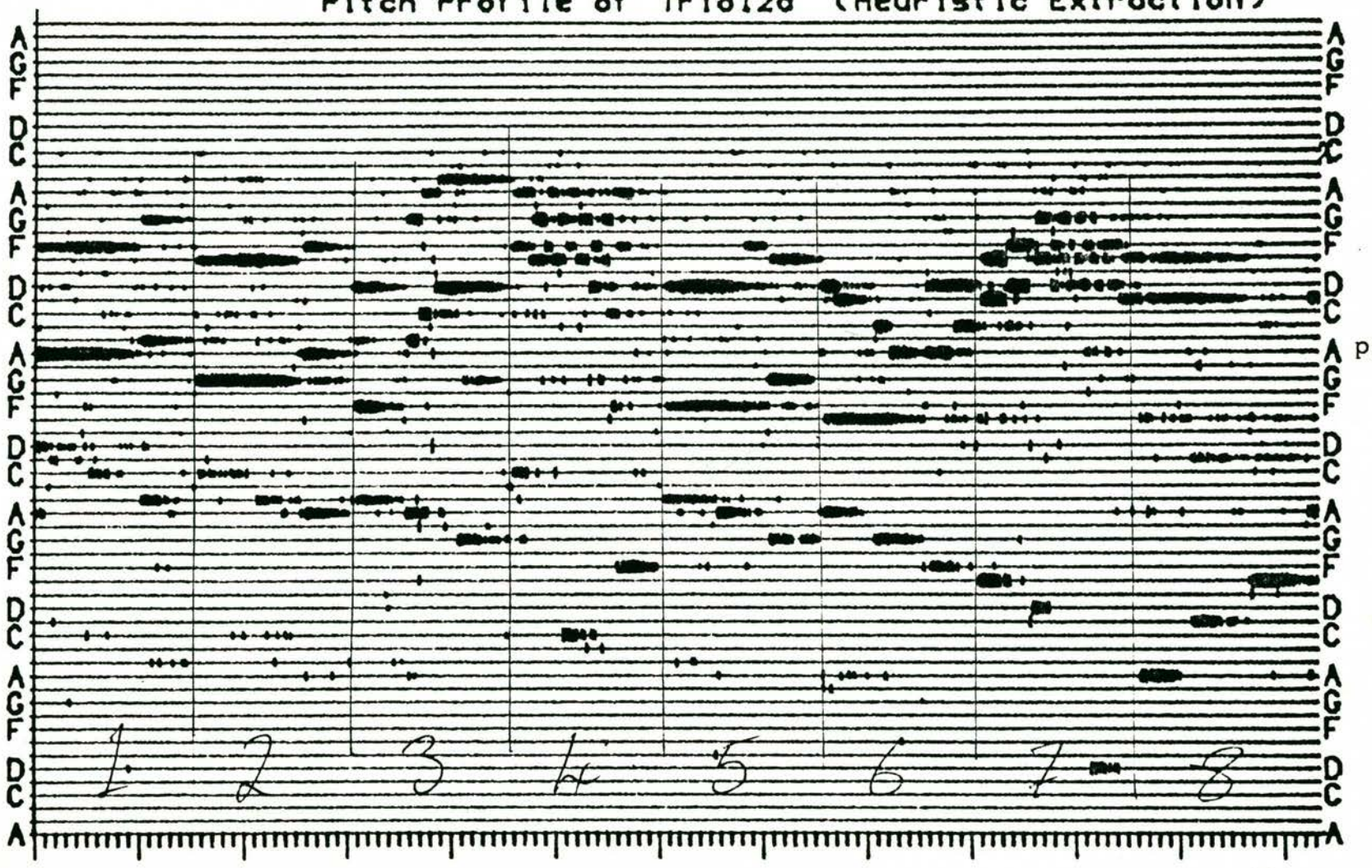
p36

Figure 7.18 plots the harmonic summing algorithm of the first 12.5 seconds of the Trio with heuristic H1 applied



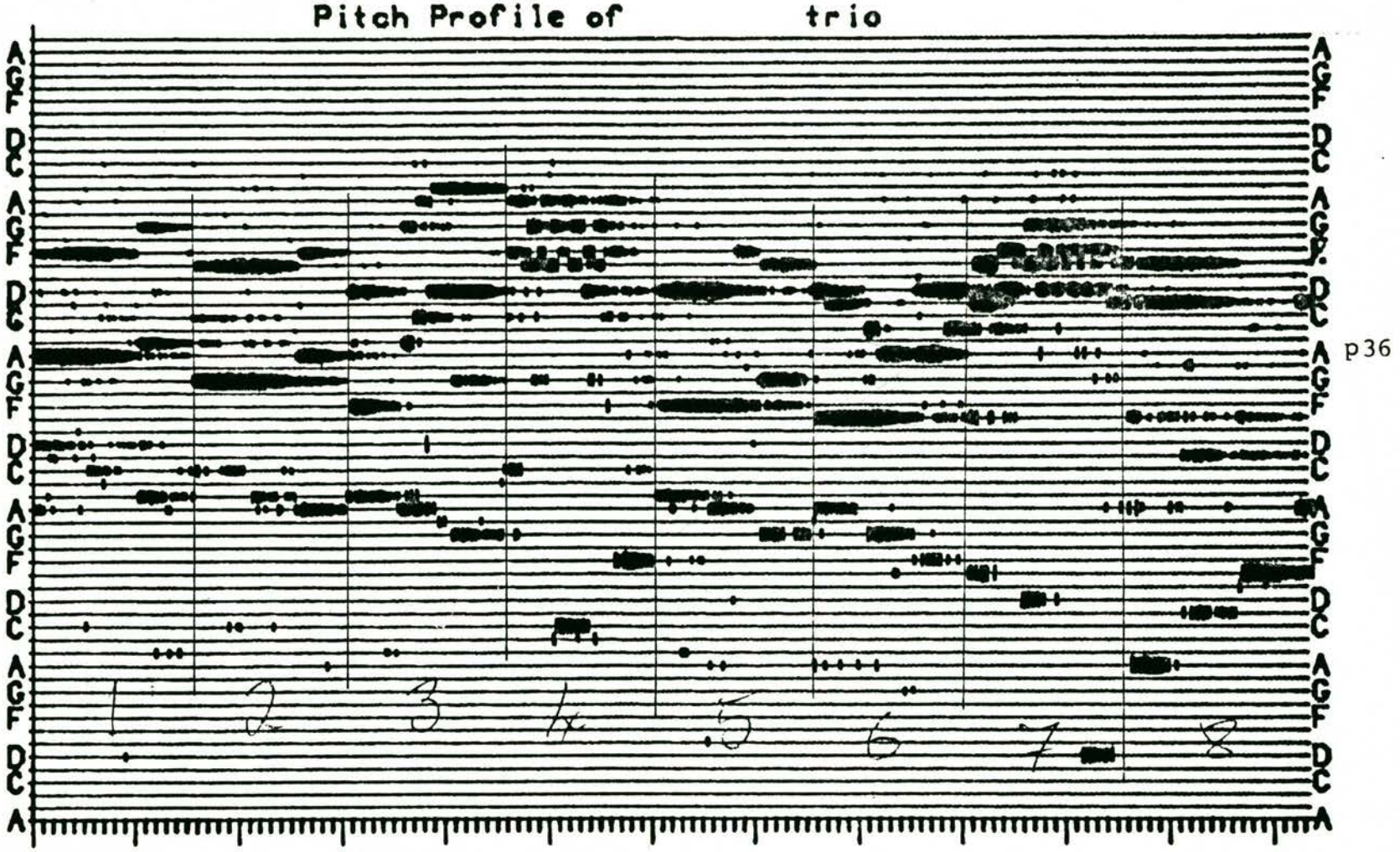
p36

Figure 7.19 plots the harmonic summing algorithm of the first 12.5 seconds of the Trio with heuristics H1, H2, H3, H4 and H5 applied (respective weightings were: 1,.3,-1,1,.1), but with no spectral extraction.



p36

Figure 7.20 plots the output of the harmonic summing algorithm for the first 12.5 seconds of the Trio with heuristics H1 to H5 applied with the best estimate iteratively attenuated.



p36

Figure 7.21 plots the output of the harmonic summing algorithm for the first 12.5 seconds of the Trio with heuristics H1 to H5 applied with the best estimate iteratively attenuated, but without the recorded reverberation deconvolved.

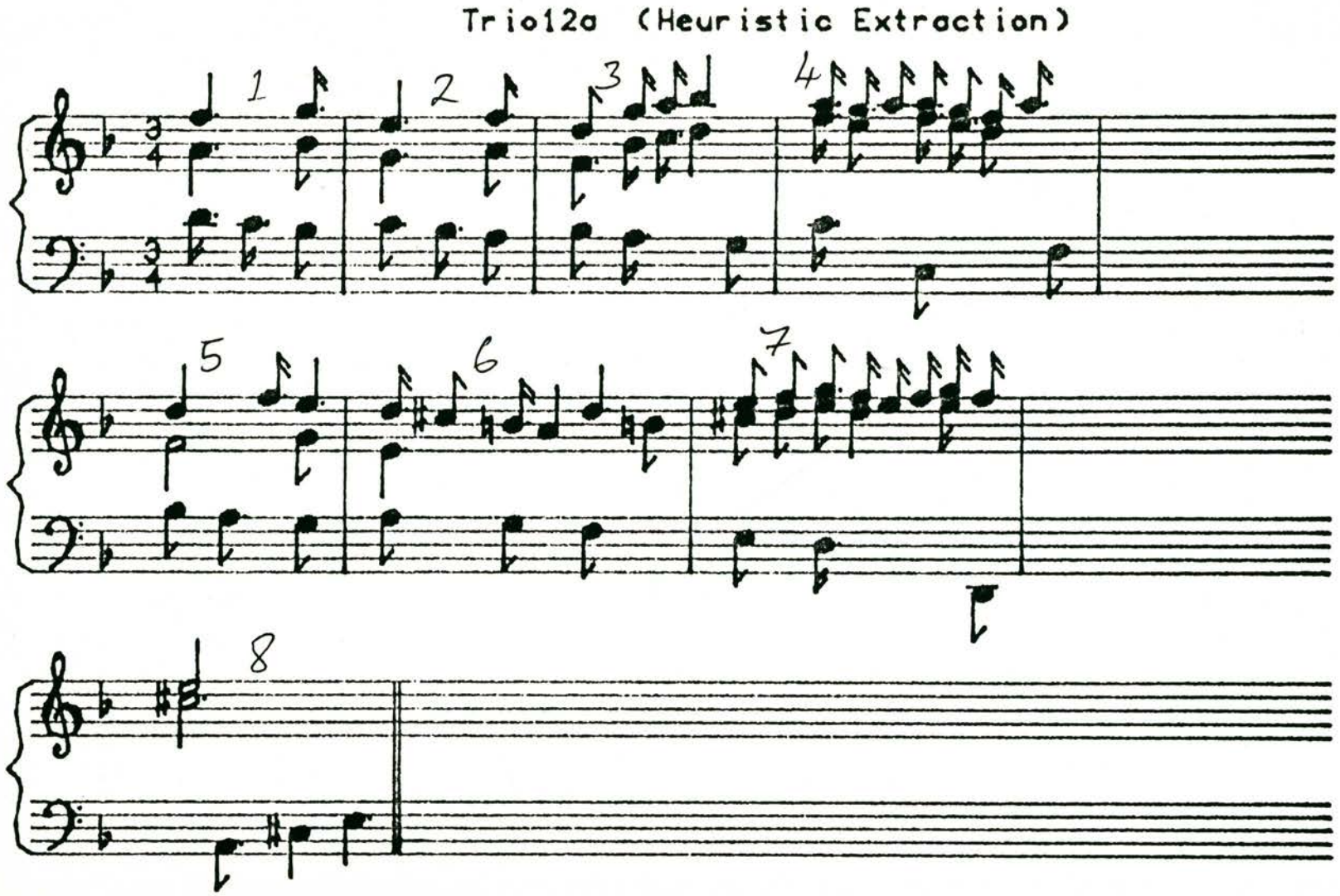


Figure 7.22 shows the music output for the Trio. Here, the general part analysis is used.

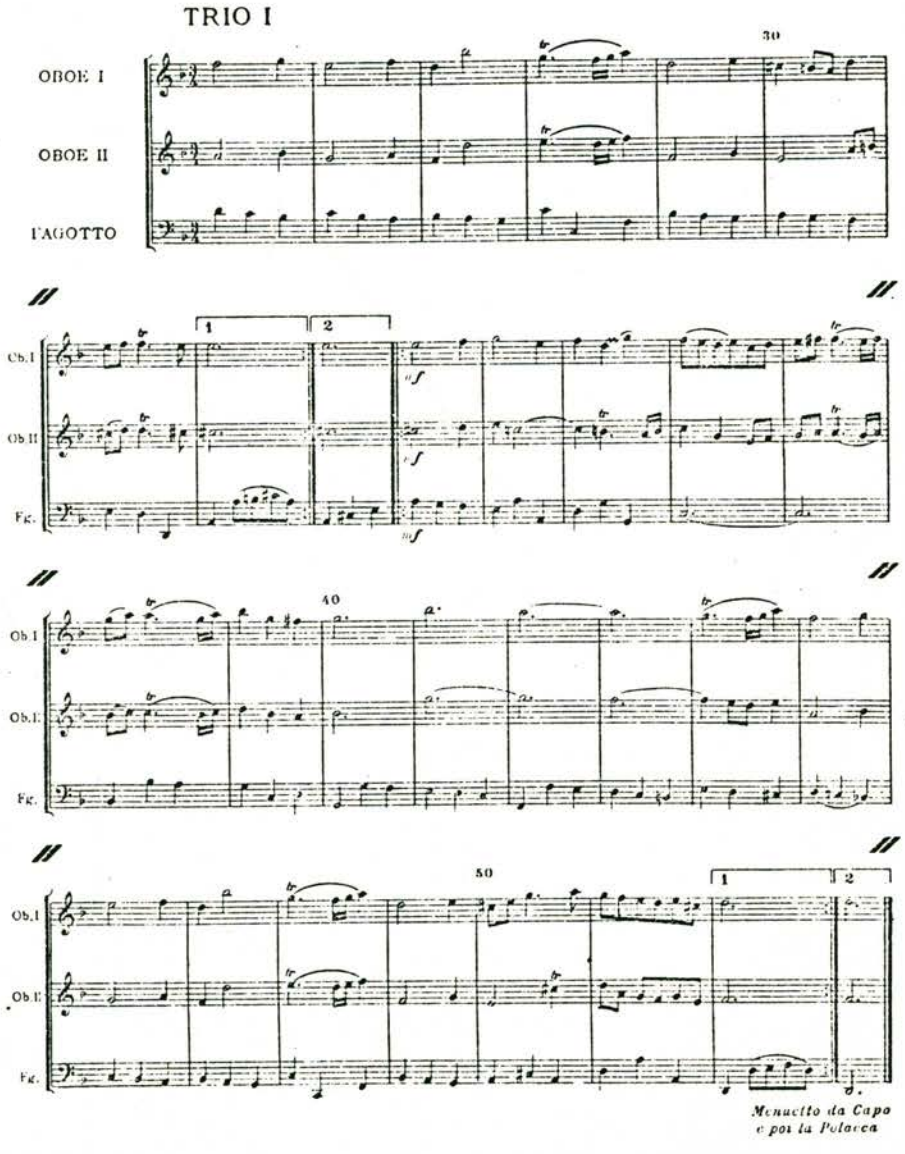
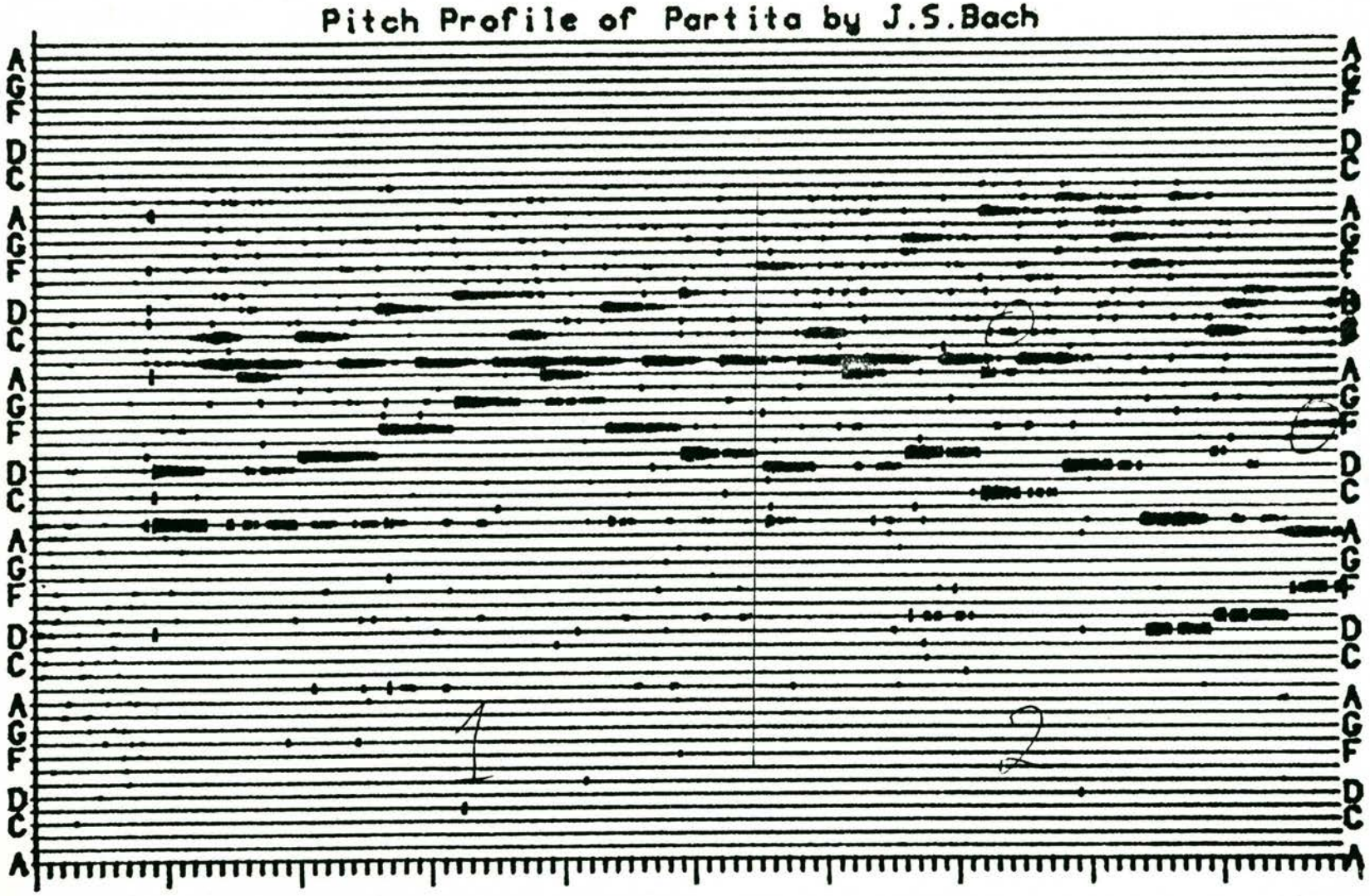


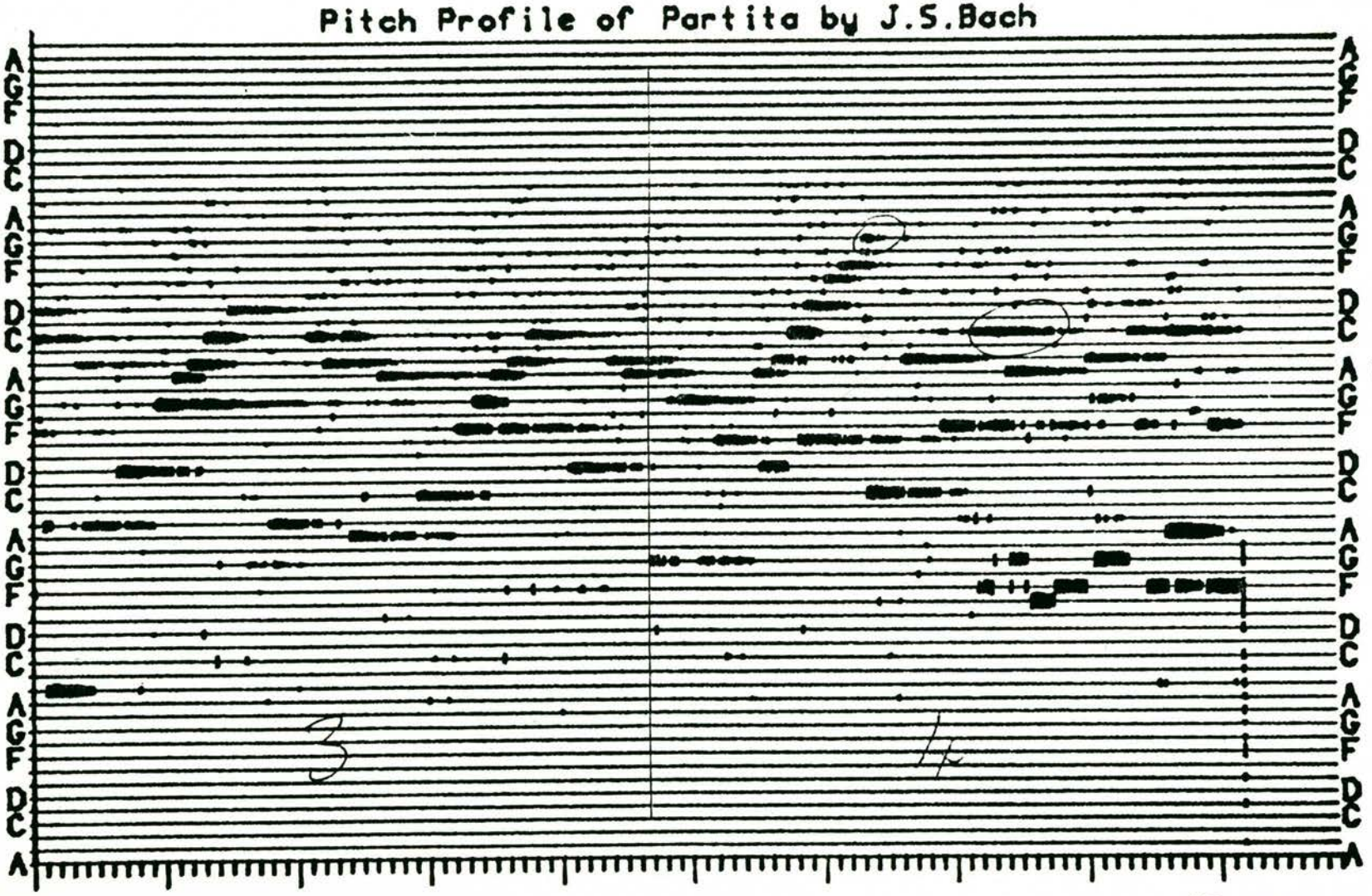
Figure 7.23 is the original written music of the Trio.

(from I Concerto Brandebourgeois, Heugel and Cie, France).



p36

Figure 7.24 plots the harmonic summing algorithm of the first 10 seconds of the Partita Prelude, with heuristics H1 to H5 applied, and with iterative extraction.



p36

Figure 7.25 plots the harmonic summing algorithm of the next 10 seconds of the Partita Prelude, with heuristics H1 to H5 applied, and with iterative extraction.

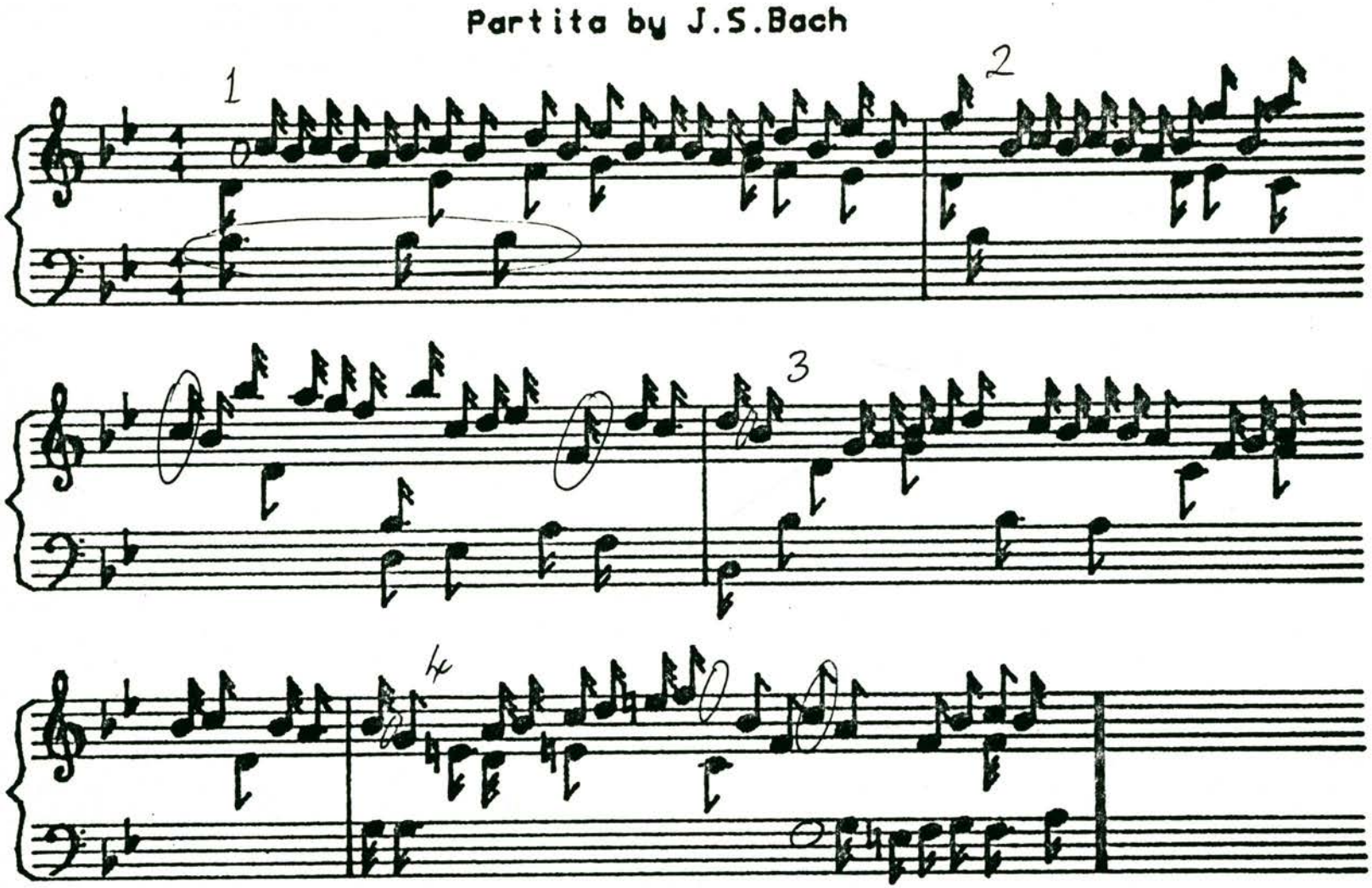


Figure 7.26 shows the music output for the Prelude.

PARTITA 1

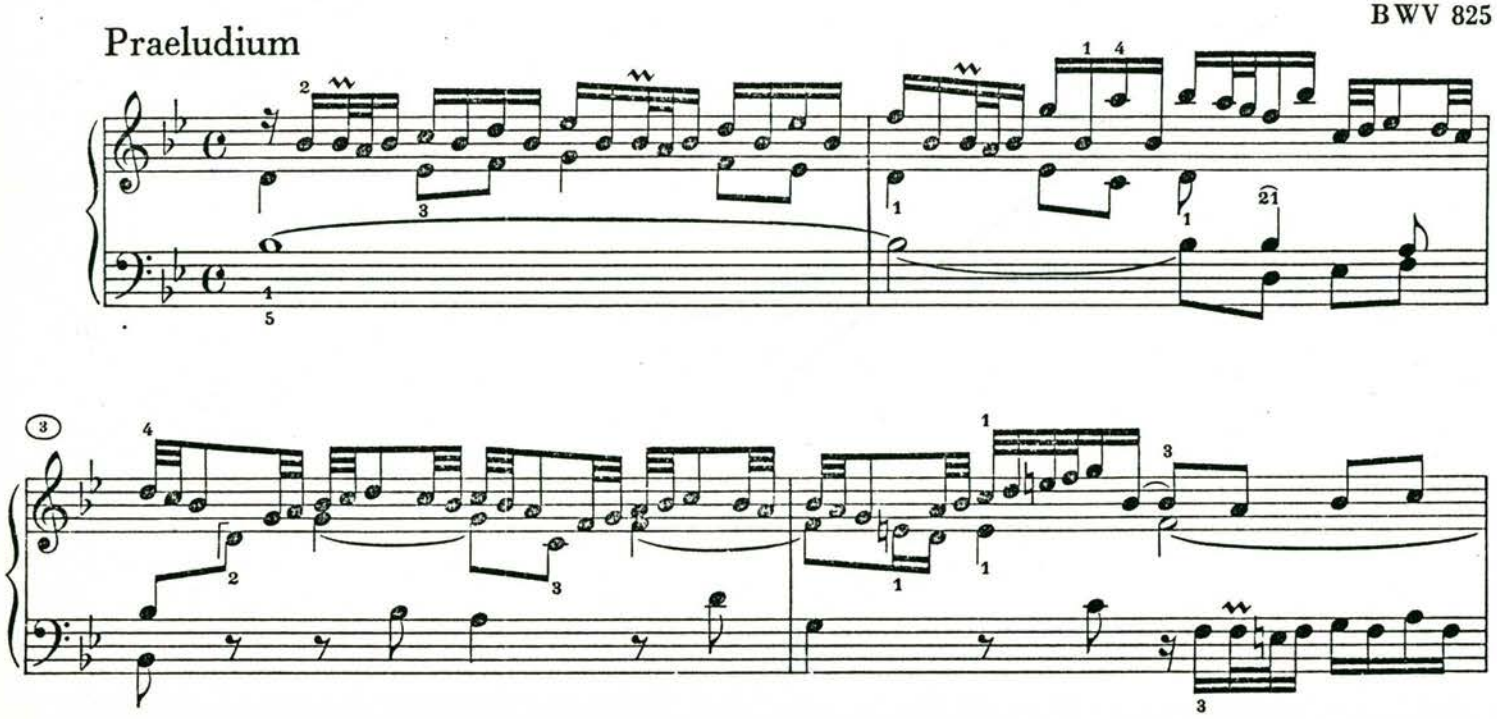
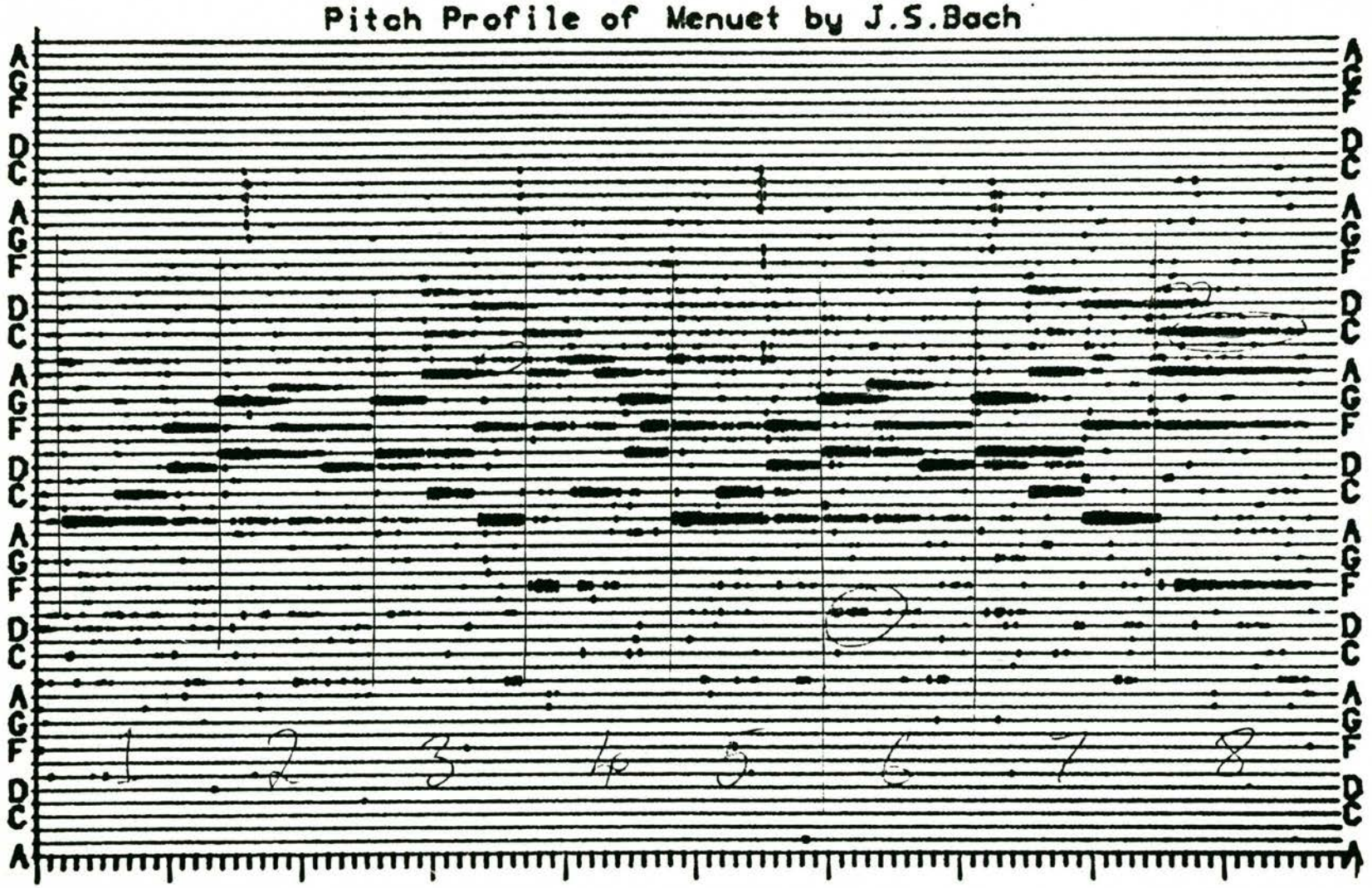


Figure 7.27 is the original written music of the Partita Prelude.
(from J.S. Bach, Partiten 1-3, Urtext Edition).



p36

Figure 7.28 displays the harmonic summing algorithm of the first 10 seconds of the Menuet, with heuristics H1 to H5 applied with iterative extraction.

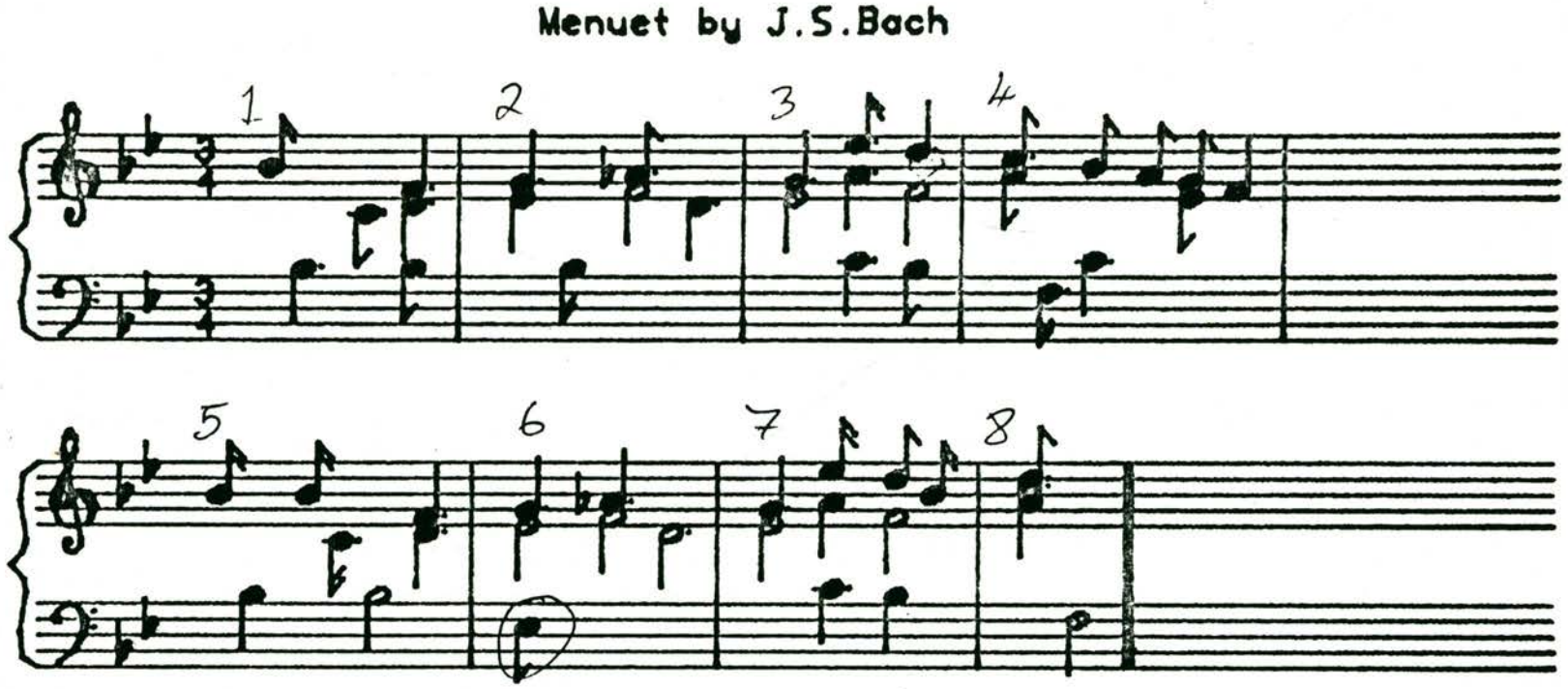


Figure 7.29 shows the music output for the first 10 seconds of the Menuet II.

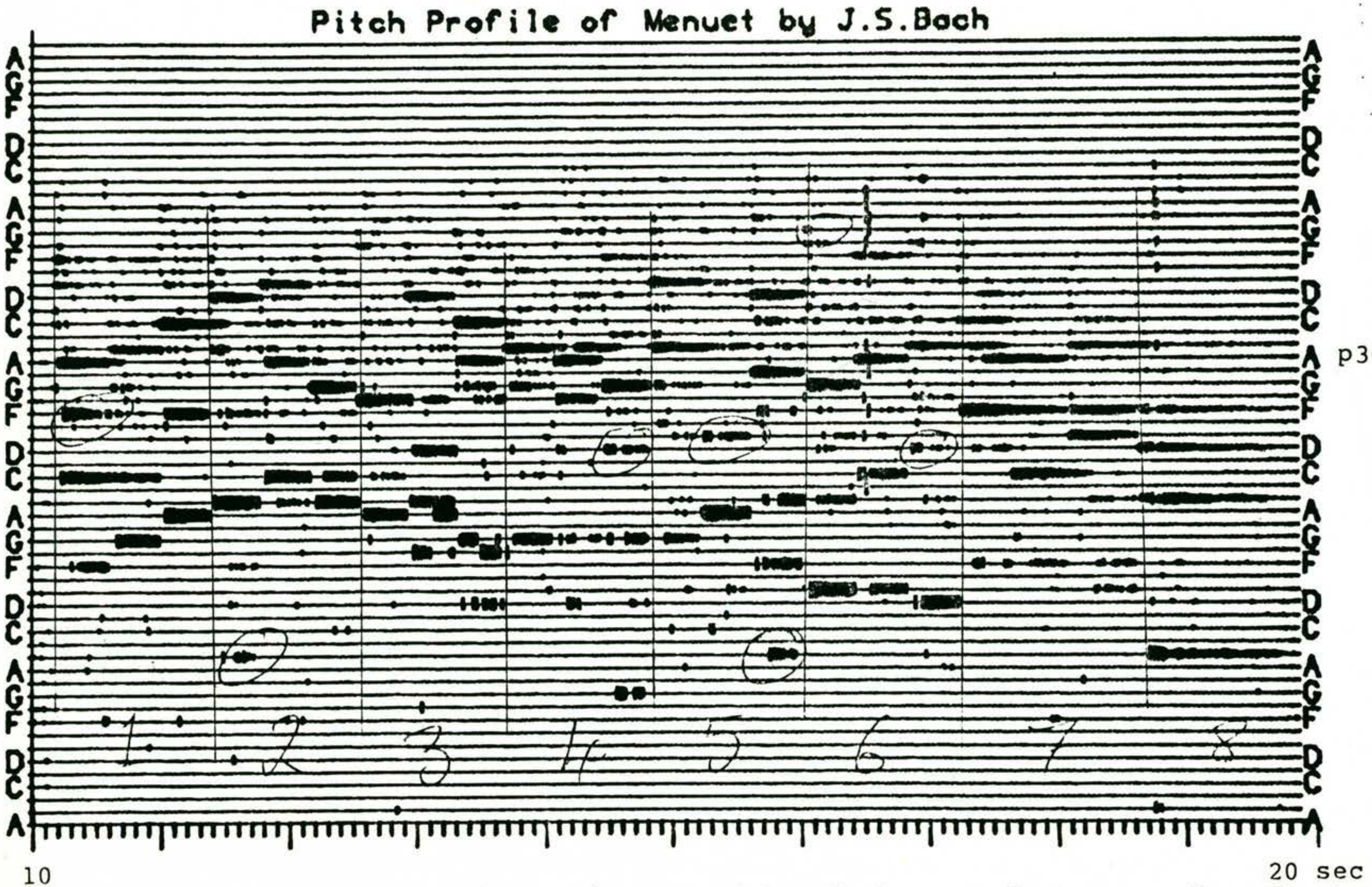


Figure 7.30 plots the harmonic summing algorithm of the second 10 seconds of the Menuet with heuristics H1 to H5 applied with iterative extraction.

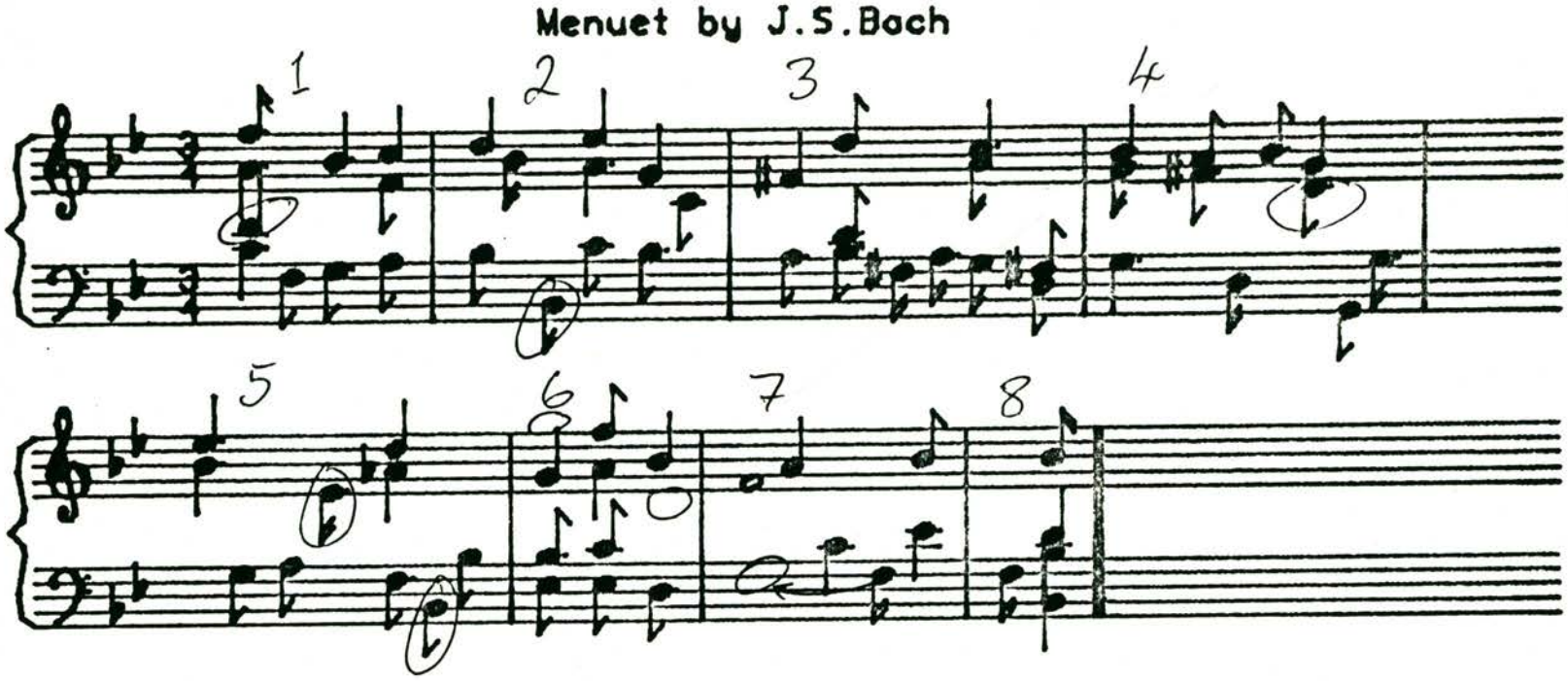


Figure 7.31 shows the music output for the next 10 seconds of the Menuet II.

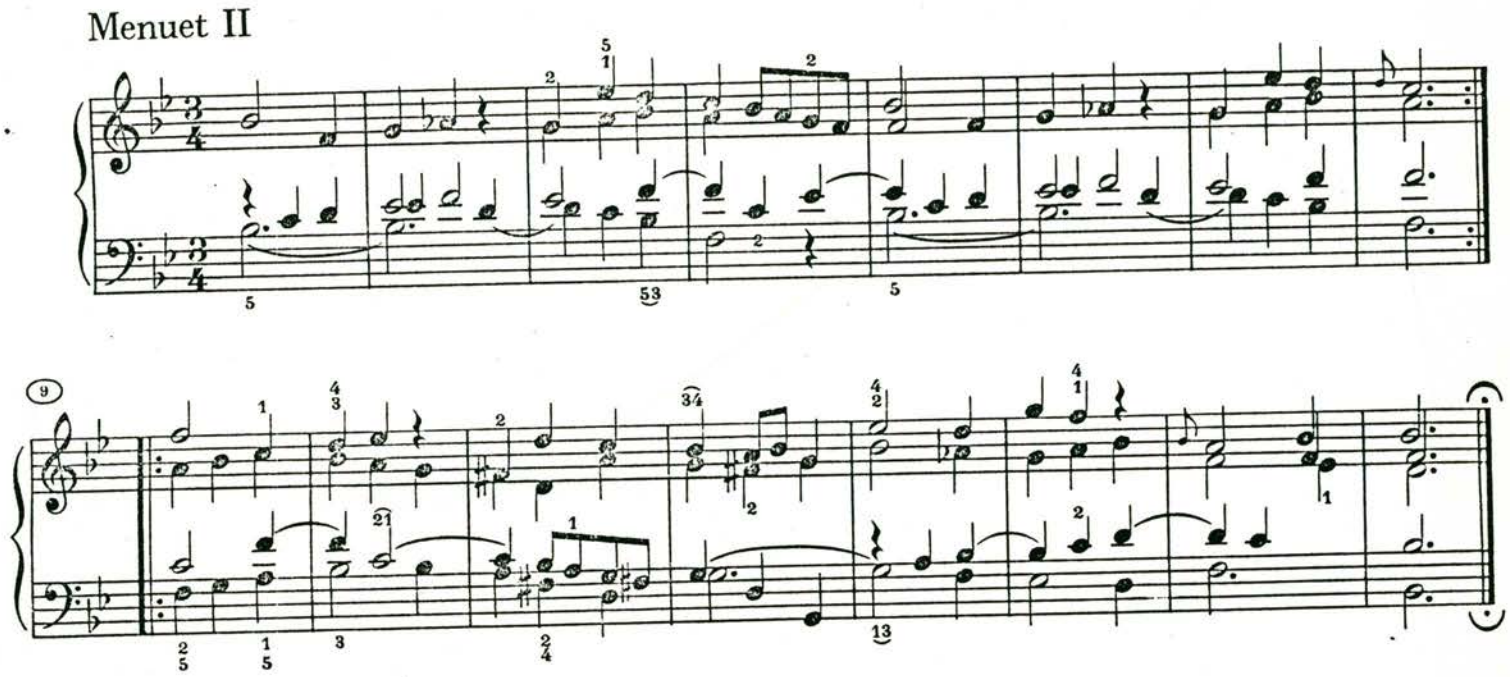
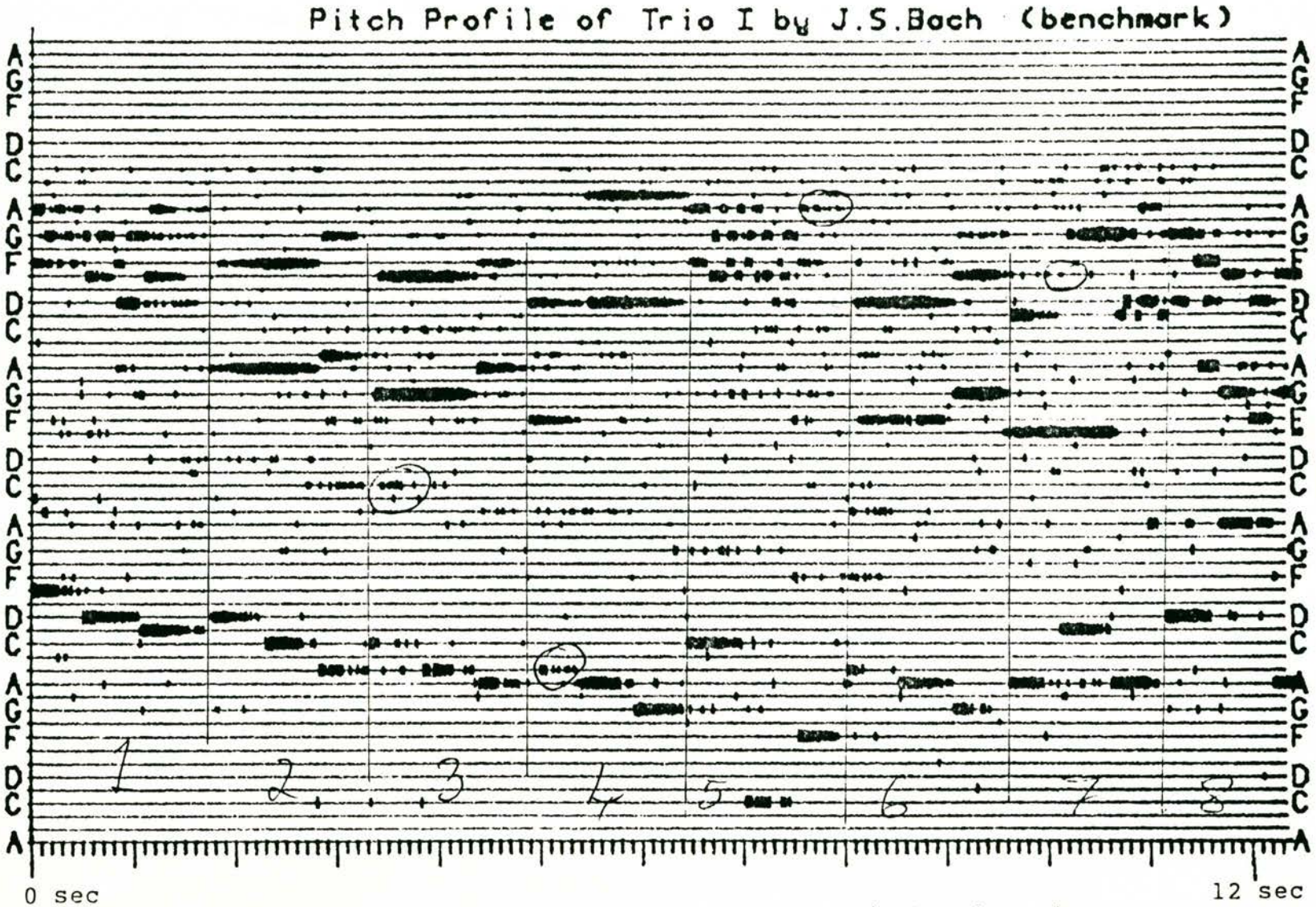


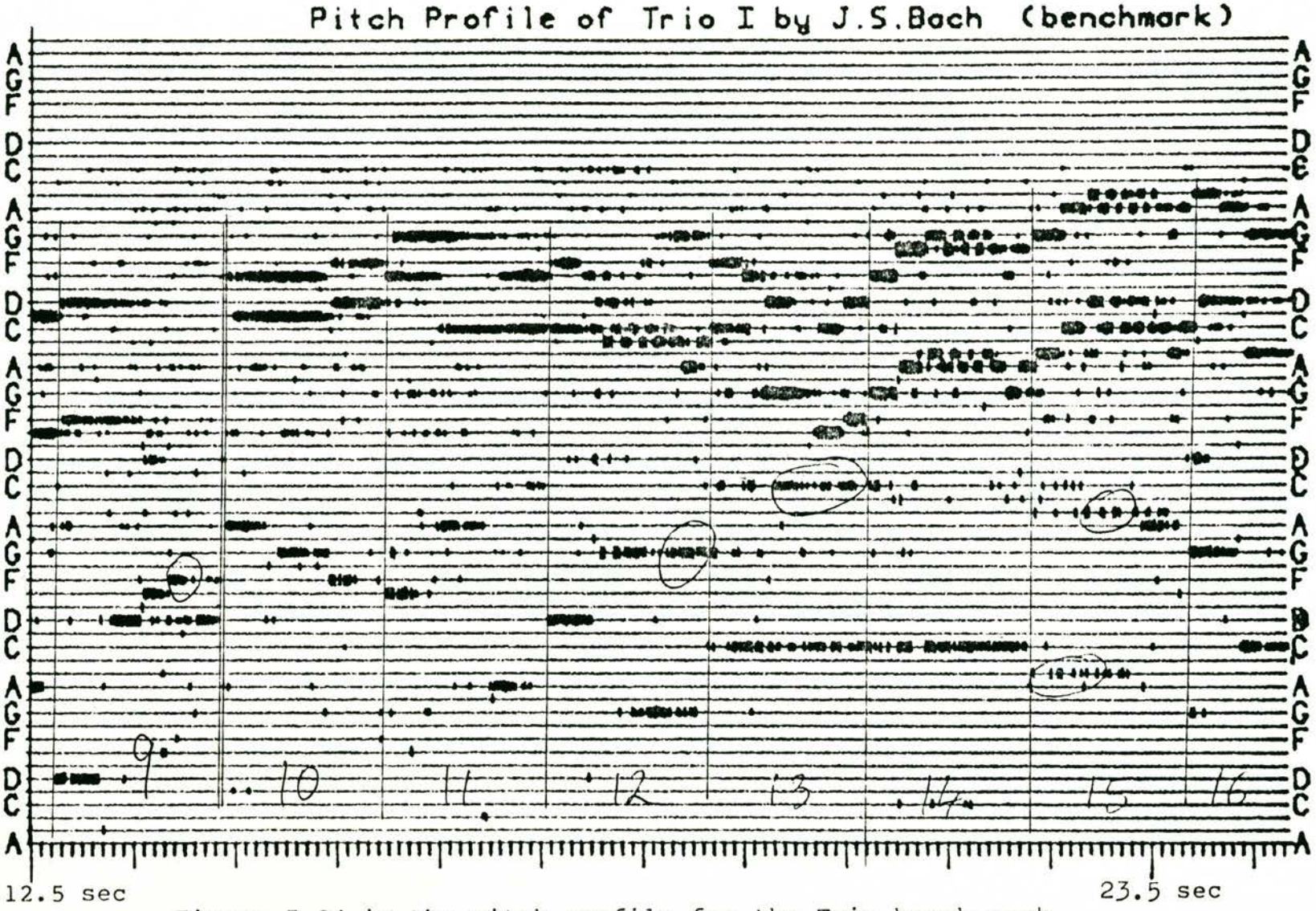
Figure 7.32 is the original written music of the Menuet.

from J.S. Bach, Partiten 1-3, Urtext Edition.



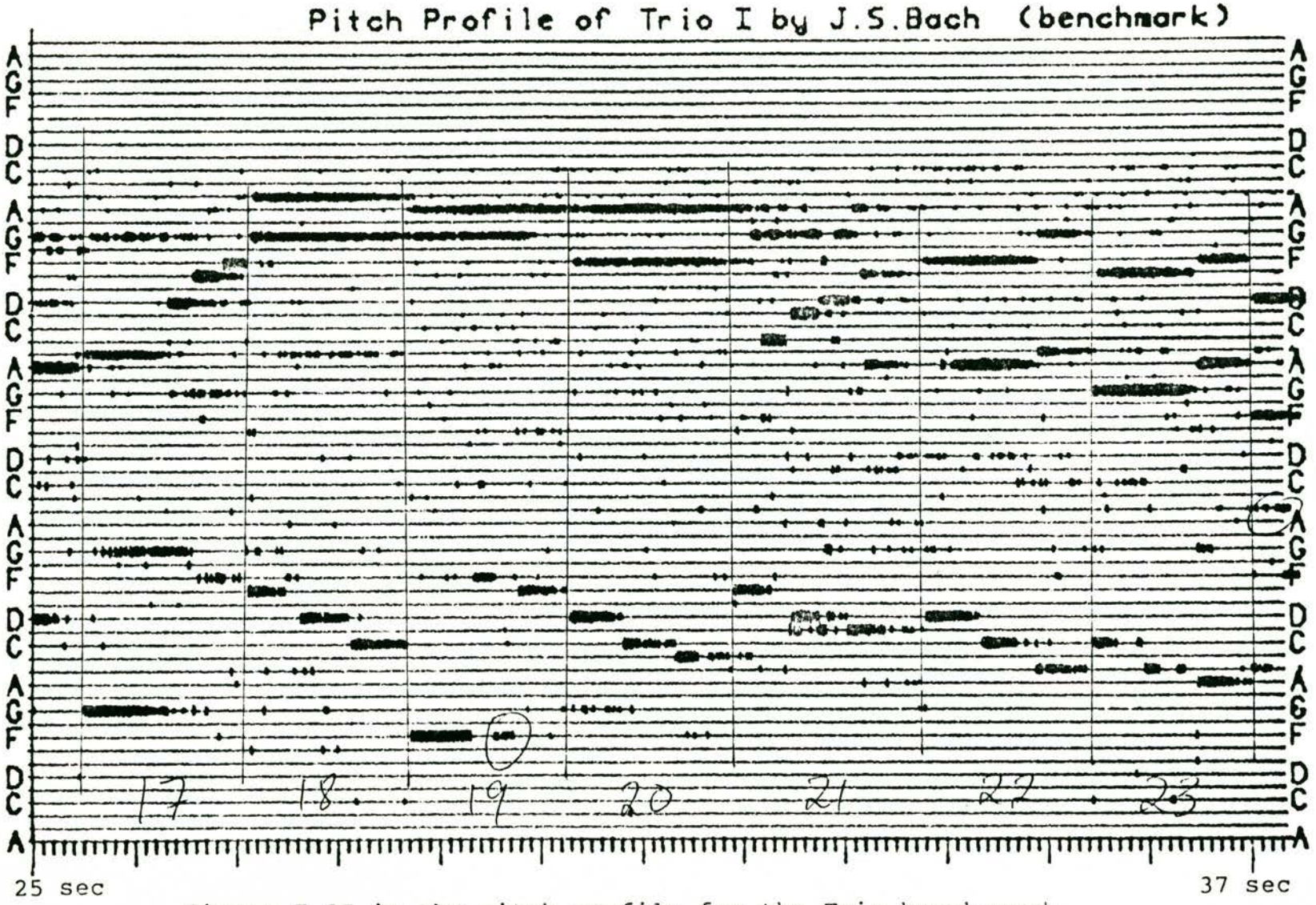
p36

Figure 7.33 is the pitch profile for the Trio bench-mark.



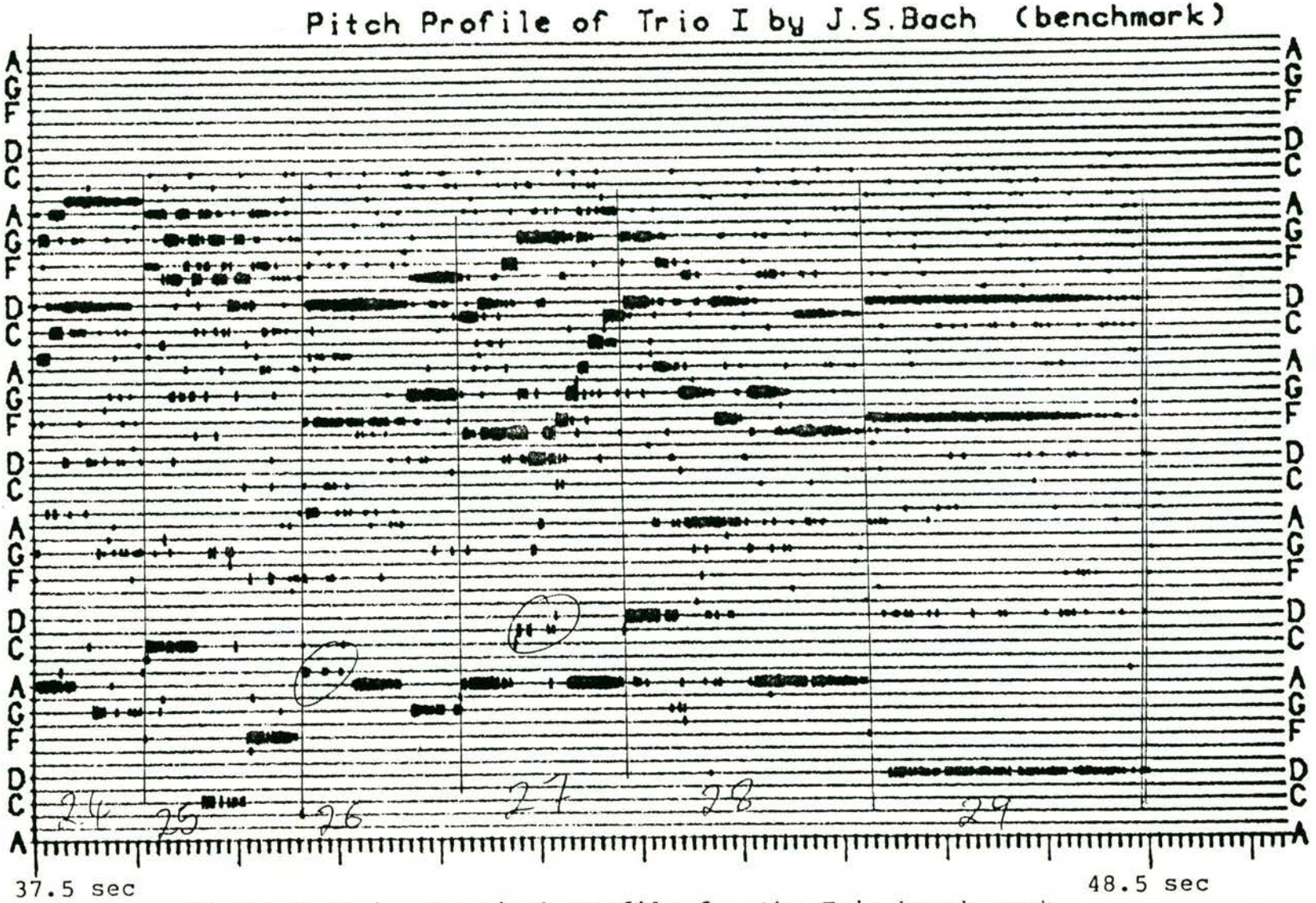
p36

Figure 7.34 is the pitch profile for the Trio bench-mark.



p36

Figure 7.35 is the pitch profile for the Trio bench-mark.



p36

Figure 7.36 is the pitch profile for the Trio bench-mark.

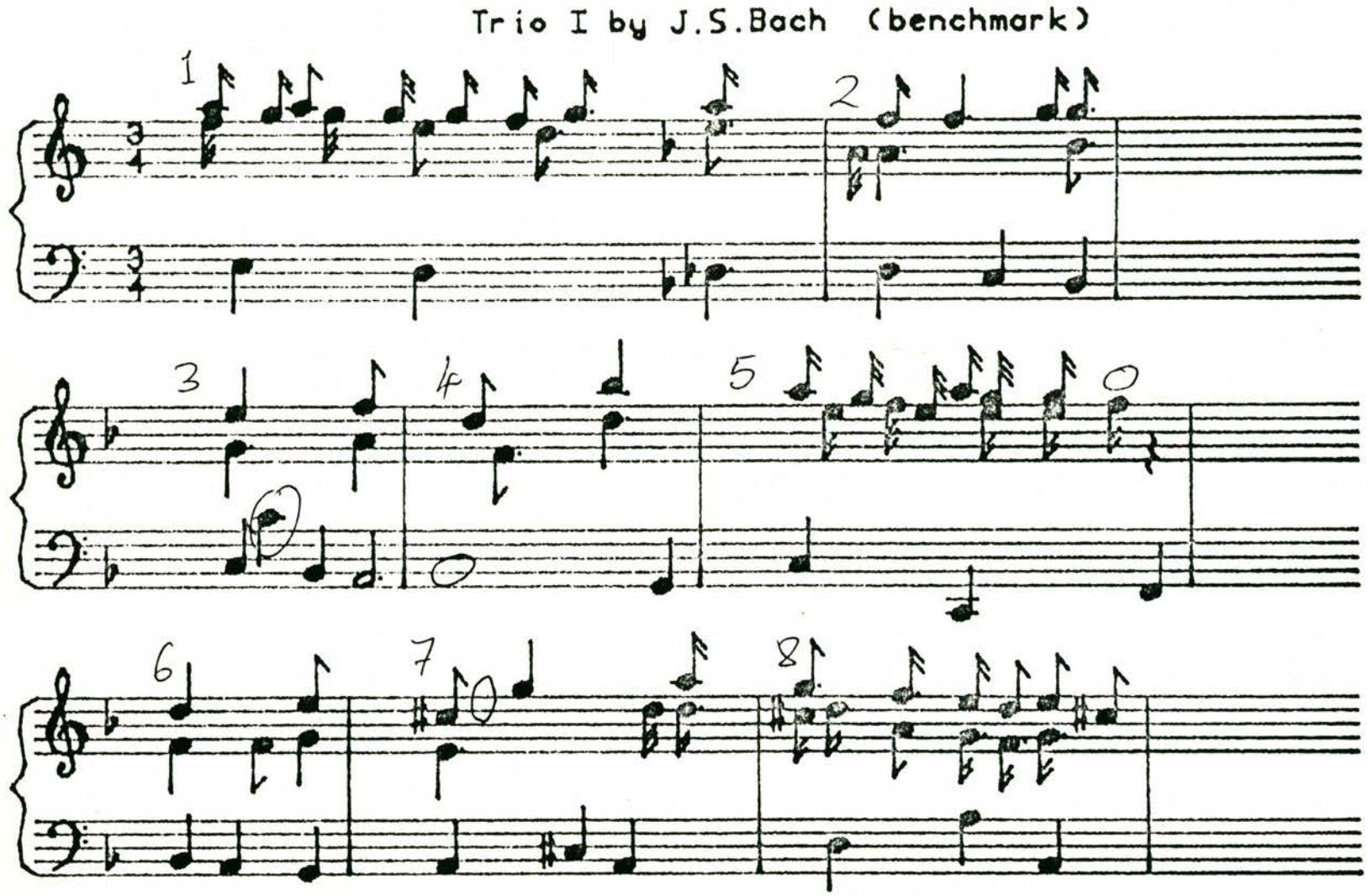


Figure 7.37 plots the music output for the benchmark (bars 1 to 8).

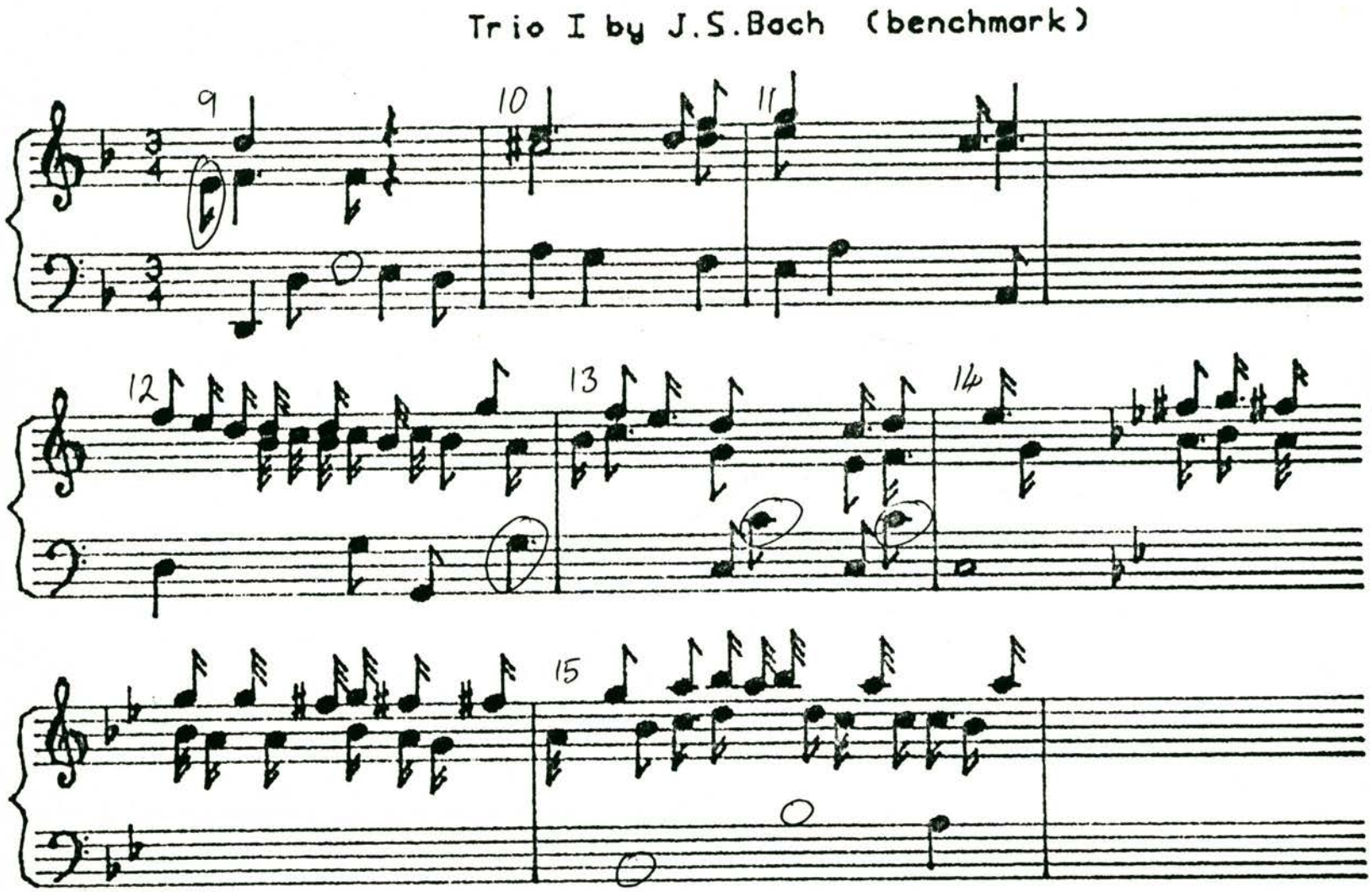


Figure 7.38 plots the music output for the benchmark (bars 9 to 15).

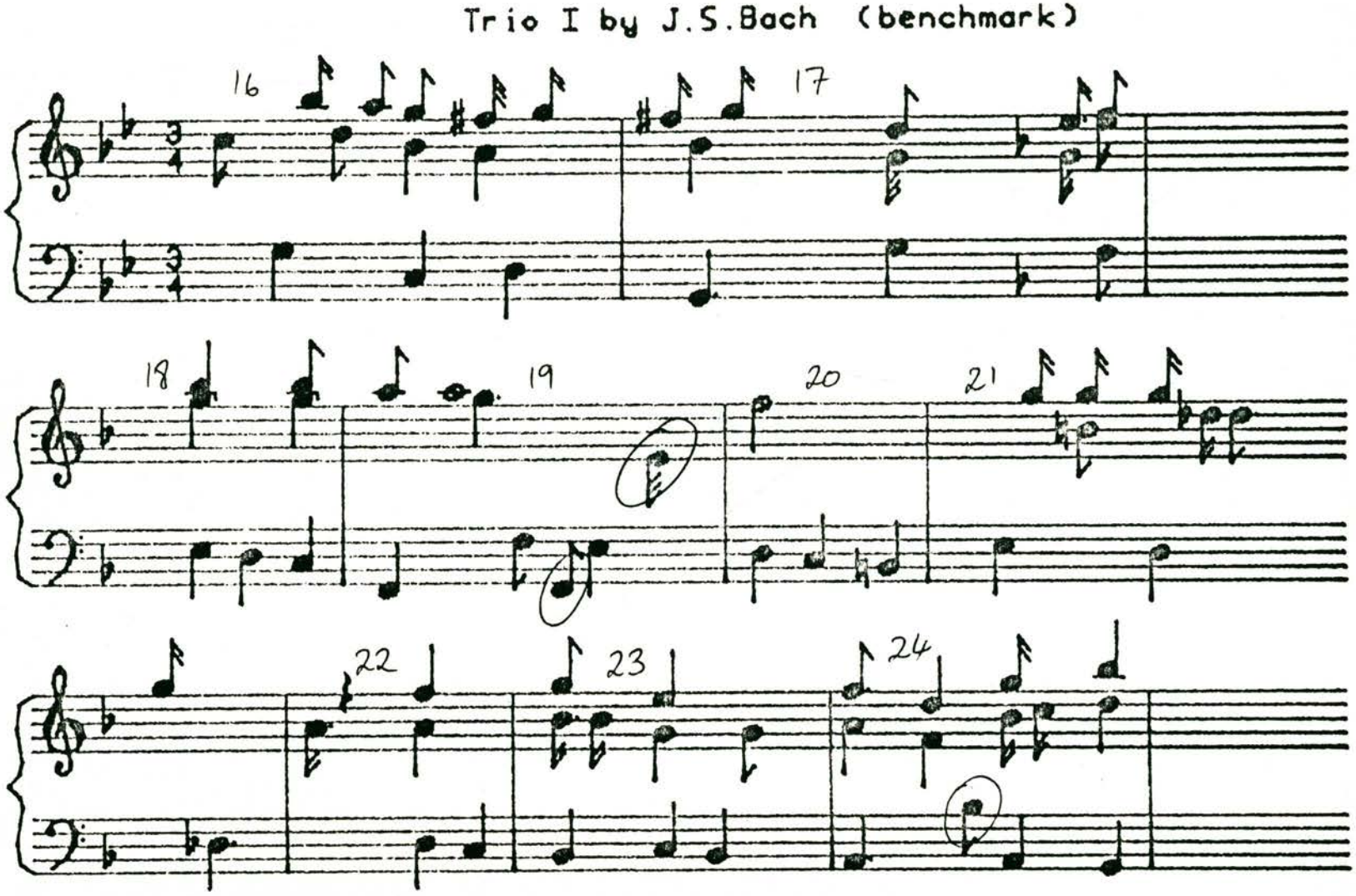


Figure 7.39 plots the music output for the benchmark (bars 16 to 24).

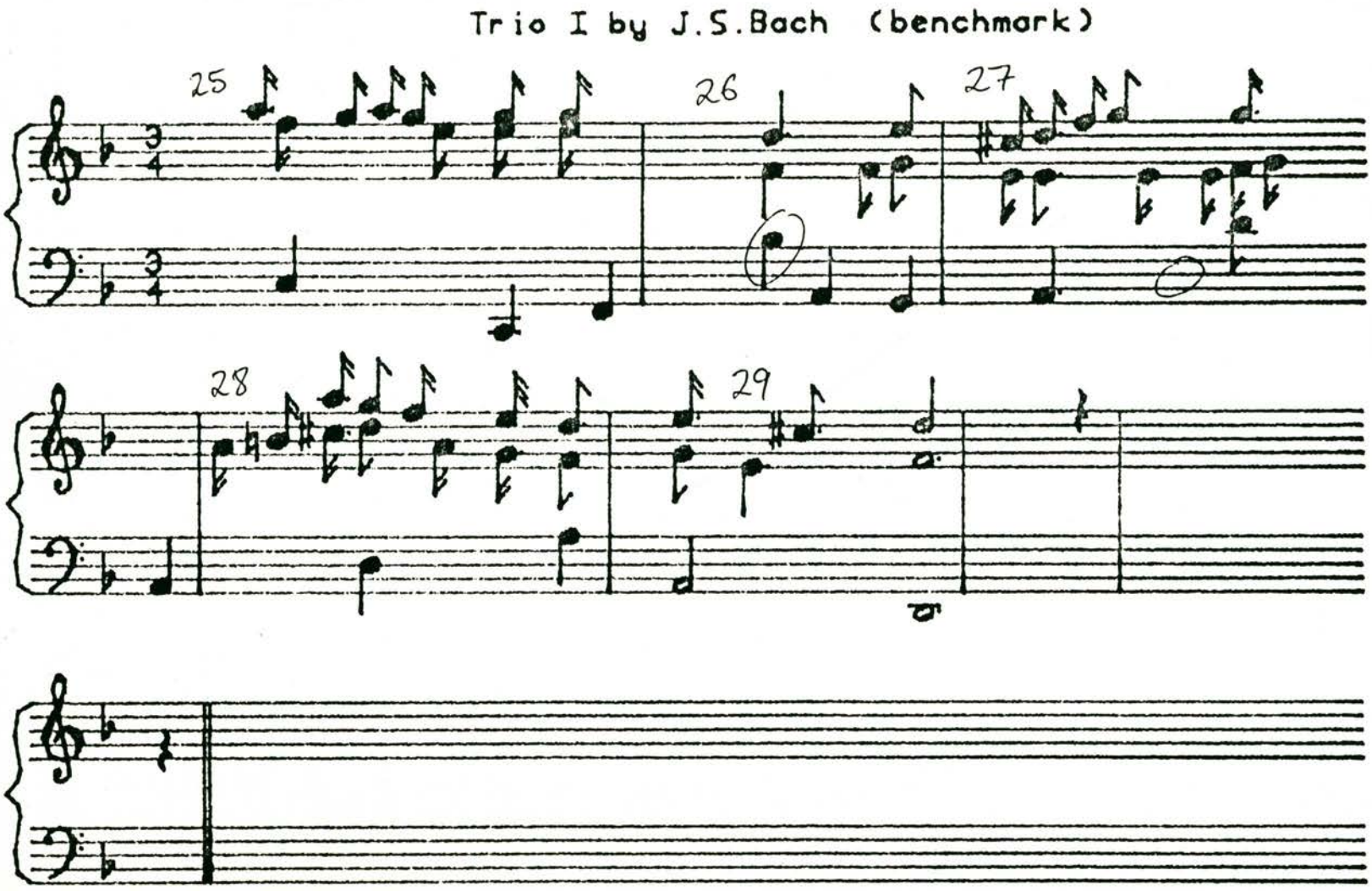


Figure 7.40 plots the music output for the benchmark (bars 25 to 29).

CHAPTER EIGHT

Error Measures for Comparing Music Analyses

8.1 Introduction

This chapter describes additional work done to establish the error rate of the music analysed in this thesis.

The problem here is to compare the musical events of a recording with the musical events produced by the analysis procedure. While visual comparison of recorded and analysed events shows a high coincidence, quantifying this similarity is non-trivial. Comparing musical events is complicated by the multi-dimensional nature of events; pitches, onset times and durations must all be compared. There are many ways of defining errors for an analysis, and no single error measure is adequate to compare analyses with the recorded music.

Section 8.2 defines some error measures and section 8.3 gives the values of these measures when applied to the music analysed in this thesis. Section 8.3.1 describes the verification of the woodwind Trio performance used as the Benchmark in section 7.5. The error measures are applied to analyses of the woodwind Trio (section 8.3.3) and synthesized music (section 8.3.4). Finally, section 8.4 presents a sensitivity analysis of the pitch determining algorithm.

8.2 Error Measures

An error measure determines the accuracy of an analysis by comparing recorded music with the analysis results. Here, measures are chosen to reflect

listeners' perceptions. Harmonically related errors are less perceptible than those which are harmonically unrelated; errors of short duration are less perceptible than those of long duration (Tobias 1970); onset errors are harder to hear than pitch errors; and errors in the finish time of notes are even less serious than onset errors (Knowlton 1971). Taking the difference in pitch as a measure of seriousness, for example, would be inappropriate; a tone played out of tune by one semitone would sound worse, to most listeners, than if it was played an octave away.

The error measures described in this chapter are applicable to any polyphonic music. They are applied here to the benchmark introduced in section 7.5; that is the final 48.5 seconds of Trio I from the *Brandenburg concerto* I, by J.S. Bach. The Trio has the form AABB. The benchmark performance begins (0 seconds) at the 12th bar of the first section B and finishes (48.5 secs.) at the end of the second section B. The repeat of section B occurs at 14.5 secs.

8.2.1 Terminology

A musical event is defined to have 3 attributes; a pitch, an onset time, and a duration. The pitch is the number of semitones above the musical note ***A*** of frequency 55Hz; onset and duration are specified in milliseconds. The onset, duration, and pitch of an event $\boldsymbol{X_i}$ are denoted $T(\boldsymbol{X_i})$, $D(\boldsymbol{X_i})$ and $P(\boldsymbol{X_i})$ respectively. The finish time of an event $\boldsymbol{X_i}$, is $T(\boldsymbol{X_i})+D(\boldsymbol{X_i})$.

Let the *performance* (denoted **X**) be the set $\{\boldsymbol{X_i}\}$ $(i = 1, \boldsymbol{N_x})$ of discrete musical events representing the analog recording. Let an *analysis* (denoted **Y**)

be the set $\{Y_i\}$ $(i = 1, N_y)$ of musical events identified by the pitch analysis algorithms given in sections 5.6 and 6.5.

An error measure E, is a function that compares **X** and **Y** and takes values in the range [0,1]. If all the events in the *analysis* are identical to those in the *performance*, then $E = 0$. Errors can be of two types; either events are in **Y** (the *analysis*) and not in **X** (the *performance*) (these are called inclusion errors) or they are not in **Y** but are in **X** (exclusion errors).

The purpose of an error measure is to provide a quantitative method of comparing analyses of musical recordings. It is therefore desirable to combine the inclusion and exclusion errors in some way.

Let N_y and N_x be the number of events in an *analysis* and *performance* respectively, and let L_y and L_x be the respective number of inclusion and exclusion errors. Then the inclusion error measure is

$$E_{incl} := \frac{L_y}{N_y}$$

and the exclusion error measure is

$$E_{excl} := \frac{L_x}{N_x}.$$

The following 3 combined measures are increasing functions of the number of inclusion and exclusion errors, which take the value zero when E_{incl} and E_{excl} are zero and the value 1 when E_{incl} and E_{excl} are 1.

$$E_{aver} := (E_{incl} + E_{excl})/2$$

$$E_{max} := max(E_{incl}, E_{excl})$$

$$E_{comb} := (L_x + L_y)/(N_x + N_y)$$

If E_{excl} is zero, and E_{incl} is near to 1, and N_y is large compared to N_x, then E_{aver} would take the value 0.5 while E_{max} and E_{comb} the value 1. This would occur if an *analysis* produced every possible note playing all the time. Every note in the *performance* would find a match in the *analysis*, and although there are no exclusion errors it is clear that the *analysis* is completely wrong so the combined error measure should have a value 1. The combined error measure E_{aver} is therefore unsuitable.

E_{max} is independent of the minimum of (E_{incl}, E_{excl}) and does not reflect variations in the lesser of E_{incl} and E_{excl}. E_{comb} is the combined error measure adopted in this thesis, because it reflects variations in the lesser of E_{incl} and E_{excl}, and has the value one in the degenerate cases where the number of events in the *analysis* (N_y) is very large or very small.

8.2.2 Error Measures Used in this Thesis

An error measure E depends on the criteria used to determine whether an element of $\mathbf{X}$ matches an element in $\mathbf{Y}$, and the weighting (or seriousness) of the error. These criteria can be combined in many ways. The Error measures used here are based either on the proportion of events that are in error, or on the proportion of time that the events are in error. The measures can also be qualified by considering only events with duration larger than some threshold; for example, the time resolution of the spectral analysis. The six error measures

described in sections 8.2.3 and 8.2.4 combine several of these criteria, and have been applied to the analyses in this thesis. They are all defined as inclusion, exclusion and combined error measures as described in the previous section.

For sections 8.2.3 and 8.2.4 the Overlap of events X_i and Y_j is defined as: $Overlap(X_i,Y_j) := \text{Min}((\text{T}(X_i)+\text{D}(X_i)),(\text{T}(Y_j)+\text{D}(Y_j)) - \text{Max}(\text{T}(X_i),\text{T}(Y_j))$, where $\text{T}(X_i)$ is the start time and $\text{D}(X_i)$ is the duration of event X_i.

8.2.3 Time Based Error Measure

$E1$:

$E1$ is the proportion of time that **X** and **Y** fail to coincide in pitch, P. Define the matching function as:

$$M(X_i,Y_j) := \begin{cases} 1, & \text{when } Overlap(X_i,Y_j) > 0, \text{ and } \text{P}(X_i) = \text{P}(Y_j); \\ 0, & \text{otherwise.} \end{cases}$$

The inclusion error for $E1$ is defined as:

$$E1_{incl}(X,Y) := 1 - \frac{\sum_{i=0}^{N_x}\sum_{j=0}^{N_y} M(X_i,Y_j)Overlap(X_i,Y_j)}{\sum_{j=0}^{N_y} Overlap(Y_j,Y_j)}.$$

The exclusion error for $E1$ is defined as:

$$E1_{excl}(X,Y) := 1 - \frac{\sum_{i=0}^{N_x}\sum_{j=0}^{N_y} M(X_i,Y_j)Overlap(X_i,Y_j)}{\sum_{i=0}^{N_x} Overlap(X_i,X_i)}.$$

The combined error for $E1$ is defined as:

$$E1_{comb}(X,Y) := 1 - \frac{2\sum_{i=0}^{N_x}\sum_{j=0}^{N_y} M(X_i,Y_j)Overlap(X_i,Y_j)}{\sum_{j=0}^{N_y} Overlap(Y_j,Y_j) + \sum_{i=0}^{N_x} Overlap(X_i,X_i)}.$$

8.2.4 Event Based Error Measures

Each of $E2, E3, E4, E5$ and $E6$ measure the error as the number of events that match in proportion to the total number of events. They differ by the matching criteria used. For $E5$, the events are weighted by their duration. The inclusion, exclusion and combined error measures for $E2, E3, E4, E5$, and $E6$ are defined as:

$$E_{incl}(X,Y) := 1 - \frac{\sum_{i=0}^{N_y} M_{incl}(Y_i)W(Y_i)}{\sum_{i=0}^{N_y} W(Y_i)},$$

$$E_{excl}(X,Y) := 1 - \frac{\sum_{i=0}^{N_x} M_{excl}(X_i)W(X_i)}{\sum_{i=0}^{N_x} W(X_i)},$$

$$E_{comb}(X,Y) := 1 - \frac{\sum_{i=0}^{N_x} M_{excl}(X_i)W(X_i) + \sum_{i=0}^{N_y} M_{incl}(Y_i)W(Y_i)}{\sum_{i=0}^{N_x} W(X_i) + \sum_{i=0}^{N_y} W(Y_i)},$$

where

$$(1) M_{incl}(Y_j) := \begin{cases} 1, & \text{when } \exists X_i, \text{ such that } M(X_i, Y_j) = 1; \\ 0, & \text{otherwise.} \end{cases}$$

$$(2) M_{excl}(X_i) := \begin{cases} 1, & \text{when } \exists Y_j, \text{ such that } M(X_i, Y_j) = 1; \\ 0, & \text{otherwise.} \end{cases}$$

(3) the weighting function is defined as:

$$W(X_i) := \begin{cases} D(X_i), & \text{for error measure } E5 \text{ and} \\ 1, & \text{for error measures } E2, E3, E4 \text{ and } E6. \end{cases}$$

The matching functions $M(X_i,Y_j)$ for $E2$ to $E6$ follow.

$E2$:

Here,

$$M(X_i,Y_j) := \begin{cases} 1, & \text{when } \mathit{Overlap}(X_i,Y_j) > 0,\ \mathrm{P}(X_i) = \mathrm{P}(Y_j), \\ & \quad [\mathrm{T}(Y_j) - \mathrm{T}(X_i)] < \text{onset\_limit}, \\ & \quad [\mathrm{T}(Y_j)+\mathrm{D}(Y_j) - \mathrm{T}(X_i)-\mathrm{D}(X_i)] < \text{finish\_limit}; \\ 0, & \text{otherwise.} \end{cases}$$

where onset_limit and finish_limit are the tolerances in determining the onset and finish times of tones.

$E3$:

For $E3$,

$$M(X_i,Y_j) := \begin{cases} 1, & \text{when } \mathit{Overlap}(X_i,Y_j) > 0,\ \mathrm{P}(X_i) = \mathrm{P}(Y_j); \\ 0, & \text{otherwise.} \end{cases}$$

$E3$ is a special case of $E2$ with onset_limit and finish_limit set to infinity. This means that the start and finish times of an event Y_j in the *analysis* can differ by any amount from those of an event X_i in the *performance*, but the events still match, provided the pitches are equal and the overlap is positive.

$E4$:

This has the same as matching criteria as E3, but only applies to events with duration greater than some limit, typically 100 milliseconds.

$E5$:

$E5$ uses the same matching criteria as $E3$ but differs in the weighting of errors.

$E6$:

Here,

$$M(X_i, Y_j) := \begin{cases} 1, & \text{when } \mathit{Overlap}(X_i, Y_j) > 0,\ \mathrm{P}(X_i) = \mathrm{P}(Y_j) \text{ modulo } 12; \\ 0, & \text{otherwise.} \end{cases}$$

An $E6$ match allows events that are several octaves apart. That is, two events match if they overlap and their pitches differ by a multiple of octaves.

8.2.5 Comparison of Error Measures

$E1$ determines the proportion of time that the events in the *analysis* and *performance* do not match, while $E2$ to $E6$ determine the number of events that match, as a proportion of the total number of events.

$E2$ requires that the difference in onset and finish times of the compared events be within the accuracy of the *performance* times and the time resolution of the spectral analysis. $E3$, $E4$, $E5$, and $E6$ measure the accuracy of pitch, but not that of onset or duration; however, a positive overlap is required to prevent the matching of events with the same pitch that occur at different times. They are more generous than the $E2$ measure, but match musicians' tolerance to errors in onset and finish times.

Consider the example where a single event of 2 seconds duration is recorded. Suppose the *analysis* gives an event with correct pitch but with a

duration of 1 second and an onset time 1 second late. This can occur if some other sound masks the acoustical signal, or if the *analysis* selects an incorrect pitch at the onset of the event. The $E2$ inclusion and exclusion measures would reject the events in the *analysis*, while all the other error measures would accept the events.

More generally, if the onset and finish times of the analysed events differ from those of the recorded events, then the $E1$ error will increase in proportion to the difference, but the other errors will remain the same, provided the matching functions for all events do not change value.

For $E4$ and $E5$, erroneous events of large duration are more significant than those of small duration. A listener may not hear the difference when a single note of 100 milliseconds duration is absent from a trill; the omission of a note of 1 second duration would be more perceptible. $E4$ applies only to events with durations greater than some limit, typically 100 milliseconds (the durations of all the notes in J.A. Moorer's analyses were greater than 200 milliseconds). The notes in the trills of the Trio *performance* are typically 100 milliseconds in duration and they account for half the $E3$ exclusion errors in the woodwind Trio *analysis*.

$E6$ matches events with pitches differing by a multiple of octaves. A listener would notice a discordant error more than a harmonically related error.

8.3 Automated Comparison of Analyses

A comparison program was written to determine the error rates for

the analyses, and the sensitivity of the analysis algorithm to variations of the parameters.

8.3.1 Verification of the Performance

The determination of onset and finish times in real music is a difficult problem. Tones can take several hundred milliseconds to attain a steady state. Sometimes the steady state is never attained. Tones can start with low signal level, and considerable noise. Some tones can change continuously in pitch to another tone, with no transients. Onset times of tones can be determined, by a human listener, more accurately than their finish times (Knowlton 1971). Also reverberation, gradual attenuation of the end of the note, and the masking by other notes makes the determination of the finish times more difficult. For the Trio, all the instruments play continuously except for breaths and phrasing. Therefore, for simplicity, the finish time of a note is taken to be the starting time of the next note, for each instrument.

The start and finish times of the *performance* were determined by listening to the recording, timing events with a stop-watch, comparing the recorded music with the written music, and using computer analyses.

The pitch of the events in the *performance* was verified by listening to the music. William James (Mus. Bach. Adel.), organist and choir master at St. James Church, Melbourne, listened to the recording repeatedly and attests to the accuracy of the notes and their pitches.

The onset times of the *performance* events were verified by stop-watch.

A further computer analysis of the pitch estimates was done; onset times were only assigned if 4 consecutive estimates of increasing strength occurred. This criterion was tested by stop-watch comparisons of the difference between a sample of onset times. This accounted for onset times for 81% of the notes in the *performance*. The onset times of the remaining notes were determined by stop-watch comparisons for notes of long duration.

It is difficult to hear the onset times of the notes in rapidly changing music, such as trills. Therefore, these times were determined by linear interpolation between known onset times. The duration of these events is less than 100 milliseconds, therefore the error from interpolating is less than this.

A stop-watch with 10 milliseconds resolution was used to determine the time difference between the onset times of notes. The recorded segment was played repeatedly, so that the observer could anticipate the onset of the events he was comparing. The variation of the measured onset times was typically 30 to 50 milliseconds.

The accuracy of determining the onset times of the *performance* from the recording is typically 50 milliseconds, and in the worst case is 150 milliseconds.

8.3.2 Differences Between the Performed and Written Music

It is interesting to compare the *performance* with the set of events representing the written music, although only the *performance* is used for the comparison of analyses in this work. The notes of the last 2 bars are played

later and for longer duration than specified by the written music (it is performed ritardando). Table 8.1 shows the passing notes, and the grace notes performed by the oboes, that are not in the written music. The bars and notes of the written music refer to figure 7.23 and the times of the *performance* refer to the horizontal axis of figures 8.1a, b, and c.

Table 8.1

Comparison of Performed and Written Trio

Performance time	Written Music part	line	bar	
25.0 trill on $F\sharp$ and G	Ob.I	3	2	$F\sharp$
25.6 shortened $B\flat$,				
26.4 passing notes D, E, F	Ob.II	3	3	$B\flat$
32.3 $B, C\sharp, D$	Ob.II	3	7	F, E, D
37.1 shortened D,				
37.7 passing notes G, A	Ob.I	4	2	D
37.1 shortened F,				
37.7 passing notes $B\flat, C$	Ob.II	4	2	F
41.8 glissando to D	Ob.II	4	5	trill on $C\sharp$ and D
42.0 passing notes D and F	Ob.I	4	5	$C\sharp, E, G$

8.3.3 Error Measures for the Analysis of the Woodwind Trio

The regular rhythm and moderate dynamic range of the woodwind Trio constrains some of the problems of analysing music, while providing a challenging example to test the pitch estimation algorithms in the face of reverberation, instrument resonances, and trills. It is therefore a good test for the algorithms described in sections 5.6 and 6.5.

Table 8.2 lists the error measures for the woodwind Trio analysis. The $E5$ error rates are less than the $E3$ rates, because many of the errors are for

notes of short duration. For completeness, the *E*3 errors are listed in tables 8.3a and 8.3b. Figures 8.1a 8.1b and 8.1c show the comparison of the woodwind Trio analysis (above the lines) with the transcription described in 8.3.1 (below the lines). Markings above the pitch lines, but not below, correspond to the errors of Table 8.3a. Markings below the lines, but not above, correspond to the errors of Table 8.3b. The pitch estimate profiles for the woodwind analysis are shown in figures 7.33 to 7.36 in the previous chapter.

Table 8.2

Comparison of Error Measures for the Woodwind Analysis

	inclusion	exclusion	combined
E1	12.0%	27.2%	19.6%
E2 (limits=100)	39.6%	47.4%	43.5%
E2 (limits=200)	24.7%	32.3%	28.6%
E2 (limits=300)	17.7%	26.0%	21.8%
E3	3.8%	15.4%	9.7%
E4 (duration>100)	3.1%	8.7%	6.0%
E4 (duration>200)	4.5%	7.5%	6.0%
E5	2.9%	7.5%	5.2%
E6	0.6%	11.7%	6.1%

Table 8.3a

E3 Inclusion Errors for the Woodwind Analysis

(* shows estimates that are also E6 errors)

Bassoon			Oboe2			Oboe1
time	dur	pitch	time	dur	pitch	(no E3 errors)
3566	530	27				
18957	795	22				
19953	264	27				
20550	264	27	26388	65	34	
			26720	65	34	
29705	264	8	30036	65	31	
37401	264	25				
40320	530	25				
*42708	264	29				

Table 8.3b
E3 Exclusion Errors for the Woodwind Analysis
(* shows estimates that are also E6 errors)

Bassoon			Oboe2			Oboe1		
time	dur	pitch	time	dur	pitch	time	dur	pitch
						*600	70	48
*4970	525	13	*6520	175	44	*6790	90	48
						*6955	95	48
						*7050	75	46
			*7330	70	41	*7350	75	44
						*7605	555	48
						9900	310	43
			10720	60	40			
			*10890	60	40			
13940	255	20				*18165	80	43
						*18320	60	43
						*18480	70	43
						*18550	270	44
						20060	245	43
						*21730	70	45
			*21765	60	37	*22150	120	43
22350	550	13						
22900	600	25	23100	125	39	*23220	65	48
			*23370	30	41	*23415	90	49
			*23490	65	41	*23545	85	49
			*23605	50	41	*23765	85	46
						*25180	70	45
						*25360	90	45
						*25450	50	43
						*32260	90	48
						*32465	70	48
						*32665	55	48
						*32875	65	44
						33190	835	48
			*33270	705	43			
37130	520	13				37775	150	48
			*38920	180	43	39240	130	48
			*39305	75	44	*39480	105	44
			*39525	95	41	*39755	565	48
			39865	455	44			
40300	490	13				42130	100	43
*42350	580	16	*42510	145	29			

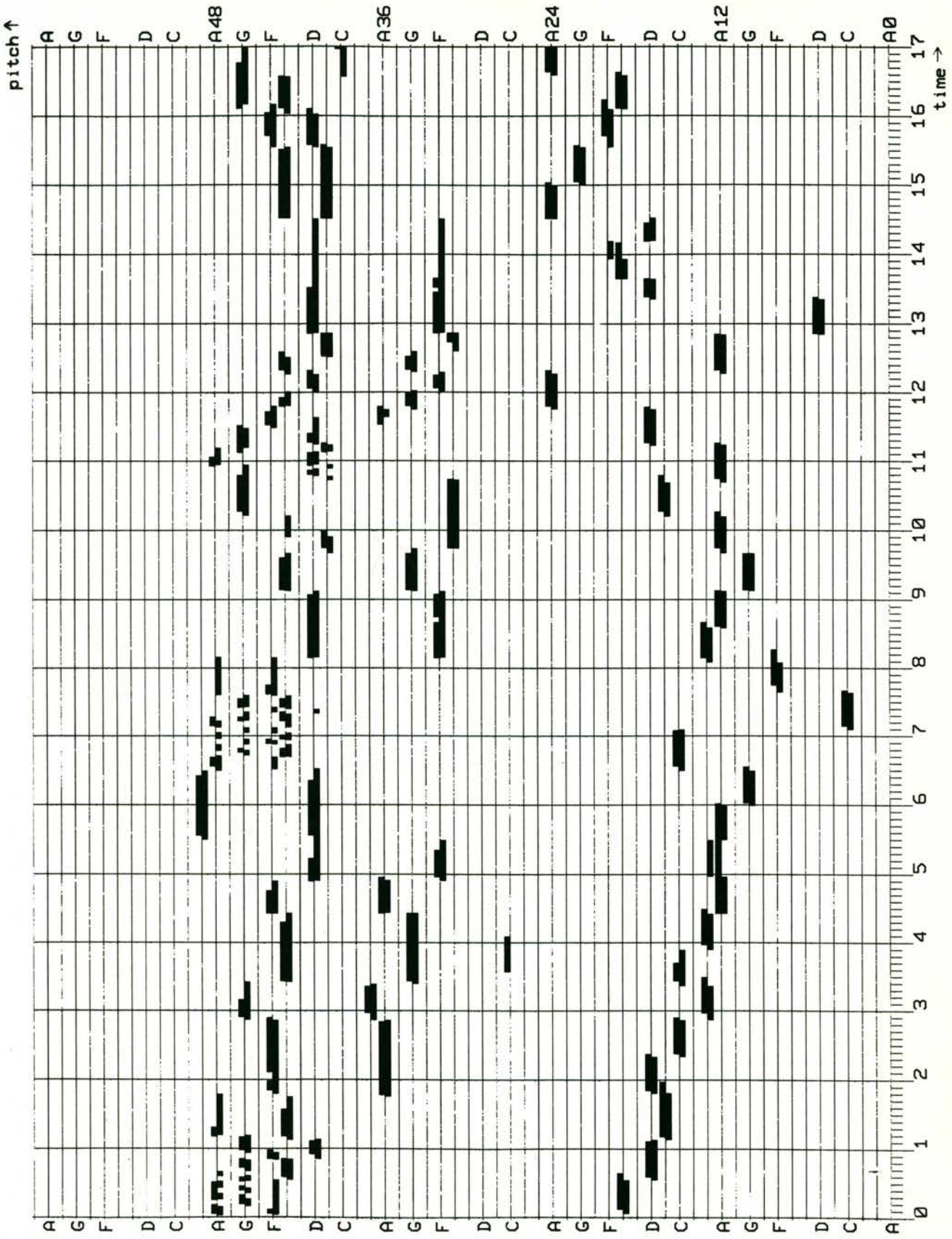


Figure 8.1a shows the first 17 seconds of the comparison of the woodwind analysis (above the lines) with the transcription described in 8.3.1 (below the lines). Markings above the pitch lines, but not below, correspond to the errors of Table 8.3a. Markings below the lines, but not above, correspond to the errors of Table 8.3b.

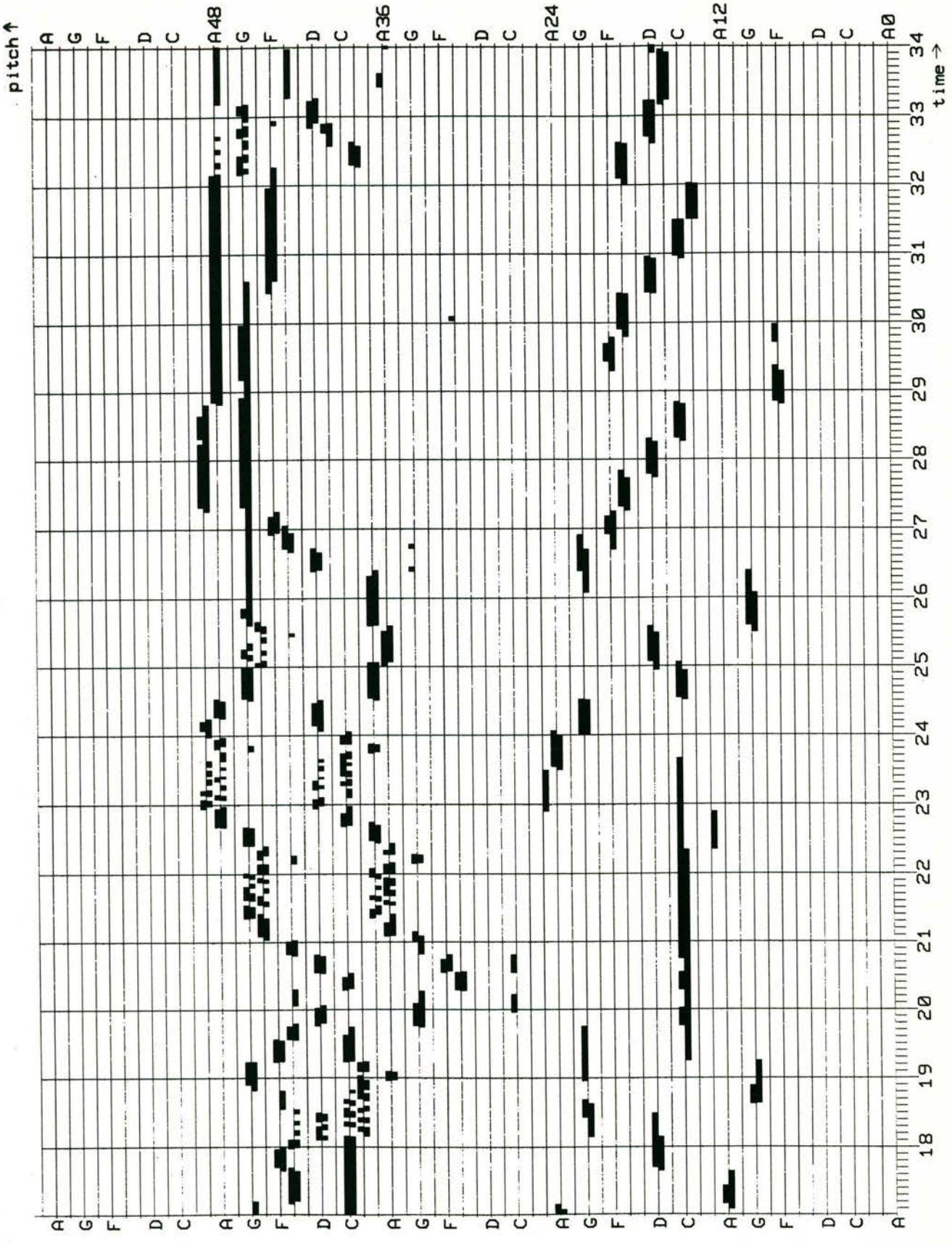


Figure 8.1b shows the second 17 seconds of the comparison of the woodwind analysis (above the lines) with the transcription described in 8.3.1 (below the lines). Markings above the pitch lines, but not below, correspond to the errors of Table 8.3a. Markings below the lines, but not above, correspond to the errors of Table 8.3b.

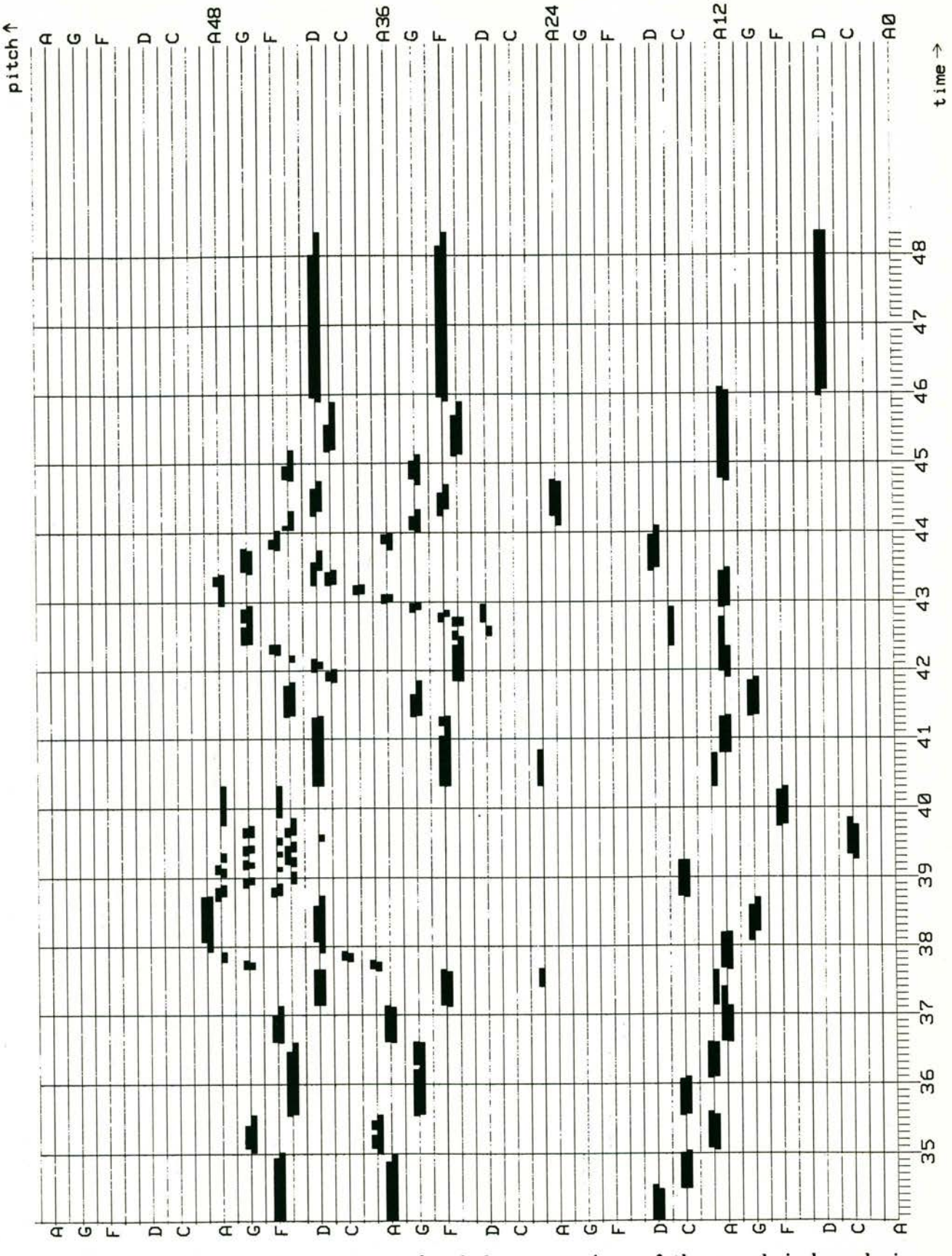


Figure 8.1c shows the final 17 seconds of the comparison of the woodwind analysis (above the lines) with the transcription described in 8.3.1 (below the lines). Markings above the pitch lines, but not below, correspond to the errors of Table 8.3a. Markings below the lines, but not above, correspond to the errors of Table 8.3b.

8.3.4 Analysis of Synthesized Music

Synthesized music was used to determine the accuracy of analysed onset and finish times. Here, there is no variation in pitch, no reverberation or decay at the finish of notes, and the notes start and finish abruptly. Therefore, the set of events that generates the synthesized music is an accurate *performance* with which to compare analyses. Using synthesized music also eliminates noise and signal distortion introduced by the recording and digitization processes.

Any music could have been synthesized, but here it was decided to make the synthesized events match the events of the Trio *performance*. Each event in the *performance* was converted to a saw-tooth wave of fixed amplitude, and superimposed as a digital recording. The saw tooth wave was chosen because it is rich in harmonics and the harmonic amplitudes decrease with increasing harmonic frequency. The durations of 19 notes (of a total of 350) were shortened to correspond to breaths and phrasing in the Trio, the remaining notes finish when the following note begins. The inserted rests provide examples of 0, 1, 2, and 3 simultaneous tones. The synthesized music was analyzed and the onset and finish times of the *analysis* compared with those of the synthesized *performance*. The standard deviation of the time differences between corresponding events in the *performance* and the *analysis* was 12 milliseconds. This compares with a standard deviation of 80 milliseconds for the onset time differences (100 milliseconds for the finish time differences) between the woodwind *analysis* (see section 8.3.3) and the *performance*.

The accuracy of analysed onset and finish times can be modelled statis-

tically. Let **Z** be a random variable being the difference between a synthesized time (onset or finish) and the matching analysed time. The precision in determining the times of events is limited by the effective width of the sampling window (i.e. 40 milliseconds) because finite-time spectral analysis is used. Assuming **Z** is normally distributed, the 68.3 percentile can be used to estimate the standard deviation. Any value of **Z** beyond 3 standard deviations is not normal with a confidence of 99.7% Applying this to the synthesized Trio means that any time difference greater than 36 milliseconds is an E2 error with a confidence of 99.7% For the *performance*, the error in determining the times is added to the uncertainty of the finite-time spectral analysis, so an onset (or finish) time difference must be greater than 240 (300) milliseconds to be an error with 99.7% confidence. There are many criteria for choosing the limit that determines whether an analyzed event is in error. It would be arbitrary to set a fixed limit for $E2$. From the discussion above, values between 40 and 300 milliseconds could be justified. It is however useful to observe the cummulative distribution of timing differences using $E2$ with various limit values. A particular $E2$ limit is also useful for comparing analyses.

Table 8.4 is a histogram of the differences between the synthesized and analyzed onset times. Here only the 2 oboe parts are synthesized, and the spectral analysis is applied every millisecond, to determine the distribution at high resolution. A negative difference means that the analyzed event begins before the matching synthesized event. The finish time differences are similarly distributed.

Table 8.4

Onset difference histogram for synthesized oboes

```
-10   *
 -9   *
 -8   *
 -7   **
 -6   ****
 -5   ***
 -4   ******
 -3   ********************
 -2   ******************
 -1   **********************
  0   ************************************************
  1   ******************************************
  2   ********************************
  3   ****************
  4   ******
  5   ********
  6   ****
  7   **
  8   *
  9   *
 10   *
```

Table 8.5 lists the error measures for the synthesized Trio, and Table 8.6a and 8.6b list the $E2$ errors with onset_limit and finish_limit set to 50 milliseconds. Onset time (time), duration (dur), pitch, onset differences and finish time differences are given for each error. Figure 8.2 shows the pitch profile of the first 17 seconds of the synthesized Trio, and figure 8.3 shows the comparison of the analysis (above the line) with the synthesized music (below the line). Markings above the pitch lines but not below correspond to the errors of Table 8.6a, and markings below the lines but not above correspond to the errors of Table 8.6b. Figure 8.4 is the music output of the analysis.

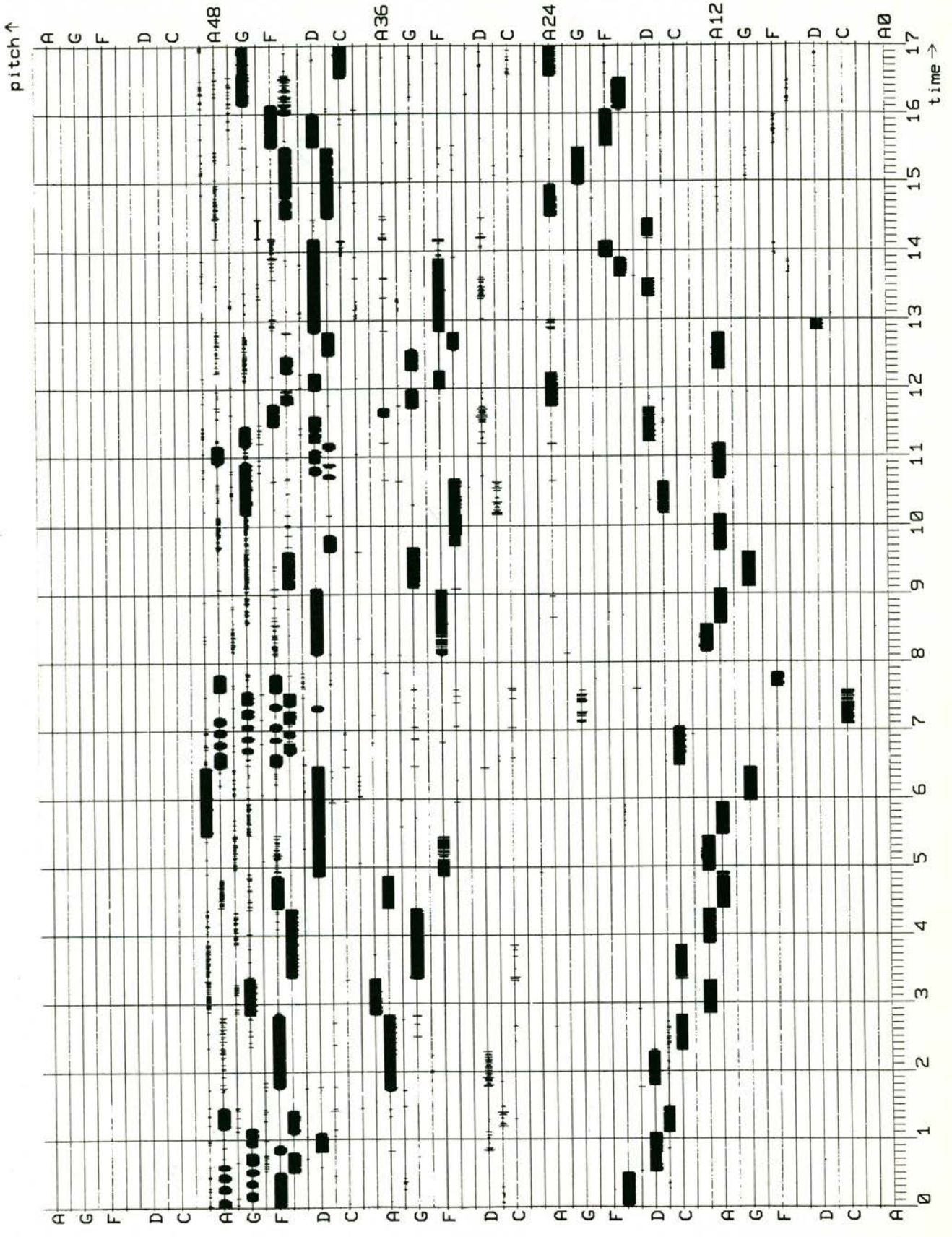


Figure 8.2 shows the pitch profile of the first 17 seconds of the synthesized Trio.

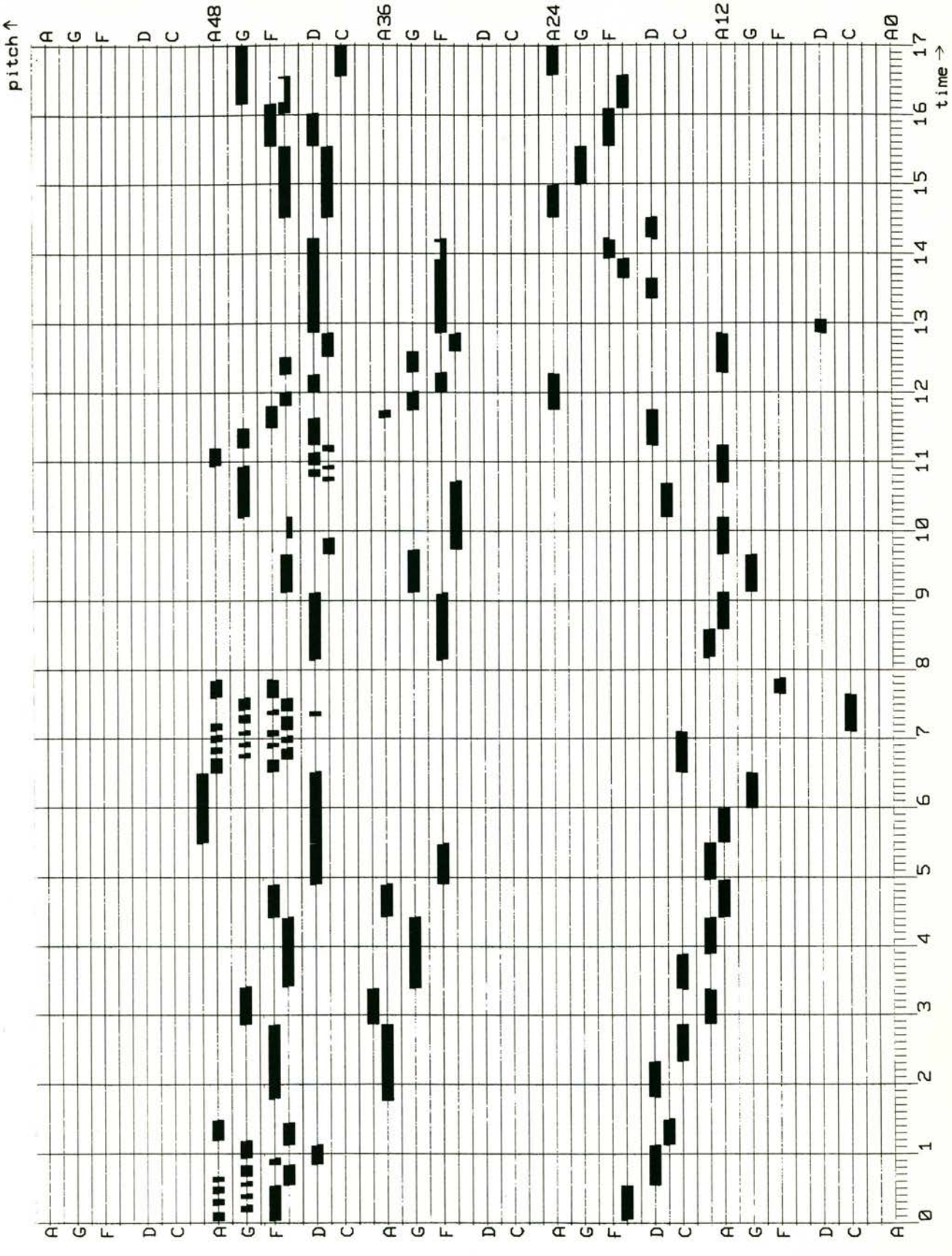


Figure 8.3 shows the comparison of the analysis (above the line) with the synthesized music (below the line). Markings above the pitch lines but not below correspond to the errors of Table 8.6a, and markings below the lines but not above correspond to the errors of Table 8.6b.

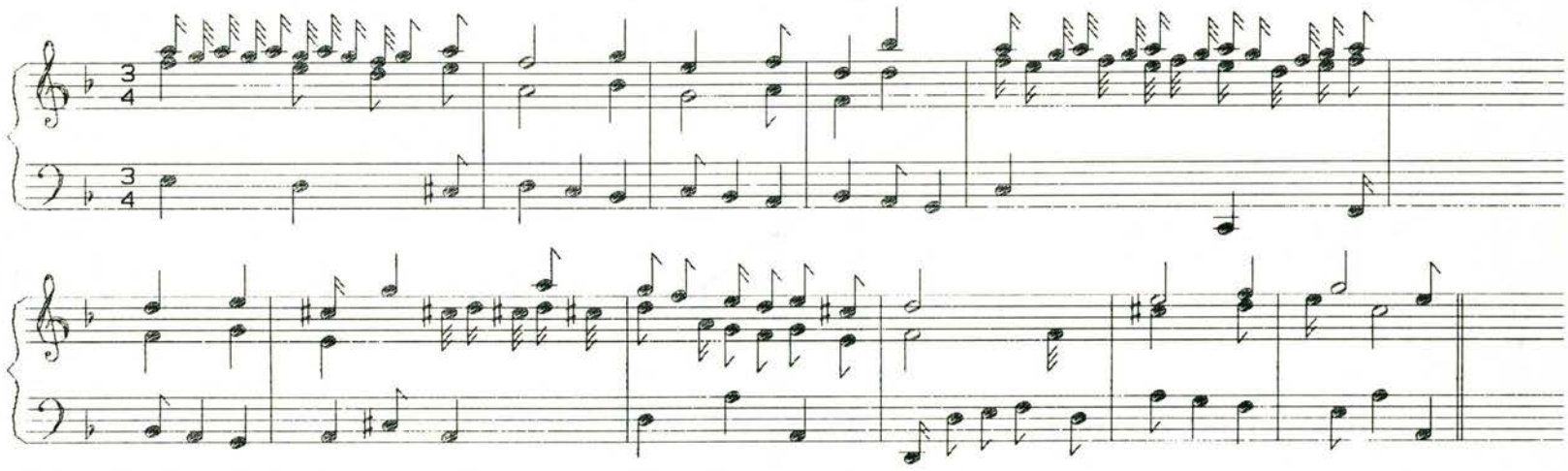


Figure 8.4 is the music output of the analysis. The note ***E*** of the top oboe is absent from the 7th bar, and the semi-demi-quaver, note ***F***, in the 9th bar should not be in the analysis.

Table 8.5

Comparison of Error Measures for the Synthesized Trio

	inclusion	exclusion	combined
E1	1.9%	3.9%	2.9%
E2 (limits=10)	31.3%	29.7%	30.6%
E2 (limits=20)	10.1%	8.3%	9.2%
E2 (limits=50)	5.9%	4.0%	4.9%
E2 (limits=100)	5.4%	3.4%	4.4%
E2 (limits=200)	4.8%	2.9%	3.8%
E3	0.6%	0.9%	0.7%
E4 (duration>100)	0.0%	0.4%	0.2%
E5	0.1%	0.3%	0.2%
E6	0.0%	0.0%	0.0%

Table 8.6a

E2 Inclusion Errors for the Synthesized Analysis

(* shows estimates that are also E3 errors)

			differences		
time	dur	pitch	onset	finish	part
12860	1060	32	-10	-300	Oboe2
14170	40	32	1300	-10	Oboe2
16020	170	43	-10	-375	Oboe2
16530	30	43	500	-65	Oboe2
20010	230	43	-50	-65	Oboe1
22650	50	49	*	*	Oboe1
23410	220	49	-5	125	Oboe1
25610	520	46	0	-830	Oboe1
26630	30	41	220	-10	Oboe2
26650	310	46	1040	0	Oboe1
37120	50	32	0	-470	Oboe2
37130	90	41	-5	-445	Oboe1
37220	30	44	*	*	Oboe1
37260	400	41	125	-5	Oboe1
37320	100	32	200	-220	Oboe2
37470	160	32	350	-10	Oboe2
40280	360	32	-40	-695	Oboe2
40730	590	32	410	-15	Oboe2
46080	2390	5	30	-80	Bassoon

Table 8.6b

E2 Exclusion Errors for the Synthesized Analysis

(* shows estimates that are also E3 errors)

time	dur	pitch	differences onset	finish	part
9900	310	43	*	*	Oboe1
12870	1350	32	-10	-300	Oboe2
16030	535	43	-10	-375	Oboe2
20060	245	43	-50	-65	Oboe1
23415	90	49	-5	125	Oboe1
23505	40	48	*	*	Oboe1
23545	85	49	-135	0	Oboe1
25610	1350	46	0	-830	Oboe1
26410	260	41	220	-10	Oboe2
37120	520	32	0	-470	Oboe2
37135	530	41	-5	-445	Oboe1
40320	1015	32	-40	-695	Oboe2
42130	100	43	*	*	Oboe1
46050	2500	5	30	-80	Bassoon

There are two *E*3 inclusion errors and three *E*3 exclusion errors and all are an octave from the corresponding correct event. Therefore there are no *E*6 errors. The *E*3 exclusion errors at time 9900 and 42130 milliseconds result from the failure to detect the higher pitched oboe tone when the oboes are playing an octave apart. At time 14000 the bassoon plays an octave below the second oboe, masking the oboe tone for the duration of the bassoon tone and splitting the analyzed oboe tone in two. This single problem causes two inclusion errors; one starting at time 12860 and finishing at 300 milliseconds early, and the other starting 1300 milliseconds late at time 14170. All the *E*2 errors with onset or finish differences greater than 100 milliseconds appear to be caused by the masking of one note by another an octave, a twelfth, or 2 octaves below (12, 19

or 24 semitones below). In each case, the note with fundamental frequency a half, a third, or a quarter of the missing note is extracted first, thus removing the harmonics of the higher note.

By comparison, J.A. Moorer's analysis of the guitar duet had one note missing of a total 57 notes, an exclusion error of 1.8% for the measures $E3, E4, E5$, and $E6$.

8.3.5 Reasons for the Errors

The steady state dynamic range (30dB) is enough to cause masking of quieter tones by louder ones. Incorrect estimates that are harmonically related to correct ones often occur, especially those an octave apart. The notes of short duration are more prone to error. The Oboe I exclusion errors at time 9900 and 42130 are the same as those described in 8.3.4 for the *analysis* of the synthesized Trio. The long onset and decay of tones, reverberation, dynamic range, and signal noise, all contribute the remaining errors in the woodwind analysis.

As an example, at time 39900 the bassoon plays a strong tone at 70 dB, which is rich in harmonics, and has a pitch of 8 (see table 8.7). The fundamentals of the oboes coincide with the eighth and tenth harmonics of the bassoon. The higher oboe harmonics suggest a dB level of 30 to 40 dB. The oboe tones are exclusion errors in the woodwind *analysis*, but are successfully detected in the synthesized Trio, where the dB level of all tones are the same.

Table 8.7

Harmonics at time 39900 msecs

Pitch	Frequency (Hz)	Harmonic dB level									
		1	2	3	4	5	6	7	8	9	10
F 8	84.7	51	42	55	61	53	68	31	64	58	57
F 44	679.9	68	63	31	25	23	17	21	12	12	5
A 48	849.5	57	55	34	27	17	8	22	5	10	7

8.4 Sensitivity Analysis

An analysis of the algorithm was done to determine the sensitivity of error rates to small parameter variations. The number of data points was limited, because each iteration required 8 CPU hours and there are many parameters. Even techniques like Nelder and Mead's simplex method (1965) are computationally prohibitive. Several analyses of the full 50 seconds were done to study the behaviour of the algorithm with different parameters and a sensitivity analysis was done for the first 17 seconds, this being a representative sample.

Table 8.8 shows the $E3$ error rates and different parameter weightings for several analyses of the woodwind Trio. The errors given are the $E3$ inclusion, exclusion, and combined (E_{comb}) errors. The parameters are the effective duration of the sampling window (width), the time between successive pitch estimates (delta), and the weightings for the extraction heuristics. H5 was not used for the analyses of Table 8.8. The first row gives the error measures and weighting for the *analysis* described in section 8.3.3.

Table 8.8

E3 errors for the woodwind analyses

Error Measures			Parameters					
$E3_{incl}$	$E3_{excl}$	$E3_{comb}$	width	delta	H1	H2	H3	H4
3.8	15.4	9.6	40	25	.6	.5	-1	1.2
4.8	12.3	8.6	40	10	.6	.5	-1	1.2
4.1	13.1	8.6	40	10	.5	.5	-.5	1.0
2.7	13.7	8.2	40	10	.5	.5	.5	1.0
4.2	16.0	10.2	40	10	.5	.3	2.5	1.0
6.2	11.4	8.8	40	10	.5	.3	2.0	1.0
5.7	11.7	8.7	40	10	.5	.2	2.0	1.0
5.1	10.3	7.7	40	10	.5	.5	2.0	1.0
6.2	13.7	9.9	40	10	.6	.5	2.0	1.2
4.4	11.1	7.8	40	10	.6	.5	1.0	1.2
4.4	11.4	7.9	40	10	.6	.6	1.0	1.2
3.0	14.9	9.0	40	10	.5	.5	1.0	1.0
3.5	11.1	7.3	40	10	.5	.5	1.0	1.0
1.8	15.1	8.9	80	10	.5	.5	.5	1.0
7.7	18.9	13.4	80	10	.5	.5	5.0	1.0
5.0	14.6	9.8	80	10	.5	.5	2.5	1.0
3.9	11.7	7.8	40	5	.4	.5	.5	1.2
7.0	14.3	10.7	40	10	.4	.15	1.5	.6

It is desirable that the analysis algorithm be insensitive; that is, the relative change in error rate should be small for small relative changes in the parameters. The condition number K, for a parameter a is defined by:

$$\left[\frac{E(a+\delta a)-E(a)}{E(a)}\right] = k\left[\frac{\delta a}{a}\right].$$

Parameters are varied independently of each other. The parameters tested are:

- the weightings for the heuristics given in chapter 5,
- the duration of the sampling window (width),

- the number of harmonics (harms) used in the harmonic summing algorithm, and for the grouping of estimates (section 6.5)

- the time to fill a pitch bucket , and

- the minimum duration (mindur)

The first 17 seconds of the synthesized Trio was used for the sensitivity analysis. The parameters used in 8.3.4 are taken as the basis for the sensitivity analysis. The base parameter values were determined by repeating the sensitivity analysis until a local minimum was found. Each parameter was varied independently above and below the base value to record the changes in error rate. The parameters were varied by 10% and 20% above and below the base values, although some discrete parameters required a greater increment. The $E1$ combined measures and condition numbers are given in Table 8.9. This error measure was chosen because it is the most sensitive to small changes in the onset and finish times of analysed notes. The optimal parameter setting depends on the measure used. For example, decreasing the weighting of H4 by 20% decreases the $E2$ combined error (limits=50msec.) from 3.5% to 3.1%, while the $E1$ exclusion error increases. Optimality is also dependent on the music being analyzed. The set of weightings found to work well for the piano music of chapter 7, is far from optimal when applied to the Trio (see the last row in Table 8.8).

Table 8.9

Sensitivity Analysis

parameter	base value of parameter	$\delta a/a$	El_{comb} decreased parameter	El_{comb} base error	El_{comb} increased parameter	K
H1	0.2	.10	2.49%	2.48%	2.49%	.04
H1	0.2	.20	2.58%	2.48%	2.49%	.20
H2	0.24	.10	2.52%	2.48%	2.58%	.40
H2	0.24	.20	2.58%	2.48%	2.58%	.20
H3	1.6	.10	2.48%	2.48%	2.51%	.12
H3	1.6	.20	2.49%	2.48%	2.52%	.08
H4	0.9	.10	2.54%	2.48%	2.52%	.24
H4	0.9	.20	2.69%	2.48%	2.53%	.42
H5	0.1	.10	2.48%	2.48%	2.48%	.0
H5	0.1	.20	2.48%	2.48%	2.49%	.02
harms	8	.20	3.28%	2.48%	3.40%	1.85
width	40msec	1.0	5.73%	2.48%	5.42%	1.31
mindur	50msec	.20	2.48%	2.48%	2.49%	.02
pitcher	60msec	.17	2.59%	2.48%	2.60%	.28

The analyser is well conditioned with respect to the parameters for this base set of parameter values. The algorithm is most sensitive to change in the number of harmonics (harms), and the width of the sampling window (width). The parameters harms and width should be considered constants of the algorithm.

8.5 Conclusion

The error measures presented allow comparisons to be made between algorithms for analysing music. These error measures may be useful to future researchers for comparing analyses of other music, and for comparing new algorithms. A sensitivity of the pitch determining algorithm showed no great changes in errors for small changes in parameters. The woodwind Trio analysis

was 90.3% accurate and the synthesized Trio analysis was 99.3% accurate, using the $E3$ combined measure.

CHAPTER NINE

A Computer Simulation of the Human Cochlea: A Model for the Discrimination of Superimposed Tones

9.1 Introduction

The human auditory system is proficient at distinguishing superimposed tones. A trained musician, for example, can hear a single instrument playing out of tune in a large orchestra. It is interesting therefore to consider what is known about human auditory signal processing, and to suggest an hypothesis to bridge the gaps in our knowledge.

Section 9.2 surveys the current knowledge of human auditory signal processing. Some new mathematical models are presented (section 9.3), and used to simulate the response of the human cochlea to polyphonic tones. (section 9.3). The observed amplitude modulation of the response is shown to be enough to distinguish superimposed tones.

9.2 Human Auditory Signal Processing: Mechanisms of Hearing

Sound waves striking the outer ear (Pinnae) pass through the external ear canal to the timpanic membrane (ear drum) at the entrance to the middle ear (Bekesy 1963). The middle ear contains three small bones (ossicles) which transmit the vibrations from the timpanic membrane to the oval window of the inner ear. The vibration of the basilar membrane (BM) in the inner ear then causes neural impulses to be sent along the auditory nerve to higher centres of

the brain (auditory cortex). Figure 9.1 is a diagram of the human ear.

9.2.1 The Outer Ear

The outer ear is important in the localization of sound. Sound waves entering the outer ear canal are filtered by diffraction at the Pinnae. This diffraction pattern differs, depending on the direction of the sound source. Schroeder (1975) states that this diffraction is responsible for the localization of sounds above, behind, and in front of a listener, when the sound impinging on each ear is identical in amplitude and phase.

9.2.2 The Middle Ear

The middle ear contains three connected bones, called the ossicles, which link the timpanic membrane to the cochlea. They are, from the outside, the incus, malleus and stapes.

The middle ear acts as an acoustical impedance transformer, to match the low impedance of the air to the high impedance of the cochlear fluid (or perilymph). This impedance transformation of a factor of 20 gives a 5 fold improvement in power transmission through the inner ear (Moller 1973).

The ossicles also act as an "automatic gain control" protecting the inner ear from overloading and possible damage (Schroder 1975). At high sound intensities the mobility of the ossicles is reduced by involuntary contraction of adjoining muscles. This compression of the incoming signal is maintained for several milliseconds. Therefore the signal is not peak clipped, but compressed

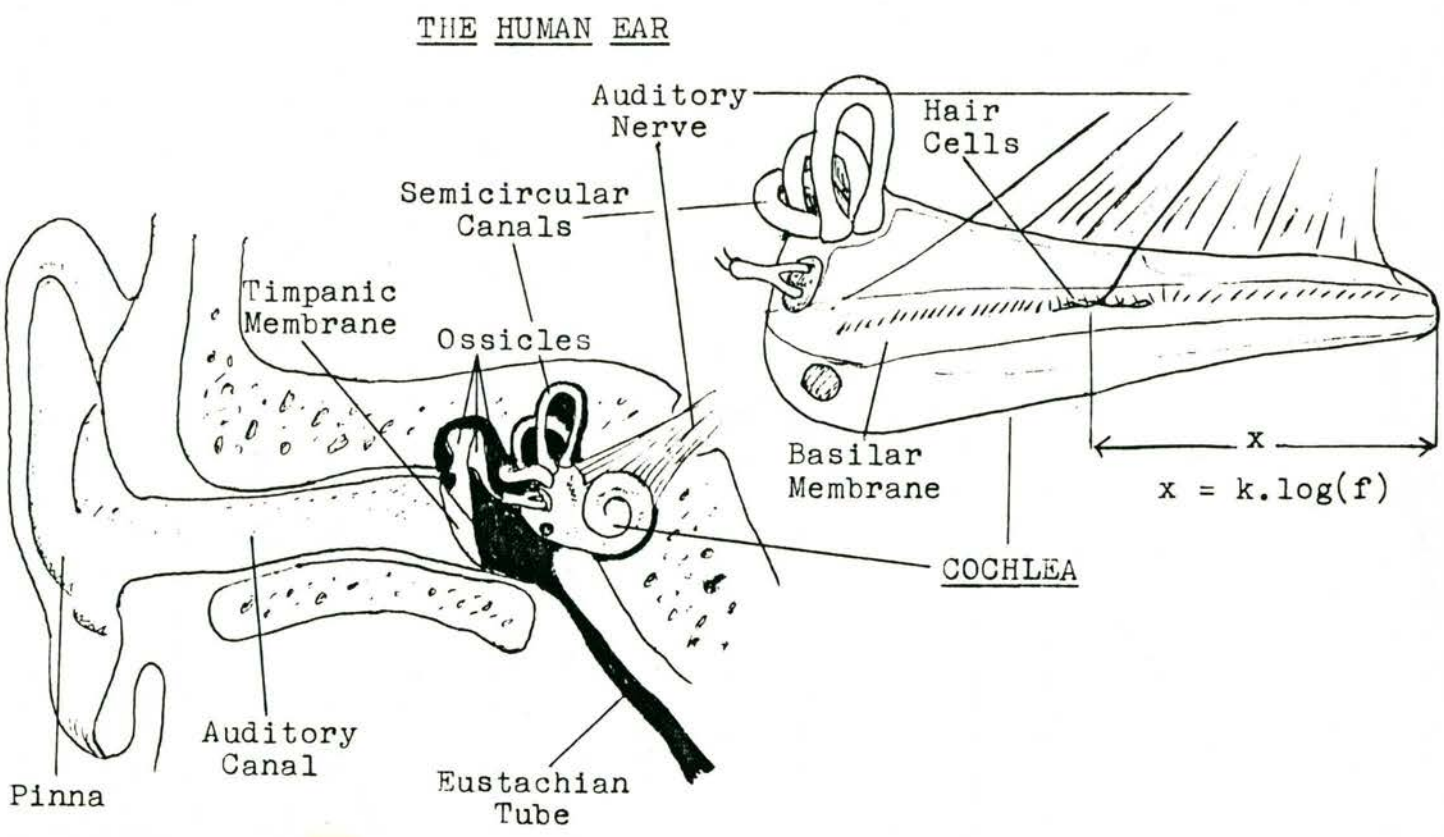


Figure 9.1 is a diagram of the human ear.

sketched from Moller (1973).

by a constant factor over many cycles. This nonlinearity of the middle ear is partly responsible for combination tones and aural harmonics at high signal levels (Allen 1977).

9.2.3 The Inner Ear

The cochlea (from the Greek word for a spiral shaped snail) is the frequency selective part of the ear and is about 35 millimetres long in humans. The cochlea is partitioned by the BM. This membrane supports the organ of Corti, where the 25,000 hair cells convert the BM displacement to neural impulses.

G. Ohm (1843) was the first to suggest the frequency selectivity of the cochlea. Hermann von Helmholtz (1863) proposed that the BM was under tension and that the varying tension accounted for the varying resonance. G. von Bekesy (1960) observed through a microscope, waves travelling along the basilar membrane from the stapes (the innermost ossicle bone) to the apex of the cochlea. The travelling waves caused by pure tones, increase in intensity as they move toward the apex, and are then abruptly attenuated. Low frequency signals travel further along the BM than high frequency signals before being attenuated; therefore the point of maximum stimulation of the BM is nearer the stapes for high frequencies, and nearer the apex for low frequencies. The characteristic frequency for any point on the BM is the frequency that produces a maximum response at that point. If x is the distance of a given point from the apex and f_c is the characteristic frequency for that point, then f_c is proportional to $\exp(kx)$, where k is constant. The band-width of the cochlear tuning curve

increases with characteristic frequency. Bekesy (1960) showed that the tuning property of the BM was caused by exponentially increasing stiffness of the BM, and not tension as Helmholtz had presumed. The tuning curves derived by von Bekesy were broader than those of normally functioning cochleas, because he worked with dead animals and applied high amplitude signals. Rhode (1971) obtained much narrower tuning curves with a wider dynamic range than Bekesy. He used a Doppler phenomenon at the nuclear level (Mössbauer effect) that enables amplitudes as small as 20 Angstroms to be measured.

9.2.4 The Auditory Nerve and Higher Centres of the Brain

Neurons are the functional units of the nervous system and are interconnected at synapses. If the total post-synaptic potential reaching the cell body at any time exceeds a threshold, the neuron fires and a spike potential is transmitted along the axon to the synapses, which then produce post-synaptic potentials in neighbouring neurons. Synapses can be inhibitory or excitatory.

The spontaneous firing rate of individual auditory nerves is typically 50 spikes per second. The maximum firing rate is limited by the refractory period of about 1 millisecond between spikes, but is rarely greater than 200 spikes per second. When the BM is vibrating, neurons are more likely to fire as the BM is moving upwards. The probability of a neuron firing is proportional to the rate of upward motion of the BM. By observing a group of neighbouring auditory nerve fibres at a point of excitation on the BM, volleys of firing occur in synchrony with the incoming signal. Rose et al. (1969) observed that the average firing

rate of a neuron is proportional to the positive-half-wave-rectified part of the incoming signal.

With continued stimulation, a given neuron becomes depleted of the chemicals required for spiking, and the probability of firing is decreased until the chemical balance is restored. This mechanism is called adaptation. Adaptation is a contributing factor in the compression of signals with different intensities. Efferent (from the brain) nerve fibres provide negative feedback from the auditory cortex, and contribute to signal compression.

Kiang et al. (1974) observed that auditory nerve fibres responded to all signals of frequency lower than the nerve's characteristic frequency, but did not respond to signals above the characteristic frequency. The response to low frequency signals was about 40 dB below the maximum response for any particular nerve. This concurs with Bekesy's observation that signals of low frequency travel through regions of high characteristic frequency before reaching their point of maximum resonance on the BM. The tuning curves of Rhode and Kiang are narrower than Bekesy's.

Although Kiang and Rhode derived their tuning curves using laboratory animals, Harrison et al. (1981) demonstrated the similarity between human and animal action potential tuning curves. Q_{10dB} is defined as the ratio between the centre frequency, and the frequency difference between spectral points at a level of 10 dB below the maximum response. By using a tone on tone masking procedure, Harrison found the Q_{10dB} to be 4.2 at 2 kHz, 6.5 at 4 kHz, and 8.5 at 8 kHz.

9.2.5 Place vs Periodicity Theories of Pitch Perception

The place theory was first proposed by Ohm and Helmholtz, who claimed that pitch was determined by the position of maximum vibration on the BM. Seebeck (1841) proposed that periodicity, and not place, was the mechanism for pitch recognition. This was based on the argument that some periodic signals without a fundamental are still perceived as having the same pitch as the fundamental, although there is no stimulation at that point on the BM. Ohm and Helmholtz argued that Seebeck's hypothesis was erroneous. Schouten et al. (1962) reinstated the periodicity theory with his experiments on residue pitch. Houtsma et al. (1972) demonstrated that pitch is not determined at the cochlea but in the auditory cortex, because residue pitches are perceived whether the components are played together through one ear or separately, but simultaneously, through both ears. Both theories are consistent with data now available on the tuning characteristics of the BM and auditory nerves (Bekesy, Kiang, Rhode). At low characteristic frequencies the harmonics are resolvable on the BM (place theory), and at high characteristic frequencies all the harmonics are superimposed to give a time domain replica of the incoming signal (Rose et al.), allowing the periodicity to be determined in the auditory cortex.

9.3 Models of the Cochlea

Several models of the cochlea have been proposed. Flanagan (1962) made a mathematical and electrical model based on Bekesy's data for the BM response. Kim et al. (1973) and Schroeder (1975) introduced nonlinearities of

the cochlea into the model. J.B. Allen (1977), among others, introduced models to account for the narrower tuning curves of Kiang and Rhode. Three models are considered here: an electrical analogue model (9.3.1), a Gaussian tuning curve (9.3.2), and a mathematical model fitted to Kiang's data (9.3.3). The first is a simplification of Flanagan's model, and the other 2 are new models proposed by the author. These models were chosen because they fit the empirical data well, but they are also easier to calculate than the more rigorous models based on cochlear mechanics (eg. Allen 1977). Figure 9.2 compares stylized tuning curves of Bekesy and Kiang with the mathematical models of this section.

9.3.1 Electrical Analogue

Taking a simplified version of Flanagan's model (1962), the mass of the BM can be considered to act in the same way as an electrical inductor, attenuating high frequency signals. Energy losses of the perilymph (cochlear fluid) and BM act in the same way as an electrical resistance. The compliance of the basilar membrane acts like a capacitor, storing the mechanical vibration. Stiffness (inverse of compliance C) was found by Bekesy (1960) to vary exponentially with the distance from the stapes. Schroeder (1975) suggested making R a quadratically increasing function of amplitude, to account for combination tones resulting from non-linearities of the cochlea.

Here the impedance of the BM is modelled as:

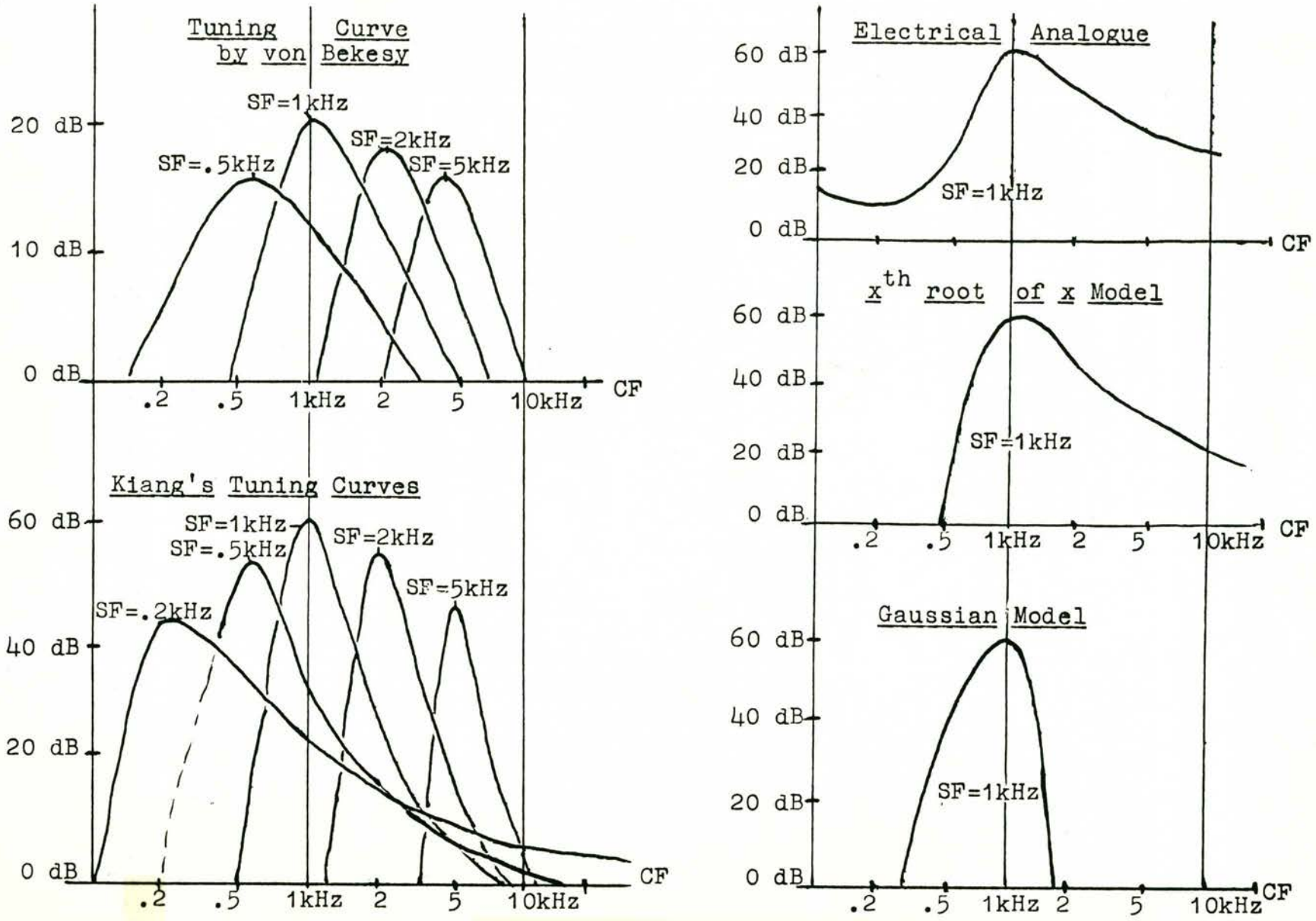


Figure 9.2 compares the empirical tuning curves of Bekesy and Kiang et al. with the mathematical models of section 9.3. (CF is the characteristic frequency and SF is the incomming signal frequency).

$$Z(\omega) = j\omega L + R + \frac{1}{j\omega C},$$

where ω is the radian frequency, $j = \sqrt{-1}$; $j\omega L$ represents the mass impedance; R, the resistive loss in the cochlear fluid and the BM; and $j\omega C$ the admittance.

The BM response is:

$$M = \frac{\frac{1}{j\omega C}}{Z(\omega)} = \frac{1}{1 - LC\omega^2 + jRC\omega}.$$

To study the behaviour of this function, consider the transformation:

$$y = \frac{1}{|M|^2},$$

$$x = \omega^2 LC,$$

$$k = \frac{R^2 C}{L}.$$

Lemma 9.1

If $y = (1 - x)^2 + kx$, then y is minimal at $x = x_0 = 1 - \frac{k}{2}$, and the minimal value for y is $y_{min} = k - \frac{k^2}{4}$.

Lemma 9.2

If $y = ay_{min}$ where a is a constant greater than 1, then

$$x = x_0 \pm \sqrt{k(1 - \frac{k}{4})(a - 1)}.$$

The proof for Lemmas 9.1 and 9.2 can be derived from elementary calculus.

M is maximal if and only if y is minimal, therefore from Lemma 9.1:

$$max|M(\omega)| = \sqrt{\frac{4}{k(4-k)}}$$

at $\omega = \omega_0 = \sqrt{\frac{1-\frac{k}{2}}{LC}}$.

Using Lemma 9.2 with $a = 100$, if $M(\omega) = 0.1M(\omega_0)$ then,

$$\omega^2 LC = \omega_0^2 LC \pm \sqrt{99k(1-\frac{k}{4})}.$$

k is small compared to 1, therefore

$$k \approx 0.01[1 - \frac{\omega^2}{\omega_0^2}]^2 \omega_0^4 L^2 C^2.$$

Table 9.1

Values of k' for the Electrical Analoque Model

input signal frequency	ω_0	$\frac{\omega_1}{\omega_0}$	$\frac{\omega_2}{\omega_0}$	k_1'	k_2'
$>$ 2 kHz	12,000	0.85	1.2	.00078	.00195
1 kHz	6,000	0.8	1.3	.00131	.0048
500 Hz	3,000	0.7	1.6	.00263	.0246

where ω_1 and ω_2 are characteristic frequencies such that:

$$M(\omega_1) = M(\omega_2) = 0.1M(\omega_0), \quad \omega_1 < \omega_2,$$

$$k_1' = 0.01\left[1 - \frac{\omega_1^2}{\omega_0^2}\right]^2,$$

and

$$k_2' = 0.01\left[1 - \frac{\omega_2^2}{\omega_0^2}\right]^2,$$

If this is an accurate model of the basilar response, then k_1' and k_2' should be nearly equal for each characteristic frequency. The average value of k_1' and k_2', over the range of required characteristic frequencies is used for simulation. To fit Kiang's data, the model should be more left-skewed for high characteristic frequencies, and more right-skewed for low characteristic frequencies.

The response of the basilar membrane M can be considered to be either a function of the frequency ω of the incoming signal, for a fixed point on the BM with characteristic frequency ω_0, or to be a function of the position on the basilar membrane, denoted by ω_0, for some incoming pure tone of frequency ω. Nonlinearities of the cochlea response are ignored, for computational economy. The BM response to a given signal can be considered to be the superposition of pure tone responses.

9.3.2 Gaussian Model

For the Gaussian Model the BM response is:

$$M(\omega) = e^{-k(\frac{\omega}{\omega_0}-1)^2}.$$

$M(\omega)$ is symmetrical in the frequency domain, and

$M_{max} = M(\omega_0) = 1.$

If ω is such that $M(\omega) = 0.1M_{max}$, then

$$k = \frac{\log_e 10}{(\frac{\omega}{\omega_0} - 1)^2}.$$

Table 9.2

Values of k for the Gaussian Model

input signal frequency	ω_0	$\frac{\omega_1}{\omega_0}$	$\frac{\omega_2}{\omega_0}$	k_1	k_2
$>$ 2 kHz	12,000	0.85	1.2	102	57
1 kHz	6,000	0.8	1.3	57	25
500 Hz	3,000	0.7	1.6	25	6

where ω_1 and ω_2 are characteristic frequencies such that:

$$M(\omega_1) = M(\omega_2) = 0.1M(\omega_0), \quad \omega_1 < \omega_2,$$

and

$$k_i = \frac{\log_e 10}{(\frac{\omega_i}{\omega_0} - 1)^2}, \text{ for } i = 1, 2.$$

In this model k is taken as the average of k_1 and k_2 for the range of frequencies of interest.

9.3.3 x^{th} Root of x Model

The tuning curves derived from Kiang's data (see figure 8.2) show a rapid increase in frequency response then an asymptotic tapering off with increasing frequency. One function that behaves like this is:

$$y = k\big(e \log_e(\sqrt[x]{x}) - 1\big) = k(\frac{e \log_e x}{x} - 1),$$

where e is the natural logarithm base, (≈ 2.71828).

The derivative of y is:

$$\frac{dy}{dx} = \frac{ek(1 - \log_e x)}{x^2},$$

therefore y is maximal at

$$x = e \text{ with } y_{max} = 0.$$

In this model the BM response is:

$$M(\omega) = e^{-k}x^{(\frac{ek}{x})} = e^{y}, \text{ where } x = \frac{e\omega}{\omega_0}.$$

Exponentiation is a monotonic function, therefore $M(\omega)$ is maximal at $\omega = \omega_0$ and $M_{max} = M(\omega_0) = 1$. As ω increases, $M(\omega)$ tends to e^{-k}. If ω is such that

$$M(\omega) = 0.1M_{max},$$

then

$$k(\frac{e \log_e x}{x} - 1) = \log_e 0.1 \approx -2.30.$$

Table 9.3

Values of k for the x_{th} root of x Model

input signal frequency	ω_0	$\frac{\omega_1}{\omega_0}$	$\frac{\omega_2}{\omega_0}$	k_1	k_2
$>$ 2 kHz	12,000	0.85	1.2	156	156
1 kHz	6,000	0.8	1.3	80	79
500 Hz	3,000	0.7	1.6	29	28

where:

$$k_i = \frac{\log_e 0.1}{(\frac{\omega_0}{\omega} \log_e(\frac{e\omega}{\omega_0}) - 1)}$$

for $i = 1, 2$.

9.4 Computer Simulation

Using the models of section 9.3, a computer model is used to simulate the BM response to various signals. The parameter k is first derived for the range of characteristic frequencies required. The inverse Fourier transform of this frequency response is calculated, and multiplied (in the complex time domain) with successive segments of the incoming signal. The Fourier transform is then determined to give the convolution of the BM response with the spectrum of the incoming signal. This is then displayed as a time-varying plot for all the characteristic frequencies of interest.

Amplitude modulation with beating frequency of the input fundamental, occurs at many points on the simulated cochlea. This phenomenon, henceforth termed inter-harmonic amplitude modulation (IHAM), was investigated to determine its applicability as a pitch-determining criterion for superimposed tones.

The IHAM detector can be an idealized neuron with an exponentially decaying threshold. When the signal for that particular point on the cochlea exceeds the threshold, the IHAM detector fires, and the threshold is reset to a level greater than the strength of the signal that caused the firing. Consequently, larger peaks can be detected, while the smaller ones are rejected. The IHAM detector also adapts to changes in the average signal level.

Alternatively, IHAM detectors can model the firing behaviour of groups of neighbouring neurons.

Figure 9.3 shows the time-varying, simulated neural firing probabil-

ity as a function of characteristic frequency. Time (spanning 40 milliseconds) increases vertically up the graph. The horizontal widths of the lines are proportional to the firing probabilities. The input signal is the steady state of a bassoon tone with fundamental frequency 494 Hz. The x^{th} root of x model is used. Signals are assumed to travel from points of high characteristic frequency (on the right of the graph) to points of low characteristic frequency (on the left). Also, the times of maximum firing probability occur at the rate of the fundamental frequency, even for the higher harmonics.

Figure 9.4 shows the signal level for one point on the cochlea (characteristic frequency = 590 Hz) as a function of time for the bassoon tone. Figure 9.5 shows the time-varying threshold level for the IHAM detector at this point. The threshold decay rate is 8.5% per millisecond. The maxima correspond to the IHAM firings. Figure 9.6 gives the time-varying firing times of all the IHAM detectors in the characteristic frequency range 0 to 1.2 kHz. The time difference between firings for most characteristic frequencies corresponds to the period of the incoming signal.

The IHAM detector can distinguish the fundamentals of two simultaneous tones. Figure 9.7 shows the time-varying, simulated neural firing probability of a two-tone signal as a function characteristic frequency. Time increases vertically up the graph. The horizontal widths of the lines are proportional to the firing probabilities. The input signal contains two superimposed saw-tooth waves with frequencies, 261.6 Hz and 207.2 Hz (a major third). The Gaussian model for the cochlear frequency response is used here.

Figure 9.8 gives a time-varying plot of IHAM firings as a function of characteristic frequency. Figure 9.9 gives a histogram of the times between firings of all the IHAM detectors in the characteristic frequency range 0 to 2 kHz. The marked peaks correspond to the periods of the two tones, and result from the periodic behaviour of many IHAM detectors during the simulation. Figure 9.10 gives a histogram of the times between firings of all the IHAM detectors in the characteristic frequency range 0 to 2 kHz for the Trio example of figure 2.10. The threshold decay rate is 4% per millisecond. The 3 highest peaks correspond to the period of the bassoon tone (197 Hz, *G* p22) and an octave below the two oboes tones (294 Hz and 456 Hz). Figure 9.11 gives a histogram of the times between firings of all the IHAM detectors in the characteristic frequency range 0 to 2 kHz for the Trio example of figure 2.10. The threshold decay rate is 21% per millisecond. The 2 highest peaks correspond to the periods of the two oboes tones (1.09 milliseconds for *B♭* p49 and 1.72 milliseconds for *D* p41).

The computer simulations of the cochlear response to superimposed tones, showed that the phenomenon of amplitude modulation was sufficient to identify the fundamental frequencies of the tones.

It is interesting that this model also matches human pitch perception of pure tones and tones with the fundamentals absent.

9.5 Cybernetic Model

Timbral cues, stereophonic location, and context play an important role in the separation of acoustical events by human listeners. However, at the level

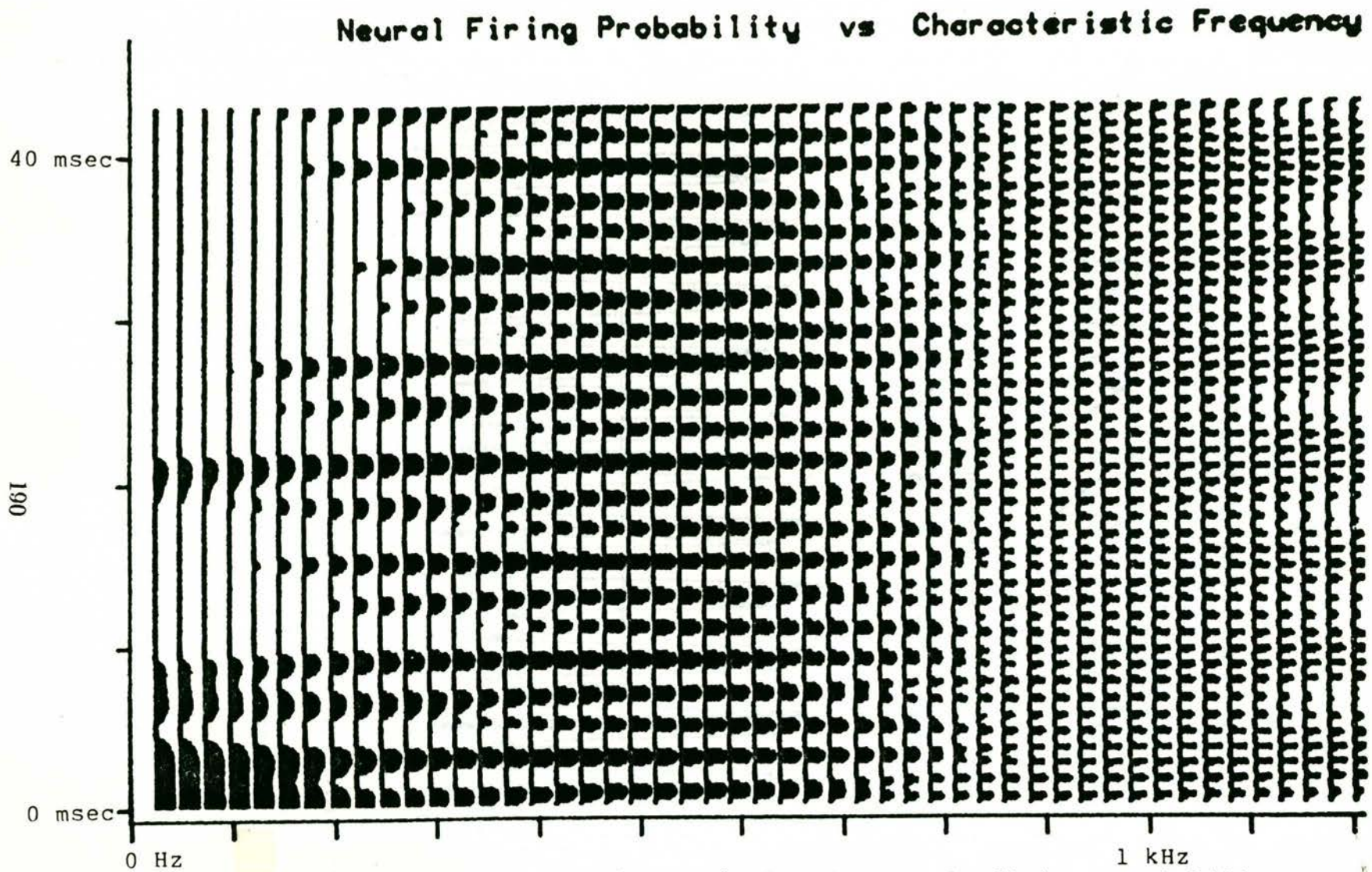


Figure 9.3 plots the time varying, simulated neural firing probability as a function of characteristic frequency. Time increases vertically up the graph (40 milliseconds). The horizontal widths of the lines are proportional to the firing probabilities. The input signal was the steady state of a bassoon tone.

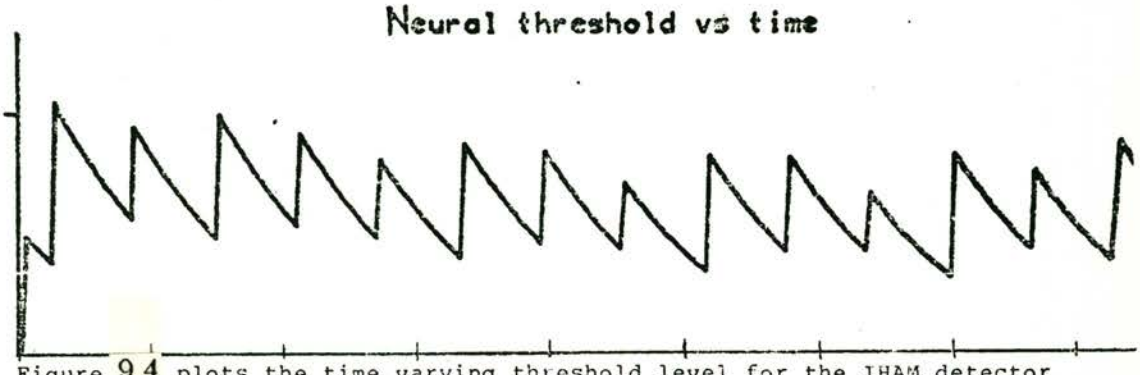


Figure 9.4 plots the time varying threshold level for the IHAM detector at this point (CF = 590 Hz).

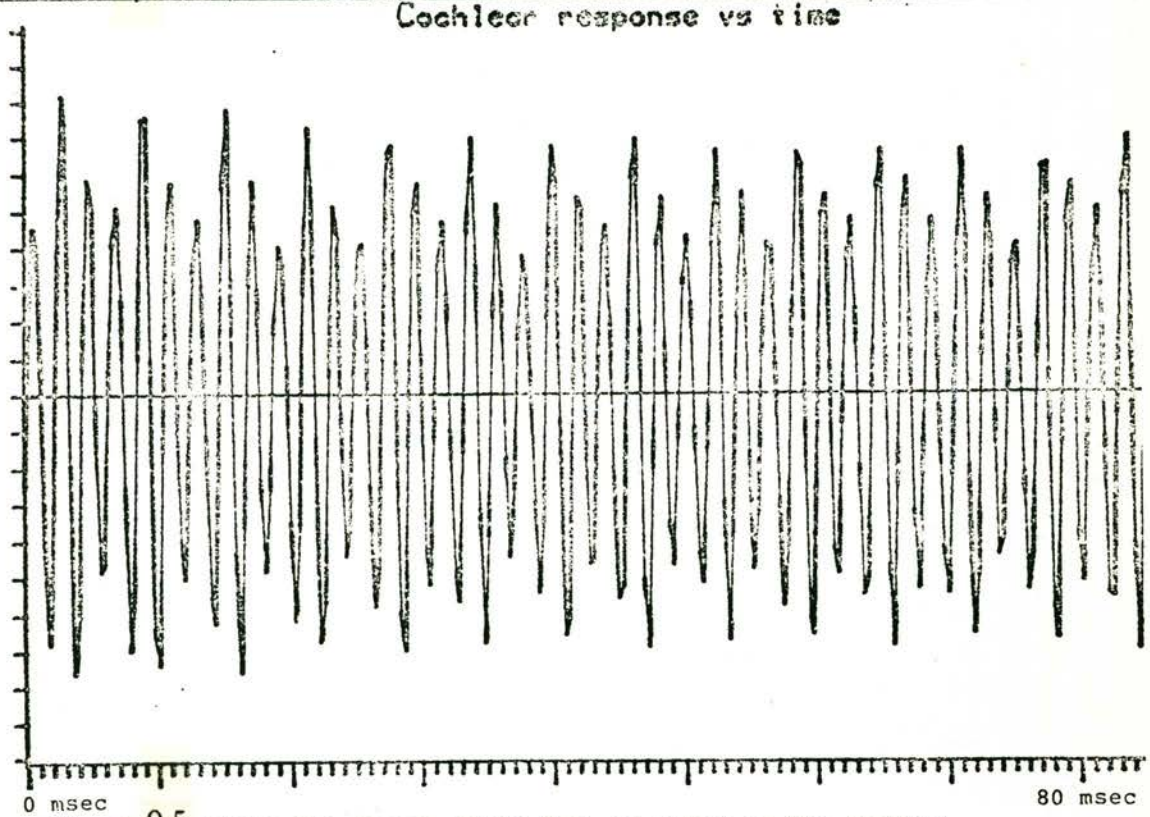


Figure 9.5 plots the signal level for one point on the cochlea (CF = 590 Hz) as a function of time for the bassoon tone

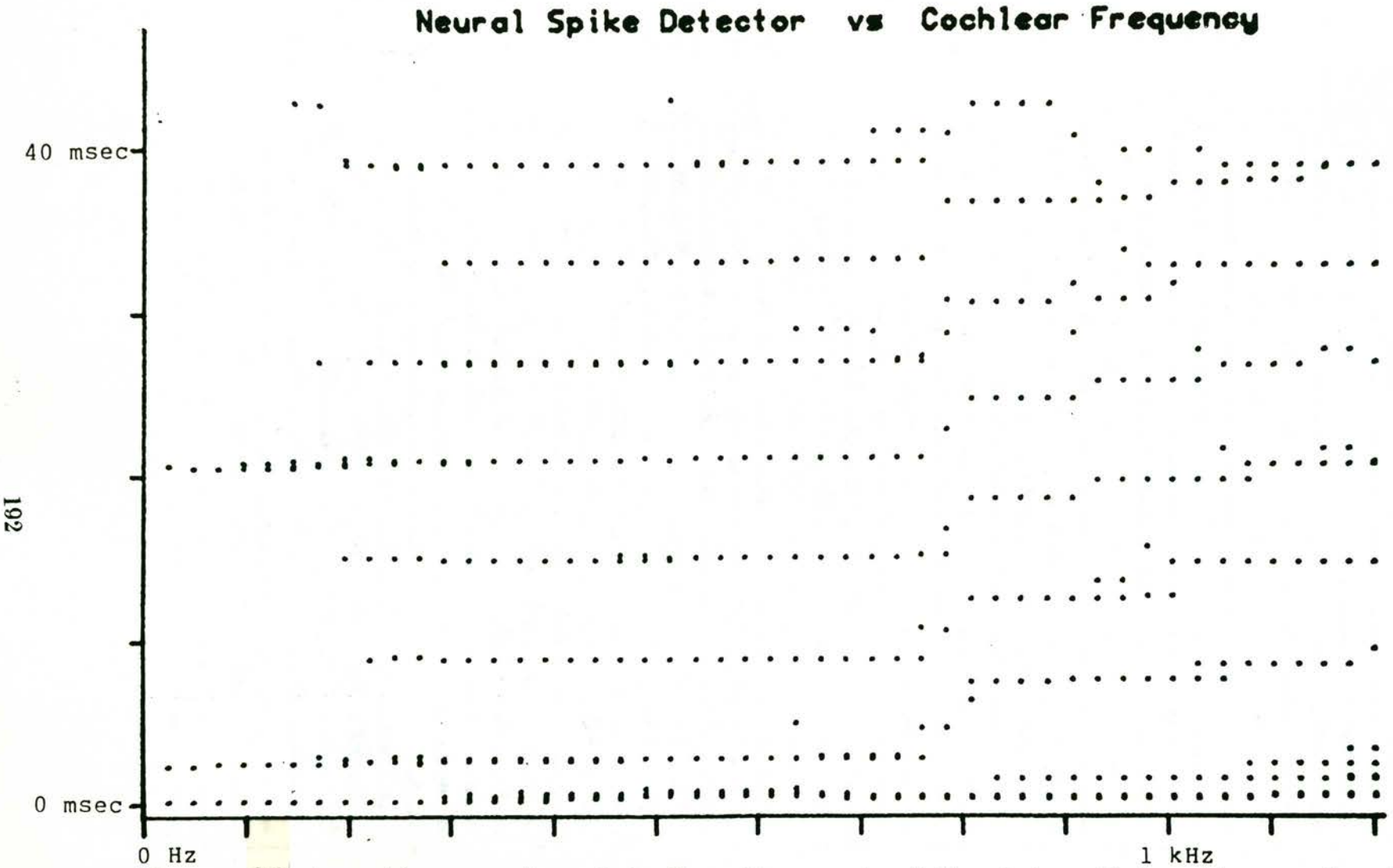


Figure 9.6 is a time varying plot (for 40 msecs) of the inter-firing times of all the IHAM detector firings in the CF range 0 to 1.2 kHz. Note the amplitude modulation of 6 milliseconds corresponding to the period of the incoming signal.

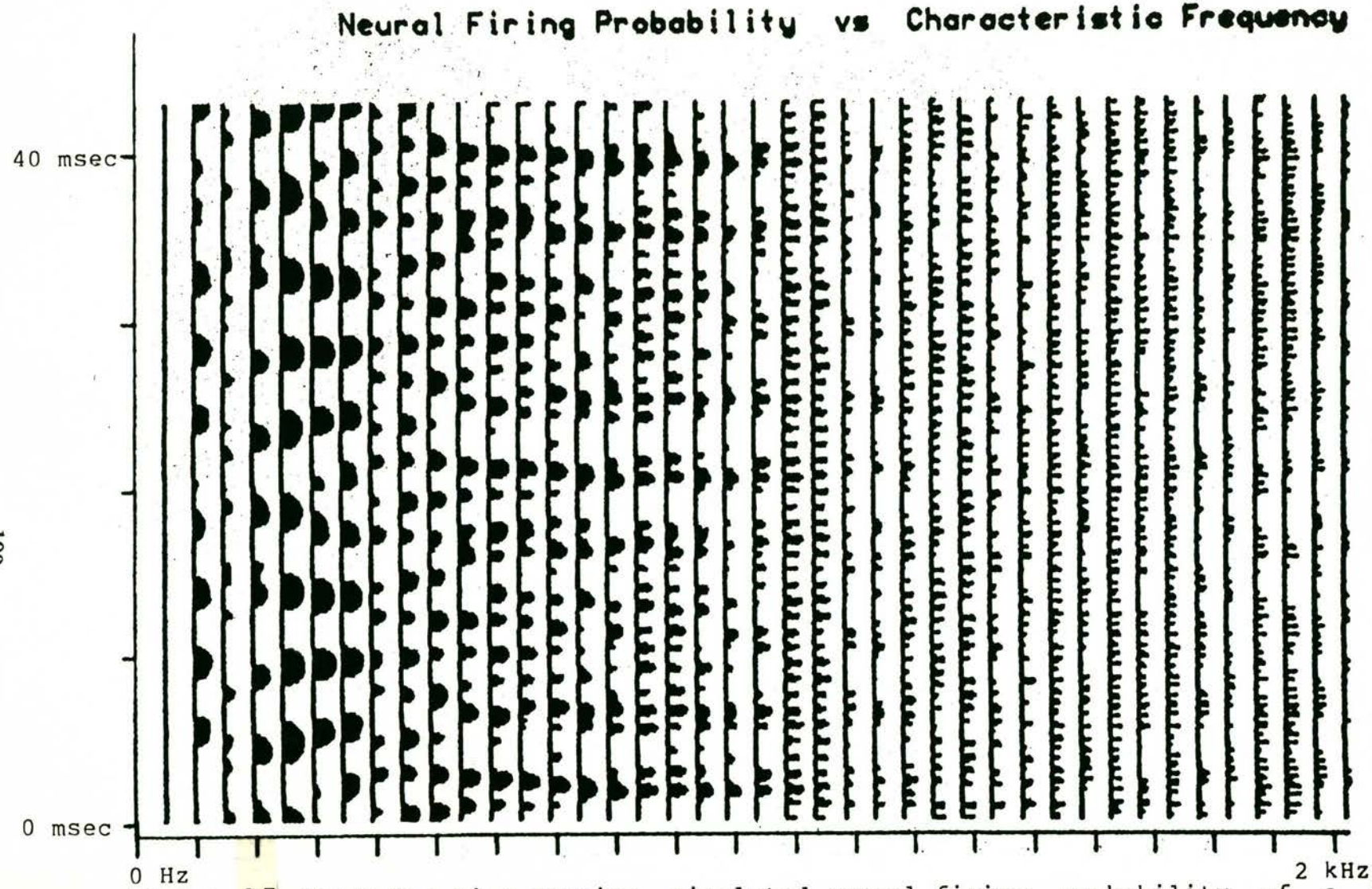


Figure 9.7 shows the time varying, simulated neural firing probability of a two tone signal.

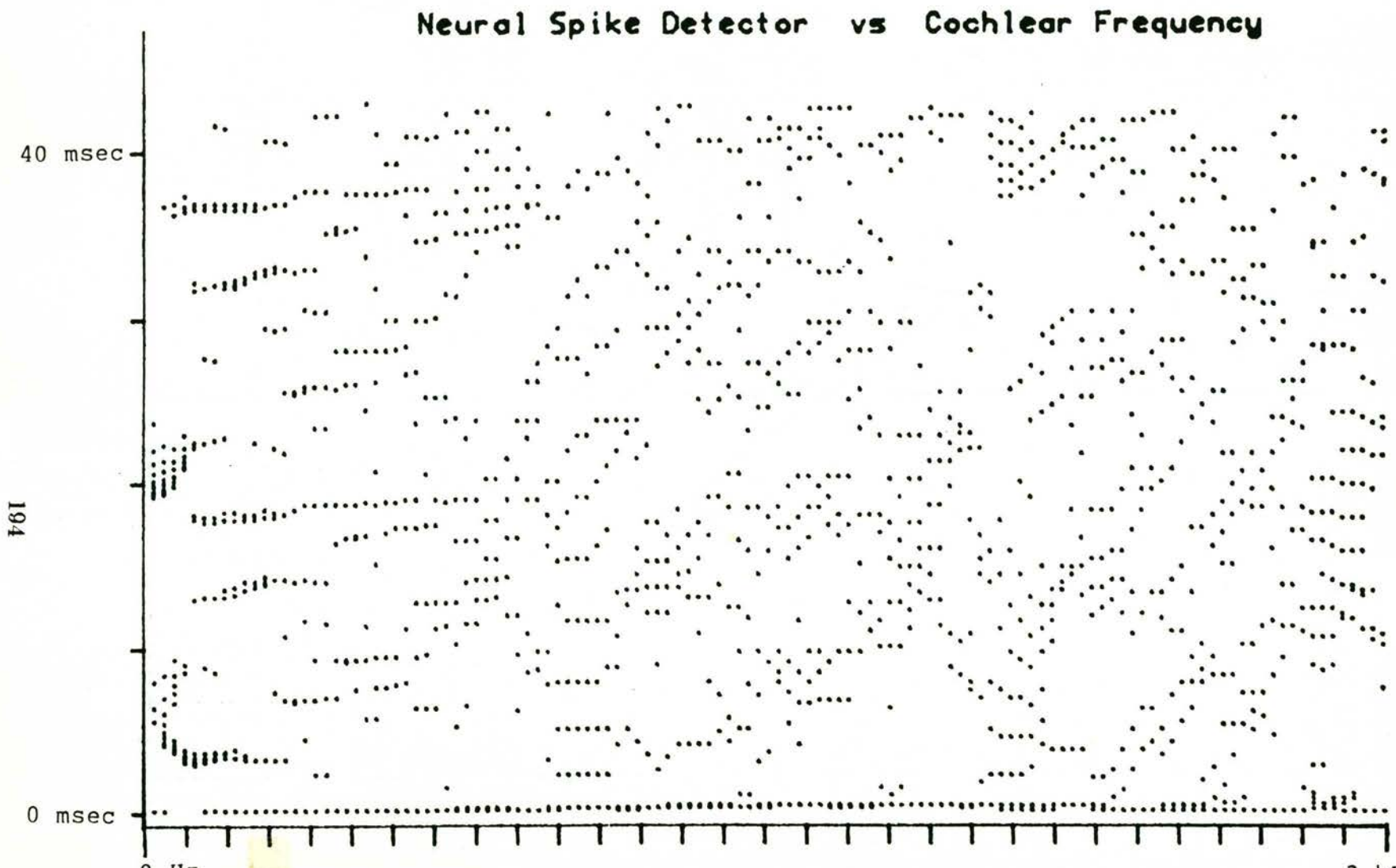


Figure 9.8 is a time varying plot of all the IHAM detector firings in the CF range 0 to 3 kHz.

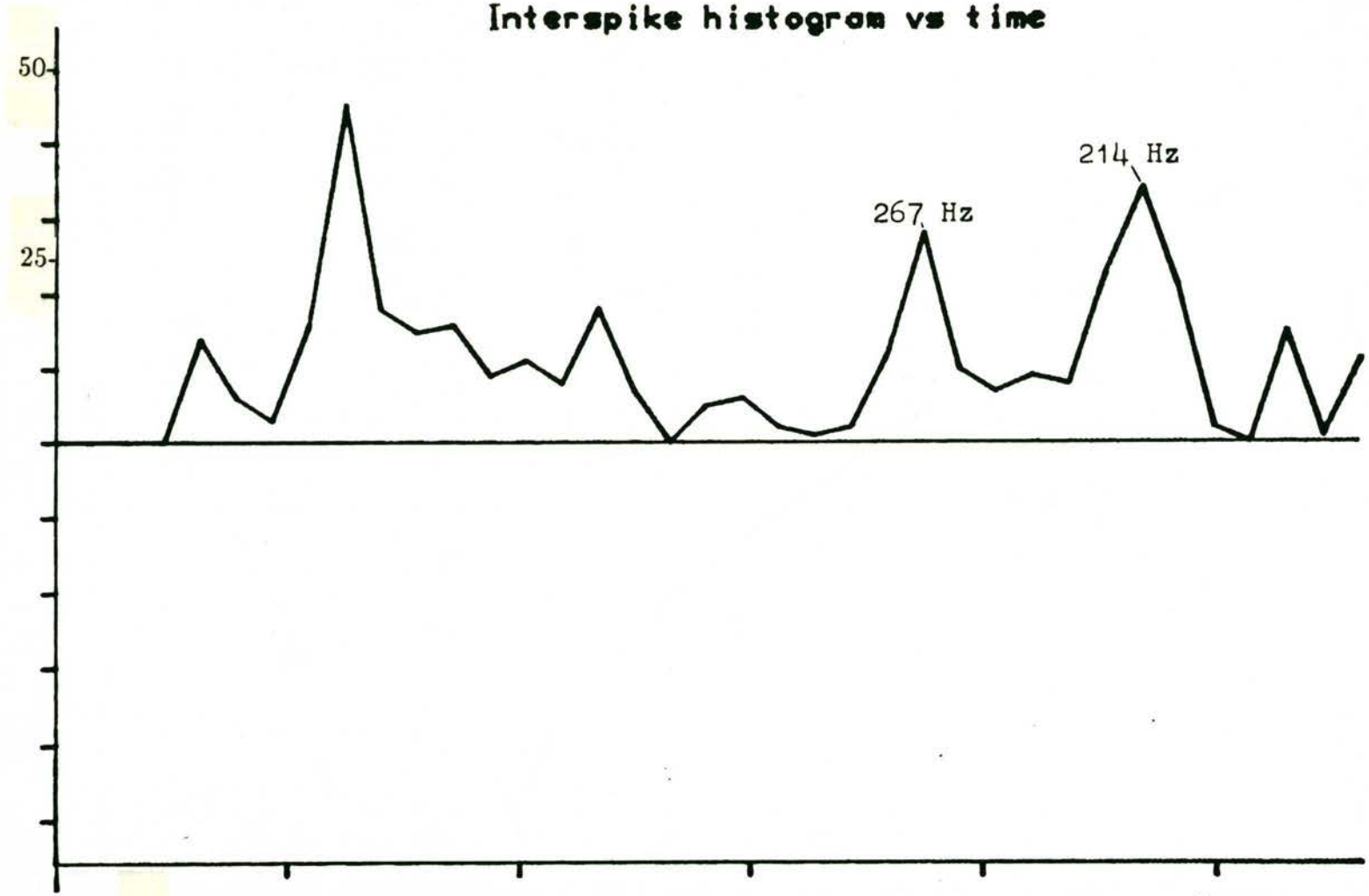


Figure 9.9 is a histogram of the inter-firing times of all the IHAM detectors in the CF range 0 to 2 kHz. Note the strong peaks corresponding to the periods of the two tones.

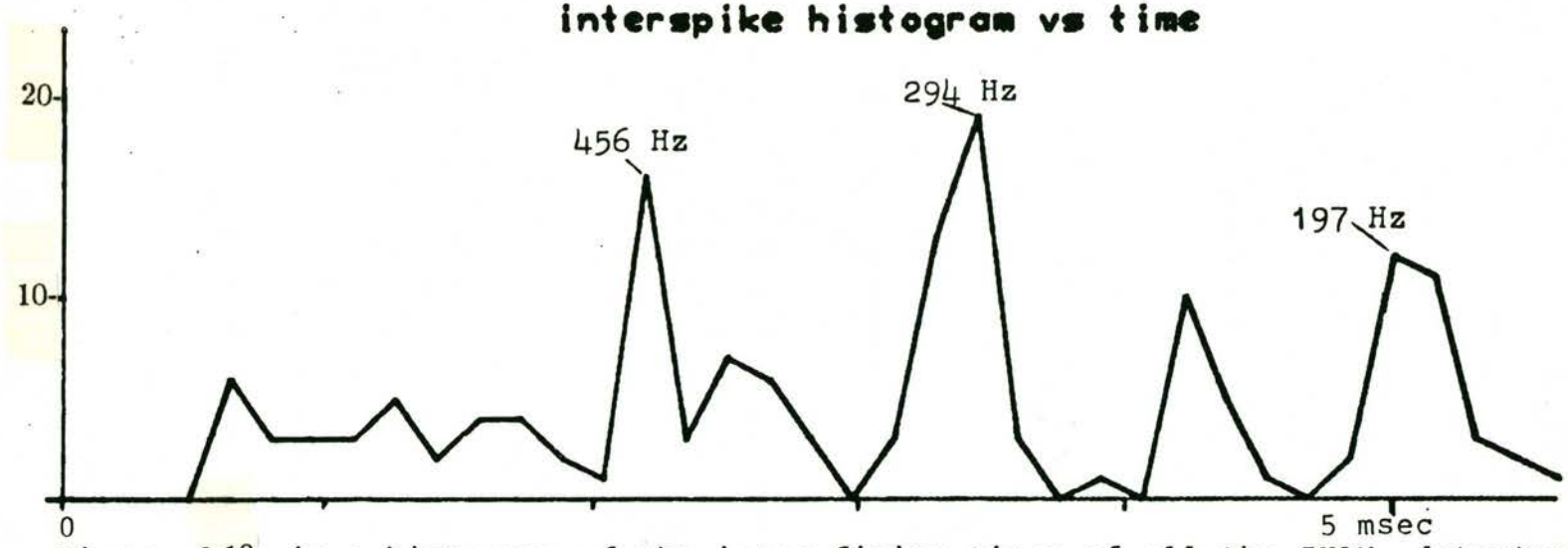


Figure 9.10 is a histogram of the inter-firing times of all the IHAM detectors in the CF range 0 to 2 kHz, for the Trio segment of figure 2.10

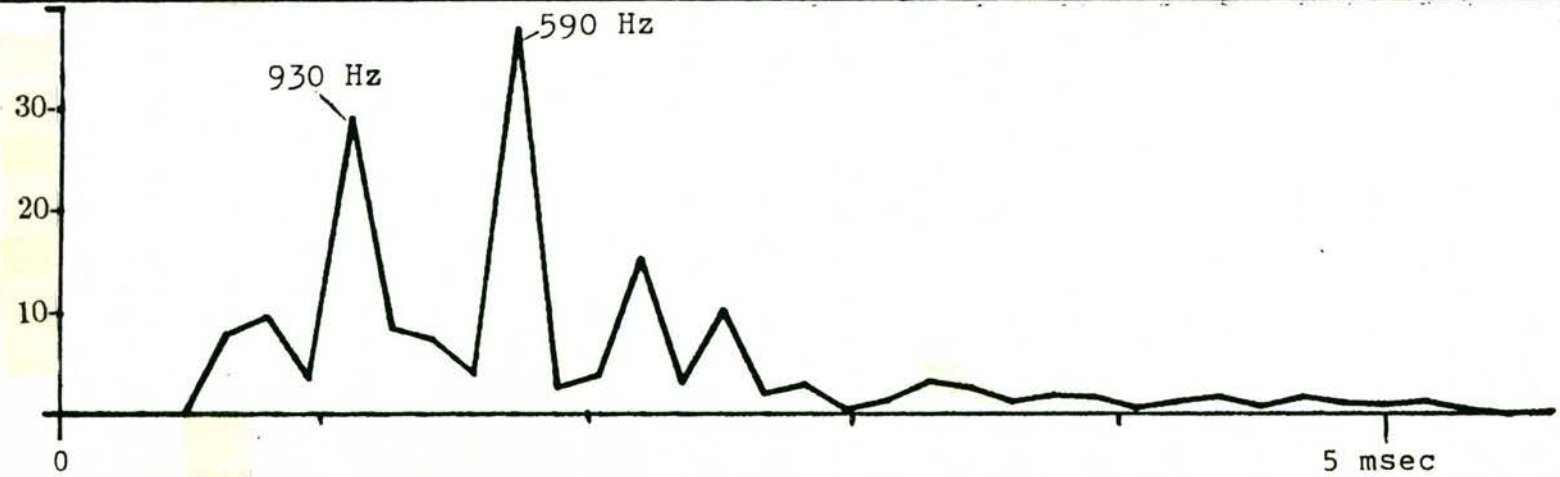


Figure 9.11 is a histogram of the inter-firing times of all the IHAM detectors in the CF range 0 to 2 kHz, for the Trio segment of figure 2.10

of pitch estimation, the IHAM detector is a reasonable hypothesis to account for the human ability of polyphonic perception.

Consider two periodic signals with fundamental frequencies; f_1 and f_2, presented to the ear. The firing pattern in the acoustic nerve would resemble that of figure 9.8. If the acoustic nerve signal is then passed through a set of IHAM detectors, with threshold decay times in the range 1 to 20 milliseconds, some will fire periodically at the rate of f_1. Others will fire at the rate of f_2, while others will fire at harmonically related frequencies or at random. If the IHAM detectors' output of period f_1 and f_2 is significant, then a neural network of positive feedback delay-loops with varying delay-times could autocorrelate these signals to determine the periodicities. The attention of the higher cortical functions could be directed to either component of the incoming signals. These cortical functions would take into account other cues (phase, timbre, amplitude and context) to determine the correct pitches. The IHAM detector could be a single neuron with a threshold decay half-life of the same order as the period it is required to detect, or a neural network serving the same purpose.

9.6 Conclusion

The response of the human cochlea to polyphonic tones was simulated. The observed amplitude modulation of the simulated response was found to be enough to distinguish superimposed tones. A neural mechanism was proposed (IHAM detector) to account for the differentiation of simultaneous tones. This mechanism can detect the pitch of sinusoidal tones as well as the pitch of periodic tones with missing lower harmonics.

CHAPTER TEN

Conclusions and Future Research

10.1 Summary

The central problem of this thesis is the automatic identification of simultaneously sounding tones from acoustical music recordings. Recorded music is entered via an analog-to-digital converter. Each digitized sample is then analysed to determine the pitches of the constituent tones. Finally the pitch estimates of the samples are grouped together in time to determine the pitches, onset times and durations of the notes.

Several signal processing algorithms were investigated to determine their applicability to the problem of separating the parts of polyphonic music, and new procedures and heuristics were developed to improve these results. The heuristics help to distinguish correct pitch estimates from harmonically related incorrect ones. An iterative procedure for extracting the harmonics of the most likely estimates from spectra gives a further improvement by reducing the effect of other tones present.

An interactive graphics signal processing system was developed by the author to test the algorithms, and to determine the optimal parameters for small segments of musical data. These algorithms were applied successively to generate pitch estimates for complete musical pieces. The pitch estimates were automatically grouped into notes and transcribed in standard music notation. The author developed software for determining key signatures and the required

accidentals to fully automate the process of generating output from the computer analyses in a form that is easy for musicians to recognize.

Error measures were defined to compare musical recordings with analysed results. These measures were used to determine the accuracy of analyses and the sensitivity of the algorithms to variations in parameters. The pitch accuracy was 99% for a three part synthesized Trio, and between 90% and 100% for piano recordings with as many as 5 notes sounding simultaneously. This is a significant advance on the work of Moorer (1975). The method used here allows a wide range of fundamental frequencies ranging from 50 Hz to 1 kHz (a musical range of over four octaves), and unlike Moorer's work, no restriction is made on the musical intervals.

The pitch grouping algorithms and plotting procedures can run in real-time, but the spectral analysis and pitch analysis procedures are much slower. For a piece of music of length 10 seconds, a pitch profile takes 6 seconds of central processing time (cpu time), the music plotting takes 9 seconds cpu time, but determination of the pitch estimates requires 500 seconds of cpu time (on a VAX 11/780).

A computer model for the response of the human cochlea to sound was developed, based on empirical research in psychoacoustics and auditory physiology, in an attempt to bridge the knowledge gap between high-level cognitive processes and low-level neural processes. A phenomenon of amplitude modulation was observed, corresponding to the fundamental frequencies of the applied tones. Some mechanisms were proposed to help explain the human ability to

discriminate simultaneous tones.

10.2 Future Research

Future research areas are proposed in this section for improving the results of this thesis. These areas include heuristics for pitch selection, improved music plotting, and pitch grouping strategies that exploit knowledge of the musical style.

The pitch grouping described in 6.5 could be improved by adaptively changing the time to empty a bucket, to match the average duration of the notes near to that time. This integrating function could also be weighted in favour of the stronger pitch estimates, instead of using a constant value. The minimum strength for accepting a pitch estimate could be varied to accommodate dynamic variations of the music. Pattern recognition techniques could be applied. For example, trills could be detected by matching an incomplete trill with a template.

Heuristics using musical knowledge and context could also improve the accuracy. For example, use could be made of the fact that most parts move by 1 or 2 semitones from one note to the next. Steps of 3, 4 and 8 semitones are the next most frequent. Notes of short duration (for example, trills and passing notes) often step by 1 or 2 semitones. Rules of harmony such as the resolution of dissonance and the expected movement of parts could also be applied.

For the algorithms described in this thesis, most errors are harmonically related to the tones being played. The identification of two tones sounding an octave apart is particularly difficult. The algorithms could be further refined to

use time varying spectral templates for the musical instruments. A harmonic of one tone of given frequency could be extracted from the signal without removing the harmonic components of the same frequency but belonging to different tones.

With the music plotting system, there are still a number of drawbacks which could be improved. Notes of duration less than a crotchet are plotted as independent short notes, and could be beamed together in groups of a crotchet duration. Notes of long duration are not split and tied across bar lines. The placement of bar-lines could be made more accurate by using adaptive beat tracking (see Harris 1982). And finally, the part allocation or voicing algorithm could be improved to track parts as independently moving melodies, by minimizing note steps from one note to the next in each part.

10.3 Applications

The computer simulation of the cochlea could be used to aid in the development of cochlear implant hearing aids. Research in Neurophysiology and Psychoacoustics could be done to investigate inter-harmonic amplitude modulation in the auditory nervous system.

This research could also find application in other signal processing areas such as speech analysis. Communication channels with multiple signals could be separated into their constituent signals.

The direct application of this research is in the area of automatic music printing, and as an automatic music scribe for musicologists. The music analysis and plotting software described in this thesis could be implemented (at a

reasonable cost) on any general purpose computer system with A/D conversion facilities and graphic output.

The analysis algorithms developed here could form the basis for a musician's assistant. The pitch profile would be a useful aid to transcription, and would provide an objective means for comparing performed music with a musical score. It would rely on the musician's visual pattern recognition, aural ability and musical knowledge. It may be possible in the future to engineer this expertise, and develop an expert system for transcribing sounded polyphonic music. This system could apply adaptive rhythm tracking (Harris 82), improved pitch grouping, and heuristics about the musical context.

By using an FFT processor with a 1024 point transform time of less than 10 milliseconds, it is feasible to build a dedicated hardware system, based on the results of this research, to transcribe polyphonic music automatically in real time. With the advent of very large scale integrated (VLSI) circuits and the ever diminishing computing costs, such a device will soon be economically viable.

BIBLIOGRAPHY

Allen J. (1975). Computer architecture for signal processing. *Proc. IEEE,* vol. 63, pp. 624-633.

Allen J. (1977). Cochlear micromechanics - a mechanism for transforming mechanical to neural tuning within the cochlea. *Journ. Acoust. Soc. Amer.,* vol.62, p.930.

Allen J. and Rabiner L.R. (1977). A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE,* vol. 65, pp. 1558-1564.

Alts R.A. (1978). The Fourier-Mellin transform and mammalian hearing. *Journ. Acoust. Soc. Amer.,* vol.63, pp.174-183.

Apel W. (1970). (Ed.) *Harvard Dictionary of Music,* Heineman, London.

Ashton A.C. (1970). *Electronics, Music and Computers.* Ph.D. Dissertation, University of Utah, Salt Lake City.

Backus J. (1970). *The Acoustical Foundations of Music.* John Murray, London.

Bariaux D., G. Cornelissen G., et al. (1975). A method for spectral analysis of musical sounds, description and performances. *Acustica,* vol. 32, pp. 307-313.

Beauchamp J.W. (1967). *A computer system for time-variant harmonic analysis and synthesis of musical tones.* Publication 992, Electrical Eng. Dept., University of Illinois, Urbana.

Beauchamp J.W. (1974). Time-variant spectra of violin tones. *Journ. Acoust. Soc. Amer.,* vol. 56, pp. 995-1004.

Beauchamp J.W. (1975). Analysis and synthesis of cornet tones using nonlinear interharmonic relationships. *Journ. Audio Eng. Soc.,* vol. 23, pp. 778-795.

Beauchamp K.G. (1975). *Walsh Functions and their Applications.* Academic Press, London.

Bekesy G. von (1963). Hearing theories and complex sounds. *Journ. Acoust. Soc. Amer.,* vol. 35, pp. 588-601.

Bekesy G. von (1960). *Experiments in Hearing.* McGraw-Hill New York.

Benade A.H. (1973). The physics of brasses. *Scientific American,* vol. 229(1), pp. 24-35.

Bengtsson Ingmar (1972). Sound analysis equipment at the Institute of Musicology in Uppsala. *Studia Instrumentorum Musicae Popularis.* vol. 2.

Benjamin R. (1980). Generalisations of maximum-entropy pattern analysis. *IEE Proc.* vol. 127, pp.341,353.

Bennett W.R. (1948). Spectra of quantized signals. *Bell Syst. Tech. Journ.,* vol. 27, pp. 446-472.

Bilsen F.A. (1973). On the influence of the number and phase of harmonics on the perceptibility of the pitch of complex signals. *Acustica,* vol. 28, pp. 60-65.

Bingham C., Godfrey M.D. and Tukey J.W. (1967). Modern techniques of power spectral estimation. *IEEE Trans. Audio Electroacoust.,* vol. AU-15, pp. 91-98.

Biondi E. and Grandori F. (1975). Modelling stimuli processing by the peripheral acoustic system. *Progress in Cybernetics and Systems Research,* vol. 1, p.321.

Blackman E.D. (1965). The physics of the piano. *Scientific American,* vol. 213(6), pp. 88-99.

Blesser B.A. (1978). Digitization of Audio: A comprehensive examination of theory, implementation, and current practice. *Journ. Audio Eng. Soc.,* vol. 26, pp.739-771.

Boker-Heil N. (1972). Plotting conventional music notation. *Journ. Music Theory,* vol. 16, pp. 72-101.

Brigham E.O. (1974). *The Fast Fourier Transform.* Prentice-Hall, New Jersey.

Byrd D. (1974). A system for music printing by computer. *Computers and the Humanities,* vol. 8, p. 161.

Childers D.G., Skinner D.P. and Kemerait R.C. (1977). The Cepstrum: A guide to processing. *Proc. IEEE,* vol. 65, pp. 1428-1442.

Chowning J.M., Grey J.M., Rush L., Moorer J.A. (1974). *Computer simulation of music instrument tones in reverberant environments.* Report STAN-M-1, Center for Computer Research in Music and Acoustics, Stanford University.

Chowning J.M. (1981). Computer Synthesis of the Singing Voice. *Proceedings of the International Conference on Music and Technology.* Melbourne, Australia.

Clark G.M., Black R. et al. (1977). A multiple-electrode hearing prosthesis for cochlear implantation. *Medical Progress through Technology,* vol. 5, pp. 127-140.

Cooley J.W., Tukey J.W. (1965). An Algorithm for the machine calculation of complex Fourier series. *Math. Computation,* vol. 19, pp. 297-301.

Connor P.M. (1977). Harmoniac - a digital audio-signal processor. *Proc. 16th*

Int. Radio & Electronics Conv.

Delosme J.M., Friedlander B. and Morf M. (1980). Source location from time differences of arrival: identifiability and estimation. *IEEE Conf. Proc.*

Deutch D. (1975). Musical Illusions. *Scientific American,* vol. 233(4), pp. 92-104.

Dubnowski J.J., Schafer R.W. and Rabiner L.R. (1976). Real time digital hardware pitch detector. *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-24, p.2.

Eccles J.C. (1979). *The Human Mystery.* Springer Int.

Eccles J.C. and Popper K.R. (1977). *The Self and its Brain.* Springer Int.

Erickson R.F. (1968). Musical analysis and the computer: a report on some current approaches and the outlook for the future. *Computers and the Humanities,* vol.3, pp.87-104.

Evans E.F. and Wilson J.P. (1977). *Psychophysics and Physiology of Hearing,* Academic Press, London.

Evans E.F. and Wilson J.P. (1975). Cochlear tuning properties: concurrent basilar membrane and single nerve fibre measurements. *Science,* vol. 190, pp. 1218-1221.

Fedor P. (1977). Principles of the design of D-neuronal networks *Biol. Cybernetics,* vol. 27, pp. 129-146.

Flanagan J.L. (1962). Computational model for basilar membrane displacement, *Journ. Acoust. Soc. Amer.,* vol. 34, p. 1370.

Flanagan J.L. (1972). *Speech Analysis, Synthesis and Perception.* Springer-Verlag, New York.

Fletcher H., Blackham E.D. and Stratton R. (1962). Quality of piano tones. *Journ. Acoust. Soc. Amer.,* vol. 34, pp. 749-761.

Fletcher H. (1934). Loudness, pitch and timbre of musical tones and their relation for the intensity,the frequency and the overtone structure. *Journ. Acoust. Soc. Amer.,* vol. 6, pp. 59-69.

Fletcher N.H. (1978). Mode locking in nonlinearly excited inharmonic musical oscillators. *Journ. Acoust. Soc. Amer.,* vol. 64, pp. 1566-1569.

Fredlund L.D. and Sampson J.R. (1973). An interactive graphics system for computer-assisted musical composition. Int. *Journ. Man-Machine Studies,* vol. 5, pp.585-605.

Freedman M.D. (1967). Analysis of musical instrument tones. *Journ. Acoust. Soc. Amer.*, vol. 41, pp. 793-806.

Gold B. and Rabiner L.R. (1969). Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journ. Acoust. Soc. Amer.*, vol. 46, p. 442.

Gold B. (1962). Computer program for pitch extraction. *Journ. Acoust. Soc. Amer.*, vol. 34, pp. 916-921.

Goldstein J.L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journ. Acoust. Soc. Amer.*, vol. 54, pp. 1496-1516.

Grey J.M. (1975). *An Exploration of Musical Timbre.* Ph.D. dissertation, Center for Computer Research in Music and Acoustics, Stanford University.

Harris C.M. and Weiss M.R. (1963). Pitch extraction by computer processing of high-resolution Fourier analysis data. *Journ. Acoust. Soc. Amer.*, vol. 35, pp. 339-343.

Harris F.J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE,* vol. 66, pp. 51-83.

Harris W. (1982). *The Automatic Analysis of Rhythm in Keyboard Music.* M.Sc. Thesis, University of Sydney.

Harrison R.V., Jean-Marie Aran, and Jean-Paul Erre (1981). AP tuning curves from normal and pathological human and guinea pig cochleas. *Journ. Acoust. Soc. Amer.*, vol. 69, pp. 1374-1385.

Helmholtz H. von (1863). *On the Sensations of Tone as a Physiological Basis for the Theory of Music.* translated and reprinted in English by Ellis, Dover, New York. (1954).

Hiller L.A. (1959). Computer music. *Scientific American,* vol. 201(6), pp. 109-120.

Hiller L.A. and Baker R.A. (1965). Automated music printing. *Journ. Music Theory,* vol. 9, pp. 129-150.

Houtsma A.J.M. and Goldstein J.L. (1972). The central origin of the pitch of complex tones: Evidence from musical interval recognition. *Journ. Acoust. Soc. Amer.*, vol. 51, pp. 520-529.

Hundley T.C., Benioff H. and Martin D.W. (1978). Factors contributing to the multiple rate of piano tone decay. *Journ. Acoust. Soc. Amer.*, vol. 64, pp. 1303-1309.

Hunt F.V. (1935). A direct-reading frequency meter suitable for high speed recording. *The Review of Scientific Instruments,* vol. 6, Feb 1935, pp. 43-46.

Jackson R. (1967). The computer as a student of harmony. Report of the *Tenth Congress of the International Musicological Society.* Ljubljana. Cvetko (Ed.), pp. 435-450.

Jain V.K., Collins W.L. and Davis D.C. (1980). DFT interpolation for estimation of tone amplitudes and phases. *IEEE Conf. Proc. Acoust., Speech & Sig. Proc.*

Kassler M. (1970). MIR - a simple programming language for music information retrieval. In: Lincoln (Ed.) (1970), pp. 299-327.

Kassler M. and Howe H.S. (1975). Computers and Music. In: *Grove's Dictionary of music and Musicians,* 6th Ed., Macmillan, London.

Kassler M. (1977). *Computer-Assisted Music Printing.* Report to the Music Board of the Australia Council.

Keeler J.S. (1972). Piecewise-periodic analysis of almost-periodic sounds and musical transients. *IEEE Trans. Audio Electroacoust.,* vol. AU-20, pp. 338-344.

Keeler J.S. (1972). The attack transients of some organ pipes. *IEEE Trans. Audio Electroacoust.,* vol.AU-20, pp.378-391.

Kemerait R.C. and Childers D.G. (1972). Signal Detection and extraction by cepstrum techniques. *IEEE. Trans. Information Theory,* vol. IT-18, pp. 745-759.

Kiang N.Y.S. and Moxon E.C. (1974). Tails of tuning curves of auditory nerve fibres. *Journ. Acoust. Soc. Amer.,* vol. 55, p. 620.

Kim D.O., Molnar C.E., Pfeiffer R.R. (1973). A system of nonlinear differential equations modeling basilar-membrane motion. *Journ. Acoust. Soc. Amer.,* vol. 54, pp. 1517-1529.

Knowlton P.H. (1971). *Interactive Communication and Display of Keyboard Music.* Ph.D. dissertation, University of Utah, Salt Lake City.

Knuth D.E. (1969). *The Art of Computer Programming,* vol. 2, Addison-Wesley, Reading, Massachusetts.

Lanczos C. (1966). Discourse on Fourier Series. In: *University Mathematical Monographs.* (Ed. Rutherford D.E.), Oliver and Boyd, London.

Lehman P.R. (1964). Harmonic structure of the tone of the bassoon. *Journ. Acoust. Soc. Amer.,* vol.36, pp.1649-1653.

Lichte W.H. and Gray R.F. (1955). The influence of the overtone structure on

the pitch of complex tones. *Journ. Exp. Psych.*, vol. 49, p. 431.

Lincoln H.B. (Ed.) (1970). *The Computer and Music*. Ithaca, New York.

Longuet-Higgins H.C. (1976). Perception of Melodies. *Nature*. vol. 263, pp. 646-653.

Longuet-Higgins H.C. and Steedman M.J. (1971). On interpreting Bach. *Machine Intelligence,* vol. 6, pp. 221-241.

Luce D.A. and Clark M. (1965). Durations of attack transients on nonpercussive orchestral instruments. *Journ. Audio Eng. Soc.*, vol. 13, pp. 194-199.

Luce D.A. and Clark M. (1967). Physical correlates of brass instrument tones. *Journ. Acoust. Soc. Amer.*, vol.42, pp. 1232-1243.

Makhoul J. (1973). Spectral analysis of speech by linear prediction. *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp.140-148.

Maksym J.N. (1973). Real-time pitch extraction by adaptive prediction of the speech waveform. *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp.149-154.

Markel J.D. (1972). The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp.367-377.

Mars P. and Cattanach J.M. (1977). Automatic transcription of keyboard music. *Proc. IEE,* vol. 124, pp. 436-440.

Mathews M.V. (1969). *The Technology of Computer Music*. The M.I.T. Press, Cambridge, Massachusetts.

Metfessel M. (1926). Technique for objective studies of the vocal art. *University of Iowa Studies in Psychology,* vol. 9, pp. 1-40.

Miller N.J. (1975). Pitch detection by data reduction. *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-23, pp.72-79.

Moller A.R. (Ed.) (1973). *Basic Mechanisms in Hearing*. Academic Press, London.

Moore M. (1974). The Seeger melograph model C. *Selected Reports in Ethnomusicology,* vol. 2, p.3.

Moorer J.A. (1974). The optimum comb method of pitch period analysis of continuous digitized speech. *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-22, pp.330-338.

Moorer J.A. (1975). *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*. Ph.D. dissertation, Center for Computer Research

in Music and Acoustics, Stanford University.

Moorer J.A. (1976). The synthesis of complex audio spectra by means of discrete summation formulas. *Journ. Audio Eng. Soc.*, vol. 24, pp. 717-727.

Moorer J.A. (1977). Signal processing aspects of computer music: a survey. *Proc. IEEE,* vol. 65, pp.1108-1137.

Nelder J.A. and Mead R. (1965). A Simplex method for function minimization. *Computer Journal,* vol.7, p308.

Noll A.M. (1964). Short-time spectrum and "cepstrum" techniques for vocal-pitch detection. *Journ. Acoust. Soc. Amer.*, vol.36, pp. 296-302.

Noll A.M. (1967). Cepstrum pitch determination. *Journ. Acoust. Soc. Amer.*, vol.41, pp. 293-309.

Noll A.M. (1968). Clipstrum pitch determination. *Journ. Acoust. Soc. Amer.*, vol.44, pp. 1585-1591.

Obata J. and Kobayashi R. (1937). A direct-reading pitch recorder and its applications to music and speech. *Journ. Acoust. Soc. Amer.*, vol. 9, pp. 156-161.

Obata J. and Kobayashi R. (1938). An apparatus for direct-recording the pitch and intensity of sound. *Journ. Acoust. Soc. Amer.*, vol. 10, pp. 147-149.

Ohm G.S. (1843). Uber die definition des tones, nebst daran geknupfter theorie der sirene und ahnlicher tonbildender vorrichtungen. *Ann. Phys. Chem.* vol. 59, pp. 513-565.

Oppenheim A.V., Schafer R.W. and Stockham T.G. (1968). Non-linear filtering of multiplied and convolved signals. *Proc. IEEE,* vol. 56, pp. 1264-1291.

Oppenheim A.V. and Schafer R.W. (1975). *Digital Signal Processing.* Prentice-Hall, New Jersey.

Papoulis A. (1962). *The Fourier Integral and its applications.* McGraw-Hill Book Co., New York.

Parker S.E. (1947). Analyses of the tones of wooden and metal clarinets. *Journ. Acoust. Soc. Amer.*, vol. 19, pp. 415-419.

Pinkerton R.C. (1956). Information theory and melody. *Scientific American,* vol. 194(2), pp. 77-86. cal sound. *Proc. 1978 Int. Computer Music Conf.* Roads (Ed.), Northwestern University Press.

Piszczalski M. (1979). Spectral surfaces from performed music. part 1. *Computer Music Journal,* vol.3, p.18.

Plomp R. (1964). The ear as a frequency analyzer. *Journ. Acoust. Soc. Amer.*, vol. 36, p. 1526.

Plomp R. (1967). Pitch of complex tones. *Journ. Acoust. Soc. Amer.*, vol. 41, p. 1628.

Plomp R. and Smoorenburg G.F. (1970). *Frequency Analysis and Periodicity Detection in Hearing.* Suithoff, Leiden.

Prerau D.S. (1971). Computer pattern recognition of printed music. *AFIPS Fall Joint Computer Conference,* pp. 153-162.

Rabiner L.R. and Gold B. (1975). *Theory and Application of Digital Signal Processing.* Prentice-Hall, New Jersey.

Rabiner L.R., Cheng M.J., Rosenburg A.E. and McGonegal C.A. (1976). A comparative performance study of several pitch detecting algorithms. *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-24, pp. 399-418.

Rabiner L.R. (1976). On the use of autocorrelation analysis for pitch detection. *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-25, pp. 24-33.

Rasch R.A. (1978). The perception of simultaneous notes such as in polyphonic music. *Acustica,* vol. 40. pp. 21-33.

Rayleigh J.W.S. (1896). *Theory of Sound.*

Rhode W.S. (1971). Observations of the vibration of the BM in squirrel monkeys using the Mössbauer technique. *Journ. Acoust. Soc. Amer.*, vol. 49, p. 1218.

Richardson E.G. (1954). The transient tones of wind instruments. *Journ. Acoust. Soc. Amer.*, vol. 26, pp. 960-962.

Rife D.C. and Vincent G.A. (1970). Use of the discrete Fourier transform in the measurement of frequencies and levels of tones. *Bell Sys. Tech. Journ.*, Feb 1970.

Rigden J.S. (1977). *Physics and the Sound of Music.* Wiley, New York.

Risset J.C. and Mathews M.V. (1969). Analysis of musical instrument tones. *Physics Today,* vol. 22, pp. 23-30.

Ritchie D.M. and Thompson K. (1975). *The UNIX Programming Manual.* Bell Telephone Laboratories.

Ritchie D.M. and Thompson K. (1975). *The C Reference Manual.* Bell Telephone Laboratories.

Ritsma R.J. (1967). Frequencies dominant in the perception of the pitch of

complex sounds. *Journ. Acoust. Soc. Amer.*, vol. 42, p. 191.

Robinson T.D. (1967). IML-MIR: a data-processing system for the analysis of music. In: Heckman (Ed.) (1967), *Elektronishe Datenverbeitung in der Musikwissenschaft.* Verlag, Regensburg. pp, 103-135.

Roederer J.G. (1977). *Introduction to the Physics and Psychophysics of Music.* Springer-Verlag.

Rose J.E., Brugge J.F., Anderson D.J. and Hind J.E. (1969). Some possible neural correlates of combination tones. *Journ. Neurophys.*, vol. 32, p. 402.

Ross M.J., Shaffer H.L., et al. (1974). Average magnitude difference function pitch extractor. *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-22, pp.353-362.

Saunders F.A. (1946). Analysis of the tones of a few wind instruments. *Journ. Acoust. Soc. Amer.*, vol.18, p.395.

Schafer R.W. and Rabiner L.R. (1973). A digital signal processing approach to interpolation. *Proc. IEEE,* vol. 61, pp. 692-702.

Schafer R.W. (1969). *Echo removal by discrete generalized linear filtering.* Ph.D. thesis, Massachusetts Institute of Technology.

Schouten J.F., Ritsma R.J. and Cardozo B.L. (1962). Pitch of the residue. *Journ. Acoust. Soc. Amer.*, vol. 34, pp. 1418-1424.

Schroeder M.R. (1970). Parameter estimation in speech: A lesson in unorthodoxy. *Proc. IEEE,* vol. 58, pp. 707-712.

Schroeder M.R. (1975). Models of Hearing. *Proc. IEEE,* vol. 63, pp. 1332-1350.

Schroeder M.R. and Atal B.S. (1962). Generalized short-time power spectra and autocorrelation functions. *Journ. Acoust. Soc. Amer.*, vol. 34, pp. 1679-1683.

Seashore C.E. (1932). The vibrato. *University of Iowa Studies in the Psychology of Music,* vol. 1.

Seeger C. (1951). An instantaneous music notator. *Journ. Int. Folk. Music. Council,* vol. 3, p 103.

Seeger C. (1957). Toward a universal music sound-writing for musicology. *Journ. Int. Folk. Music. Council,* vol. 9, p 63.

Siebert W.M. (1970). Frequency discrimination in the auditory system: place or periodicity mechanisms. *Proc. IEEE,* vol. 58, pp.723-730.

Smith L.C. (1973). Editing and printing music by computer. *Journ. Music*

Theory, Fall 1973, pp. 292-308.

Smoorenburg G.F. (1970). Pitch perception of two-frequency stimuli. *Journ. Acoust. Soc. Amer.,* vol. 48, p. 924.

Solomon L.N. (1958). Semantic approach to the perception of complex sounds. *Journ. Acoust. Soc. Amer.,* vol. 30, pp. 421-425.

Sondhi M.M. (1968). New methods of pitch extraction. *IEEE Trans. Audio Electroacoust.,* vol. AU-16, pp. 442-448.

Stockham T.G., Cannon T.M. and Ingebretsen R.B. (1975). Blind deconvolution through digital signal processing. *Proc. IEEE,* vol. 63, pp. 624-633.

Strong W. and Clark M. (1967). Perturbations of synthetic orchestral wind-instrument tones. *Journ Acoust. Soc. Amer.,* vol. 41, pp. 277-285.

Styles B.C. (1974). Describing music to a computer. Int. *Journ. Man-Machine Studies,* vol. 6, pp.125-134.

Sundburg T. (1977). The acoustics of the singing voice. *Scientific American,* vol. 236(3), pp. 82-91.

Terhardt E. (1974). Pitch, consonance and harmony. *Journ. Acoust. Soc. Amer.,* vol. 55, p. 1061.

Tobias J.V. (Ed.) (1970). *Foundations of Modern Auditory Theory. Vol. 1,* Academic Press, New York.

Tove P.A., Norman B., et al. (1966). Direct-recording frequency and amplitude meter for analysis of musical and other sonic waveforms. *Journ. Acoust. Soc. Amer.,* vol. 39, pp. 362-371.

Tucker W.H. and Bates R.H.T. (1978). A pitch estimation algorithm for speech and music. *IEEE Trans. Acoust. Speech, and Signal Proc.,* vol. ASSP-26, pp. 597-604.

Tucker W.H., Bates R.H.T. et al. (1977). An interactive aid for musicians. *Int. Journ. Man-Machine Studies,* vol. 9, pp. 635-651.

Tucker W.H. (1977). *Interactive Computer Based Music Systems.* Ph.D. Dissertation. University of Canterbury. Christchurch, New Zealand.

Ward W.D. (1954). Subjective musical pitch. *Journ. Acoust. Soc. Amer.,* vol. 26, pp. 369-380.

Ward W.D. (1970). Musical perception. In: Tobias (1970).

Weiss M.R., Vogel R.P. and Harris C.M. (1966). Implementation of a pitch

extractor of the double-spectrum-analysis type. *Journ. Acoust. Soc. Amer.*, vol.40, pp.657-662.

Weyer R.D. (1977). Time-varying amplitude-frequency-structures in the attack transients of piano and harpsichord sounds - I. *Acustica,* vol.35, p. 232.

Weyer R.D. (1977). Time-varying amplitude-frequency-structures in the attack transients of piano and harpsichord sounds - II. *Acustica,* vol.36, p. 241.

Whittaker E.T. and Robinson G. (1946). *The Calculus of Observations, A Treatise on Numerical Mathematics.* Blackie & Son, London.

Wiener N. (1974). Extrapolation, Interpolation and Smoothing. *American Scientist,* vol. 62, pp. 208-215.

Winograd T. (1968). Linguistics and the computer analysis of tonal harmony. *Journ. Music Theory,* vol. 12, pp. 2-49.

Winograd S. (1978). On computing the discrete Fourier transform. *Mathematics of Computation,* vol. 32, p. 175.

Wise J.D., Caprio J.R. and Parks T.W. (1976). Maximum likelihood pitch estimation. *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-24, pp. 418-423.

Wood A. (1940). *Physics of Music.* Academic Press, London.