Discrimination of spectral density

W. M. Hartmann
Department of Physics, Michigan State University, East Lansing, Michigan 48824[a] *and Institut de Recherche et Coordination Acoustique/Musique, 31, rue Saint-Merri, F-75004 Paris, France*

Stephen McAdams, Andrew Gerzso, and Pierre Boulez
Institut de Recherche et Coordination Acoustique/Musique, 31 rue Saint-Merri, F-75004, Paris, France

(Received 30 July 1985; accepted for publication 27 February 1986)

Experiments were performed to determine the ability of human listeners to discriminate between a sound with a large number of spectral components in a band, of given characteristic frequency and bandwidth, and a sound with a smaller number of components in that band. A pseudorandom placement of the components within the band ensured that no two sounds were identical. The data suggested that discrimination is primarily based upon the perception of temporal fluctuations in the intensity of the sound and secondarily upon resolved structure in the spectrum, perceived as tone color. Experiments using clusters of complex harmonic sounds showed that listeners are able to use the information in upper harmonic bands to discriminate spectral density.

PACS numbers: 43.66.Ba, 43.66.Fe, 43.66.Mk [RDS]

INTRODUCTION

The work presented in this paper is an attempt to discover how human listeners discriminate between sounds with different spectral densities when the spectral density for both sounds is high. Experimentally, we posed the following question: How many discrete spectral components must there be in a given frequency band so that the resulting sound is indistinguishable from a noise with an arbitrarily large number of components in that band? Schafer *et al.* (1950) studied a related question. They found the minimum density of sine components for which the masking of a sine tone was the same as the masking produced by thermal noise. These authors concluded that in order to provide masking equivalent to a band of thermal noise 32 Hz wide, a spectral density of one sine wave component per Hz was required.

The question of discrimination between large and small spectral densities was posed by one of us (Gerzso, 1980) in connection with the synthesis of dense spectra by a digital synthesizer with a large number of oscillators. Experiments were done using recorded stimuli with varying numbers of sine waves in a critical band. An experimental trial included three sounds, two of which were the same, with the remaining sound having a different spectral density. The tape was played to panels of listeners who were asked to decide which of the three sounds was different from the other two. The results showed that spectral densities which were just discriminable from thermal noise were considerably smaller than those found in the masking study by Schafer *et al.* (1950). The results also showed that the required spectral density was smaller when the sounds were heard in a reverberant environment compared to a drier environment.

When the experimental tape used in Gerzso's experiments was heard repeatedly through headphones, however, an unexpected result occurred: The ability to distinguish between thermal noise and a less dense sound turned out not to be a monotonically decreasing function of the number of components in the less dense sound. We tentatively attributed this effect to the ability of the listener to learn to recognize certain intensity fluctuation patterns in the constant-waveform stimuli and to make responses based upon those particular patterns. To eliminate this artifact, the experiments reported in the present paper were done in a way which made it impossible for the listener to learn patterns associated with a particular spectral density.

I. EXPERIMENT 1: SINE WAVEFORMS

A. Task

Subjects were presented with two sounds in succession; one of them, the standard, had 60 sine components, the other had a variable number N of sine components, $3 < N < 25$. The subject's task was to choose the sound which had the larger number of components.

B. Stimuli

The sounds were produced by the 4C synthesizer at the Institut de Recherche et Coordination Acoustique/Musique. By using the synthesizer, we were able to make each sound different from every other sound. The set of frequencies and initial phase angles for the components of each successive sound were determined by a different set of random numbers. This approach prevented subjects from learning a pattern of fluctuations which they could associate with a given number of components.

In order to minimize the effect on the bandwidth of randomly choosing the component frequencies, we put the components into bins, for both the standard and the test sound. For a given number N of components, we divided the band into N bins of equal width in hertz; we put one component at a random frequency within each bin, using a rectangular

[a] Permanent address and address for correspondence.

 0001-4966/86/061915-11$00.80

probability distribution. All N components ($N = 60$ or N variable) had the same amplitude, proportional to the inverse square root of N, so that all sounds had equal intensity, 75 dBA, independent of N.

The synthesizer patch allowed 60 oscillators, which imposed an upper limit of 60 components for the most dense sound. The sample rate was 16 000 Hz, and the resolution with which frequencies could be defined was $16\,000/2^{24} = 0.001$ Hz. The digital waveform was converted to an audio signal by a 16-bit DAC and an 8-KHz low-pass filter.

C. Procedure

The procedure was essentially a method of constant stimuli. The number of components in the variable sound took on 12 different values: 3,5,7,...,25. In each experimental block, there were ten repetitions of each value of N, presented in random order, for a total of 120 trials, lasting about 7 min.

Sounds were presented diotically by Beyer DT 48 headphones to subjects seated in a sound-treated room. Each trial consisted of five intervals: 600-ms warning interval, first sound interval of duration T($T = 500$, 1000, or 2000 ms), gap of 500 ms, second sound interval of duration T, and a response interval which was subject controlled. Sounds were turned on and off with a raised cosine envelope of 10-ms duration.

Subjects learned to identify the sound with the larger number of components during training runs in which feedback was given after each response by means of colored lights. When subjects believed that they had learned how to do the task, the feedback was turned off and testing continued until after it appeared that subjects had reached a stable level of performance.

D. Parameters

The primary parameters in the study were the band bottom frequency f_b and the bandwidth W. The bottom frequency took on values $f_b = 500$, 1000, and 2000 Hz. Some additional blocks were done at values of 250 and 3000 Hz. The bandwidth values were $W = 50$, 100, 200, 400, and 800 Hz. Experimental blocks with different parameters were done in a haphazard order. Training blocks with feedback were done when a listener was first introduced to new values

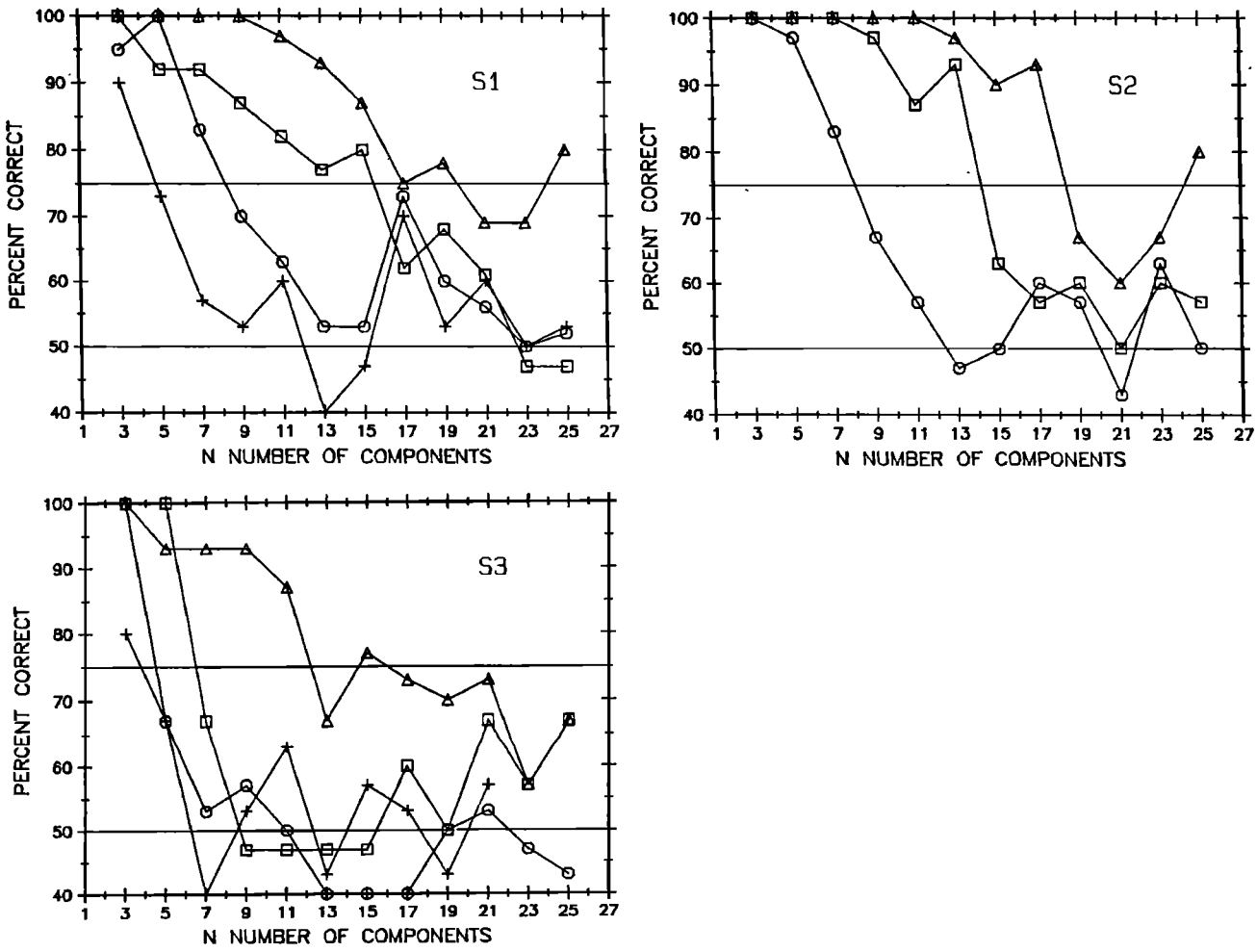

FIG. 1. Typical psychometric functions showing the percentage of correct identification of the signal with a larger number of components as a function of N, the smaller number of components, for three subjects, for four values of the band bottom frequency and the bandwidth: + $f_b = 250$ Hz, $W = 50$ Hz; circles, $f_b = 500$ Hz, $W = 100$ Hz; squares, $f_b = 1000$ Hz, $W = 200$ Hz; and triangles, $f_b = 2000$ Hz, $W = 400$ Hz. These conditions correspond to $Q = 5.5$. The duration was $T = 500$ ms.

of the parameters. To obtain final data, we averaged the results of the last three blocks (30 trials per value of N for each subject) done without feedback, for each of the sets of parameters in the study.

E. Subjects

Three subjects, S1, S2, and S3, participated in the experiments. Subjects S1 and S3 were authors and experienced listeners; all subjects had normal hearing according to their own reports.

F. Results

Typical data are presented in the psychometric functions in Fig. 1. The abscissa is N, the number of components in the variable sound, which was always compared with a 60-component sound. The ordinate is the percentage of trials where the subject judged correctly which sound had the larger number of components.

As expected, performance generally decreases as N increases. It was not evident *a priori*, however, that performance would decrease monotonically with N. It seemed possible that several different cues would become available at different values of N and that performance might show a minimum or a second peak at higher values of N. For most of the values of the experimental parameters, the upper limit of $N = 25$ was well above the value of N for which performance had fallen to the chance level of 50% correct, giving us the opportunity to look for such nonmonotonic behavior. Second peaks did appear quite often in our data; Fig. 1 is typical in that respect. However, when additional blocks of trials were done, up to a total of ten blocks, the additional structure disappeared. Further, there was little agreement among the subjects as to the value of N for second peaks or unusually large dips.

We concluded that there was no systematic evidence that performance is not a monotonically decreasing function of N. The structure in the psychometric functions for individual listeners which suggested the contrary was attributed to our limited sampling. We therefore fitted the psychometric functions by eye with a smooth monotonically decreasing curve, and then found the 75% correct point, $N(75)$, to describe a threshold. Threshold values for the various experimental parameters are given in the tables testing the hypotheses discussed below.

The tables show that values of $N(75)$ for listener S3 were usually smaller than those for the other two listeners. This fact will not affect our tests of the hypotheses, which are done within subjects and across conditions.

G. DISCUSSION

1. General

Apparently, the technique of equating the intensity of sounds with different N was successful in eliminating any usable loudness cue. Subjects reported informally that they were unable to discern any loudness differences for any set of parameters tested.

Apparently, also, subjects learned the correspondence between the acoustical cues and the correct response from the feedback on training runs. The nature of the task was such that 0% correct would have indicated equally as good discrimination as 100% correct. However, the psychometric functions rarely fell much below 50% correct.[1]

For all conditions tested, except for bandwidths of 400 and 800 Hz, performance fell to the chance level with increasing N well before the maximum value, $N = 25$. This implies that, on the average, 25 components and 60 components sound the same.[2] From this, we infer that the sound with 60 components is actually asymptotically dense, i.e., that our results would not have been different had our comparison sound included an arbitrarily large number of components. For a bandwidth of 400 Hz, near $N = 25$, the performance was either below threshold or heading rapidly towards threshold. It therefore seems reasonable to regard the 60-component case as infinitely dense perceptually for bandwidths less than or equal to 400 Hz.

2. Hypotheses

A primary goal of this study was to elucidate the nature of the cues used by subjects to judge spectral density. A first attempt at consolidating the data and understanding the nature of the cues centered upon several hypotheses.

a. Hypothesis 1: Constant performance occurs for constant Q. Analogous to filter theory, the Q is defined as the center frequency divided by the bandwidth. The hypothesis says that for the conditions of center frequency/bandwidth equal to 550 Hz/100 Hz, 1100 Hz/200 Hz, and 2200Hz/400 Hz, all corresponding to $Q = 5.5$, performance should be constant. Figure 1 shows psychometric functions for $Q = 5.5$. The corresponding threshold values $N(75)$ are given in Table I, which also shows thresholds for $Q = 1.75$, $Q = 3$, and $Q = 10.5$. If the hypothesis is correct, then all the entries in a given column, for a given Q, should be the same. The figure and the table show that the hypothesis is unlikely to be successful. Performance always increases for increasing bandwidth, even if the center frequency increases proportionally.

To test the hypothesis more formally, we needed to compute a measure of performance for each of the different conditions (of f_b and W) and to compare those values, taking into account the variability in the performance for each condition. For the measures of performance, we chose to average the percent correct over the 12 values of N presented in each run (120 trials). This measure is equivalent to the area under the psychometric function for each run. From the three runs for a given condition, we found a mean and a standard deviation, $N - 1 = 2$ weight. To compare two conditions, we calculated the difference between the two means, as measured in units of the corresponding standard deviation, which was the square root of the sum of variances for the two means being compared. We called this statistic D. When there were more than two conditions being compared, we performed a round-robin calculation of Ds among all the conditions. The mean of the absolute value of these Ds, as computed for a given subject, was called $\bar{D}$, and this is shown in the lower half of Table I. Because the values of $\bar{D}$ are mostly greater than two, we conclude that hypothesis 1 fails.

b. Hypothesis 2: Constant performance occurs for con-

TABLE I. Test of hypothesis 1: Constant performance occurs for constant Q, $N(75)$ is the number of components in a band with bottom frequency f_b, and width W for 75% correct performance. If the hypothesis is correct, then values of $N(75)$ within a given column for a given value of Q should all be the same and the values of D should not be greater than unity ($T = 500$ ms).

			$N(75)$		
	f_b (Hz)	W(Hz)	S1	S2	S3
$Q = 1.75$					
	500	400	>25	>25	16
	1000	800	>25	>25	22
$Q = 3.0$					
	500	200	17	15	9
	1000	400	23	21	13
	2000	800	>25	>25	20
$Q = 5.5$					
	500	100	8	8	4
	1000	200	15	13	9
	2000	400	19	19	16
	3000	600	21	25	18
$Q = 10.5$					
	500	50	4	4	4
	1000	100	7	7	5
	2000	200	11	11	9
	Q			$\bar{D}$	
			S1	S2	S3
	1.75		1.5	3.6	5.1
	3.		2.3	2.0	4.0
	5.5		2.8	3.9	5.0
	10.5		1.4	2.3	1.9

TABLE II. Test of hypothesis 2: Constant performance occurs for constant bandwidth, independent of band bottom frequency f_b. If the hypothesis is correct, then the values of $N(75)$ in a given column for a given value of W should all be the same and the values of D should not be greater than unity ($T = 500$ ms).

		$N(75)$		
W(Hz)	f_b (Hz)	S1	S2	S3
50				
	250	4	...	4
	500	4	4	4
	1000	4	4	3
100				
	500	8	8	4
	1000	7	7	5
	2000	8	8	4
200				
	500	17	15	9
	1000	15	13	9
	2000	11	11	9
400				
	500	>25	>25	16
	1000	23	21	13
	2000	19	19	16
800				
	1000	>25	>25	22
	2000	>25	>25	20
W(Hz)			$\bar{D}$	
		S1	S2	S3
50		0.8	0.3	0.3
100		0.9	0.2	1.1
200		1.3	0.9	0.9
400		1.6	2.9	1.3
800		4.5	...	2.0

stant bandwidth, independent of band bottom frequency. By the Fourier integral theorem, a signal with many components of equal amplitude is completely characterized by the set of frequencies $\{f_i\}$ and the set of initial phase angles $\{\varphi_i\}$ for the components. The temporal fluctuations in the intensity of the signal are caused by beats and are thus completely characterized by the sets of differences $\{f_i - f_j\}$ and $\{\varphi_i - \varphi_j\}$ $(i<j)$. Therefore, a statistical description of the intensity fluctuations depends only upon the number of components, the bandwidth, and the rule by which components are put into the band; it is independent of the band bottom frequency f_b (see the Appendix). If subjects base their judgments only upon the perceived temporal fine structure of the intensity, then one would expect that performance for a given bandwidth should be independent of f_b.

We tested this hypothesis by comparing performance for f_b values of 500, 1000, and 2000 Hz. There were five tests, corresponding to different bandwidths. The threshold values $N(75)$ are shown in Table II. If the hypothesis is correct, then, for a given bandwidth, the entries in a given column should be the same. Comparison of these values suggests that the hypothesis has some merit. Conditions of constant bandwidth certainly lead to more constant performance than do conditions of constant Q (hypothesis 1). However, there is a tendency for performance as measured by $N(75)$ to decrease as the band bottom frequency increases. Further, this decrease appears to be more pronounced for wider bandwidths than for narrower ones. These two points concerning the deviation from the hypothesis play an important role in our conclusions in Sec. IV. Typical psychometric functions showing the success of hypothesis 2 at 100 Hz and its failure at 400 Hz are given in Figs. 2 and 3, respectively.

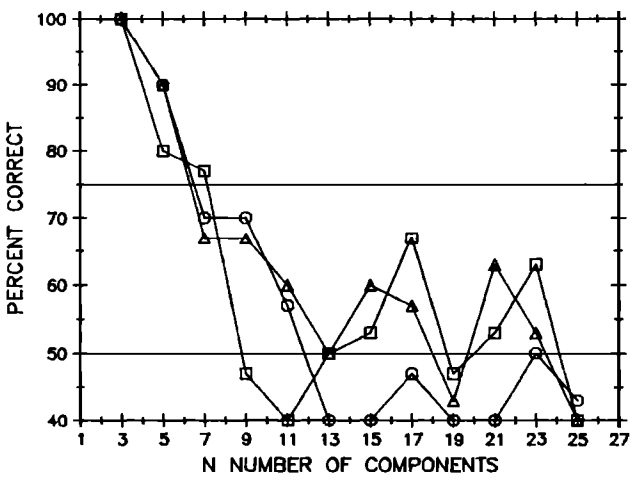

FIG. 2. Same as Fig. 1, but for subject S2 for a bandwidth of $W = 100$ Hz and three values of the band bottom frequency: circles, $f_b = 500$ Hz, squares, $f_b = 1000$ Hz, and triangles, $f_b = 2000$ Hz.

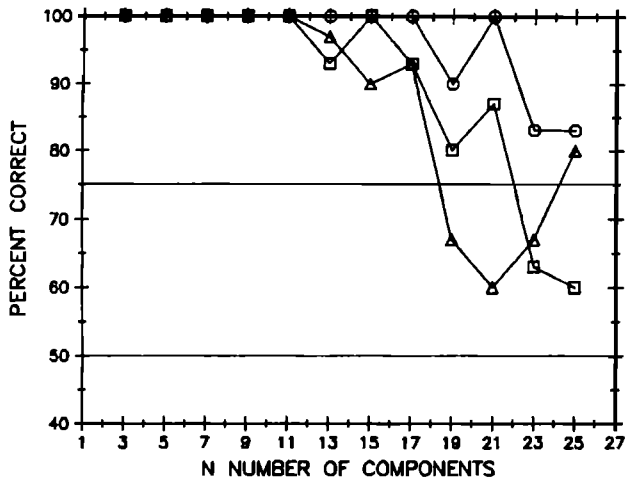

FIG. 3. Same as Fig. 2 but with a bandwidth of $W = 400$ Hz.

The hypothesis was tested statistically as for hypothesis 1. The values of $\bar{D}$ in the lower part of Table II show that the hypothesis can be accepted for small bandwidths but not for large ones. At large bandwidths, the improvement in performance with decreasing band bottom frequency causes the hypothesis to fail.

c. Hypothesis 3: Constant performance occurs for constant spectral density. This hypothesis means that the primary variable N affects judgments only through the value of N divided by the bandwidth. According to this hypothesis, if a cluster of four components in a band of 50-Hz width is barely distinguishable from a cluster of 60 components in a 50-Hz band, then a cluster of eight components in a 100-Hz band is barely distinguishable from a cluster of 120 components in a 100-Hz band. Because we regard 60 components in a band as asymptotically dense for bandwidths of 400 Hz or less, the number 120 could be replaced by 60, as it was in our experiments.

We tested this hypothesis against data from experiments with bandwidths of $W = 100$, 200, 400, and 800 Hz and band bottom frequencies of $f_b = 500$, 1000, and 2000 Hz. The thresholds are given in Table III. If the hypothesis is correct, then the densities, given in the last three columns, should be the same within a single column. It is not necessary for the success of hypothesis 3 that the equality hold across different values of f_b. However, to the extent that hypothesis 2 is correct, equality *will* hold across different values of f_b.

From the values of $N(75)$, it appears that hypothesis 3 is only partially successful; the last three columns are similar for a given subject, but there is a tendency for the threshold density to decrease for increasing bandwidth, especially for larger values of f_b. For a statistical test of hypothesis 3, one needs to compare the percent correct responses for different bandwidths over the same range of spectral densities. It is not possible to do a round-robin comparison because the bandwidths differ by as much as a factor of eight. With only five components in a 50-Hz band the density is greater than with 25 components in a 400-Hz band. Therefore, our comparison was limited to only neighboring bandwidths in the table. For example, we compared $W = 50$ and 100, $W = 100$ and 200, and $W = 200$ and 400. In each case, we compared average performance for all 12 values of N for the larger bandwidth with the average performance for the six smallest values of N for the smaller bandwidth. Otherwise, the computation of $\bar{D}$ was the same as for hypothesis 1 and 2. The values of $\bar{D}$ shown in the lower part of Table III are mostly less than unity and none of them is much larger than unity. The success of the hypothesis in the statistical test is partly due to the fact that the test could only be done for adjacent values of the bandwidth and partly due to the fact that it is a rather good hypothesis.

d. Hypothesis 4: Constant performance occurs for constant values of the product of the bandwidth and the duration. The temporal structure in the intensity depends upon the bandwidth; it becomes less rapidly varying for smaller bandwidths because beat frequencies between pairs of components are smaller. On the average, halving the bandwidth also halves the number of peaks and valleys in the instantaneous intensity, giving the listener fewer chances to judge the structure. Hypothesis 4 says that this effect can be overcome by making the sounds correspondingly longer.

We tested this hypothesis for two values of the bandwidth-duration product, (200 Hz × 0.5 s) = (100 Hz × 1 s) = (50 Hz × 2 s) = 100, and (100 Hz × 0.5 s) = (50 Hz × 1 s) = 50. The results are shown in Table IV. The hypothesis appears to give a good account of the values of $N(75)$ for the product of 50. For the product of 100, however, there is a clear tendency for the performance with wider bandwidth and shorter duration to be better than with narrower bandwidth and longer duration. The statistical test

TABLE III. Test of hypothesis 3: Constant performance occurs for constant spectral density N/W. If the hypothesis is correct, then values of $N(75)/W$ within a given column should all be the same and the values of D should not be greater than unity ($T = 500$ ms).

		$N(75)$			$N(75)/W$(KHz)		
f_b(Hz)	W(Hz)	S1	S2	S3	S1	S2	S3
500							
	50	4	4	4	80	80	80
	100	8	8	4	80	80	40
	200	17	15	9	85	75	45
	400	>25	>25	16	>62	>62	40
1000							
	100	7	7	5	70	70	50
	200	15	13	9	75	60	45
	400	23	21	13	58	53	32
	800	>25	>25	22	>31	>31	28
2000							
	100	8	8	4	80	80	40
	200	11	11	9	55	55	45
	400	19	19	16	48	48	40
	800	>25	>25	20	>31	>31	25

f_b(Hz)	$\bar{D}$		
	S1	S2	S3
500	0.7	0.8	0.8
1000	0.3	1.1	0.5
2000	1.1	1.0	0.9

TABLE IV. Test of hypothesis 4: Constant performance occurs for constant values of the product of the bandwidth and the duration WT. If the hypothesis is correct then, for a given value of the product WT, the values of $N(75)$ within a given column should all be the same and the values of D should not be greater than unity.

		$N(75)$		
W(Hz)	T(s)	S1	S2	S3
(a) $f_b = 1000$, $WT = 50$				
100	0.5	7	7	5
50	1.0	6	6	3
(b) $f_b = 1000$, $WT = 100$				
200	0.5	15	13	9
100	1.0	12	12	5
50	2.0	9	10	5
WT			$\bar{D}$	
		S1	S2	S3
50		0.5	0.5	0.6
100		1.8	1.5	0.8

of this hypothesis was the same as for hypothesis 1 and 2; the results of the test, shown in Table IV, agree with the expectation based upon the values of $N(75)$. The hypothesis can be accepted for a bandwidth-duration product of 50, but not for a product of 100.

3. Observations

There are two classes of perceptual cues, loosely described as spectral and temporal, which might be involved in discrimination of spectral density. A spectral cue would correspond to differences in tone color or even differences in pitch associated with a small density compared to a large density. For small numbers of components in wide bands, e.g., three components in a 400-Hz band versus 60 components, such tone color and pitch changes can easily be heard, and they probably contribute to making the performance 100% correct in such conditions. Whether or not spectral cues play a role near threshold, however, is not immediately clear. The salience of spectral cues depends upon the spectral resolving power of the auditory system. The impressive failure of hypothesis 1, where constant Q approximates constant resolving power, argues against spectral cues as primary near the threshold.

Temporal cues in the present context are principally intensity fluctuations. The relative success of hypothesis 2, and, less directly, that of hypotheses 3 and 4, leads us to suspect that threshold density discrimination is mainly determined by such temporal cues. The following section describes calculations, based upon a model of the perception of intensity fluctuations, which are used to predict the threshold value of N for comparison with the experimental data.

II. RMS FLUCTUATION MODEL CALCULATIONS

Our model of the perception of intensity fluctuations is a simple one, in which the auditory system responds to the fluctuations in the stimulus power as weighted by an exponentially decaying moving window with time constant τ. The sounds used in the experiment are only defined statistically; therefore, it is appropriate to calculate a configuration-averaged value of the internal power fluctuation, i.e., an average over all possible frequency distributions consistent with the rule that components are placed into frequency bins.

The mathematical development of the model is given in the Appendix. It turns out that the integrals involved can actually be done analytically to give a closed-form expression for the rms power fluctuation [cf. Appendix Eq. (A7)]. As noted above, this expression is independent of the band bottom frequency (or band center frequency), it depends upon the bandwidth W and upon the integration time τ, and it depends upon them only through their product $W\tau$. Figure 4 shows the rms fluctuation as a function of the number of components in the band for various values of $W\tau$.

Applied to our experiment, the model says that decisions are reached by comparing the fluctuation for 60 components with the fluctuation for N-variable components. For actual calculations, one must hypothesize a basis of comparison, a threshold criterion within that basis, and an integration time τ.

As possible bases for comparison, we considered a simple difference and a simple ratio. Computations using fluctuation differences produced nonsensical predictions; therefore, we used the ratio basis, as suggested by the log scale on the ordinate of Fig. 4. As criterion values for the ratio, we believe that 0.95 or 0.9 are reasonable choices; i.e., we assume that a 5% or 10% decrease in rms fluctuation corresponds to a just noticeable difference. [We note that Terhardt (1974) found that the perceived roughness of amplitude modulated tones was halved when the power fluctuation was reduced by a factor of 0.7. A reasonable choice for a criterion ratio must, in any case, be rather larger than this value.]

Figure 4 shows that, for a given bandwidth, it is evidently to the system's advantage to use the longest possible inte-

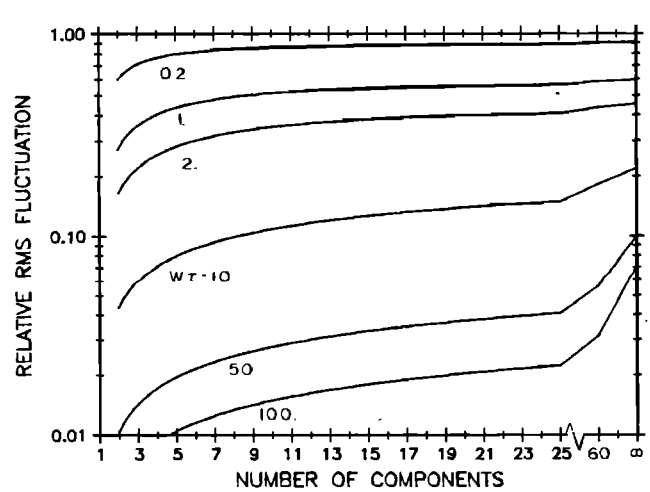

FIG. 4. Values of the rms internal power fluctuation as a function of the number of components in the band for various values of the product of the bandwidth and the integration time $W\tau$, as predicted by the rms fluctuation model.

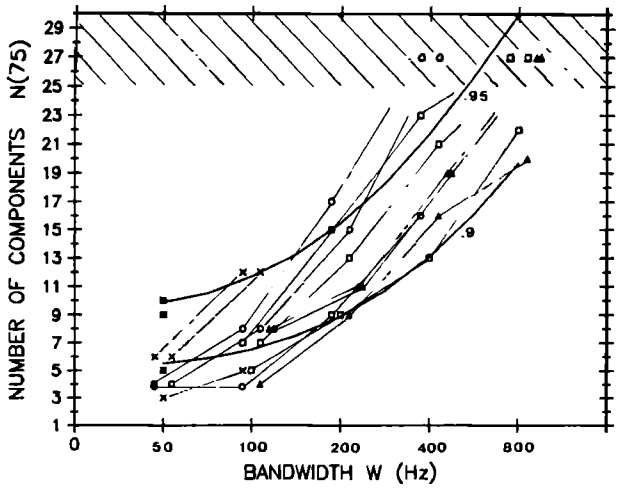

FIG. 5. Threshold values for the number of components in the band for 75% correct performance as a function of the bandwidth. The curves marked "0.95" and "0.9" show the predictions of the rms fluctuation model for $\tau = 3$ ms and criterion ratios of 0.95 and 0.9. Other symbols show experimental thresholds for various conditions: circles, $f_b = 500$ Hz, $T = 500$ ms; squares, $f_b = 1000$ Hz, $T = 500$ ms; triangles, $f_b = 2000$ Hz, $T = 500$ ms; ×, $f_b = 1000$ Hz, $T = 1000$ ms; *, $f_b = 1000$ Hz, $T = 2000$ ms. A straight line connects the data points for a single listener and condition. True values of $N(75)$ are not known for data points plotted above the line $N(75) = 25$.

gration time, because the ratio between the rms fluctuation for 60 components and the rms fluctuation for a smaller number of components always increases for increasing WT. But typical maximum integration times, 100 to 300 ms, cannot be used in the model to provide a satisfactory fit to the data. Further, as shown in Fig. 4, a long integration time corresponds to small fluctuations, but listeners were always aware of large fluctuations in instantaneous loudness. Therefore, a minimum integration time, as opposed to a maximum integration time, seems appropriate in this context. We used a value of 3 ms (Viemeister, 1979) for τ.

The results of the model calculation are shown in Fig. 5. The two labeled lines show the predicted values of the threshold $N(75)$ as a function of the bandwidth for both 0.95 and 0.9 criterion ratios. The various points on the plot represent the experimental values of $N(75)$ for all subjects for all the experimental band bottom frequencies and sound durations. Most of the data fall between the theoretical curves corresponding to the two criterion ratios, a result which supports the model. However, the lines which connect experimental points for a given subject and condition show that performance increases with increasing bandwidth faster than the model predicts. The lower the band bottom frequency, the greater is the discrepancy between the rate of increase observed experimentally and the rate predicted by the model.

III. EXPERIMENT 2: COMPLEX WAVEFORMS

One of the goals of the present study was to gain some insight into the perception of instrumental choruses. We asked: How many instruments must there be in a unison chorus so that the resulting sound is indistinguishable from a very large number of instruments, given that the intensity is constant? Experiment 1 showed that, if the instruments are sine oscillators, then the threshold value is, for example, four, for oscillators playing C5 and almost a whole tone out of tune ($f_b = 500$ Hz, $W = 50$ Hz). Such a small number is not consistent with one's ordinary musical experience with the complex tones of musical instruments.

To extend our study to sounds which are somewhat realistic musically, we performed density discrimination experiments using clusters of complex harmonic tones.

A. Stimuli and procedure

The single tone, which was the basis of the cluster, had the spectrum of a violin, taken from the IRCAM sound library (Gerzso *et al.*, 1978). For different fundamental frequencies, different violin spectra were used: for a fundamental of 500 Hz C5, for 1000 Hz C6, for 2000 Hz C7, all including harmonics up to 6000 Hz. The relative harmonic levels, measured at the output of the power amplifier, are given in Table V. The harmonic components were added in sine phase to make the waveform. To construct the cluster of tones, we followed the same procedure as for the sine waveform in experiment 1. Fundamental frequencies were random within a bin of width W/N. Waveform amplitudes were scaled by the inverse square root of N to provide constant intensity for the clusters. The procedure and the subjects were the same as for experiment 1; all sounds were 500 ms in duration.

B. Results

Threshold data from the experiments with clusters of tones having a violin spectrum are shown in Table VI, for various fundamental band bottom frequencies and bandwidths. Comparison with the corresponding graphs for the sine waveform show that performance is considerably better for the violin waveform than for the sine waveform with the same frequency. Thus the data for violin spectrum clusters are somewhat more in line with expectation based upon musical experience.

TABLE V. Relative intensities of the harmonics of the violin waveform used for complex components.

Harmonic	C5 ($f_b = 500$ Hz)	C6 ($f_b = 1000$ Hz)	C7 ($f_b = 2000$ Hz)
1	0 dB	0 dB	−1 dB
2	−5	−18	0
3	−7	−17	−12
4	−10	−31	
5	−17	−37	
6	−27	−38	
7	−29		
8	−36		
9	−45		
10	−40		
11	−46		
12	−45		

TABLE VI. Threshold values for tones with a violin spectrum compared with thresholds for the individual harmonics. The data for the sine waveform were taken from experiment I ($T = 500$ ms).

	f_b(Hz)	W(Hz)	$N(75)$ S1	S2	S3
violin	250	50	12		7
sine 1	250	50	4		4
sine 2	500	100	8		4
sine 3	750	150	...		...
sine 4	1000	200	15		9
...					
sine 8	2000	400	19		16
violin	500	50	11	15	6
sine 1	500	50	4	4	4
sine 2	1000	100	7	7	5
sine 3	...	...	...	...	...
sine 4	2000	200	11	11	9
violin	500	100	19	21	9
sine 1	500	100	8	8	4
sine 2	1000	200	15	13	9
sine 3	...	...	...	...	...
sine 4	2000	400	19	19	16
violin	1000	100	10	17	8
sine 1	1000	100	7	7	5
sine 2	2000	200	11	11	9
violin	1000	200	25	>25	11
sine 1	1000	200	15	13	9
sine 2	2000	400	19	19	16
sine 3	3000	600	19	21	18
violin	2000	100	18	12	4
sine 1	2000	100	8	8	4

C. Discussion

One can imagine ways in which the auditory system can use the information in the harmonic bands to improve performance. The simplest model is one in which listeners are able to listen selectively to the different harmonic bands and can make their decisions based upon the one band which provides the most information. Because the sine waveform experiments reported in Sec. I were done at octave values of f_b, it was possible to do a limited test of this model by comparing the thresholds for the violin spectrum sound with the threshold for the individual sine components. The comparison is done in Table VI.

The table shows that for subjects S1 and S2, and for fundamental frequencies of 500 Hz and above, the performance with the violin spectrum is approximately as good as the best performance measured for any harmonic of the violin spectrum sound, in agreement with the model. For example, for the 500-Hz fundamental, the threshold value of N for the violin spectrum is quite close to that for the 4th harmonic alone. The data for subject S3, however, do not show the expected improvement.

The experiments with a fundamental frequency of 250 Hz were done in response to comments on an earlier version of this paper; listener S2 was no longer available. The values of $N(75)$ shown in Table VI show that performance for the violin tone is now less good than the best performance measured in any harmonic band. The model cannot explain that result unless it is supplemented with the qualification that a harmonic band cannot be used if it is within a critical bandwidth of another harmonic band. Using critical bandwidth given by Scharf (1970), we checked this qualified model against the other data in Table VI and found that the qualified model fails for the two experiments done with a fundamental frequency of 1000 Hz.

IV. CONCLUSION

The above sections have presented data, hypotheses, and a model concerning the discrimination of spectral density in dense clusters of tones, where the tones have either a sine waveform (Secs. I and II) or a complex waveform (Sec. III). This section presents our conclusions for these two cases.

Our experiments found that for clusters of sine tones the largest spectral density which was ever distinguishable from a very large density was 80 components per kHz. This value of the discrimination threshold is considerably less than the value of 1000 components per kHz found in the masking experiments by Schafer *et al.* (1950). A similar conclusion was reached by Gerzso (1980).

A comparison might also be made with the spectral densities used in profile analysis studies (Green and Mason, 1985, and references therein). The largest density to be found in those studies is 43 components logarithmically spaced between 200 and 5000 Hz. The largest spectral density occurs at the bottom of the band where it is 63 components/kHz. At the signal frequency of 1000 Hz, it is a factor of 5 smaller and at the top of the band it is yet another factor of 5 smaller. A comparison with Table III shows that these densities are considerably smaller than those which are indistinguishable from noise. The profile analysis experiments and our density discrimination experiment are operating in rather different regimes of spectral density.

The ability to discriminate spectral density appears to be mediated primarily by the discrimination of loudness fluctuations in the sound. The primacy of this cue was suggested by the partial success of hypothesis 2, which says that performance in density discrimination depends upon the bandwidth and is independent of the band center frequency, and by the rather good overall agreement between the data and the rms fluctuation model of Sec. II.

However, for large bandwidths, hypothesis 2 and the rms fluctuation model tend to fail. As the bandwidth is increased, the observed performance improves faster than the model predicts. This is particularly true for smaller band bottom frequencies. As shown in Fig. 5, the disagreement between the data and the model increases as the band bottom frequency decreases from 2000 to 1000 to 500 Hz.

One possible explanation for the discrepancy is that tone-color differences, clearly audible for small values of N, continue to be usable cues as N increases to $N(75)$. Because tone-color discrimination depends upon spectral resolution within the band of components, one expects it to be of increasing importance as the bandwidth increases and, accord-

ing to this explanation, to result in performance which is increasingly better than is predicted by the rms fluctuation model alone. Further, because critical bandwidths (and frequency jnds) decrease with decreasing frequency, tone-color cues should be more important for small band bottom frequencies than for large. The explanation is thus in qualitative agreement with the nature of the discrepancy between the experimental values of $N(75)$ and the prediction of the rms fluctuation model. It is also in agreement with the nature of the failure of hypothesis 4 when the bandwidth-duration product is 100.

An alternative explanation for the discrepancy between the experimental data and the rms fluctuation model raises the question of the number of components of the signal which contribute to the perceived fluctuations. The rms fluctuation model began with the physical waveform and thus included all spectral components. There is reason to believe, however, that temporal fluctuations in the auditory system are generated only by those components in a single critical band (Zwicker, 1952). Our experiments, in fact, included some bandwidths which were smaller than the critical bandwidth and some which were larger, for every value of f_b.

A simple approach to the case in which the band of components is larger than a critical band is to assume that fluctuations in the auditory system are mainly due to components within a critical band centered somewhere within the band of components. Therefore, to maintain constant performance when the bandwidth is increased would require that the number of components in the band increases proportionately. This expectation is actually hypothesis 3, which says that constant performance occurs for constant spectral density. *In toto*, this approach predicts that the plot of $N(75)$ as a function of bandwidth should follow the rms fluctuation curve, as shown in Fig. 5, from small bandwidths up to the critical bandwidth, and should thereafter be directly proportional to the bandwidth. This variation on the rms fluctuation model actually fits the experimental values of $N(75)$ shown in Fig. 5 somewhat better than does the model based upon the physical signal because it has a steeper slope for large W. However, there are several arguments against it. First, this alternative model predicts a slope which is too steep. (The densities shown in Table III are not constant but decrease with increasing bandwidth.) Second, there is a conceptual difficulty with the critical-band-limited fluctuation variation in that a critical band centered somewhere within the band of components is not the most advantageous critical band for performing the task of density discrimination. More advantageous would be a critical band centered well outside the band of components so that the total number of effective spectral components would be small for both the dense standard and the less dense sound. But using such a critical band would seem to lead to a performance which is enormously better than observed experimentally. In the end, we believe that this alternative explanation for the deviation of the data from the predictions of the rms fluctuation model raises more problems than it solves.

The rms fluctuation model involves only the magnitude of the fluctuation. It does not include the possibility that the perception of particular patterns in the fluctuations may play a role in the discrimination task. A major feature of our experimental procedure was the randomization of the frequencies in each sound to try to avoid such cues. We cannot prove, however, that the randomization procedure was entirely successful in doing so. Indeed, informal comments by subjects as they groped for a criterion to use in performing the task indicated that imagined characteristic patterns, particularly for the 60-component sounds, may have contributed to the decisions. It is not known, however, whether decisions based upon such a pattern criterion were right more often than they were wrong. When subjects subsequently performed the three-interval task,[1] with two different 60-component sounds in each trial, they concluded that different 60-component sounds produced quite different patterns.

The rms fluctuation model also does not allow for the possibility that the perception of fluctuation rate may contribute to discrimination. An estimate of the distribution of frequencies from which the fluctuation is composed is given by the configuration-averaged power spectrum of the power fluctuations, in Eq. (A8) of the Appendix. The spectrum is basically a low-passed version of the distribution of frequency differences in the signal, so that the most probable rate depends upon the number of components in the band. However, informal comments by listeners did not suggest that fluctuation rate was an important cue in performing the task. Introspective listening suggested that the fluctuation rate is indeed correlated with the fluctuation magnitude, but that it is less salient. We suspect that the spectral distribution of fluctuation frequencies is too broad to make the rate a useful cue near threshold.

In summary, there are a number of different perceptual cues which could serve as a basis for discrimination of spectral density. The hierarchy of cues which seems most plausible to us is one in which the magnitude of loudness fluctuations is of primary importance and the perception of tone color is second in importance. The tone-color cue becomes increasingly important for large bandwidths and low band center frequencies.

For complex sounds, we have persuasive evidence that listeners can use information in the harmonic bands to discriminate spectral density. For tones with a violin spectrum, listeners can use harmonic bands at least as high as the fourth, though we were not able to account for this fact in terms of a single simple model. Performance is better with complex sounds than with sine tones because bands with increasing harmonic number are increasingly wide. However, we believe that the threshold densities obtained for the violin spectrum cluster are still smaller than would be obtained for musical instruments in a unison chorus. The reason is probably that the tones of musical instruments are individually complicated with characteristic attacks, vibrato, jitter, and time-dependent structure in the spectrum. The studies of McAdams (1984) suggest that such individual features can cause individual instruments to be heard individually. It is possible that asymptotic spectral density occurs when the number of instruments is large enough for these individual variations to be obscured.

Further experiments may place a limit on the number of harmonic bands which can be used to discriminate density

and may determine whether information may be accumulated across different harmonic bands. To decide the basis for density discrimination for the complicated tones of musical instruments will require further work using sounds with all the properties of such complicated tones including asynchronous attacks, vibrato, and jitter, and individual intensity and spectral variations.

ACKNOWLEDGMENTS

We are grateful to Dr. L. A. Wheeler for useful discussions, to Dr. Laurent Demany for comments on the manuscript, and to B. K. Smith, who programmed the 4C synthesizer. The early work on this project benefited from discussions with Dr. J. A. Moorer, Dr. M. V. Mathews, and Dr. D. L. Wessel. Supported, in part, by the National Institutes of Health and by the International Program of the National Science Foundation in conjunction with the Centre Nationale de Recherche Scientifique, France.

APPENDIX: RMS FLUCTUATION MODEL

This Appendix describes the development of the rms fluctuation model, applied above in Sec. II. The model assumes that the auditory system responds to fluctuations in the instantaneous power of the stimulus, as passed through an averaging window, moving in time. The effective, or perceived, power fluctuations are calculated for the general stimulus as a function of time. To allow a comparison between the model and the experimental data, the root-mean-square fluctuation is computed, where the mean represents a configuration average, an average over all possible stimuli for a given number of components in the band. The computed rms value of the fluctuation is time independent and can be calculated analytically. The details of the development are as follows.

To calculate the envelope of the stimulus with N components in a band, one describes the signal in phasor form:

$$x(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e^{i\omega_i t} e^{i\varphi_i}, \tag{A1}$$

where ω_i and φ_i are the angular frequency and the phase of the ith component.

The power P is the square of the absolute value in Eq. (A1), given by

$$P(t) = 1 + \frac{2}{N} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \cos[(\omega_i - \omega_j)t + \varphi_i - \varphi_j]. \tag{A2}$$

Because the signal is normalized by the number of components, the mean power, given by the first term, is unity. The double sum in Eq. (A2) represents the fluctuation about the mean power. The fluctuation depends upon the frequencies and phase angles of the components only through differences. Therefore, the statistical nature of the fluctuations depends only upon the number of components and the bandwidth; it is independent of the band center frequency.

Equation (A2) gives the exact value of the power at any time. The fluctuations in time are caused by beats among the components, and some of these may be very rapid. We allow for the possibility that the auditory system may not precisely follow rapid beats by folding the instantaneous power with an exponential window (sometimes called a "memory function") of time constant τ, to obtain energy E:

$$E(t) = \int_{-\infty}^{t} e^{(t'-t)/\tau} P(t')dt'. \tag{A3}$$

The lower limit in the integral involves the assumption that τ is much less than the stimulus duration T.

The effective power, $\bar{P} = E(t)/\tau$, is then given by

$$\bar{P}(t) = 1 + \frac{2}{N} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \frac{\cos(y_{ij}) + \tau\Delta\omega_{ij} \sin(y_{ij})}{1 + (\tau\Delta\omega_{ij})^2}, \tag{A4}$$

where

$$y_{ij} = (\omega_i - \omega_j)t + \varphi_i - \varphi_j$$

and

$$\Delta\omega_{ij} = \omega_i - \omega_j.$$

The double sum in Eq. (A4) is the fluctuation F. Its time-average value is zero for any nondegenerate configuration of frequencies and phases, as is its configuration average for any value of time. We, therefore, study the configuration average of the *square* of the fluctuation $\langle F^2 \rangle$:

$$\langle F^2 \rangle = \left\langle \frac{4}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \frac{\cos^2(y_{ij}) + (\tau\Delta\omega_{ij})^2 \sin^2(y_{ij})}{[1 + (\tau\Delta\omega_{ij})^2]^2} \right\rangle. \tag{A5}$$

Terms in F^2 involving products of cosine functions and sine functions of y_{ij} and $y_{i'j'}$, ($i \neq i'$ and $j \neq j'$) average to zero and do not appear in Eq. (A5). The squared cosine and sine functions which remain can be replaced by their configuration-average values of 1/2. Then the mean-square fluctuation is given by the time-independent equation,

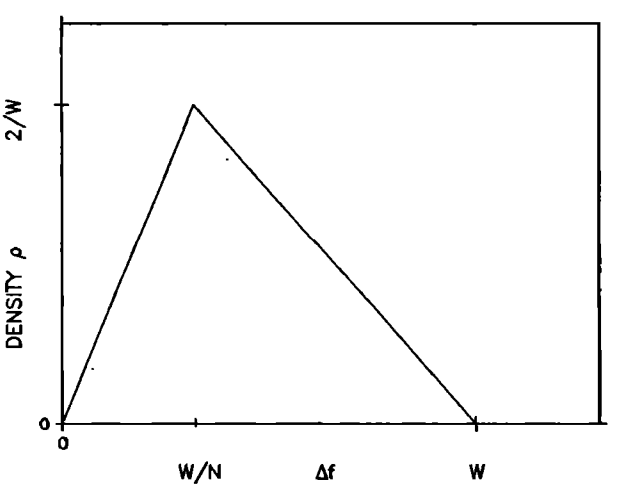

FIG. A1. Probability density for the frequency difference between two components when the frequency band, of width W, is filled by placing one component in each of N bins.

probability density for frequency differences, multiplied by $N(N-1)/2$. For the binned frequencies used in our experiment, the density of differences is a triangle, as shown in Fig. (A1). The density is zero when the frequency difference is equal to zero or when it is equal to the bandwidth W (angular frequency difference equal to $2\pi W$). The most probable value of the difference is the bin width, W/N. The integral can be done analytically to give a final expression for $\langle F^2 \rangle$,

$$\langle F^2 \rangle = \frac{2}{W\tau}\left[\tan^{-1}(W\tau) - \tan^{-1}\left(\frac{W\tau}{N}\right) + \frac{N}{2W\tau}\ln\left(1+\frac{W^2\tau^2}{N^2}\right) - \frac{1}{2W\tau}\ln(1+W^2\tau^2)\right]. \tag{A7}$$

The function in Eq. (A7) has the following properties:

(1) It is always less than or equal to 1; the fluctuation in power can never be greater than the average power.

(2) It depends upon the bandwidth W and upon the integration time τ only through their product $W\tau$.

(3) Its limit for $W\tau = 0$ is $1 - 1/N$. The interpretations of the individual limits, $W = 0$ and $\tau = 0$, are, however, quite different. For $W = 0$, the limit is simply wrong; the configuration-averaging assumptions above are invalid for the highly degenerate case of zero bandwidth. For $\tau = 0$, the limit is correct; for a given N, this limit gives the largest possible mean-square fluctuation.

(4) It is zero for $N = 1$; there is no fluctuation for only one component.

(5) Its limit for an infinite number of components is obtained by summing the first and the last terms on the right-hand side of Eq. (A7).

(6) The fluctuation increases for increasing number of components N and for decreasing bandwidth W. However, it is not a simple function of the spectral density N/W.

The rms value of the fluctuation, computed by taking the square root of Eq. (A7), is shown in Fig. 4 for several values of $W\tau$.

An impression of the statistical nature of the *rate* of the fluctuation can be gained by taking the Fourier transform of the time-dependent internal power in Eq. (4), omitting the constant term. Because the phase differences among components are random, a configuration average of the Fourier transform is zero. The configuration average of the power spectrum of the power fluctuation is finite, however, and it is simply related to the density of differences $\rho(\omega)$.

$$\langle |\Delta P^2(\omega)| \rangle \alpha \frac{N-1}{N}\frac{\rho(\omega)}{1+(\tau\omega)^2}. \tag{A8}$$

Density $\rho(\omega)$ is the triangle given in Fig. A1, with $\omega = 2\pi\Delta f$.

[1]We have also used a three-interval paradigm which does not require the listener to have any preconceived notion about how a densely packed band should sound. The first interval contained a version of the 60-component sound. The next two intervals contained a different version of the 60-component sound and the sound with a variable number of components, randomized in order of presentation over trials. The subject's task then became one of deciding which of the latter two sounds was like the first. Performance on the three-interval task was not better than performance on the two-interval task; it actually appeared to be somewhat worse. Worse performance might be understood in the following way: In the two-interval task, the listener could use accumulated experience to form a general image of the 60-component sound. The image could be used as a general template for comparison with any of the infinite variety of 60-component sounds presented during the task. In the three-interval experiment the listener was required to compare one version of the 60-component sound with another. Since these were often dissimilar, the listener made errors. In any event, the two-interval paradigm is the more efficient, and all data reported in this paper were obtained by using it.

[2]The three listeners each did three experimental blocks with $f_b = 1000$ Hz, $W = 200$ Hz, and $T = 500$ ms for large values of the variable number of components, $N = 25,30,35,\ldots,60$. Performance never reached threshold for any value of N, suggesting that there is no additional cue, for large number of components, which subjects can use to do the task.

Gerzso, A. (1980). "Density of spectral components: Preliminary experiments," IRCAM Rep. 31 (unpublished).

Gerzso, A., Moorer, J. A., and Wessel, D. L. (1978). "Spectral data base of orchestral instrument sounds," IRCAM Rep. (unpublished).

Green, D. M., and Mason, C. R. (1985). "Auditory profile analysis: Frequency, phase, and Weber's law," J. Acoust. Soc. Am. 77, 1155–1161.

McAdams, S. (1984). "Spectral fusion, spectral parsing, and the formation of auditory images," Ph.D. thesis, Stanford University, Stanford Department of Music Report STAN-M22 (unpublished).

Schafer, T. H., Gales, R. S., Shewmaker, C. A., and Thompson, P. O. (1950). "The frequency selectivity of the ear as determined by masking experiments," J. Acoust. Soc. Am. 22, 490–496.

Scharf, B. (1970). "The critical band," in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias (Academic, New York), pp. 159–202.

Terhardt, E. (1974). "On the perception of periodic sound fluctuations (roughness)," Acustica 30, 202–213.

Viemeister N. (1979). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. 66, 1364–1380.

Zwicker, E. (1952). "Die Grenzen der Horbarkeit der amplitudenmodulation und der frequenzmodulation eines tone," Acustica 2, AB 125–133.