# Perceptual effects of spectral modifications on musical timbres

John M. Grey and John W. Gordon

*Center for Computer Research in Music and Acoustics, Artificial Intelligence Laboratory, Stanford University, Stanford, California 94305*
(Received 1 August 1977; revised 3 January 1978)

An experiment was performed to evaluate the effects of spectral modifications on the similarity structure for a set of musical timbres. The stimuli were 16 music instrument tones, 8 of which were modified in pairs. This modification consisted of exchanging the shape of the spectral energy distribution between the two tones within each pair. The three-dimensional spatial representation of similarities among the 16 tones was obtained by multidimensional scaling techniques and compared to a previous scaling of the original 16 unmodified tones [J. M. Grey, J. Acoust. Soc. Am. **61**, 1270–1277 (1977)]. The pairs of tones which had exchanged spectral shapes in fact exchanged orders on the spatial axis which had been previously interpreted as relating to spectral shape, thereby supporting the earlier interpretation. The two remaining axes of the spatial solution also retained their original interpretations, relating to various temporal details of the tones. A set of formal quantitative models for the spectral dimension was constructed and tested, and the results further supported the interpretation of this perceptual axis.

PACS numbers: 43.66.Jh, 43.66.Ba, 43.66.Lj, 43.75.−z

## INTRODUCTION

A number of investigators in the last several years have sought to apply multidimensional scaling techniques to the study of musical timbre perception (Plomp, 1970; Wessel 1973, 1974; Miller and Carterette, 1975; Grey, 1975, 1977). The typical study first obtains from listeners a rating of relative similarity between all pairs of stimuli. Then, using some particular multidimensional scaling algorithm, the investigator constructs a geometric *map* that graphically represents the perceived similarity relationships among the stimuli. The stimuli are plotted as points in this geometric space, and the distances between the stimulus points correspond to their degrees of dissimilarity. The closer any two points are in their psychological distance, that is, the more similar they are to one another, the closer they will be on the map. The multidimensional scaling algorithm will optimize the relationship between the similarity ratings and the spatial distances in order to construct a map having a given number of dimensions. The utility of this spatial representation is to help guide the investigator in interpreting the bases underlying the relational judgments of the stimuli. A direct analogy is usually made between the axes of the map and the dimensions or factors of perceptual similarity. Therefore, typically an interpretation of the map is made with respect to stimulus attributes that best correlate to the orderings of stimuli along the various axes, or dimensions, of the geometric space.

Various related computer algorithms have been used to construct the optimal map for the investigator to view, reducing the similarity data to a specific number of dimensions (Shepard, 1962a, 1962b; Kruskal, 1964a, 1964b; Carroll and Chang, 1970). In timbre research, either two- or three-dimensional maps have been found to be appropriate for viewing the similarity relationships among a set of tones. Most of the research on timbre perception has been oriented towards finding the physical properties of the stimulus tones which would explain the dimensions of the map, thereby enumerating psychophysical relationships for timbre.

Among the several attempts to scale timbre perception performed in the last decade, various types of results have been obtained depending upon the type and number of stimuli used. The earliest study by Plomp (1970), employing steady-state segments of musical tones, found that the resultant three-dimensional map of the similarity structure could be well interpreted entirely in terms of the amplitude pattern of the harmonics, reduced to a classical formant-type model.

Wessel, in both his own study (1973, 1974), and in a reanalysis of the data of Wedin and Goude (1972), found that a two-dimensional map was appropriate for interpreting the similarity relationships among sets of temporally complete instrument tones, that is, tones which include the attack and decay portions in addition to the steady state. One axis was found to relate to the spectral distribution of energy in the tones. The second axis showed clustering according to the instrument families of the tones: strings, woodwinds and brass. This second axis was related to some conglomerate of temporal features of the tones that characterize family membership, such as the relative timings for the entries of the various harmonics.

Grey (1975, 1977) obtained a three-dimensional map which best explained a larger set of temporally complete instrument tones (16 stimuli instead of 9) and two of the dimensions were similar to those found by Wessel. Again, one axis clearly related to the spectral energy distribution of the stimuli. The clustering of stimuli by instrument family was also found, whereby the three families clustered in the remaining two dimensions, forming a cylinder about the spectrally related axis. The possible physical bases of the second and third dimensions were uncovered, as one axis seemed to relate to the degree of temporal synchronicity in the attacks and decays of the upper harmonics and the corresponding degree of spectral fluctuation

 0001-4966/78/6305-1493$00.80  

throughout the signal. The third dimension appeared to relate to the presence or absence of high-frequency, possibly inharmonic, energy in the initial attack segments of the tones.

Miller and Carterette (1975) used totally synthetic tones, not fully representative of natural instrument sounds, but rather having a determined subset of timbral attributes: the number of harmonics present, the shape of the temporal energy envelope, and the pattern of the relative timings of the entries of the harmonics. A three-dimensional similarity structure suggested that listeners based their judgments on the first two features, ignoring the latter property of the stimuli.

Note that all the above studies have uncovered at least one common attribute of tone underlying the similarity structures for steady-state tones, temporally complete natural tones, and synthetic tones. That attribute has to do with the *spectral energy distribution* of the signals. For steady-state tones this tonal property serves as the sole basis for interpreting the perceptual relationships, and it manifests itself as a complex three-dimensional structure. With temporally complete natural tones the spectral component would seem to be a single dimension in the similarity space. The interpretation of this axis becomes more ambiguous, taking into consideration many factors to characterize the energy distribution such as the bandwidth of the signal, the balance of energy in the low harmonics, and the existence of upper formants. In the case of the synthetic tones, the number of harmonics, or the bandwidth of the signal, was found to underly two of the three dimensions of the similarity structure.

In the following study, we are interested in the effects of modifications on the spectral energy distributions for a set of natural tones that had been scaled previously for similarity (Grey 1975, 1977). A comparison of the perceptual similarity structure of the modified tones with that of the original tones would show predictable alterations if the interpretation of the spectral axis was correct. To simplify the testing of this interpretation, we modified only half of the original stimuli by trading their spectral shapes in pairs. Hence eight of the original stimuli were unaltered and the eight modified tones were modified in pairs, leading to the simple prediction that only the latter eight tones would exchange positions on the spectral axis, by pairs, in the new scaling solution. Note, however, that spectral modifications will necessarily alter the temporal characteristics of the signals—hopefully, to a lesser extent—due to alterations in the relative onset and offset slopes of the partials. Hence we would expect (minor) perturbations along the temporally related axes in the scaling of the new tones.

## I. STIMULI

The stimuli were derived from 16 instrumental notes played near the pitch of $E\flat$ above middle $C$, approximately 311 Hz, whose durations ranged between 280–400 ms (see Grey, 1977, for a more complete description of the recording process). The tones were digitized at a 25.6-kHz sampling rate and the high-order 12 bits were stored in digital form for computer analysis or playback. The 16 instrumental tones consist of two oboes (different instruments and players), English horn, basson, $E\flat$ clarinet, bass clarinet, flute, two alto saxophones (one instrument, played at $p$ and $mf$), soprano saxophone, trumpet, French horn, muted trombone, and three celli (one instrument, played normally, muted *sul tasto* and *sul ponticello*).

The tones were digitally analyzed with a heterodyne filter technique (see Grey and Moorer, 1977). This technique produced a set of time-varying amplitude and frequency functions for each harmonic of the instrumental tone. Digital additive synthesis was used to produce a synthetic tone, where a set of harmonics, each controlled in amplitude and frequency through time, were added together. The functions obtained from the analysis were considerably simplified before being used for synthesis. Where the original functions were quite complex, consisting of 300–500 sampling points, the data-reduced functions consisted of 4–8 line segments. The simplified tones were almost indiscriminable from their complex versions (Grey and Moorer, 1977). The stimuli were equalized in an experiment for loudness, pitch, and perceived duration (see Grey, 1975).

These 16 tones were used as described above in our previous scaling study (Grey 1975, 1977).

For the following study, however, eight of the tones were modified spectrally in four pairs: trumpet–trombone, oboe–bass clarinet, bassoon–French horn, and two celli (normal and *sul ponticello*). The spectral envelopes for each pair of tones were traded, so that the resulting peak amplitudes for the harmonics of one tone were made to correspond to the original peaks of the other tone, and vice versa. Figure 1 presents the original and modified versions of the trumpet and trombone. An exchange of spectral envelope was done by resetting the maximum amplitude reached by some particular harmonic of the trumpet to the level of the corresponding harmonic in the trombone, and vice versa. This switching operation was performed only between the harmonics that existed in both tones; if one of the tones had more harmonics than the other, its extra upper harmonics were not changed in amplitude. Therefore, the original bandwidths of the signals were not effected by the spectral shape exchange, as one can observe with the trumpet–trombone exchange. Only the shape of the distribution of spectral energy within the common set of harmonics was traded.

## II. LISTENERS AND PROCEDURE

This experiment was run on two separate occasions. On the first occasion, 19 listeners at Stanford University were employed, with one listener repeating the experiment, giving a total of twenty data sets. On the second occasion, we attempted to replicate our first results with 20 additional runs, using five repeating listeners from the first set, plus 11 newly recruited listeners, four of whom repeated the experiment. This gave us a grand total of 40 data sets. Listeners were musically sophisticated, some actively involved in ad-

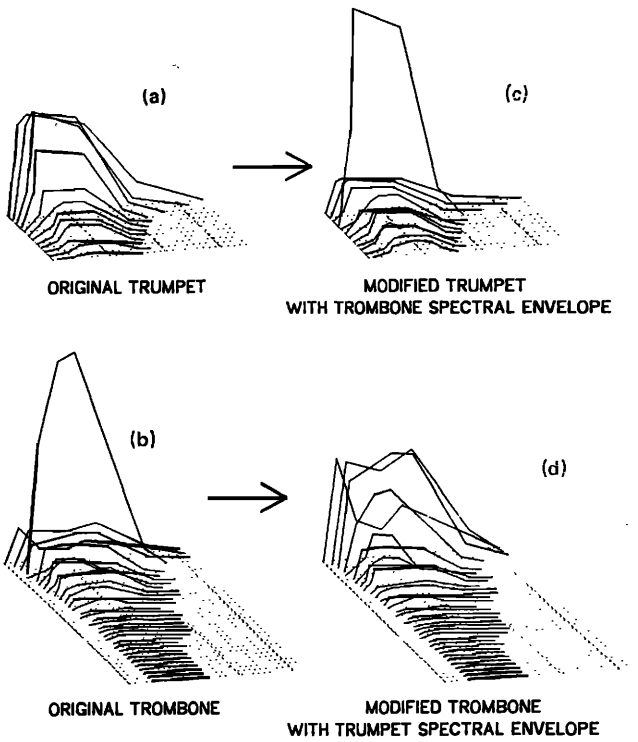


FIG. 1. Exchange of spectral envelopes between the trumpet (a) and trombone (b) to form the modified trumpet (c) and modified trombone (d). Note that only the common harmonics trade peak amplitude values, and hence the original frequency bandwidths (number of harmonics) are retained.

vanced instrumental performance and others in conducting, musical composition, and/or music synthesis.

Data sets were collected in an hour session. Each trial consisted of a warning knock, the two tones for comparison, and a decision interval. The warning knock preceded the first tone by 2.5 s, 1.5 s separated the two tones, and 6 s were given for the listener to make a judgment. There were a total of 270 trials, 30 of which were practice trials. The remaining 240 trials consisted of the $n(n-1)$ possible pairs of 16 tones, given in both directions. Trials were presented in a random order.

The only difference in presentation of the stimuli between the two runs of the experiment was a change in location. The first run was presented in a large and relatively reverberant room, whereas the second run was presented in a smaller and considerably less reverberant room. In both cases, listeners were located approximately 12 feet from the speaker, an Altec Lansing model 604.

Listeners were told to rate the *similarity* of the two tones relative to that of all other pairs of tones heard. They were instructed that the first 30 pairs were practice, and that they could change their rating strategies during that time. The similarity rating was made on a scale of 1–30, and this scale was presented to listeners as having three general ranges: (1) 1–10, *very dissimilar*; (2) 11–20, *average level of similarity*; and (3) 21–30, *very similar*, relative to all pairs.

## III. RESULTS AND DISCUSSION

The similarity judgments for each listener were stored as a $16\times16$ matrix of data, recording responses with respect to the exact order of presentation for any pair of stimuli. An examination was made for the existence of consistent order-related response differences across all listeners for any pair of stimuli. A half-matrix of order-related response differences was generated for each listener by subtracting the upper from the lower half of the response matrix. A student $t$-test was made for each cell of the half-matrix across the 40 data sets for a consistent direction of differences. None of the 120 cell means for the 40 half-matrices were significantly different from zero, and they were all within 0.8 standard deviations of zero. Therefore the original response matrices were transformed into half-matrices for each listener by averaging the two responses to a pair of stimuli presented in different orders.

The 40 averaged half-matrices were treated with a multidimensional scaling algorithm that takes individual differences into account: INDSCAL (Carroll and Chang, 1970). Spatial representations were obtained in two, three, and four dimensions. Goodness-of-fit measures, defined as the correlation between the scalar products of the actual and predicted distances, for the solutions were 0.81 for four dimensions, 0.78 for three dimensions, and 0.70 for two dimensions.

In order to compare the similarity structures between the first and second runs of this experiment, the two respective sets of 20 half-matrices were given to INDSCAL for three-dimensional spatial solutions. The spatial solution for the first run was rotated to best fit that of the second run and Pearson product-moment correlations were found for spatial coordinates. The correlation between the two configurations was 0.99, suggesting that there were no major differences between the two runs.

In addition to subjecting the similarity matrices to multidimensional scaling analysis, they were also treated with a hierarchical clustering algorithm (HICLUS by Johnson, 1967). A group matrix was formed by averaging the rank orders of the ratings in the 40 individual matrices (since HICLUS works on rank orders of responses). The clustering algorithm produced an analysis of the similarity data which was independent of the spatial-dimensional reduction generated by multidimensional scaling. The *compactness*, or diameter, method of clustering was found to give the most interpretable results. The analysis grouped the most similar stimuli into clusters; then grouped such clusters into higher-order clusters, continuing this way until the whole set of stimuli were in one cluster. Cluster strength reflected degree of similarity, so that the lowest level clusters were the strongest.

The INDSCAL and HICLUS analyses were used in conjunction with one another to interpret the data (see Shepard, 1972). The three-dimensional INDSCAL solution was found to be the most useful for interpreting the similarity structure of the stimuli. The two-di-

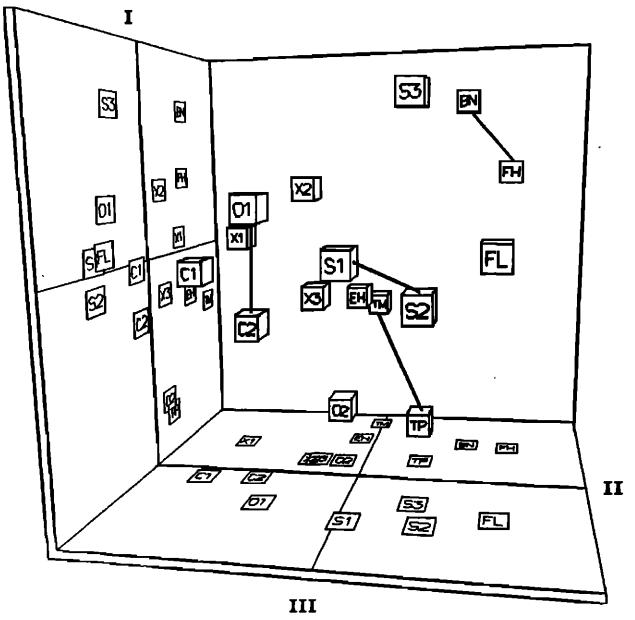


FIG. 2. Three-dimensional spatial solution for 40 similarity matrices generated by multidimensional scaling program INDSCAL (Carroll and Chang, 1970). Two-dimensional projections of the configuration appear on the wall and floor. Lines connect pairs of tones that traded spectral envelopes. Abbreviations for stimulus points: O1 and O2, oboes; C1 and C2, clarinets; X1, X2, and X3, saxophones; EH, English horn; FH, French horn; S1, S2, and S3, strings; TP, trumpet; TM, trombone; FL, flute; BN, bassoon.

mensional spatial solution presented several discrepancies with the clustering analysis, and as a spatial solution was difficult to interpret. The three-dimensional solution overcame the problems of clustering and seemed more interpretable. However, there was no benefit found for interpreting the data by increasing the number of dimensions to four.

The three-dimensional INDSCAL solution is shown in the perspective plot of Fig. 2. Dimension I is the vertical axis, II is the horizontal axis, and III is the depth axis. The abbreviations adopted for the 16 tones are O1 and O2, the two oboes; EH, the English horn; BN, the basson; C1, the $E\flat$ clarinet, and C2, the bass clarinet; X1 and X2, the two saxophone tones (*mf* and *p*, respectively); X3, the soprano sax; FL, the flute; TP, the trumpet; FH, the French horn; TM, the muted trombone; S1, S2, and S3, the cello tones (labeled strings: *sul ponticello*, normal bowing, and muted *sul tasto*, respectively). The distances of the stimuli are given by their relative sizes, and the two-dimensional projections of the configuration on the wall and the floor.

The same interpretation of the axes arrived at for the original scaling study is applicable to this solution [see Grey (1977) for complete details]. Dimension I can be related to the spectral energy distribution of the signals, in terms of the combined effects of bandwidth, balance of energy in the lower harmonics, and the prominance of upper-formant regions. We present a quantitative model for this axis in Sec. IV. Dimension II appears to relate to the form of the onset–offset patterns of tones, especially with respect to the presence of synchronicity in the collective attacks and decays of upper harmonics; closely related to synchronicity is the degree of spectral fluctuation found throughout the tone, where less synchronicity in the attack and decay groupings is accompanied by greater overall spectral fluctuation through time. Dimension III is also interpretable in terms of temporal patterns of the signals, in this case the presence or absence of high-frequency, low-amplitude energy, most often *inharmonic energy*, during the attack segment. Note also that the projection of the tones along the latter two dimensions (on the floor of Fig. 2) shows a clustering by family relationships that resembles the previous results.

Of central interest in this study is the comparison of the configuration obtained using these modified stimuli with the perceptual configuration based on the original stimuli. The earlier scaling solution (Grey 1975, 1977) is shown in Fig. 3. The tone pairs that traded spectral shapes in the present experiment were TP–TM, FH–BN, S1–S2, and C2–O1. In accordance with our previous interpretation of the spectral axis, these pairs actually exchanged orders on that axis. The tone pairs that exchanged spectral envelopes for this experiment are joined by lines in Figs. 2 and 3.

Slight alterations in position on the other two axes may be noted, along with small alterations in the overall structure of the configuration. Changes along these axes were expected, since spectral modifications also modified the temporal characteristics of the tones. For instance, TP, moved from the previous position of an extreme on dimension III, in Fig. 3, to a more central position in the present scaling, in Fig. 2. This may have been a result of the amplification of the inharmonic frequency shifts during the attack segment in the upper harmonics given by the spectral exchange

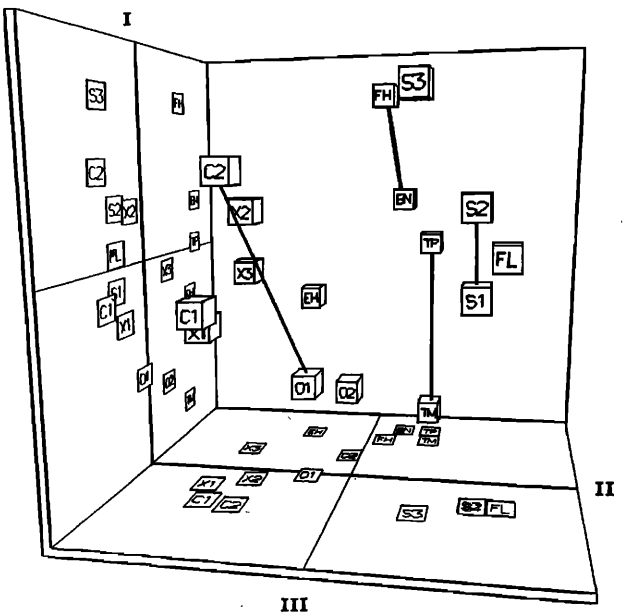


FIG. 3. Three-dimensional spatial solution for original 16 tones (Grey, 1977), as in Fig. 2. Lines connect original unmodified pairs that traded spectral envelopes in this experiment. Abbreviations as in Fig. 2.

with TM. The FH appears to have made a large shift along dimension II, and the modified signal indeed displays more spectral fluctuation than the original. We have not attempted to formally explain all shifts on the temporally related axes, however, in that this experiment was primarily oriented towards spectral effects. The local perturbations found in the new configuration obviously reflect alterations in the perceived relationships among the eight unchanged tones and the eight modified tones due to the effects of the modifications. An analysis of these changes along with a formal modeling of a dimension has been limited in this study to the spectral axis.

## IV. QUANTITATIVE MODELS OF SPECTRA

In order to test further the interpretation of the spectral axis, we have attempted to mathematically model the spectral energy distributions of our stimuli. The modeled spectra can then be compared with one another and with reference to the ordering of stimulus points along the axes interpreted to relate to spectrum in the scaling studies. A number of closely related models were constructed; the common goal of the models was to represent the spectrum of each stimulus with a single number. We wanted to do this so that a statistical correlation could be made between the distribution of modeled stimulus spectral values and the arrangement of the stimuli along the spectral axes in the perceptual spaces. The success of any particular quantitative model was measured by the relative strength of its correlation with the perceptual axis. Our primary goal in constructing various mathematical models of stimulus spectra to be correlated with the relevant perceptual axis was to formally verify the interpretation of that axis. A secondary goal was to examine the effects of using different methods for characterizing the physical spectra of the stimuli.

There were several steps in the numeric characterization of a stimulus spectrum. In the first place, a static spectral energy distribution, commonly represented by a *line spectrum*, had to be derived from the time-variant spectral information used to synthesize any stimulus. Three different methods were used to derive a *line spectrum* from a set of time-variant amplitude envelopes of the harmonics of a tone: in the first, each harmonic level was set by the peak amplitude reached by that harmonic through time; the second technique set each harmonic level equal to the temporal average of its amplitude envelope; the third and final alternative employed in our modeling set each harmonic level equal to the temporal average of the energy of that harmonic. At this point, the given line spectrum might be represented in the units specified above, or transformed into *decibels*. Further transformations of the line spectrum that related to perceptual processing were also sought. We decided to apply the model of loudness perception constructed by Zwicker and Scharf (1965) and derived functions describing both the peripheral *excitation pattern* of a line spectrum as well as the more central *loudness function*.

The static characterization of a stimulus spectrum, following one particular set of choices above, is presented graphically in Fig. 4. The time-variant amplitude envelopes of the harmonics of the clarinet tone C1 are shown in Fig. 4(a). This information has been transformed into an average spectral representation of the tone, shown as the line spectrum in the middle of the figure, by taking the temporal average of each amplitude envelope. This line spectrum was then transformed by Zwicker and Scharf's model for loudness perception of a steady-state tone, shown at the bottom of the figure.

Given one of the various possible static representations of a stimulus spectrum, the final stage was to compute a numeric characterization of that spectral pattern. This finally would represent the stimuli on a one-dimensional vector to be correlated with the perceptual axis. We chose two methods to characterize numerically a spectral distribution: the mean (centroid, or "balance point") and the median. As an example, the centroid of the loudness distribution for the

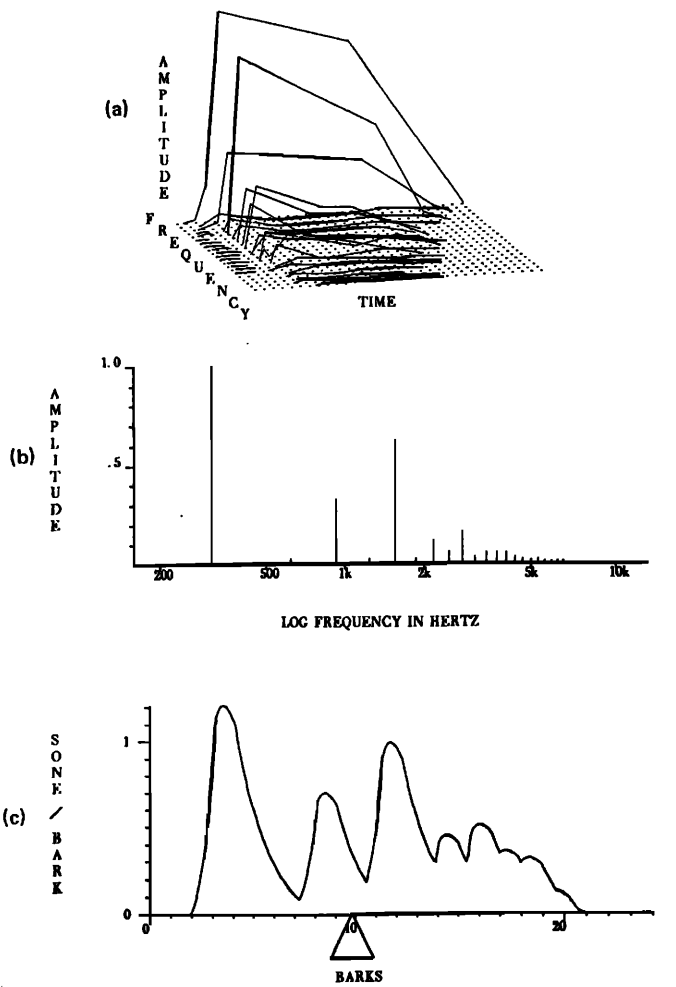


FIG. 4. Model for spectral energy distribution applied to clarinet tone. The time-variant amplitudes of each harmonic (a) are averaged over the total duration of the tone to produce a line spectrum (b) which is then transformed by a function representing the perceived loudness of a complex tone (c) according to the model of Zwicker and Scharf (1965). The balance point of (c) is taken to represent the spectral distribution, shown by the fulcrum placed at 9.9 Barks for this tone.

TABLE I. Correlations between modeled stimulus spectra and the spectrally related axis for each of three perceptual experiments [original tones (Grey, 1977), spectrally modified tones (the present study), and interpolated tones (Grey, 1975)]. Rows represent the method used to derive a harmonic level from the time-variant amplitude envelopes (peak amplitude, average amplitude, or average energy). Columns represent the form of the static spectral pattern [line spectrum in linear units or decibels, excitation pattern, or loudness function (Zwicker and Scharf, 1965)]. The correlation using the centroid of spectral patterns is entered first, with the correlation using the median following in parentheses.

| | | Static spectral pattern | | | |
|---|---|---|---|---|---|
| Harmonic level | Experiment | Line-spectrum (linear units) | Line-spectrum (decibels) | Excitation pattern | Loudness function |
| Peak Amplitude | Original tones | 0.90 (0.86) | 0.76 (0.76) | 0.88 (0.88) | 0.92 (0.88) |
| | Spectral mod. | 0.81 (0.82) | 0.46 (0.50) | 0.77 (0.79) | 0.89 (0.86) |
| | Interpolated | 0.86 (0.80) | 0.77 (0.76) | 0.83 (0.81) | 0.86 (0.84) |
| Avg. Amplitude | Original tones | 0.93 (0.91) | 0.83 (0.79) | 0.91 (0.91) | 0.94 (0.93) |
| | Spectral mod. | 0.87 (0.90) | 0.61 (0.60) | 0.80 (0.82) | 0.92 (0.90) |
| | Interpolated | 0.91 (0.84) | 0.87 (0.84) | 0.90 (0.90) | 0.92 (0.91) |
| Avg. Energy | Original tones | 0.85 (0.66) | 0.80 (0.78) | 0.90 (0.90) | 0.94 (0.90) |
| | Spectral mod. | 0.90 (0.77) | 0.55 (0.56) | 0.79 (0.81) | 0.91 (0.88) |
| | Interpolated | 0.80 (0.61) | 0.84 (0.82) | 0.87 (0.87) | 0.90 (0.88) |

clarinet tone is shown at the bottom of Fig. 4. A one-dimensional vector of stimulus points according to this model was correlated to the perceptual axis believed to relate to spectrum; all of the possible numeric characterizations of spectral energy distributions were similarly tested. Correlations were computed not only with the perceptual axis obtained in this experiment, but also with the spectrally related axes from two earlier experiments—one was the original experiment from which the present study was derived (Grey, 1977) and the other a scaling of a mixed set of natural and interpolated tones (Grey, 1975).

The results of correlating the outputs of the different models above with the perceptual axes of the three experiments are given in Table I. The arrangement of the models in the table has the rows represent the technique used to characterize a harmonic level (peak amplitude, average amplitude, or average energy) and the columns represent the various forms for characterizing the static spectral pattern (line spectrum in linear units or decibels, excitation pattern, or loudness pattern). The correlations based on the centroids of the spectral patterns are given for each of the three experiments (original tones, spectrally modified tones, and interpolated tones), and the correlations of the medians follow in parentheses.

The set of spectral models constructed was not, of course, an exhaustive set. However, the generally high correlations of all models with the perceptual results did add formal support to the interpretation of the perceptual axis. This was our primary goal. Secondarily, however, we may discriminate between the various models constructed by making the following general observations. First, the linear representations of physical line spectra worked better than representations in terms of decibels. Similarly, the loudness function was better than the excitation pattern, although in some cases the excitation pattern achieved high correlations as well. We may also note that the loudness function was slightly better than the linear physical representation, a result we would expect in correlations with a perceptually derived axis. While the loudness function gave high correlations in all rows, overall we find that the average amplitude characterization of a harmonic level worked slightly better than the other two methods. Finally, we note that the centroid was a better numeric representation for most static spectral distributions than the median.

The most successful of the set of models employed here was the centroid of the loudness function for the time-averaged amplitudes of the stimulus harmonics, as shown in Fig. 4. With slight rotations ($< 10°$) of the spatial axes obtained from the three experiments, the correlations improved slightly to 0.95, 0.94, and 0.94. We feel that this particular model is an adequate representation of spectral energy distribution, in that it simultaneously takes into consideration the many factors which may be important: overall bandwidth, balance of levels in the lower harmonics, and the existence of strong upper formants. Furthermore, it attempts to model the distribution perceptually, in this case by use of the loudness transformation. We would expect that alternate models that took the above factors into consideration in their construction could possibly surpass the models thus far presented. However, the high correlations given here do indeed support, and begin to formalize, the interpretation of the spectral axis.

## V. CONCLUSIONS

The multidimensional scaling of a modified set of music instrument tones performed in this study has supported an interpretation of one of the axes from earlier scaling studies. The axis under examination had been related to the spectral energy distribution of the signals: a composite of the effects of bandwidth, balance of energy in the lower harmonics, and upper-formant regions. The stimuli constructed for the pres-

ent experiment were derived from those of an earlier experiment by modifying the spectral shapes of half of the signals. These modified signals were constructed in pairs, by a process of trading their spectral shapes.

In the resulting three-dimensional configuration that represented the structure of the listeners' similarity judgments, one axis again appeared to relate to the spectral energy distribution of the signals. The positions of the modified pairs of points exchanged orders on the axis, thereby supporting the original interpretation of this dimension as relating to spectrum.

Since the spectral modifications did not change the bandwidths of the signals (only the amplitudes between common harmonics were traded) and since the resultant signals did in fact exchange orders on the spectral axis, we may conclude that bandwidth per se is not as important a spectral factor as the actual shape of the energy distribution. The case which shows this most clearly is TM vs TP, where the dominance of the fourth harmonic seems to be the most critical factor.

A set of models were constructed to represent numerically a time-variant spectrum, and the relationships between modeled stimulus spectra generally corresponded well to the positions of the stimulus points along the spectral axes of the perceptual similarity spaces from three separate experiments. The most successful model of the set measured the level of each harmonic by its time-averaged amplitude envelope and then transformed the resulting static spectral pattern by a function relating to the perception of loudness (Zwicker and Scharf, 1965). The centroid of this transformed spectrum, that is, the mean of the distribution, was taken to characterize numerically the input physical spectrum. The high correlations found with most of the models more formally supported the interpretation of the spectral axis; additionally the differential successes of various models suggested components that may be important in the construction of such models.

The earlier interpretations of the remaining two dimensions of the three-dimensional timbre space also seemed consistent with our present results. These two dimensions related to the temporal pattern of the attack and decay of the tones, namely, the presence of low-amplitude, high-frequency energy (perhaps inharmonic) in the initial attack segment and the presence of synchronicity in the attacks and decays of the higher harmonics along with a corresponding degree of spectral fluctuation throughout the signal.

Previous observations had noted the possibility that stimuli clustered according to musical instrument family membership in the similarity space (Wessel 1973, 1974; Grey 1975, 1977), and that this clustering took place in the dimensions of the space relating to temporal characteristics of the tones. In looking at both three-dimensional configurations shown above, one may observe very similar clusters of tones in the two-dimensional projection of the temporal axes. The overall structure might be compared to a cylinder, where the three family clusters (represented in two dimensions) are spread along the spectral axis. It is hoped that our preliminary interpretations of the two temporal axes in fact may relate to features of tone which characterize family membership.

The dimension that relates to spectral fluctuation and synchronicity is apparently divided into two on the basis of belonging to the woodwind family or not. The two exceptions to this observation (BN and FL) have different temporal patterns from the other woodwinds, making them more similar to the non-woodwinds. This may indicate that the clustering is based more upon perceived features of tone than to some cognitive recognition or class-membership naming function. From this point of view, the temporal pattern of woodwind seems to include synchronous attacks and decays of upper partials with little spectral fluctuation in the interim. The subjective correlate to this dimension of timbre has been informally observed as the quality of being *static* vs *dynamic* through the duration of the tone. The woodwinds, while often having inharmonic attacks (see below), tend to quickly obtain a spectral quality and hold it constant through most of the tone; within this pattern, the fundamental usually precedes the upper harmonics, which enter as a group. Contrasting to this, the brass, strings, FL, and BN have much more noticeable spectral transitions in their attacks and decays. Often, especially in the strings and FL, the tone never does settle down to a static condition.

A subjective correlate observed for the other temporal axis, relating to the existence of high-frequency inharmonicity during the attack, was the noiselike character of the attack. Those instruments which have perceptible inharmonicity show the quality of a "buzz-like" or slightly grating attack. This is in contrast to attacks that are relatively clean, or which are dominated by low-frequency energy. Another strong correlate (pointed out by Wessel, 1977) is the "explosiveness" or "hardness" of the attack. If there is much precedent high-frequency energy in the attack, the tone is heard to have a longer, softer attack; more time is taken to reach maximum amplitude. If the lower harmonics seem to come in quickly and dominate the initial spectral shape, then the attack sounds faster, harder, and more explosive. Note that the two exceptions to the family clustering, FL and BN, are appropriately located at opposite extremes on this dimension and are apparently clustered within respective family groups that characteristically do or do not display precedent high-frequency inharmonicity, namely the strings and the brass. The ordering of tones within the woodwinds along this dimension reflects differences in the amount or salience of their high-frequency precedent energy.

The spectral modifications performed on tones for this study caused very strong changes in the perceived attributes of the tones. The spectral distribution in speech gives vowels their particular color. With musical tones the same sort of percept occurs. Comparing a pair of tones before and after spectral exchange, one hears the tones switch to each other's vowel-like color but maintain their original articulatory pattern of attack and decay. The TM–TP switch took the muted quality from the former and put it upon the latter (the high

fourth harmonic is a correlate); the O1–C2 switch tended to cause a reversal in identification of the instruments (the spectral cues for these two instruments are particularly strong); the switch between the strings formed the analogy of two playing styles performed on each of two instruments; and the switch of BN–FH suggested reversal of identification, but accompanied by the original temporal controls, presented paradoxical attacks: a bassoon with the highly explosive attack of a brass instrument and a horn with the rounder attack of a reed instrument.

## ACKNOWLEDGMENTS

This work was supported in part by NEA grant C 50-31-282 and by NSF contracts BNS 75-17715 and DCR 75-00694. We would like to thank the many musicians who participated in this experiment. Also, we are indebted to the Stanford Artificial Intelligence Laboratory, whose computer facilities were used to perform this research.

Beauchamp, J. W. (1969). "A computer system for time-variant harmonic analysis and synthesis of musical tones," *Music by Computers*, edited by H. von Foerster and J. W. Beauchamp (Wiley, New York).

Carroll, J. D., and Chang, J. J. (1970). "Analysis of individual differences in multidimensional scaling via an *N*-way generalization of 'Eckart-Young' decomposition," Psychometrika **35**, 283–319.

Chowning, J. M. (1973). "The synthesis of complex audio spectra by means of frequency modulation," J. Audio Eng. Soc. **21**, 526–534.

Freedman, M. D. (1967). "Analysis of musical instrument tones," J. Acoust. Soc. Am. **41**, 793–806.

Freedman, M. D. (1968). "A method for analyzing musical tones," J. Audio Eng. Soc. **16**, 419–425.

Grey, J. M. (1975). "An exploration of musical timbre," Ph.D. dissertation (Dept. of Music Report No. STAN-M-2, Stanford University, CA) (unpublished).

Grey, J. M. (1977). "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**, 1270–1277.

Grey, J. M., and Moorer, J. A. (1977). "Perceptual evaluations of synthetized musical instrument tones," J. Acoust. Soc. Am. **62**, 454–462.

Johnson, S. C. (1967). "Hierarchic clustering schemes," Psychometrika **32**, 241–254.

Kruskal, J. B. (1964a). "Mutlidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika **29**, 1–27.

Kruskal, J. B. (1964b). "Nonmetric multidimensional scaling: a numerical method," Psychometrika **29**, 115–129.

Licklider, J. C. R. (1951). "Basic correlates of the auditory stimulus," *Handbook of Experimental Psychology*, edited by S. S. Stevens (Wiley, New York).

Luce, D. A. (1963). "Physical correlates of nonpercussive musical instrument tones," Ph.D. dissertation (MIT, Cambridge, MA) (unpublished).

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

Miller, J. R., and Carterette, E. C. (1975). "Perceptual space for musical structures," J. Acoust. Soc. Am. **58**, 711–720.

Moorer, J. A. (1975). "On the segmentation and analysis of continuous musical sound by digital computer," Ph.D. dissertation (Stanford University, CA) (unpublished).

Plomp, R. (1970). "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (A. W. Sijthoff, Leiden).

Risset, J. C. (1966). "Computer study of trumpet tones," Bell Telephone Labs, Murray Hill, NJ.

Shepard, R. N. (1962a). "The analysis of proximities: multidimensional scaling with an unknown distance function. I," Psychometrika **27**, 125–140.

Shepard, R. N. (1962b). "The analysis of proximities: multidimensional scaling with an unknown distance function. II," Psychometrika **27**, 219–246.

Shepard, R. N. (1972). "Psychological representation of speech sounds," *Human Communication*, edited by E. E. Davis and P. B. Denes (McGraw–Hill, New York).

Wedin, L., and Goude, G. (1972). "Dimension analysis of the perception of instrumental timbre," Scand. J. Psychol. **13**, 228–240.

Wessel, D. L. (1973). Report to Mathematical Psychology Meetings, San Diego, CA (unpublished).

Wessel, D. L. (1974). Report to C. M. E. University of Calif., San Diego, CA.

Zwicker, E., and Scharf, B. (1965). "A Model of Loudness Summation," Psych. Rev. **72**, 3–26.