Canadian University Music Review
Revue de musique des universités canadiennes

Canadian University Music Review

Some Perceptual Aspects of Timbre

Campbell L. Searle

Numéro 3, 1982

URI : https://id.erudit.org/iderudit/1013829ar
DOI : https://doi.org/10.7202/1013829ar

Aller au sommaire du numéro

Éditeur(s)

Canadian University Music Society / Société de musique des universités canadiennes

ISSN

0710-0353 (imprimé)
2291-2436 (numérique)

Découvrir la revue

Citer cet article

Searle, C. L. (1982). Some Perceptual Aspects of Timbre. *Canadian University Music Review / Revue de musique des universités canadiennes*, (3), 80–101. https://doi.org/10.7202/1013829ar

SOME PERCEPTUAL ASPECTS OF TIMBRE*

Campbell L. Searle

In the study of how we perceive musical timbre, we encounter a number of different *representations* of the musical information, several of which have appeared in preceding papers in the symposium. The first is the musical score, which in engineering terms is a graph of frequency versus time, with appended annotations about loudness, scoring, timing, etc. The next representation is sound in a concert hall, created from the score by an orchestra or computer: technically, sound pressure as a function of time. The representations of interest in this paper are those produced by the mechanical and neural systems in our heads in response to this sound.

The First Auditory Representation

Properties of the Ear

Experiments involving auditory masking (see Patterson 1976), critical bands (summarized in Tobias 1970), basilar membrane motion (see Johnstone & Boyle 1967; Rhode 1971; & Evans & Wilson 1973) and tuning curves of primary auditory nerve fibers (see Kiang 1965 & 1974), all indicate that the first neural representation of sound in our heads results from a frequency analysis of the incoming sound by a fluid-filled bony structure called the

*Various aspects of this work were supported by the National Research Council, the Medical Research Council, the Defence and Civil Institute of Environmental Medicine through the Department of Supply and Services, Canada; and the Vinton Hayes Fund at the Massachusetts Institute of Technology. The author gratefully acknowledges the help of M.M. Taylor, B.J. Frost, W. Richards, H. Secker-Walker, and L. Schwalm. Portions of this article appeared previously in *The Humanities Association Review*, XXX/1-2, (1979), 93-103, and in the *Canadian Journal of Psychology*, XXXVI/3, (1982), 402-19. Reprinted by permission.

cochlea. The frequency resolution of this system is somewhat less than one-third of an octave (above 400 Hz), as can be seen from Figure 1, the so-called critical bandwidth data derived from psychophysics. A corresponding set of psychophysical experiments indicate that the temporal resolution of the system is at best a few milliseconds (e.g., see Viemeister 1979). The phase sensitivity of the ear is still a controversial topic, but most studies indicate that the system is only minimally sensitive to the relative phase of harmonics. Of course, the ear is exquisitely sensitive to the *interaural* phase of any sine wave component below 500 Hz. But this gives rise to auditory localization, and not perception of timbre. Hence in the remainder of this paper we will ignore phase.

It may at first glance seem quite incongruous that the auditory system, which has excellent pitch discrimination (two or three hertz for 1000 Hz pure tones), should analyze sound with filters as broad as one third of an octave, that is, three or four notes on the chromatic scale. But it is easy to show that if one uses not just one filter, but several overlapping filters, then accurate information concerning pitch is available, limited only by the *slope* of the filter characteristic, and not the filter bandwidth.

The Model

To obtain a better idea of this first neural representation of music and speech, we have constructed a "model ear" with properties approximating those of the human ear discussed above. Our model, shown in Figure 2, consists basically of a bank of 1/3 octave filters, covering from 125 Hz to 6.3 kHz, followed by envelope detectors. To simulate the roughly constant critical bandwidth below 400 Hz, we added together the detector outputs of the 125- and 160-Hz channels, and also the 200- and 250-Hz channels (see Dockendorff 1978).

The detector time constants were chosen to produce fast rise time consistent with low ripple. In filter systems such as this one which have wider bandwidths at higher frequencies, the rise time of the filters decrease with increasing center frequency. Hence we chose detector time constants to correspond, such that the overall rise times of the filter-detector units were inversely proportional to frequency. Specifically, the 1 kHz channel has an overall rise time of 6 milliseconds, the 2 kHz channel, 3 milliseconds, and so forth.

As noted in Figure 2, the detectors are connected to a 16-channel CMOS multiplex switch, which samples the output of

each channel every 1.6 milliseconds. (This rate is appropriate for the high-frequency channels, but oversamples the low channels.) The multiplexed output is then passed through a logarithmic amplifier to match the logarithmic nature of perceived loudness in the ear. There are several important aspects of the human auditory system that are not modeled by this system, such as two-tone inhibition, the limited dynamic range of the neural system, etc., but it appears to us to be a reasonable starting point for research. (For more details, see Rayment 1977, or Searle, *et al.* 1979.)

Examples of the First Representation

Let us examine the output of our model ear when speech and music are applied. The log amplitude outputs from each of the filter-detector channels for three seconds of piano music are shown in Figure 3. The pianist played two ascending octaves of the C major scale — from C_4 to C_6 — at a fairly rapid tempo. The note names are shown at the top of the figure, with arrows indicating the time when the notes were struck. The figure is rich in detail, as we should expect, because to the extent that the system in Figure 2 models the peripheral auditory system, all details of pitch, rhythm, melody, timbre, etc. must be represented somewhere in the plot.

This particular way of plotting emphasizes the *temporal* aspects of the music. For example, we see that each piano note has an abrupt onset: in the high-frequency channels at the top of the figure, the intensity of the sound may increase a hundred-fold in a few milliseconds. Also, each note reaches maximum intensity shortly after onset, and thereafter gradually dies down until another note is struck. Both of these features are characteristics of percussive instruments. Figure 4 shows the quite different temporal plot for an alto flute. Note, for example, the much more gradual attack for each note, lasting for 40 or 50 milliseconds.

In contrast to the original two-dimensional score representation, this first neural representation, of which Figures 3 and 4 are examples, is a three-dimensional display: amplitude versus frequency and time.[1] The third dimension is required because the notations on the score about amplitude, voicing, etc. must be coded into the neural representation. Displaying a three-dimensional plot on the two dimensions of a printed page is a challenge of the visual arts quite unrelated to auditory reception. In Figures 3 and 4 we chose to plot the three dimensions as amplitude versus time

for each of sixteen different frequencies. To help the visual system interpret our data, we can replot the same data as amplitude versus frequency, for many different times. Figure 5 shows the piano passage replotted in this manner. This plot, which we call the "running spectrum," emphasizes the relative amplitudes of the overtones rather than the time course of the notes. For example, the figure shows clearly one aspect of timbre, namely the change in the relative size of the harmonics as we progress up the scale. At C_4, the second harmonic is substantially larger than the fundamental, at D they are roughly equal, and from F_4 to F_5, the fundamental becomes increasingly dominant. Also evident is the progression of the "melody" up the scale, information that was difficult for the visual system to discern in Figure 4.

The corresponding running spectrum for the flute is shown in Figure 6. As expected, we see a very different harmonic structure, and a different timbre change as we move up the scale.

The first representation must be musically complete, because we have no other path for the music to reach the brain. Therefore all information of cadence, melody, rhythm, tonality, timbre, brightness, fullness, presence, openness, etc. must be somehow coded into these plots. It is almost insulting to say that all the beauty of a performance of Moussorgsky's *Pictures at an Exhibition* can be reduced to a collection of lines on these graphs, but disquieting as it is, such a statement follows logically from the above argument.

The concept of *spectral envelope* introduced by Wayne Slawson in the preceding paper (see Fig. 1, p. 68) is quite compatible with this first neural representation. Instruments such as the flute have a fairly fixed spectral envelope regardless of what note is being played, especially within one register (see Luce & Clark 1967). This envelope can be obtained by averaging together spectra derived from the instrument when many different notes are played. Thus simply averaging the running spectra of Figure 6 will yield a close approximation to the spectral envelope of the flute. Figure 7 shows the flute spectral envelope so derived, with the corresponding plots for a piano and a viola. The acoustic resonances and coupling effects which give rise to these spectral envelopes are responsible for the characteristic change in harmonic structure of instruments, as discussed above. Hence the spectral envelopes are another important aspect of the complex concept of timbre of musical instruments.

The spectral envelope of a vowel sound can also be obtained

from our first neural representation. Because the pitch of the male voice (about 100 Hz) is substantially lower than the musical pitches represented in the preceding figures, our filter bank model cannot resolve the individual harmonics of the voice. Hence it smears adjacent harmonics together, to directly produce an approximation to the spectral envelope. An example of this is shown in Figure 8, which shows the running spectrum for the first half of the word "beer." Each line represents the filter outputs during a particular 1.6-millisecond sampling interval. The first eight lines in the figures thus represent thirteen milliseconds of "silence," that is, tape recorder noise, etc. The rise or cliff seen in the ninth and tenth lines is the release of the burst. The upper two-thirds of the figure corresponds to the steady-state vowel, and hence is an approximation to the spectral envelope for /i/. The two strong formants of 400 Hz and 1700 Hz are plainly visible.

An example of a "running spectrum" for conversational speech is shown in Figure 9, which has been plotted in perspective to emphasize the basically three-dimensional structure of all of this data. The plot corresponds to the italicized portion of the sentence "*The watchdog* gave a warning growl." The time markers on the right represent blocks of sixteen spectra, hence give a time scale of roughly twenty-five milliseconds per division. The changing spectral envelope of the sonorants corresponding to the vowel in "The," and the *wa* in "watch" (blocks 18 through 29) arises from a change in tuning of the vocal tract by motion of the jaw and tongue while articulating the sentence.

What we have shown thus far is that as a first approximation, the first neural representation can be characterized as a filter-bank analysis of the incoming music or speech, with roughly one-third octave frequency resolution, and temporal resolution at high frequencies of the order of a few milliseconds. It is clearly appropriate for displaying many important aspects of musical timbre, such as the attack time, the latency of attack of higher harmonics, and the spectral envelope. The representation is basically three-dimensional in character: amplitude as a function of frequency and time. Of significance is the fact that the particular choice of filter properties in the ear takes a two-dimensional representation of the signal in the concert hall (sound pressure versus time) and generates a three-dimensional representation in which most of the aspects of music that musicians would call temporal, such as rhythm, attack, cadence, etc., are in one dimension, and most aspects that are considered spectral, such as tone color, are in the other dimension.

This section has reviewed work we have done in analyzing music and speech in a way roughly analogous to the auditory system. We now turn to a much more speculative discussion: an attempt to draw together several diverse papers on auditory and visual psychophysics and to suggest an interesting and possibly important simplified representation of speech and music which closely parallels color vision.

A Possible Second Neural Representation

Thus far we have achieved a clever repackaging of the incoming sound wave which has preserved spectral envelopes and temporal attack profiles. Three papers (see Yilmaz [1967 & 1968]; Richards [1979]; & Pols [1977], also working with M.F. Taylor) suggest that a substantial simplification in this representation may be possible. The central question is: how much spectral detail do we need to comprehend and appreciate music and speech? Do we need 1500 points on our spectra, as suggested by the 1500 inner hair cells in the cochlea, or will 256 points be sufficient, as suggested by the loudspeaker spectral data of Toole (see pp. 49-66)? Or will far fewer numbers suffice?

Yilmaz (1967 & 1968)

In the visual system, we do not measure the color spectrum in great detail. Instead we measure only three quantities: the amount of redness, blueness, and greenness reflected from an object. These three numbers are then converted by trivial algebra to brightness, hue, and saturation, which leads to the familiar color triangle of color perception. Huseyin Yilmaz claimed that speech perception should have a similar organization. There should be a three-dimensional vowel space, with loudness, "hue," and "saturation"; three "primary" vowels from which all vowels can be constructed; and complementary vowels just as we have complementary colors. Vowels should be displayed in two dimensions as a "vowel circle" similar to the color triangle, by deflecting the X and Y axes of an oscilloscope with sine-weighted and cosine-weighted averages of the spectrum:

$$X = \sum_{n=1}^{N} S(n) \sin 2\pi \frac{n}{N} \qquad (1)$$

$$Y = \sum_{n=1}^{N} S(n) \cos 2\pi \frac{n}{N} \qquad (2)$$

where S(n) is the log spectral magnitude from the nth critical-band filter, and N is the total number of filters.

By representing speech in this way, Yilmaz is implicitly stating that vowels and vowel-like sounds can be represented by substantially fewer parameters than were used to represent the original spectrum. He is suggesting that the brain does not have to pay attention to all the nuances of spectral shape shown in Figure 9, for example, in order to understand the speech. All we need in any given instant are two numbers, specifically the weighted averages calculated in Equations 1 and 2 above.

To illustrate, we have analyzed the speech spectra of Figure 9 using Yilmaz's "color" method. Figure 10 shows the effect of transforming the spectral data in Figure 9 in accordance with Equations 1 and 2, and forming the Yilmaz "vowel circle" plot, in direct analogy to the color triangle plot. To further elucidate this diagram, we have labeled around the circle the locations corresponding to various pure-tone inputs ("saturated sounds"), analogous to the location of the saturated colors on the color triangle.

The succession of spectra in Figure 9 are converted to a succession of points in the new space, that is, a *trajectory*. To facilitate intercomparison, the numbers shown beside the trajectory are the same time markers shown in Figure 9. The trajectory starts at the neutral point ("white"), 14, 15, then a brief frication for the /ð/, 16, a transition to the /ə/ at 17, flowing directly into /w/ at 19 and 20, and /ɔ/ at 24, 25, and 26. The silent period preceeding the /tʃ/ forces a long transition back to the neutral point, 28, 29, 30, then an attack of the /tʃ/ 31, followed by the steady state /tʃ/, 32 and 33. Approximate positions of other vowels are also indicated in the plot. Similar two-dimensional plots can be found in Schouten and Pols (1979a & 1979b) & Cote (1981).

The Yilmaz plots are obviously closely related to the formant plots discussed in Wayne Slawson's paper (see his Fig. 2, p. 70, for example). The "Lax" position is roughly in the center of the vowels in our figure, and the vowels have a similar topological relationship, with the exception of a left-right inversion. Slawson's discussion of inversion and complementarity are entirely consistent with the Yilmaz color analogy. Hence it would appear

that the concepts conceived by Slawson to aid in the composition of music have an important counterpart in a perceptually-related neural representation, as proposed by Yilmaz.

Richards (1979)

Richards and his students have been studying various aspects of visual perception, such as texture, visual orientation, flicker, etc., and have concluded that all of the systems resemble the color-perception system in that at some level in the neural processing chain the information appears to be represented by a limited number of perceptual channels, usually three or four. He shows, for example, that a striped "rug" made up of three shades of gray is almost undistinguishable from one made from sixty-four shades of gray. Further, a textured "rug" made up from three spacial frequencies is difficult to distinguish from one made up of random noise of many spacial frequencies. Richards postulates that all sensory processes, including the various auditory processes, should in some sense resemble the color system in that they should be representable in terms of a very limited number of "primaries" or primatives. From this perspective, Yilmaz's ideas on vowel perception become a special case of Richards's broader theories of "generalized colorimetry."

Taylor and Pols (1978)

The work of Pols (1977) and the unpublished work of Taylor and Pols (1978) lend considerable experimental support to the theories of Yilmaz and Richards. Pols used a 17-channel filter bank modelled after the auditory system to analyze conversational speech. He generated log-magnitude spectra for a minute of speech (one spectrum every ten milliseconds, or 60,000 spectra) for each of ten speakers (two languages, English and Dutch). For data reduction, he did not simply assume sine and cosine basis vectors as did Yilmaz (see Equations 1 and 2). Instead, he applied principal components analysis[2] to his 17-point spectra to derive a new picture of the data in which the maximum amount of the variance has been forced into the first component, the maximum remaining variance into the second component, etc. The transform was surprisingly effective in forcing the information into a few components. In Pols's original running spectrum data, the variance is more or less uniformly distributed throughout the filter channels, with no one channel accounting for more than 11% of the variance. After the principal components transformation,

50% of the variance is in the first component, 29% in the second, 8% in the third, and the other 13% in decreasing amounts of the remaining fourteen components. Also, the basis vectors derived from the analysis are quite speaker independent and language independent for English and Dutch (see Taylor 1978).

The experiments of Pols and Taylor lend considerable credibility to the "vowel circle" theory of Yilmaz. As noted above, roughly 90% of the variance can be accounted for by the first three components in the new space. Careful examination of their data indicates that the first component in their new representation is the intensity or loudness of the speech, the second corresponds roughly to Yilmaz's sine weighting, and the third to his cosine weighting. Thus this analysis provides a solid experimental basis for construction of a vowel circle which is topologically very similar to that proposed by Yilmaz.

Following the lead of Pols and Taylor, we have applied principal components analysis to our Watchdog sentence, Figure 9, in an attempt to generate a second neural representation which is more compact than the one discussed in the first section. The basis vectors generated by this analysis turned out to be very similar in form to those obtained in other studies on completely different material (see Pols 1977 & Zahorian 1978). When our basis vectors are used to transform the data of Figure 9, we obtain the new representation shown in Figure 11. The time dimension in this plot remains unchanged, but the horizontal axis is now just "component number," because of the strange coordinate rotation introduced by the basis vectors. Clearly something quite dramatic has occurred. The spectral information now appears to be heavily concentrated in the first few components of this proposed second neural representation, and the "higher" components seem to be only very weakly correlated with the speech.

Conclusion

Music is initally represented by a score. An orchestra or a computer with an appropriate sound system then converts the score to a new representation, that of sound pressure in air as a function of time. In our heads we must form neural representations of this sound. Two possible representations begin to emerge from studies of auditory physiology and psychophysics and color perception. The evidence suggests that an initial representation is derived from spectral analysis of the sound, describable either

in terms of a set of critical-band filters and detectors, or as a short-term constant Q Fourier Transform. These spectra may be then subjected to a second linear transform, such as principal components analysis, to produce a neural representation containing only a minimum number of perceptually important channels or dimensions. All of the aspects of music we have been discussing — timbre, rhythm, cadence, melody, etc. — must somehow be encoded in each of these representations.

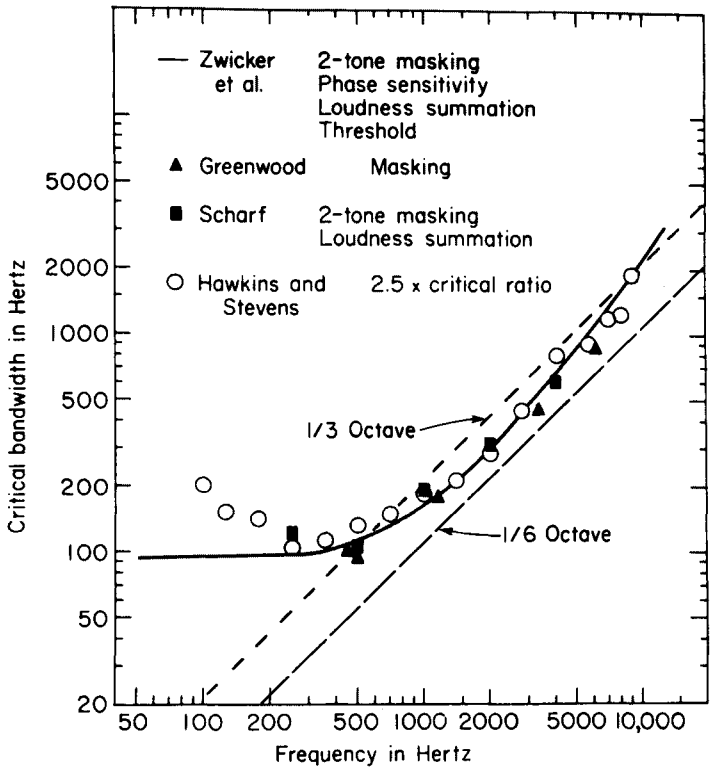

Figure 1
Critical bandwidth as a function of frequency (from Tobias 1970)

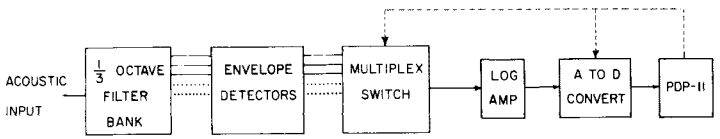

Figure 2
Block diagram of the system for analyzing music and speech

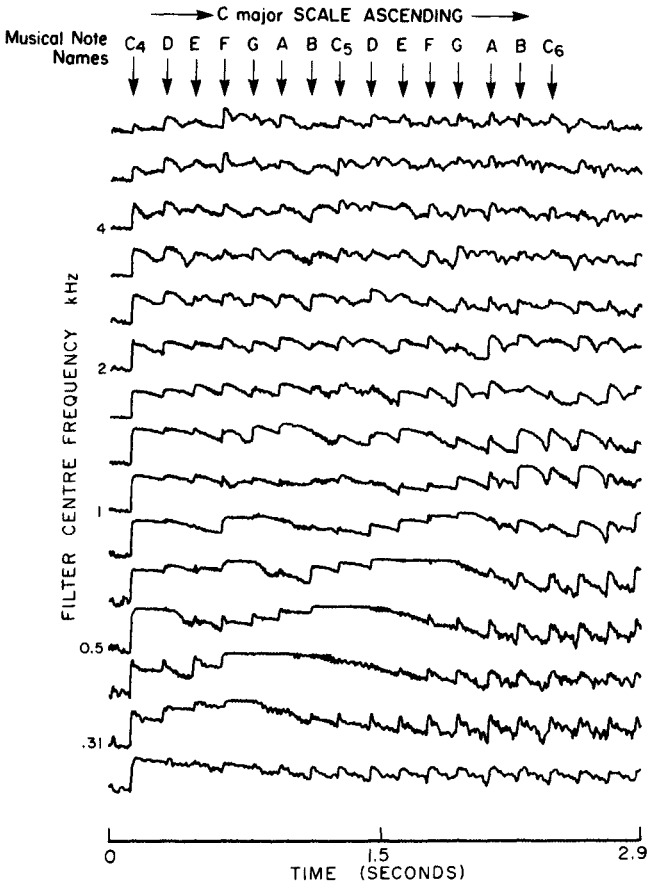

Figure 3

Log magnitude outputs of fifteen filter-detector channels for three seconds of piano; bottom trace, 250 Hz center frequency, top trace, 6.3 kHz center

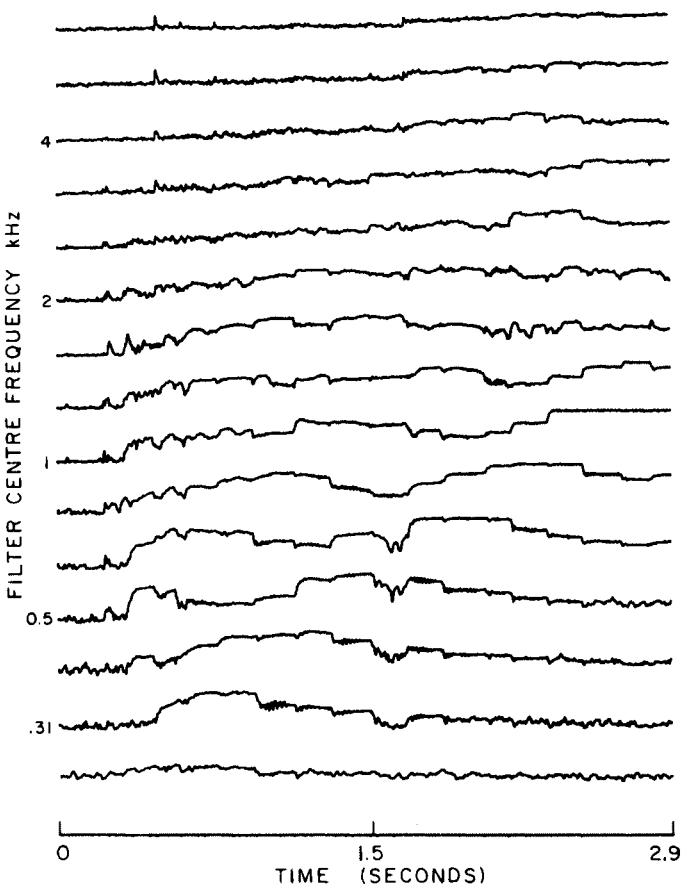

Figure 4
Similar plot to Figure 3, except for an alto flute

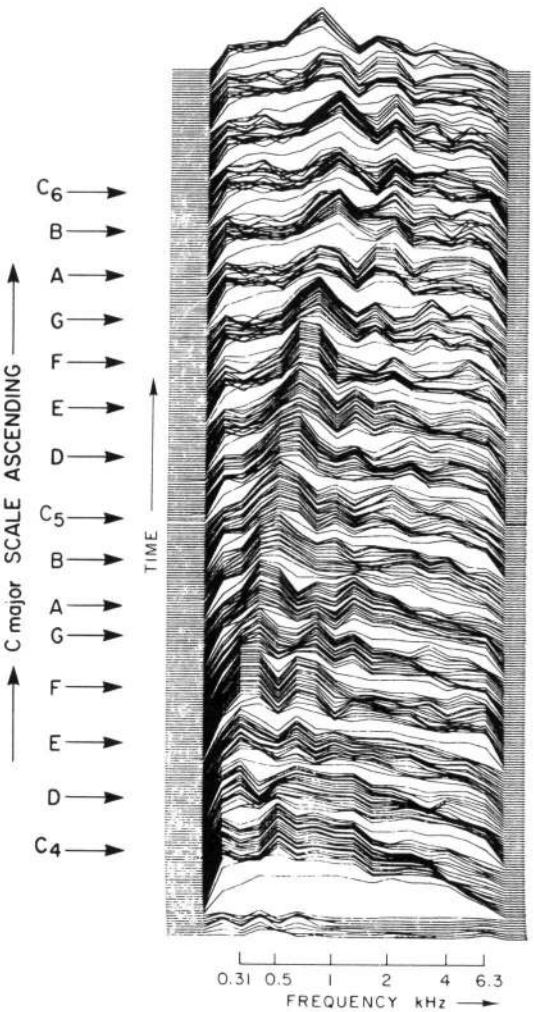

Figure 5

The same piano scale as in Figure 3, except replotted as "running spectra" to emphasize the spectral detail

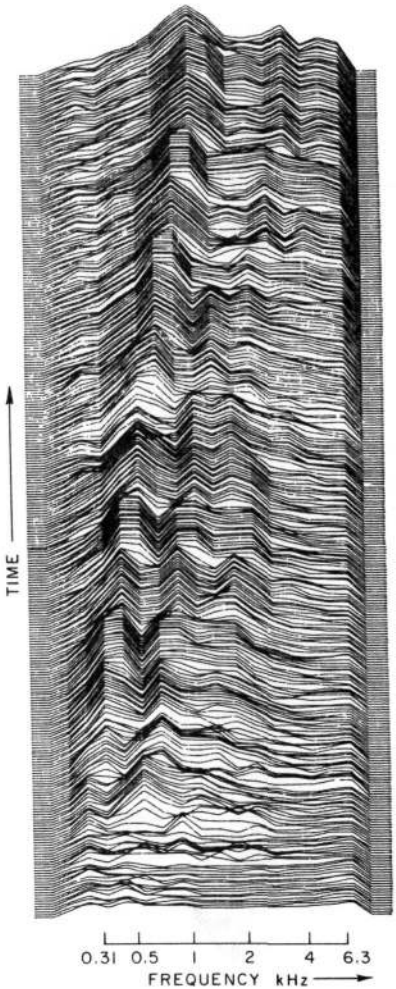

Figure 6
Running spectrum for the alto flute passage

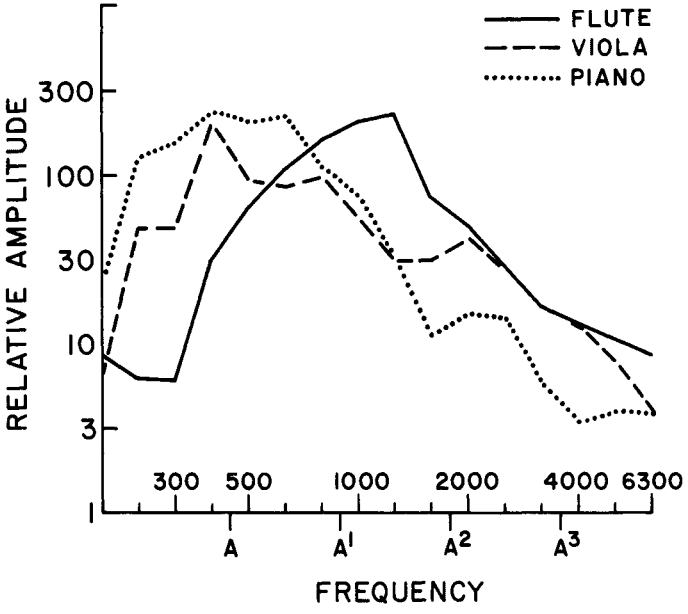

Figure 7
Spectral envelopes for the alto flute, piano, and viola

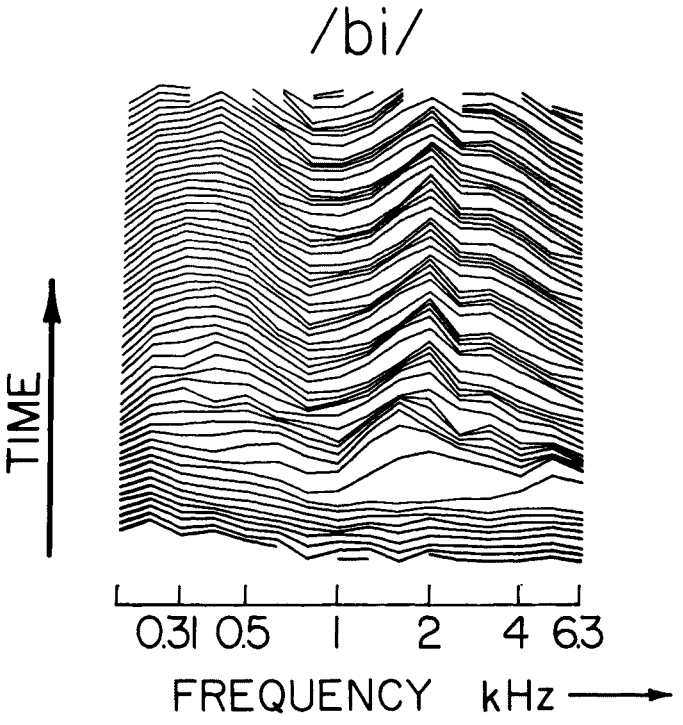

Figure 8
Running spectrum for the first half of the word "beer"

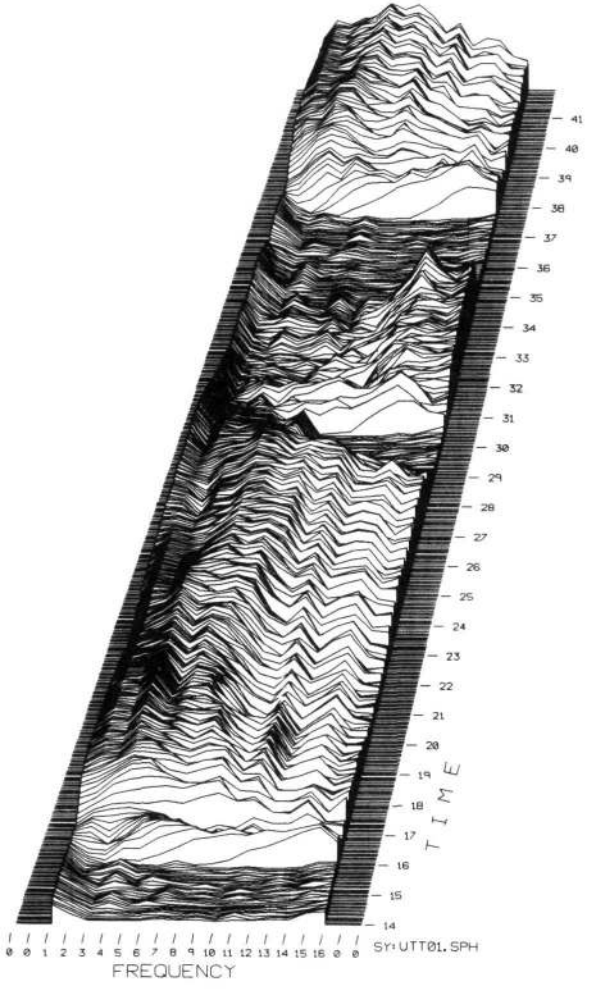

Figure 9

Running spectrum for the italicized portion of the sentence The watchdog *gave a warning growl"*

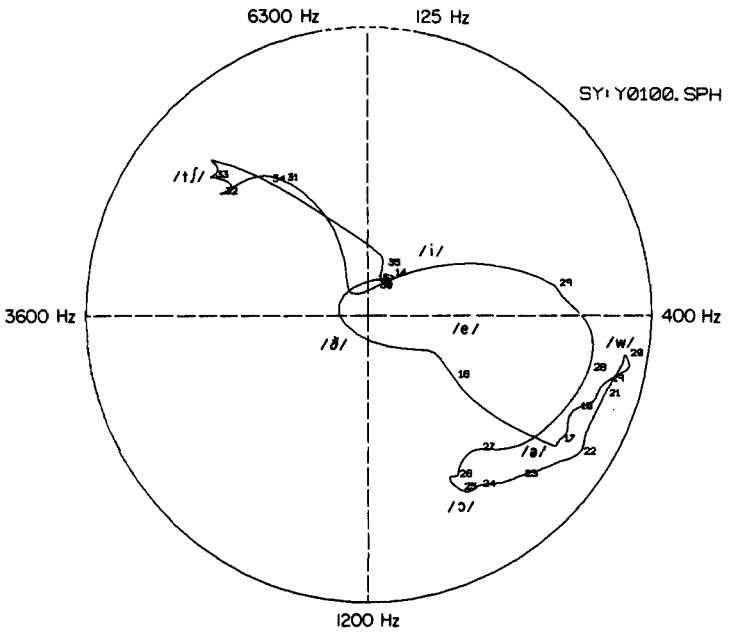

Figure 10
Yilmaz "vowel circle" for "The watch. . ."

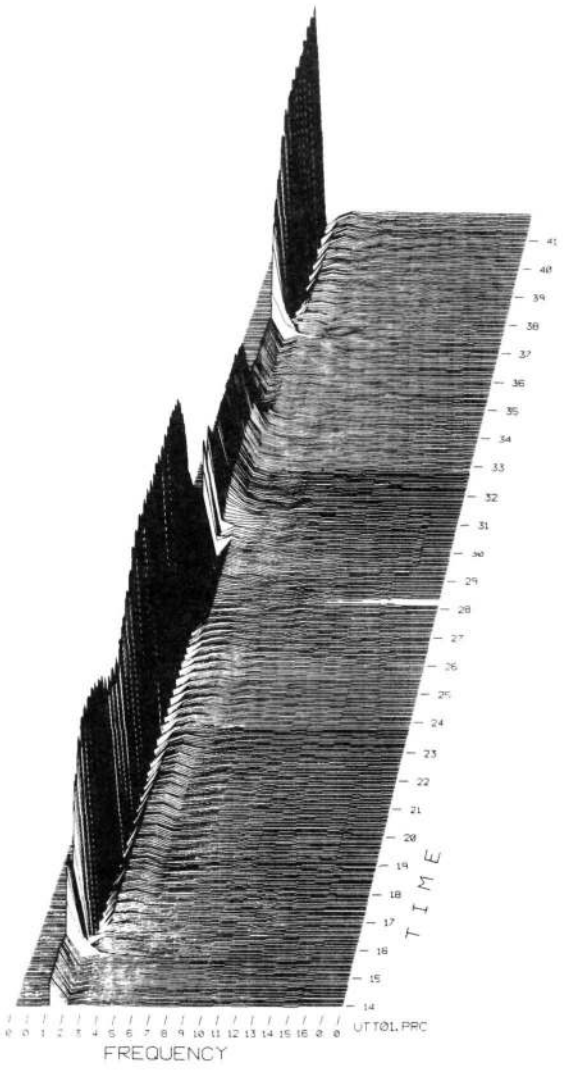

Figure 11

The new representation of "The watchdog. . .", derived from the plot in *Figure 9 by principal components analysis*

NOTES

1. In engineering terms, it is the short-term Fourier Transform (see Flanagan 1972), with the added constraint of constant Q (see Youngberg & Boll 1978).

2. Principal components analysis is similar to factor analysis except that factor analysis uses sums and products of the *input* variables to account for the variance in the output data, whereas principal components analysis uses linear combinations of the *output* variables.

REFERENCES

COTE, A.J.

1981: "A Relative Portrait of Some Vowel Sounds." (Private communication from the author).

DOCKENDORFF, D.D.

1978: "Application of a Computer-Controlled Model of the Ear to Multiband Amplitude Compression." M. Sc. thesis, Queen's University.

EVANS, E.F. and WILSON, J.P.

1973: "Frequency Selectivity in the Cochlea," in Moller, A.R., ed., *Basic Mechanisms in Hearing.* New York: Academic Press, 519-54.

FLANAGAN, J.L.

1972: *Speech Analysis, Synthesis and Perception.* New York: Springer.

JOHNSTONE, B.M. and BOYLE, A.J.F.

1967: "Basilar Membrane Vibration Examined with the Mössbauer Effect," *Science*, CLVIII/3799, 389-90.

KIANG, N.Y.S., WATANABE, T.E.C., and CLARK, L.F.

1965: *Discharge Patterns of Single Nerve Fibers in a Cat's Auditory Nerve.* Cambridge, Mass.: The MIT Press.

KIANG, N.Y.S. and MOXON, E.C.

1974: "Tails of Tuning Curves of Auditory-Nerve Fibers," *Journal of the Acoustical Society of America*, LV/3, 620-30.

LUCE, D. and CLARKE, M.

1967: "Physical Correlates of Brass-Instrument Tones," *Journal of the Acoustical Society of America*, XLII/6, 1232-43.

PATTERSON, R.D.

1976: "Auditory Filter Shapes Derived with Noise Stimuli," *Journal of the Acoustical Society of America*, LIX/3, 640-54.

POLS, L.C.W.

1977: *Speech Analysis and Identification of Dutch Vowels in Monosyllabic Words.* Soesterberg, The Netherlands: Institute for Perception, TNO.

RAYMENT, S.G.
1977: "Phoneme Recognition based on Feature Extraction from a Model of the Auditory System." M. Sc. thesis, Queen's University.
RHODE, W.S.
1971: "Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys using the Mössbauer Technique," *Journal of the Acoustical Society of America*, XLIX/4, 1218-31.
RICHARDS, W.
1979: "Quantifying Sensory Channels: Generalizing Colorimetry to Orientation and Texture, Touch and Tones," *Sensory Processes*, III/3, 207-29.
SCHOUTEN, M.E.H. and POLS, L.C.W.
1979a: "Vowel Segments in Consonant Contexts: A Spectral Study of Coarticulation—Part I," *Journal of Phonetics*, VII/1, 1-23.
1979b: "CV- and VC-transitions: A Spectral Study of Coarticulation—Part II," *Journal of Phonetics*, VII/3, 205-24.
SEARLE, C.L., JACOBSON, J.Z., and RAYMENT, S.G.
1979: "Stop Consonant Discrimination Based on Human Audition," *Journal of the Acoustical Society of America*, LXV/3, 799-809.
SEARLE, C.L.
1979: "Analysis of Music from an Auditory Perspective," *The Humanities Association Review*, XXX/1-2, 93-103.
TAYLOR, M.M. and POLS, L.C.W.
1978: Unpublished data (personal communication).
TOBIAS, J.V.
1970: *Foundations of Modern Auditory Theory*, Vol. I. New York: Academic Press.
VIEMEISTER, N.F.
1979: "Temporal Modulation Transfer Functions Based Upon Modulation Thresholds," *Journal of the Acoustical Society of America*, LXVI/5, 1364-80.
YILMAZ, H.
1967: "A Theory of Speech Perception, I," *Bulletin of Mathematical Biophysics*, XXIX/4, 793-824.
1968: "A Theory of Speech Perception, II," *Bulletin of Mathematical Biophysics*, XXX/3, 455-79.
YOUNGBERG, J.E. and BOLL, S.F.
1978: "Constant-Q Signal Analysis and Synthesis," *IEEE International Conference on Acoustics, Speech & Signal Processing*. New York: Institute of Electrical and Electronics Engineers, Inc., 375-78.
ZAHORIAN, S.A.
1978: "Principal Components Analysis for Low Redundancy Encoding of Speech Spectra." TR-78-10, Electrical Engineering Department, Syracuse University.