CENTER FOR COMPUTER RESEARCH IN MUSIC AND ACOUSTICS
MARCH 1990

Department of Music
Report No. STAN-M-64

Music from Machines:
Perceptual Fusion & Auditory Perspective
- for Ligeti -

John M. Chowning

CCRMA
DEPARTMENT OF MUSIC
Stanford University
Stanford, California 94305

Music from Machines:
Perceptual Fusion & Auditory Perspective
-for Ligeti-

John M. Chowning
The Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University, Stanford, California

Ligeti spent six months at Stanford from January to June, 1972. He came as a guest composer having no knowledge of the work in computer music that we had been pursuing over the previous eight years. At that time we were but a small part of the Artificial Intelligence Laboratory, with no support other than limited access to the computer, requiring that we work at nights and on weekends. Ligeti's first visit to the lab led to far ranging discussion of the capabilities offered by the computer in projecting sound in space, transformations of timbre, the fine control of pitch and time, and precisely constructed tuning systems. On his return to Europe he spoke to his colleagues of the work he had seen in computer music in California. Ligeti became an advocate for the medium. His understanding and vision were great indeed. They still are.

Loudspeakers controlled by computers form the most general sound producing medium that exists, but there are nonetheless enormous difficulties that must be overcome for the medium to become musically useful. Music does not come easily from machines. This is true whether the machine is a musical instrument of the traditional sort or a computer programmed to produce musical sound. In the case of musical instruments, years of training are required of the performer and instrument builder and in the case of the computer, substantial knowledge about digital processing, acoustics, and psychoacoustics is required of the composer/musician. It is with this newest instrument, the computer, that we confront new problems whose solutions have led to insights that transcend the medium, increase our knowledge, and enrich our experience in the grandest sense.*

There are two issues that are addressed in this paper: 1) the auditory system's sensitivity to minute fluctuations, a significant characteristic that is little known but which has important implications, and 2) auditory perspective, with some insight regarding the multi-dimensionality of perceived loudness.

Much of what is discussed surrounds phenomena that are well-known to musicians and scientists, such as periodic waves, vibrato, loudness, etc. In the course of this discussion I question the common understanding of some of these phenomena. What is of interest pertains to subtleties of perception that require a more comprehensive understanding of these phenomena. For example, periodic is a

* *It is perhaps important to note that the precision required in constructing the sounds that led to the topics of this paper was not available before computers were first programmed to produce sound by Max V. Mathews at the Bell Telephone Laboratories in 1957.*

term frequently used by scientists/engineers to describe a large class of natural tones whose component parts fall in the harmonic series, but in fact they are not strictly speaking periodic, "the ear" knows this to be so and that is the point!

Perceptual Fusion and Quasi-Periodicity

The Limits of Perfection- We may have thought that one of the purposes of both the performer and the instrument builder was to reach ever greater degrees of perfection, that the finest instrument and the finest performer could be superseded by some even finer yet. The great violins of the 17th and 18th centuries might be replaced by new superior instruments, having strings of ever greater constancy in mass, played by performers whose bow arms, through perhaps better training, could maintain ever more even pressure and velocity while in contact with the string. Curiously, there are degrees of perfection in acoustic signals beyond which the auditory system responds in quite surprising ways; it can become confused in regard to what instrument or source might have produced the sound, or in regard to assignment of the constituent parts or partials of a sound to their proper source in the case of simultaneously occurring sounds. Faced with such perfection the auditory/cognitive system can exercise a kind of *aesthetic rejection*. It is curious, and perhaps fortunate, that such degrees of perfection are well-beyond the capabilities of both acoustic instrument building and human performance, now and probably forever. This order of perfection exists only in sound generated electronically, especially by means of digital devices (computers, synthesizers, etc.).

Periodicity and Quasi-periodicity- A perfectly regular recurrent pattern of pressure change in time is periodic. A recurrent pattern of pressure change in time that has *small variations* in period and/or pressure is quasi-periodic, as compared in Fig. 1. Acoustic waves that appear to the auditory system to be periodic, having

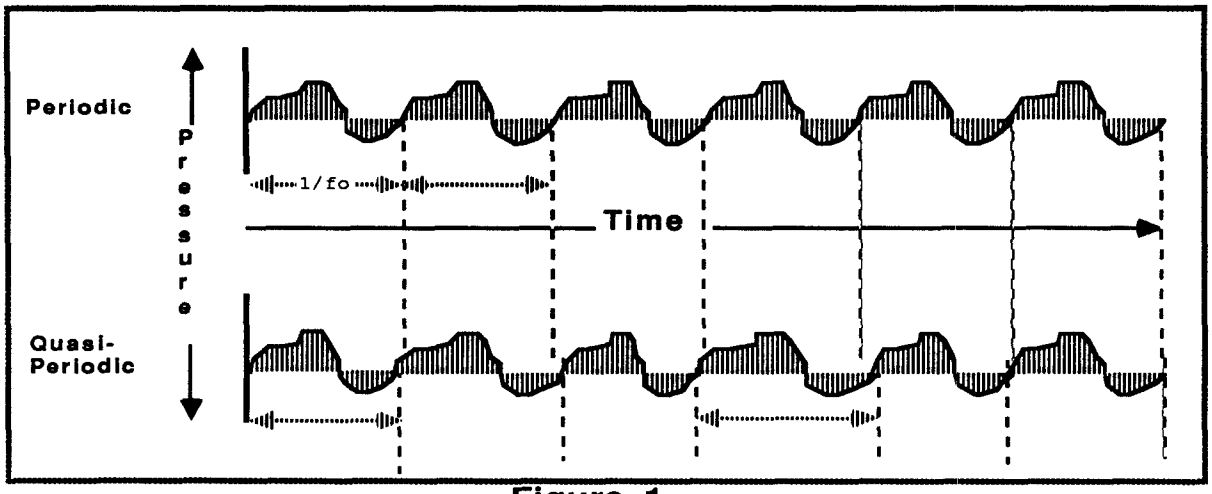

Figure 1.

undetectable variation, are little known in nature but can be produced by loudspeakers whose signals have been generated electronically. Quasi-periodic waves however are typical in nature. The auditory system is extraordinarily sensitive to quasi-periodicity as it is able to detect a variation in period of a small fraction of a percent. These small continuous variations are imposed by nature in the form of random pitch and in many cases an additional variation is consciously imposed by the performer in the form of vibrato and/or tremolo.

Random pitch variation occurs even when there is no vibrato imposed by the performer. This variation is caused by small imperfections in both the performer and the instrument. In the case of a singer there are small variances in pressure of the air from the lungs as it is forced through the vocal folds, small changes in muscular tension of the vocal folds themselves, non-linearities resulting from turbulence at the vocal folds coupling with the acoustic wave in the vocal tract, etc. The set of harmonics composing the waveform are modulated by a common random variation pattern, also referred to as 'jitter'[1].

Vibrato is a more or less regular variation in pitch that results from a small modulation of string length, length of the air column, or tension of a vibrating reed or lips. Singers produce vibrato by a variation in tension of the vocal folds. *Tremolo* is a similar variation, but one of loudness, resulting from a variation of bow pressure and/or velocity in the case of strings and air pressure in the case of winds. Singers produce a tremolo by varying the breath pressure through the vocal folds. (Organ pipes and recorders are constrained to tremolo modulation alone because of their sound-producing mechanisms, whereas most other instruments, including the voice, are capable of both.) Both kinds of modulation, but especially vibrato, serve a variety of musical, acoustic, and perceptual functions.

<u>*Source Identification*</u>- It was hardly ten years ago that we performed experiments that for the first time revealed the special significance of such small amounts of variation in pitch[2]. The experiments were based on modeling the voice of a singer (the only musician who is the instrument, its maker, and its performer). A soprano tone lasting 15 seconds was synthesized in three stages, as seen in Fig. 2a:

1) A sinusoid at the frequency of the fundamental, $f_0 = 400Hz$,
2) Harmonics are added appropriate to a sung vowel, $2f_0 \cdots nf_0$,
3) A mixture of random pitch variation and vibrato is added to the total signal.

Stages 2) and 3) evolve continuously from the previous stage. At stage 2, all of the spectral information is present that is required for the singing voice. However, not only is the character of 'voice' unidentifiable during stage 2, but the added harmonics do not even 'cohere' with the fundamental as an entity. Not until the

small amount of random deviation and vibrato are added at stage 3 do the harmonics fuse, becoming a unitary percept and identifiable as a voice. Without the variation in pitch, the sound of the simulated singer (whose control over especially pitch has reached perfection) does not have a source 'signature' or contain information that is essential to her identification as a source.

Perceptual fusion is dependent upon a wave or signal being in a condition of quasi-periodicity where component partials through common motion or variation in the pitch space, define themselves as "belonging together"[3]. The random frequency (and/or amplitude) variation seems to be present in all sources, while the particular pattern of variation differs according to the source class.

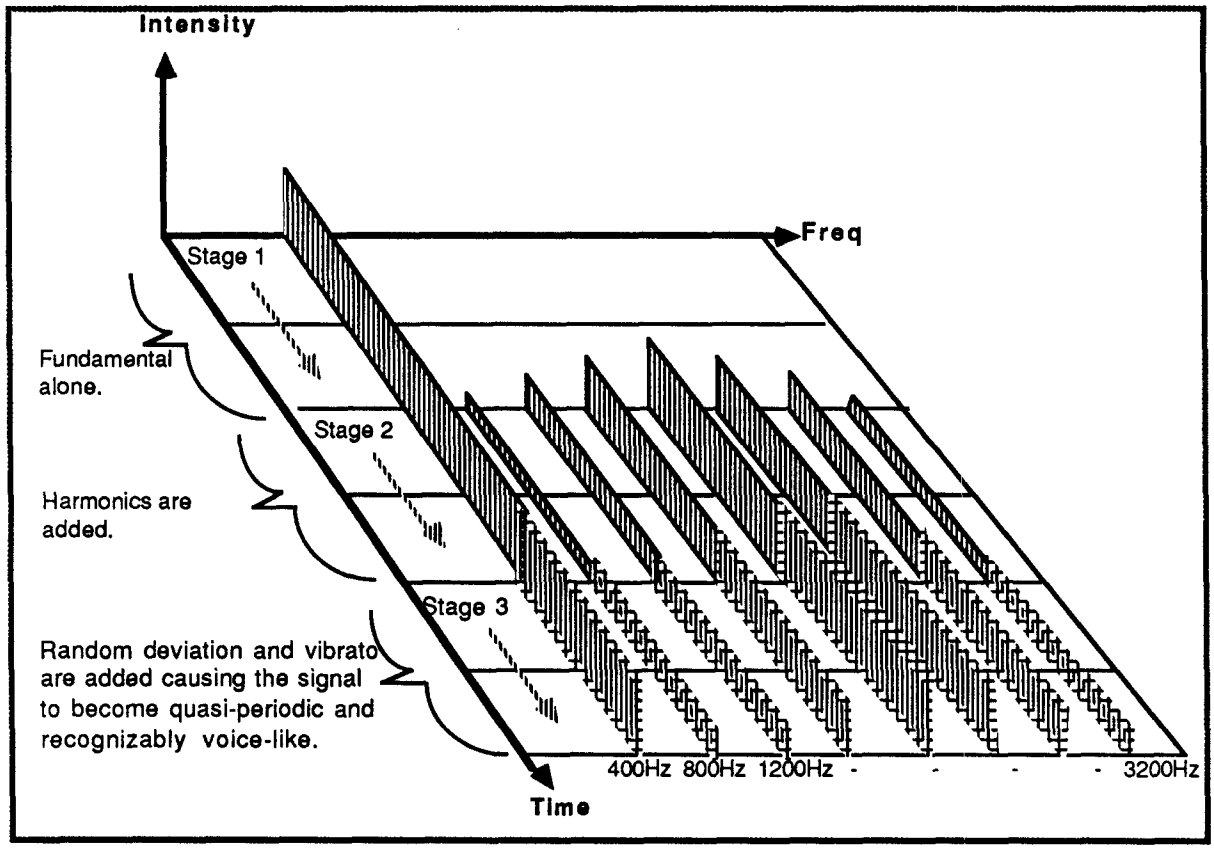

Figure 2a.

Source Segregation- The fusion of the constituent partials of a sound is a requirement for the auditory system to segregate sources or perceive sources as being separate from one another. If we were to listen to the experiment shown in Fig. 2a with the addition of two more sinusoids at 500Hz and 600Hz followed by their associated harmonics we would expect to hear a purely tuned triad having pitches at 400Hz, 500Hz, and 600Hz. At stage 2, the triad is not easily heard since the partials of all three groups form a harmonic series over a missing fundamental

could neither identify its nature (*source identification*), nor hear parts (*source segregation*), nor recognize that there are more than one source per part (*chorus effect*). Thus, the auditory system is utterly dependent upon acoustic imperfections.

More About Vibrato- In addition to being an expressive device, vibrato serves a variety of acoustic, perceptual, and musical functions. Vibrato can complement the natural random pitch variation of critical importance to source identification and source segregation. The timbral richness (identity) of a source is much enhanced by even a small amount of vibrato as partials oscillate under resonant envelopes causing a complex asynchronous amplitude modulation. In solo/ensemble contexts instruments having a limited dynamic range such as the violin use vibrato to help segregate their sound from that of the ensemble which would otherwise mask the solo instrument. This is analogous to the visual system's ability to segregate an object hidden in a background only when the object moves. And finally, vibrato frequency and depth are used expressively to support pitch and dynamics in the articulation of a musical line.

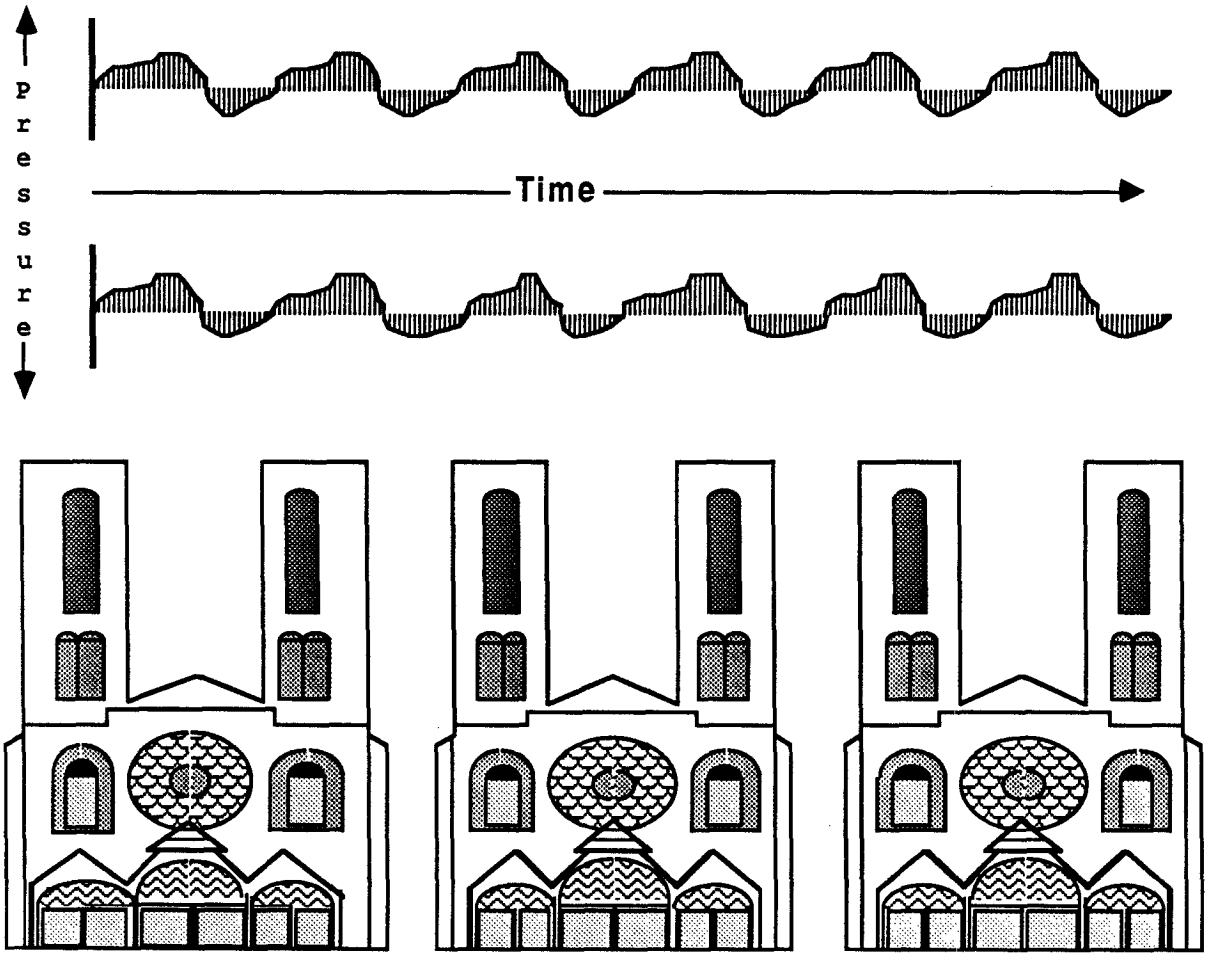

Figure 3.

Periodicity and Symmetry- The auditory and visual systems seem to treat periodicity and symmetry in a similar manner, but differ in degree. While the eye does not detect immediately the quasi-periodicity in Fig. 3 without the aid of the lines indicating the periods, nor the one image of the three that is least symmetrical about the center axis, the ear can readily detect a fraction of a percent of deviation from periodicity, as noted above.

Both systems seem to become inattentive or 'turn off' when periodicity/symmetry is perceived over even a rather short time, failing to extract critical information (especially in the case of the auditory system).

The making of music using machines demands that attention be given to the requirements of the perceptual system. Unlike acoustic instruments, electronic 'instruments' do not have the inherent imperfections upon which the auditory system depends.

Auditory Perspective

The perception of sound in space remains a critical issue in music composed for loudspeakers, whether prerecorded or from real-time digital synthesizers. In the simplest case a listener localizes the emanating sound from points defined by the position of the loudspeakers. In all other acoustic settings the listener associates a sound source with horizontal and vertical direction and a distance. The auditory system seems to map its perceived information to the higher cognitive levels in ways analogous to the visual system. Acoustic images of great breadth reduce to a point source at great distances, as one would first experience listening to an orchestra at a distance of 20m and then at 300m, equivalent to converging lines and the vanishing point. Sounds lose intensity with distance just as objects diminish in size. Timbral definition diminishes with distance of a sound from a listener just as there is a color gradient over large distance in vision. Therefore perspective is as much a part of the auditory system as it is of the visual system. It is not surprising that the two systems should have evolved in a way that avoids conflict of sensory mode in comprehending the external world since many visually perceived objects can also be sound sources. These sources can be especially important to survival, for example the mother's voice or the growl of a lion at a distance or close at hand, or the approach of a fast moving automobile. While not perceived with great precision, the perceived position of sound in space, auditory perspective, is composed of important acoustic and psychoacoustic dimensions [4].

Loudness

Commonly thought to be the perceptual correlate of physical intensity[5], loudness is a more complicated percept involving more than one dimension. In order to reveal this we can imagine the following experiment:

> A listener faces two singers, one at a distance of 1m and the other at a distance of 50m. The closer singer produces a **pp** tone followed by the distant singer who produces a **ff** tone. Otherwise the tones have the same pitch, the same timbre, and are of the same duration. The listener is asked which of the two tones is the louder (See Fig. 4)? Before speculating about the answer, we should consider the effect of distance on intensity.

Sound emanates from a source as a spherical pressure wave (we are ignoring small variances resulting from the fact that few sources are a point). As the pressure wave travels away from the source the surface area of the wave increases with the square of the distance (as the area of a sphere increases with the square of the radius). The intensity at any point, then, decreases according to the inverse square law: $1/d^2$, as seen in Fig. 5.

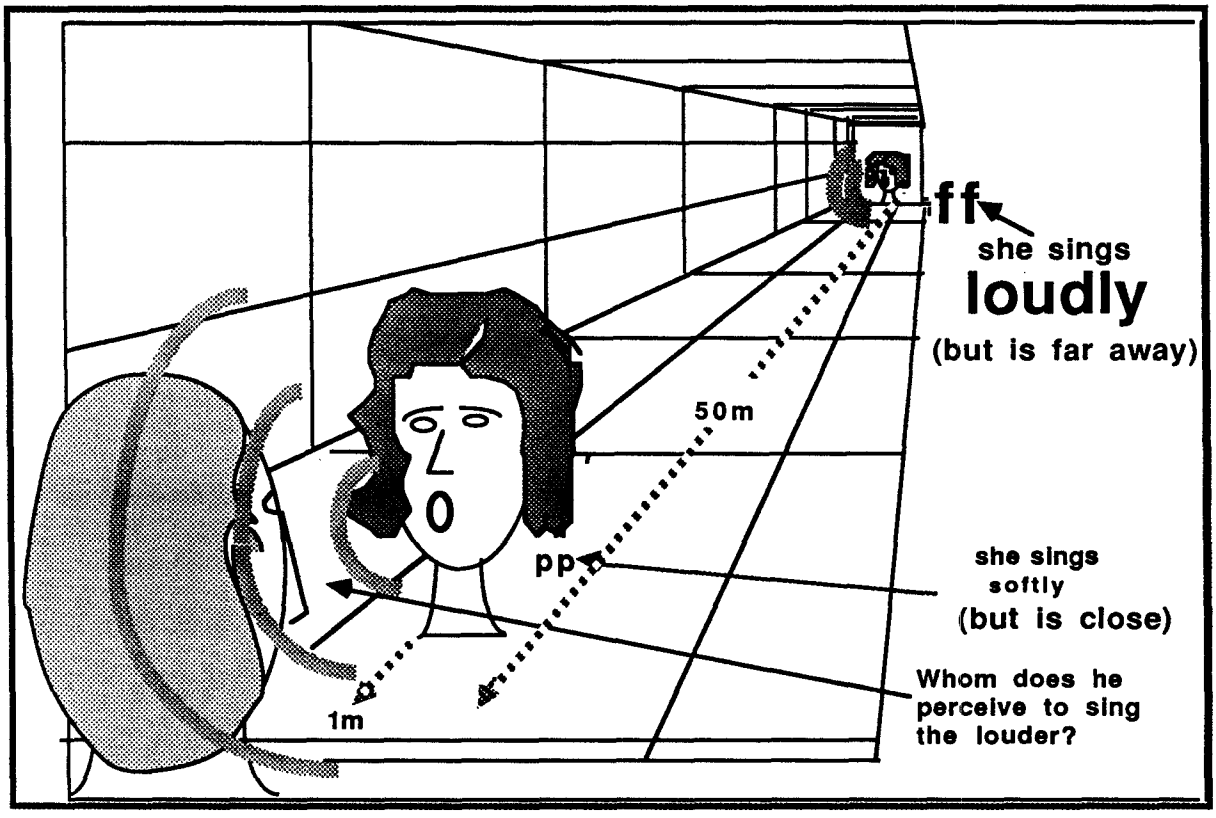

Figure 4.

The distance in the experiment is 50m which will result in a decrease of intensity of $1/50^2$ or 1/2500 the intensity of the same **ff** tone sung at a distance of 1m. The listener, however, is asked to judge the relative loudness where the closer tone is a **pp** rather than **ff**. Let us suppose that the intensity of the **pp** is 1/128 that of the **ff**. The greater of the two intensities then is the closer **pp** and by a large amount. If loudness is indeed the perceptual correlate of intensity then the answer to the question is unambiguous. However, the listener's answer is that the second tone at 50m is the louder even though the intensity of the closer tone is about 20 times greater. How can this be so?

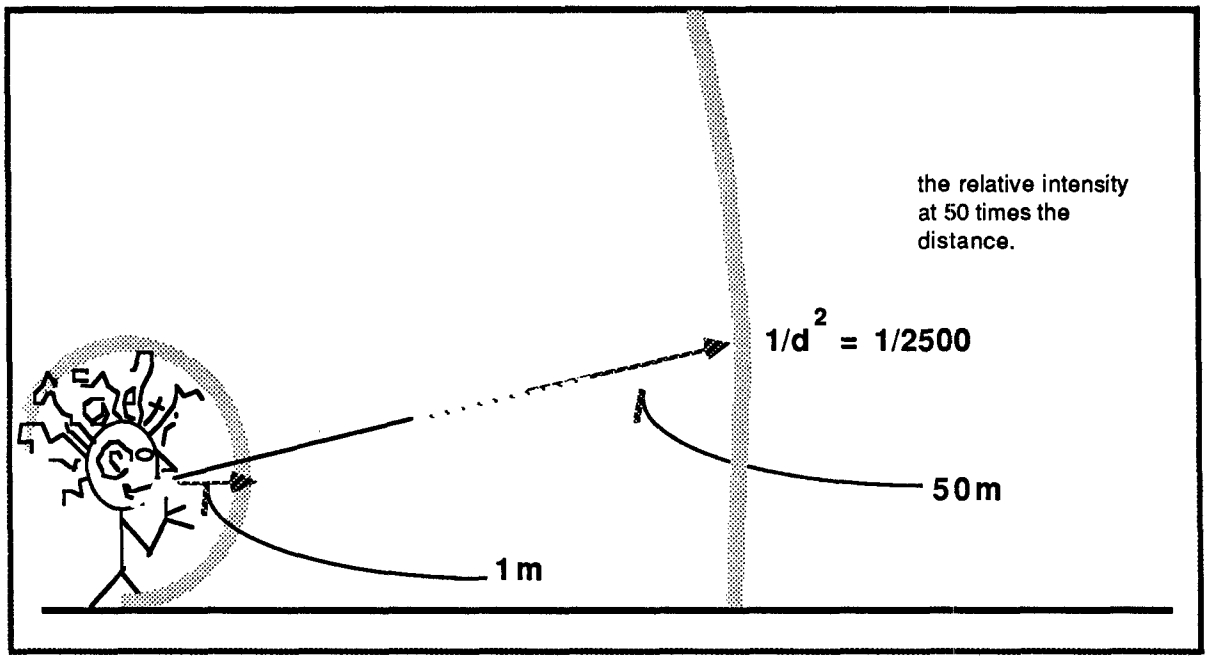

Figure 5.

Spectral Cues- In the definition of the experiment it is stated that the timbre of the two tones is the same. The listener perceives the tones to be of the same timbral class: soprano tones that differ only in dynamic or vocal effort. In natural sources the spectral envelope shape can change significantly as pitch and energy applied to the source changes. In general, the number of partials in a spectrum decreases and the spectral envelope changes shape as pitch increases, that is the centroid of the spectrum shifts toward the fundamental. Similarly, the spectral envelope changes shape favoring the higher component frequencies as musical dynamic or effort increases, the centroid shifts away from the fundamental. Fig. 6 represents a generalization of harmonic component intensity and spectral envelope change as a function of pitch, dynamic (effort), and distance. Because of the high dimensionality involved, a representation is presented where two dimensional spaces (instantaneous spectra) are *nested* in an enclosing three dimensional space.

The position of the origins of the two dimensional spaces are projected onto the 'walls' of the three dimensional space in order to see the relative values. Nesting spaces can allow visualization of dimensions greater in number than three, an otherwise unimaginable complexity*. Here we see the difference in overall intensity and spectral envelope between the tone that is soft and close and the tone that is loud but far.

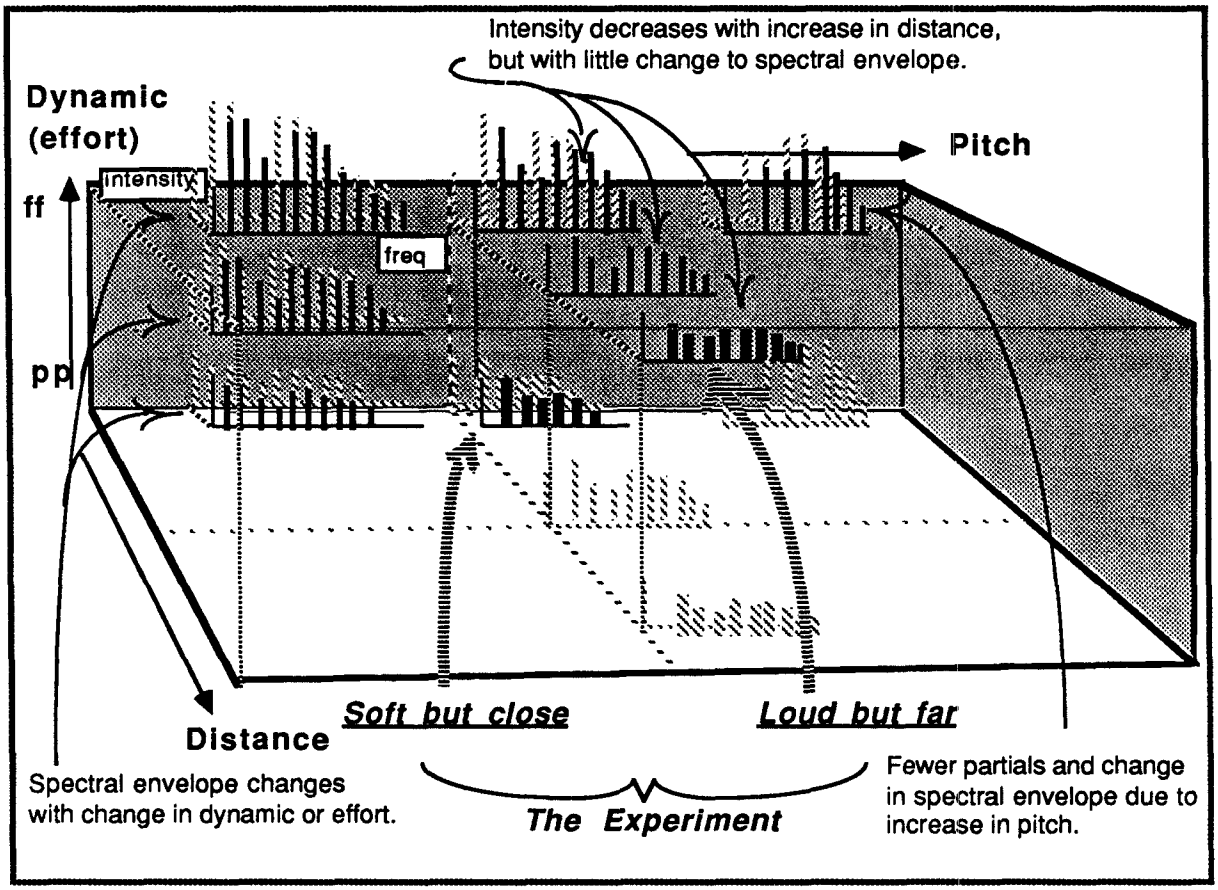

Figure 6.

Now we can understand how the listener in the experiment was able to make a judgment regarding loudness that controverts the dominant effect of intensity on perceived loudness. Knowing the difference in timbral quality between a loudly or softly sung tone, reflecting vocal effort, the listener apparently chose spectral cue over intensity as primary. But what if the two tones in the experiment were produced by loudspeakers instead of singers and there were no spectral difference

**One's ability to assemble the enormous collection of spectra resulting from a single instrument class along the loudness and pitch dimensions and designate it a continuum "soprano" or "violin" is a considerable accomplishment of the perceptual/cognitive systems and even more so were we to consider the additional dimensions of articulation. Timbral continuity, then, is first of all dependent upon perceptual fusion (signal coherence) and source identification, and secondly placing of a tone in the perceptual timbre space.*

as a result of difference in effort? Again, the answer is most probably the distant tone even though its intensity is the lesser of the two - **if** there is reverberation produced as well.

Distance Cue and Reverberation- The direct signal is that part of the spherical wave that arrives uninterrupted, via a line of sight path, from a sound source to the listener's position. Reverberation is a collection of echos, typically tens of thousands, reflecting from the various surfaces within a space arriving indirectly from the source to the listener's position. The intensity of the reverberant energy in relation to the intensity of the direct signal allows the listener to interpret a cue for distance. How does our listener in the experiment use reverberation to determine that the distant tone is the louder?

If, in a typical enclosed space, a source produces a sound at a constant dynamic or effort, but at increasing distances from a stationary listener, approximately the same amount of reverberant energy will arrive at the listener's position while the direct signal will decrease in intensity according to the inverse square law, see Fig.7.

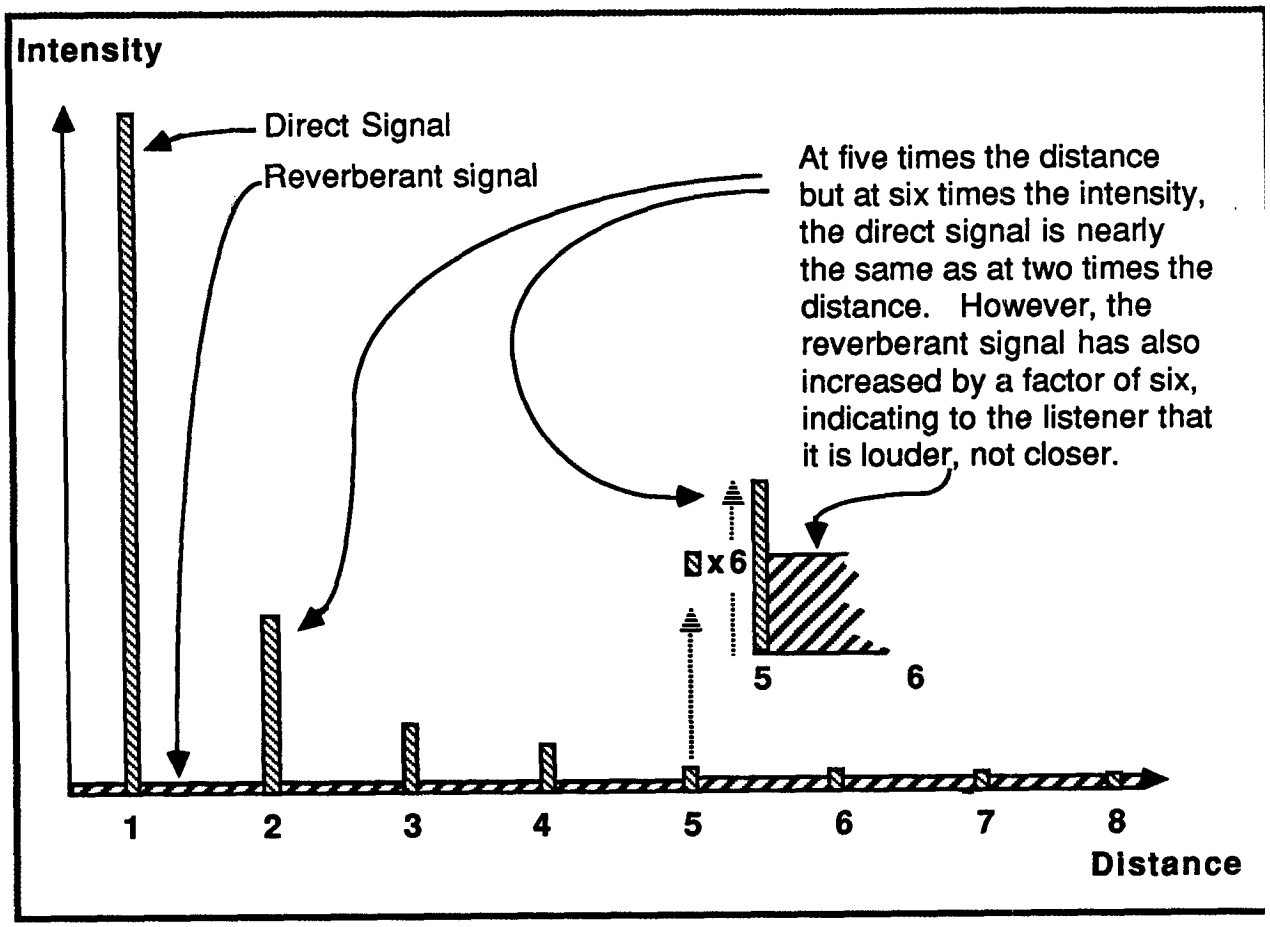

Figure 7.

If at a distance of 5, the sound is produced having six times the intensity then the reverberant signal increases by the same factor. It is for this reason that the listener does not confuse the location of the source with a distance of 2 whose direct signal intensity is approximately the same. The listener in the experiment determined that the reverberant energy associated with the distant loudspeaker was proportionally greater than was the reverberant energy from the softly sounding close loudspeaker leading him to infer that there was greater intensity at the source.

A sound having constant intensity at the source will be perceived by a stationary listener to have constant loudness as its distance increases from 1, 2, 3 .... etc. As seen in Fig. 7, it is the constant intensity of the reverberant energy which provides this effect of **loudness constancy** when there are no spectral cues. A similar phenomenon occurs in the visual system. **Size constancy** depends upon perspective and allows judgments to be made about size that do not necessarily correlate with size of the retinal image. In Fig. 8, we can see what is required to produce constant image size at the retina and constant intensity for the listener. The distant image is the same size as the closest.

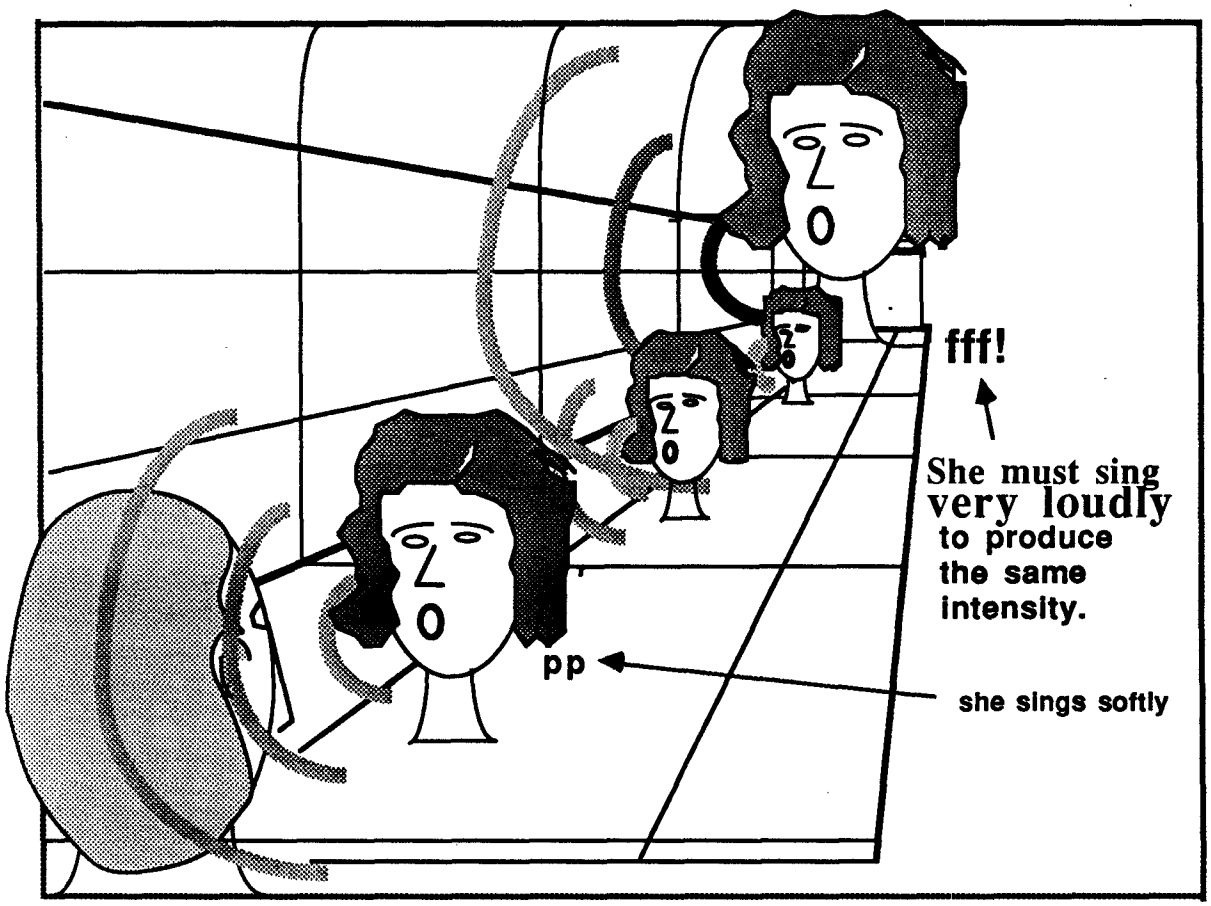

Figure 8.

As the distant singer must sing very much louder to produce the equivalent intensity as the nearest singer, so must she also become bigger in order to produce the same size image at the retina of the listener. It must be noted that 'loudness constancy' is complicated in a way that 'size constancy' is not: few images in the visual world are expected to increase or decrease in actual size in short time, thus facilitating the perceptual task, whereas auditory sources are commonly expected to vary in loudness in very short times, especially in music where many different loudnesses can occur in quick sequence without confusing the listener.

"Auditory perspective," is not a metaphor in relation to visual perspective, but rather a phenomenon that seems to follow general laws of spatial perception. It is dependent upon loudness (subjective!) whose physical correlates we have seen to include spectral information and distance cue, in addition to intensity. Further, the perception of loudness can be affected by the 'chorus effect' and vibrato depth and rate in a very subtle but significant manner.

The listener in the experiment, then, used all the information available, spectral and distance cues in addition to intensity, to make a determination of loudness at the source. When deprived of spectral cues then the distance cue sufficed. Were there no reverberation present in the latter case, then intensity alone would be the cue and the answer to the question would then be that the closer of the two is the louder[6].

Computers can be programmed with some care to extend the dimensions of loudness beyond intensity thereby providing the composer with a control of loudness vastly more subtle, **musical**, than that provided by intensity alone. However, only recently have synthesizers offered the composer spectral and intensity change as a function of effort (key velocity), and distance as a function of constant reverberant signal in relation to a varying direct signal (the latter has been possible since the first spring reverberators became available). The musical importance of these dimensions of loudness can not be over-emphasized, yet their use in either general purpose computers or synthesizers is not widespread.

The issues surrounding perceptual fusion, including quasi-periodicity and source identification, segregation, chorus effect, can still only be fully addressed with computers and large general purpose synthesizers. To be sure, there may be reasons of economy why generally available synthesizers can not provide such capabilities. However, there may also be some insensitivity to the importance of perceptual domains in which musicians find their reality.

Finally, these issues of perceptual fusion and auditory perspective are of general interest because they bear upon the very basis of music perception. The domain of sounds to which these issues are relevant is not constrained to those similar to

natural sounds, but may include all imaginable sounds. In fact, the understanding and exploration of these issues suggests somewhat magical musical/acoustic boundaries that cannot be a part of our normal acoustic experience yet which can find expression through machines in ways that are consonant with our perceptual/cognitive systems.

REFERENCES

[1] McAdams, S. *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images*, Stanford University, Dept. of Music (CCRMA) Technical Report. STAN-M-22 (1984).

[2] Chowning, J. M., "Computer Synthesis of the Singing Voice," in *Sound Generation in Winds, Strings, and Computers*, Johan Sundberg, editor (Royal Swedish Academy of Music, Stockolm), 4-13 (1980).

Chowning, J. M., "Frequency Modulation Synthesis of the Singing Voice," in *Current Directions in Computer Music Research*, Max V. Mathews and John R. Pierce, editors, MIT Press, 57-63 (1989). (This is the 1980 article rewritten with a slightly different emphasis.) A compact disc containing sound examples described in the text is available from MIT Press.

[3] Bregman, A. S. "Auditory Scene Analysis," Proc. IEEE Conf. on Pattern Recognition, July 1984, 168-175.

[4] Chowning, J. M. "The Simulation of Moving Sound Sources," J. Aud. Eng. Soc. 199, 2-6, 1971.

Moore, F. R. "A General Model for Spatial Processing of Sounds," Computer Music J. Fall, 6-15, 1983.

[5] Zwicker, E. & Scharf, B. "A Model of Loudness Summation," Psychological Review, **72**,3-26,1965.

[6] Gardner, M. "Distance Estimation of 0° or Apparent 0°-Oriented Speech Signals in Anechoic Space," J. Acoustical Society of America, **45**: 47- 1969.

** Excellent reference books for the field which provide the means for implementation of that which is discussed in this paper.

Dodge, C. & Jerse, T., Computer Music, Schirmer, 1985.

Moore, F. R., Elements of Computer Music, Prentice Hall, 1990.