Some Initial Ideas on the Control of Digital Sound Synthesis Through AI Techniques

Charles Dierbach
Dept. of Computer and Information Sciences
University of Delaware

INTRODUCTION

Anyone who has played any of the digital synthesizers available today is at first impressed by the fine adjustments available to him in generating various waveshapes. This enthusiasm can quickly fade, however, if one tries to then sit down and produce a particular sound in mind. The user is soon overwhelmed by all the low-level specification of control, and usually proceeds to the more common interaction with the machine, that of trial and error, possibly from a jumping-off point of some known synthesis technique. Thus, the discovery of interesting sounds is somewhat a game of chance, and not the result of a methodological search.

This low-level access problem is not unfamiliar to computer scientists. They have a natural tendency to work away from the machine level of specification towards a level of description more appropriate for the problem at hand. It can be said that the programmability of sound objects with digital synthesis systems is still at the "assembly level" of the evolutionary ladder of programming language design. Higher-level languages do exist for other aspects of computer music generation, e.g., score description. This is not unexpected, since the problems here are more tractable. However, in dealing with the perception of sound waves, problems of appropriate control become much more subtle, due to our psychoacoustic interpretation of these physical signals.

Currently, the low-level specification available to, and required by, the user leaves him in a position of needing to know many of the details of the physics of sound and psychoacoustics. For example, he may specify a perfectly linear amplitude rise for a portion of the amplitude envelope of a particular waveshape, but must realize that the listener would perceive that attack as logarithmic [Roederer, 1979]; he may specify a phase shift of a harmonic in a given constructed sound, but must realize that any noticeable perceptual change, if any at all, is minimal, as Helmholtz experimently discovered [Grey, 1975].

Thus, a goal would be to create an overall system with sufficient knowledge and flexibility to allow the user to interact in a more *subjective* way with the system, not requiring him to know details of the physics of sound or psychoacoustics. The fact that for a given synthesis system having 16 controllable harmonics, with relative strengths of 0-100, gives more possible timbral combinations than can be searched in a lifetime, raises some serious concerns about the way to go about reducing this search space of sounds. A user would need to have discovered, or have been told, a set of *rules* to keep in mind in helping to guide him toward the appropriate synthesis choice for attaining a desired effect. Helmholtz began such a set of rules by relating physical waveshape properties to subjective descriptions :

> "... simple tones sound sweet and pleasant, without any roughness, but dull at low frequencies; complex tones with moderately loud lower harmonics, up to the 6th, sound more musical and rich than simple tones, but they are still sweet and pleasant if the higher harmonics are absent; complex tones consisting of only odd harmonics sound hollow and, if many harmonics are present, nasal; predomination of the fundamental gives a full tone, in the reverse the tone sounds empty; complex tones with strong harmonics beyond the 6th or 7th sound sharp, rough and penetrating." [Grey, 1975]

Expert systems, a conceptual framework for program development of *knowledge engineering* problems, are receiving some attention of late. They provide a way to program-in sets of rules ("expertise") for solving specific problem domains, and can apply this knowledge in an automated way. Thus, the acquired knowledge of appropriate synthesis techniques by human experts might be captured in such a system for the more naive user to make use of, in the way of offering subjective control, relieving the user of detailed knowledge of each synthesis technique. We shall explore this more fully later.

CONCEPT OF INSTRUMENT

The concept of *instrument* , in the traditional sense, has served as a way for dealing with particular classes of timbre. Despite the vast range of timbres potentially generated by any one instrument, the "instrument character" of each is essentially preserved and recognizable throughout its "timbre subspace". This has aided in providing conceptual coherency to the timbral world. With the infinite timbral modifications available to the digital synthesis user today, though, should we abandon this notion of instrument and proceed by *unrestricted* note-by-note sculpturing of timbre, or adopt this paradigm, placing on ourselves some set of *constraints* for which any one musical part may draw from the "timbral universe"? I subscribe to the thesis that the creative process flourishes more within a given bounded selection space than within an unbounded universe of choice.

How are we to proceed, then, in creating for ourselves an *instrument definition* that will serve us well? We seek to construct a class of *timbre definitions* which are all functionally related in some way, thus providing us with a timbral subspace which is consistent with our notion of "instrument".

We can take a cue from the natural instruments themselves. Changes of timbre, by any instrument, are a function of loudness, pitch-step (e.g., "register quality"), or performer-invoked. These nuances provide the qualities which make natural instruments inherently interesting. (Of course, acoustical environment is also responsible for timbral alterations, but that will not concern us here.) We can examine these as to which are most appropriately part of the concept of instrument definition, if we are to create "instruments" with as much inherent character as our natural models.

It seems reasonable to include register quality and volume-related timbral modifications in our instrument definitions. However, changes which are too discontinuous within the range of pitch or volume will be troublesome. For example, good melody has been described as containing a certain ratio of *expectedness* to *unexpectedness*. (Of course, the measure of "expectedness" is somewhat dependent upon the listener's musical sophistication.) It would be disconcerting for the listener, making it difficult for him to abstract any musical form from what he is hearing, if melody had too much of the unexpected in the melodic intervals. This can also be said of timbre. A musical passage which changes timbral quality in discrete jumps as a result of change in pitch or volume will cause trouble in the listener perceiving that as one melodic idea. This fact, called *melodic fission* , was exhibited by Wessel[1979]. To have timbre change from that of flute-like to bassoon-like because of a half-step change in pitch is too erratic. Not that this effect might not have legitimate creative structure, but the point is that the listener might not "understand" the resulting passage as intended, or the composer himself might loose some of his own conceptual understanding of the piece by having such erratically behaved instrument definitions.

The subtle performer-invoked qualities of sound are not aspects of musical thought that we wish to build into an instrument definition. The slight alteration of a note's duration, the added "push" at the attack of a particular rhythmic point, are examples of modifications which account for musical style and good taste. One might envision, though, control of these during play-back, to be captured as a permanent recording of the piece, or simply considered as one unique performance.

PROBLEM BREAKDOWN

Figure 1 shows all of the interfaces where interpretation of information occurs during the sound synthesis process. The existence of many levels here is indicative of the fact that *natural* control of such a process can be elusive.

The man/machine interface has associated with it the usual problems of man-machine communication. What is desired here is the most natural mode of communication achievable for the user. With the subtle qualities of timbre, however, we need to offer a means of "timbral exploration" in both large and small increments of change. Thus, the need for sufficient control in order to aid the user in arriving at *subjectively* pleasing results puts extra emphasis on the importance of the user interface design.

The machine/wave-propagation-medium interface concerns itself with digital sound synthesis techniques and the process of digital-to-analog (D/A) conversion. There have been an abundance of techniques developed for the digital synthesis of waveforms. The concern here is in choosing the appropriate one for the effect desired. The design of D/A converters, on the other hand, has essentially remained unchanged for the last 20 years, the retention of information here being based on the bit-precision of the DAC. Therefore, this problem is well understood.

The wave-propagation-medium/psychoacoustic-interpretation interface introduces many modifications upon the signal processing which greatly affect our eventual interpretation of sound. Things such as our non-linear response to amplitude in the perception of loudness, the ability we have to "supply the missing fundamental", and the slight dependence of pitch-perception on loudness are a few examples. There are many empirically derived results on subjective interpretation here, but many have yet to be

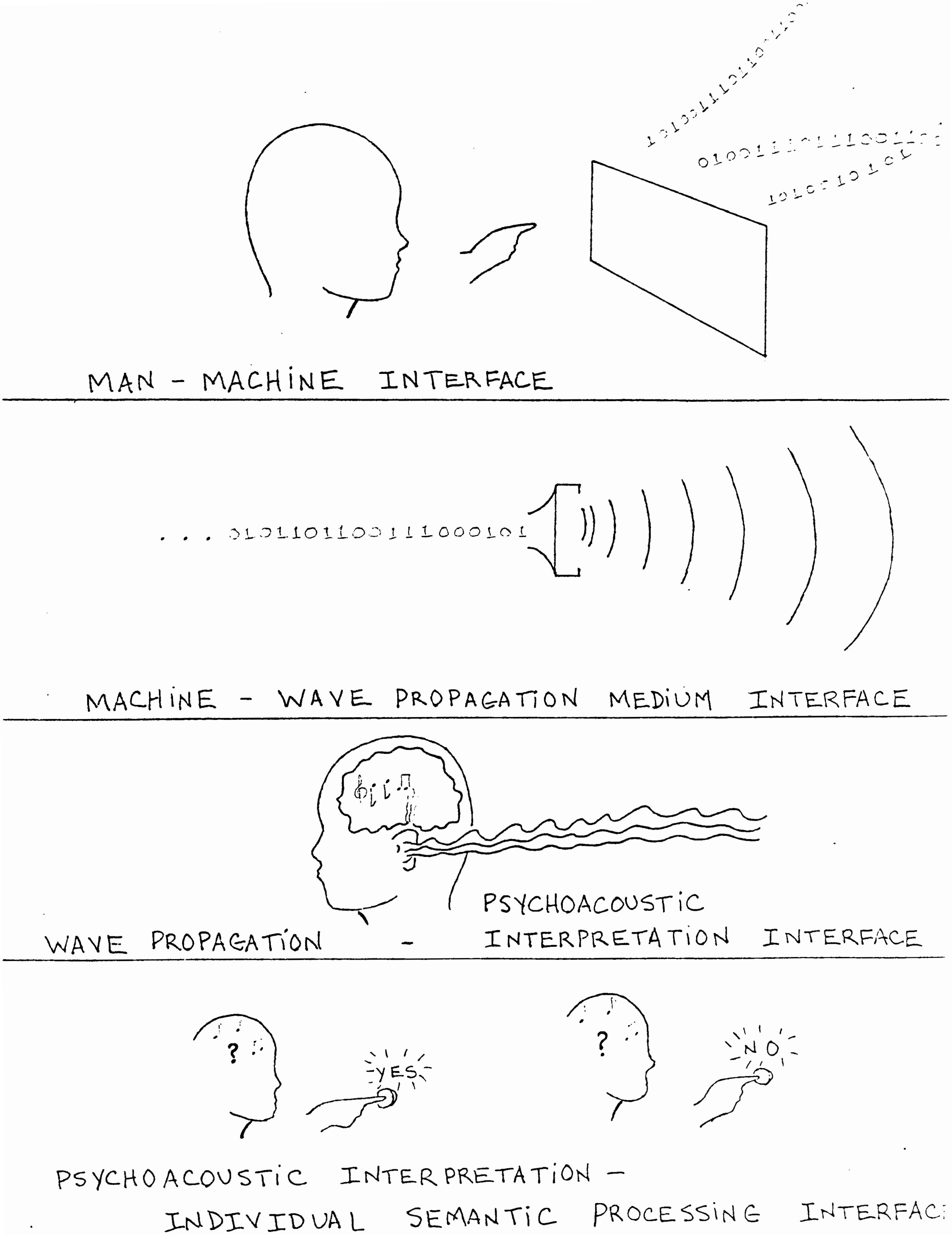
MAN - MACHINE INTERFACE
MACHINE - WAVE PROPAGATION MEDIUM INTERFACE
WAVE PROPAGATION - PSYCHOACOUSTIC INTERPRETATION INTERFACE
YES
NO
PSYCHOACOUSTIC INTERPRETATION - INDIVIDUAL SEMANTIC PROCESSING INTERFACE

Figure 1

fully explained in terms of the underlying physiological and neurological basis for these [Roederer, 1979].

The last interface level, the psychoacoustic-interpretation/ individual-semantic-processing level, has to do with "personal preference". Given all that has been said so far, if two people were to perceive the same subjective response to a given sound, one may still be pleased with what he is hearing, and the other not. This has to do with "learning experience", preconceived notions of what important qualities particular sounds should have. This again points to the need for allowing flexibility of control at the user interface, since these qualitative judgements can be based on quite subtle aspects of the physical waveform, relative to the signal processing as a whole.

AID OF AI

The field of Artificial Intelligence (AI) has been in existence essentially since the development of the computer itself. The tasks it took out to solve have proved to be difficult indeed. (Natural language understanding is one such area.) Its problems made clear the need to develop more appropriate models of intelligence, prompting people to "think about thinking".

The programs that attempt to solve these problems do not "calculate" in the numerical sense, but rather by drawing *inferences* on a given set of believed facts. These facts are stored as rules in a data base (hence, *knowledge base*), and inferences are drawn on these, generating more believed facts in the knowledge base. In most situations, all possible inferences cannot be made, due to the resulting "combinatorial explosion" of the number of facts in the knowledge base. This aspect is controlled by the *inference engine* , and is generally considered distinct from the knowledge base. (The inference engine contains the *heuristics,* or "rules of thumb", of how to go about applying the facts in the knowledge base in generating a new set of facts which satisfy a *goal.)* In programs such as these, the knowledge base tends to be the larger portion of the program, whereas the inference engine can be rather small. The "cost" of such systems is due to "memory-intensive", rather than "processor-intensive", aspects of computing [Michie, 1982].

A more immediately fruitful result of this knowledge engineering has been a programmed approach called *expert systems.* It is a natural first step in building intelligence into a program in that the decision-making rules in the inference engine are simply modeled after the decision-making of a human expert. Furthermore, the problem domain attacked by these programs is a very specific one, further reducing the design complexity.

The first expert system, DENDRAL, was developed in 1965 at Stanford University. Its task was to aid organic chemists in determining the molecular structure of chemical compounds. It is still in use today. Other systems have been built for such problem areas as medicine, geology, education, and knowledge engineering itself [Webster and Miner, 1982].

We can look at the characteristics of expert systems today to see how they may aid us in our particular problem domain. One of these features is that of *flexibility* and *expandability* [Webster and Miner, 1982]. They can be easily expanded or corrected simply by the addition or deletion of rules within the knowledge base. This would be useful as new sound synthesis techniques became available and the system had to be "taught" about that technique's usefulness.

A second characteristic is their ability to deal with *incomplete data* [Davis, 1982]. Just as human experts must make the best decision given only partial information to a problem, so should an expert system. Thus, if a user gave only partial specification in his sound construction dialogue, the system would be able to provide *some* kind of output, allowing the user to then provide further directions relative to what he is currently hearing.

Many problem-solving situations benefit from taking different views of the given information. For example, Stefik[1983], uses the example of road-map information for traveling from a location in San Diego to New York. One would first want to have a detailed map of San Diego for the sake of city street information, before looking at a continental interstate road map. However, both views are ultimately needed. For our problem domain, we can benefit from looking at our signal processing in different views also. An obviously needed one is the *physical view.* This is the view that describes the actual physical aspects of the waveshape being processed. Another view would be the *subjective view.* It is the view taken by the user in perceiving and describing the sound quality desired. A last view would be a *methodological view.* This is the view taken by the expert system as a goal state. Since its task is to select an appropriate synthesis method, it ultimately selects one of the methodological views, and further uses the subjective view in determining the parameter values needed for the proposed method (physical view).

A final characteristic of expert systems is what is known as *transparency* [Webster and Miner, 1982]. It is the ability of the system to explain its reasoning process. This, of course, is of most importance in something such as a medical diagnosis system, providing the doctor with a way of "overseeing" the reasoning behind a diagnosis. It is useful here, however, as a learning tool. In case the user *is* interested in the synthesis technique chosen for his subjective request, he can follow the selection process. Of course, the thrust of the whole system design is to relieve him of these details.

PROPOSED DESIGN

We shall first look at the intended user interface made available by this subjectively-controlled sound synthesis system and then look at some specific requirements of each system component.

It has not yet been explicitly stated, but the user interface envisioned would rely mainly upon natural language sentence fragments for input, without aid from any graphical representation or formal notation. The fragments would serve as means of *interpolating* between some "fixed-points" in the timbre space.

At first thought, this might seem a limiting approach to systems which have within them the potential for very subtle information processing. However, most appropriate control would be one which lends itself to our, as human beings, most natural mode of conveying these subtleties. In fact, this idea is not unique. Grey[1975] has shown the feasibility of this idea to the extent of developing an algorithm for interpolating between waveshapes. The waveshapes are represented by linear approximation for the sake of data reduction, but retain the same perceptual similarity as the original waveshapes. The interpolations, in the realm of additive synthesis, produce smooth, continuous perceptual transitions of timbre. The fact that the algorithm was not based on a simple linear mixing of the interpolation points, but was able to take into account time variant aspects of the signals, was even more encouraging. (One aspect of control not approached in this method, though, was the *inharmonic* portions of waveshape.) He concludes that the timbre space has more the nature of a continuous transition of timbre between points rather than the classical thought of timbre falling into distinct categories.

Further conclusions by Grey also support the possibility of successful interpolation in more coarse ways (as far as the physical signal) than might be expected, without by-passing many perceptually distinct timbres. From his experiments in the data reduction of waveshapes by line-segment approximations, the success of the approximations with respect to retained perceptual similarity to the original, leads him to conclude that in most cases, microfluctuations within waveshape have little perceptual significance.

There is not much variation in how timbral description is currently done by most of those involved with this process - either by an *implicitly defined* sound quality (e.g., "that of a clarinet"), or by *relativistic references* (e.g., "less harsh"). A musician has hardly, in the past, needed to know of Fourier analysis in order to convey a desired sound quality to a colleague. These modes of communication, most necessary for musicians, have sufficed within the context of the world of traditional instruments. When synthesized sounds came to the fore as potential instruments in themselves, their underlying method of sound production generated such relativistic references as "square wave - like" and "white noise - like", which, because of the steady-state character these initial systems generated, became just one new relativistic reference for popular use, that of "electronic".

Of course, these analog systems have been surpassed in "timbral potential" by the emergence of digital synthesis. We now have the ability to generate many more transient aspects of a signal, such as spectral envelope, which is of most importance in producing sounds with as much inherent interest as the natural instruments. This ability far surpasses the capabilities of the analog world. The universe of sounds that can be (potentially) produced by these binary wonders is impressive indeed, in fact overwhelming, (in fact, too overwhelming?). The exploring electronic musician might easily waive his option to go with the "digital stream of things" and keep his more limiting, but tractable, approach of "patching together" sounds.

This is, in fact, an acknowledged problem in digital synthesis. The point is, can we leave behind the techno-terminology of square waves and indices of modulation for a descriptive world that all, especially technology-intimidated musicians, can relate to?

Other important results of Grey[1975] concern the applicability of viewing the physical aspects of waveshapes in three dimensions. This has lead Wessel[1979] to suggest that the most natural way to move around the timbre space is by attaching "handles of control" directly to these dimensions, allowing one to interpolate between defined points in the space. Thus, ideas on the feasibility of interpolating through the timbre space have been firmly planted. The attempt here, in this paper, is to lead towards *subjective* control for all the interpolations of timbre and remaining aspects of sound. What follows is an overall system design that is hoped to support some of these ideas.

Figure 2 shows an overview of one possible system design. Its basic components are a *user interface* allowing the user to have dialogue with the system in simple sentence fragments; a set of *subjective interpreters* , used to translate subjective specifications into quasi-physical quantities; and an *expert system* , which uses expertise in the form of built-in rules to select the most appropriate synthesis technique based on the given *timbre vector, subjective qualities* and *subjective curves,* which will be explained shortly.

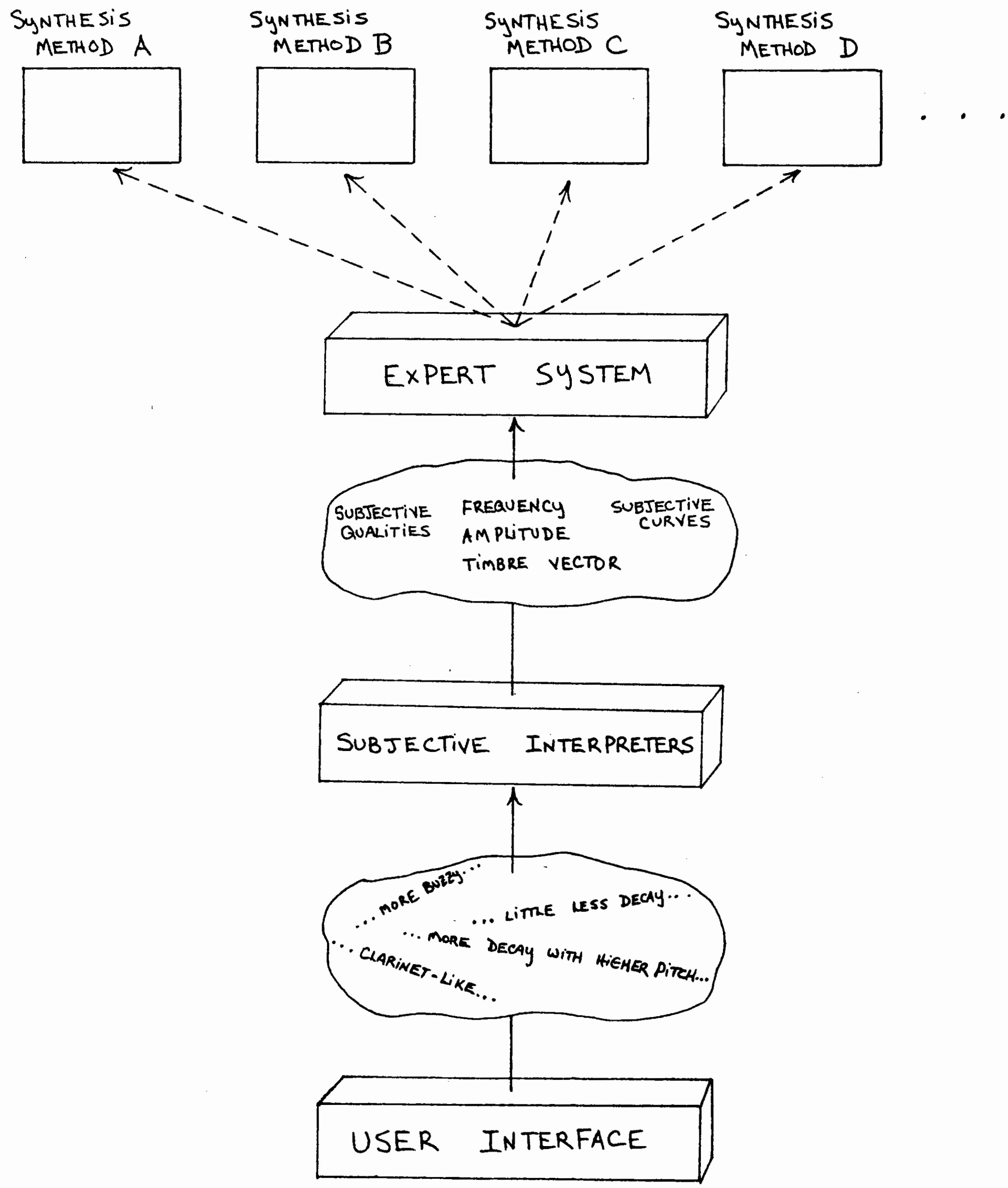

Figure 2

We begin by looking at the details of the user interface. As was stated earlier, timbral descriptions would consist of simple sentence fragments. Figure 3 gives a mock user session for the sake of demonstrating the general flavor of the expected user interaction. It serves by showing the redundancy of sentence fragments that would probably occur in such a domain-specific dialogue. The task of the natural language processing required here can probably be severely restricted in sentence construction options, without forcing unnatural sentence-building on the user. In fact, some directions, such as "much more decay", do not seem to have many alternative sentence constructs within even an unrestricted language context.

> OBOE-LIKE

> LESS NASAL

> LESS

> MORE

> SLIGHTLY MORE

> MORE TRUMPET-LIKE

> MUCH MORE DECAY

> MORE

> LITTLE LESS

> LITTLE MORE TRUMPET-LIKE

> MORE TUBA-LIKE

> INCREASINGLY MORE BRIGHTNESS ABOVE Eb’’’

> F’’ and BELOW, INCREASINGLY MORE DULLNESS

> MORE DRAMATIC CHANGE IN BRIGHTNESS FROM F’’ TO C’

> MORE GRADUAL CHANGE IN DULLNESS BELOW C’

> LITTLE MORE EDGINESS WITH INCREASING LOUDNESS

> LITTLE MORE

> LITTLE MORE

> CAPTURE ’instr4.temp’

Figure 3

It would seem, then, that sentences would be composed of two main sentence parts: a *quantitative* or *directional* fragment, and a *subjective* fragment. Fragments such as "more ...", "little less ...", and "more like a ..." would serve as examples of the first type; "clarinet-like", "buzzy", and "bright" would be examples of the second type.

The first fragment type would seem to be a small enough set of possibilities that not many constraints would have to be imposed. However, fragments of the second type can obviously lead to a very large set of possibilities. Grey[1975], and others, have applied psychological distance measures on the subjective qualities of timbre. This was the way a workable three-dimensional space was derived for the timbre space. One could take such an idea and apply the same techniques to that of the actual vocabulary generally used for sound description in everyday language. (Grey[1975] mentions the fact that attempts have been made in this direction.) Thus, we would obtain a large set indeed if we tried to collect all such words as rough, brassy, metallic, bright, electronic, hollow, flute-like, etc. We would want to reduce these to some primitive set of words where their timbral models are as well-spaced throughout the timbre space as possible. Since distance in the timbre space is a subjective measure, an experiment could be constructed which collects, by survey, the degree of similarity of word pairs as would be used in the context of sound description. Then, some scaling technique, such as cluster analysis, could be used to reduce this word set to a smaller number of word clusters with close meaning. A representative word from each cluster (one of most common usage) would be used as an element of the primitive set. The system, then, might consider "metallic" and "brassy" as both representing the same subjective quality in the timbre space, but would recognize either as entered by the user. In this way, the user can feel less restricted in his descriptions of timbre, and in the ideal case, the system would be able to interpret any word proposed. However, all of the descriptions would be mapped into a rather small, but adequate, number of *primitive timbre descriptors*.

What is strived for here is not necessarily a model which offers the potential to interpolate in an absolutely continuous way through the timbre space, but some reasonable compromise which offers the user a way to "poke around" the space. The possible number of timbres achieved would be a subset of the timbral universe, but yet many orders of magnitude greater than the number offered by the primitive timbre descriptors themselves.

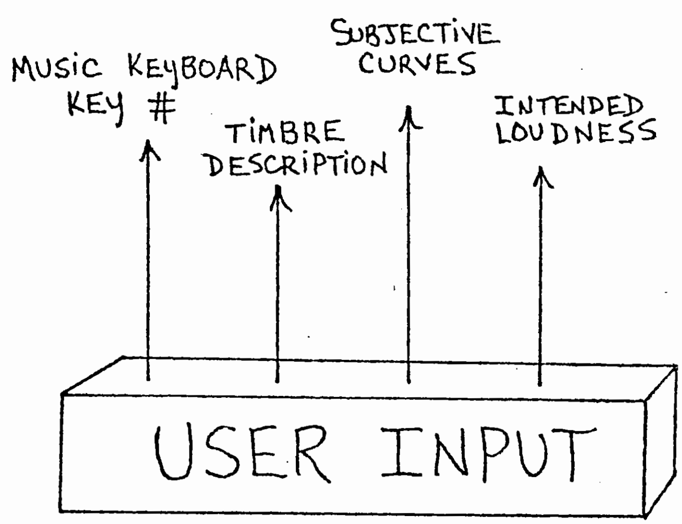

Figure 4

Figure 4 shows the detail of the user input. A combination terminal and music keyboard would provide the means for user interaction. The intent is that the user would provide information in sentence fragment form for the timbre description, transient qualities, and intended loudness. (If a pressure-sensitive keyboard were available, the intended loudness information could be directly read from it.). The transient qualities, or *subjective curves,* here apply to any sound character that varies in a functional way. This is intended to be used in a general way, in that it might refer to the degree of timbral variation as a function of pitch-step (e.g., "register quality"), pitch variation over time (e.g., vibrato), decay as a function of pitch-step, etc. Note that these curves would be *descriptively* defined. In light of the discussion of the requirement of continuity in the change of sound character, the restriction of discontinuities in the curves would make this approach more feasible than might seem at first. Also note that this approach allows the specification of the variation of one subjective curve by another. This can be seen, for example, if the user specifies how decay varies with pitch. In that amplitude envelopes are specified by subjective curves, and decay is a part of amplitude envelope, we have the possibility of creating "second-order" subjective curves. This further extends the power of the notion of subjective curves (Figure 7).

Having said the above, the subjective fragments can now be seen to break-down into two distinct types: *fixed-point* fragments and *quality* fragments. The fixed-points are those values interpolated between in the *descriptive timbre space.* These consist of the descriptions of clarinet-like, tuba-like, etc. , and contain all the multi-dimensional and transient aspects of timbre of those objects which they define (i.e., register quality changes, variation of timbre vs. loudness, etc.). Therefore, interpolation through the descriptive timbre space is an interpolation through all three dimensions of timbre. The quality fragments, on the other hand, allow alteration of only one dimension of subjective perception, consisting of words such as buzzy, dull, hollow, etc. I.E., those fragments which of themselves completely define a timbre object (e.g., clarinet-like) are seen to belong to the interpolation environment of the descriptive timbre space. However, those fragments which of themselves do *not* completely define a timbre object (e.g., bright) are seen to be modifications allowed in the timbre construction process.

The rest of the user input, i.e., pitch and duration, would be entered via the music keyboard. Note that duration is not diagrammed as part of the input in Figure 4. This is because duration is not seen to be part of the inherent quality of sound here. It is simply a value providing an on/off switch of the sound output, having no other functional relationship to the generated sound than that.

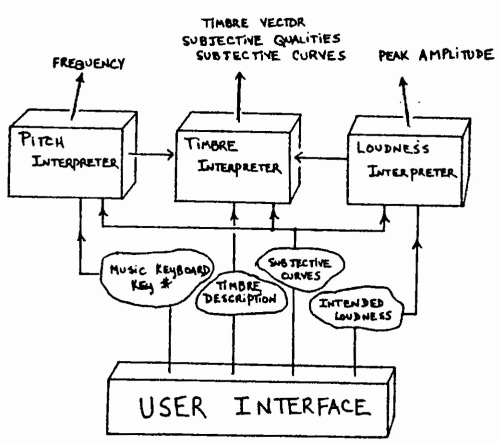

Figure 5

The means by which all these subjective fragments become translated into physical quantities of waveshape is by a set of *subjective interpreters* (Figure 5). These are passed the appropriate sentence fragments to be processed into one of frequency, amplitude, or timbral aspects of the sound. Here is where some of the knowledge of physics of sound and psychoacoustics is built-in, in that the subjective interpreters know what physical attributes of sound to control in providing the subjective interpretation requested.

The pitch interpreter would provide rather straight-forward translation of references of pitch into actual frequency. It would simply need be aware of the current "tuning" of the music keyboard, and what key is pressed. Any first-order subjective curves applied here would be interpreted as a vibrato effect. The frequency value would be passed over to the timbre interpreter for use with any subjective curves applied to timbre as a function of pitch.

The loudness interpreter would also provide rather straight-forward translation of references of volume into actual peak amplitude values. Any first-order subjective curves applied here would be interpreted as amplitude envelopes. The peak amplitude value would be passed over to the timbre interpreter for use by any subjective curves applied to timbre as a function of volume.

The timbre interpreter is the most complex of the subjective interpreters. Whereas pitch and loudness are attributes which have only one dimension of scaling, timbre has been seen to be multi-dimensional. Therefore, we perceive the timbre descriptions as placing us at a particular point within a descriptive timbre space relative to out last point of reference (Figure 6). The *value* we derive from this space is a vector value, relating the relative distance our current timbre definition is from each of the primitive timbre descriptors (the "fixed-points" of the space). Therefore, only a subjective data structure is produced by the interpreter, and not actual physical values of the timbre. (Note that this is a *bounded* space, and any point within it is not specified by reference to a set of basis vectors, but rather by relative distance measure from a somewhat arbitrarily chosen number of fixed points. This approach would not be appropriate in an *unbounded* space.) The physical interpretation of this subjective definition of timbre remains to be translated into actual physical values of the waveshape by the next component of the system, the expert system. The first-order subjective curves here, as mentioned above, are interpreted as timbral variation as a function of pitch-step or loudness.

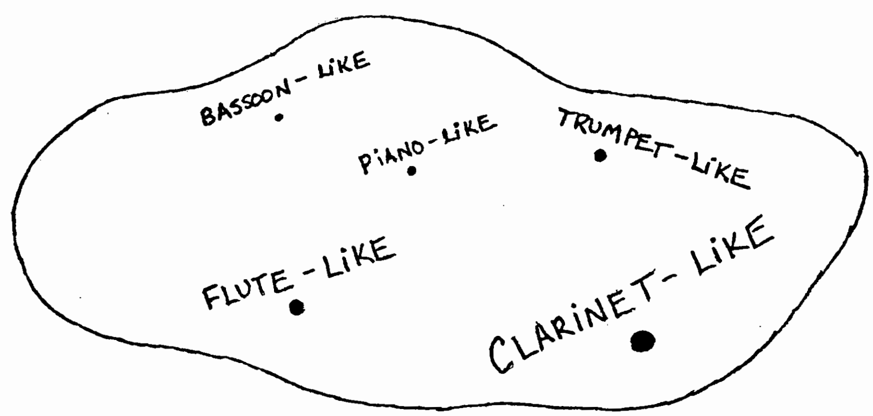

Figure 6

It should be said here that these subjective interpreters would evaluate any given sentence fragment in a *context-sensitive* way. During the user's dialogue (Figure 3), the successive requests of "more decay","more decay" ..., should be interpreted as increases in decay in not fixed steps, but in exponentially growing ones. The proper interpretation of a constant request for "more" should be taken to be that the current quantitative value of the attribute being modified is *far* from sufficient. Likewise, if a user constantly requests "little less...", the corresponding interpreted change in attribute should be exponentially *decreasing* , to allow for the ability of "fine tuning", and prevent the clumsiness of constantly "overshooting" and "undershooting" the goal. A following request of "little more ..." would begin at the current size of incremented step, and would again grow exponentially as the command to do so were repeated. The ability of this method to traverse a wide range of values and yet center-in on any one desired point in a reasonable amount of time is effective for the same reason that the classic binary search algorithm of general computer science is so effective. Figure 7 summarizes the interpretation of some subjective curves according to use.

SUBJECTIVE CURVES

1st order

VIBRATO
frequency as a function of time

AMPLITUDE ENVELOPE
amplitude as a function of time

REGISTER QUALITY
timbre as a function of pitch-step

LOUDNESS CHARACTER
timbre as a function of amplitude

2nd order

ATTACK CHARACTER
attack segment of amplitude envelope as a function of pitch-step

DECAY CHARACTER
decay segment of amplitude envelope as a function of pitch-step

VIBRATO CHARACTER
degree of vibrato as a function of pitch-step

VIBRATO ENVELOPE
degree of vibrato as a function of time

Figure 7

We now come to the heart of the proposed system. As described in the above section, an expert system can offer us a means of building modifiable intelligence into our synthesis system. Its goal is to choose the most appropriate synthesis technique and provide the parameter values that that technique requires. Therefore, each time it is passed different subjective data structures from the timbre interpreter (i.e., timbre vector, subjective quality or subjective curve), it must seek another goal.

In the ideal case, the expert system would have in its knowledge base information of the totality of digital synthesis techniques, and know which synthesis techniques are available to it in the current hardware configuration. Thus, if the actual synthesis system being directed by the expert system had 32 controllable harmonics with relative strengths of 0-100, this information becomes part of its knowledge base. If the synthesis environment is upgraded, the expert system need only have the appropriate rules of its knowledge base altered to reflect the change. The inference engine part of the system need not be changed, in that it has the potential to seek the best goal with *any* given set of circumstances.

Figure 8 shows the relationship of the expert system to what can be called the *physical attribute space*. This space consists of a set of nodes, one node for each of the actual parameter values of the synthesis system in use. Around the space we can envision "windows" which are nodes containing slot values for a subset of the possible attributes in the space. These windows represent the *methodological view* of the space, and are precisely what are chosen by the expert system. Therefore, if the expert system decides to employ FM synthesis to satisfy a given timbre vector, it need only specify the set of parameter values that are included as slot values in the FM synthesis window to the space. Thus, these windows provide an appropriate abstraction of all parameter values of the synthesis system.

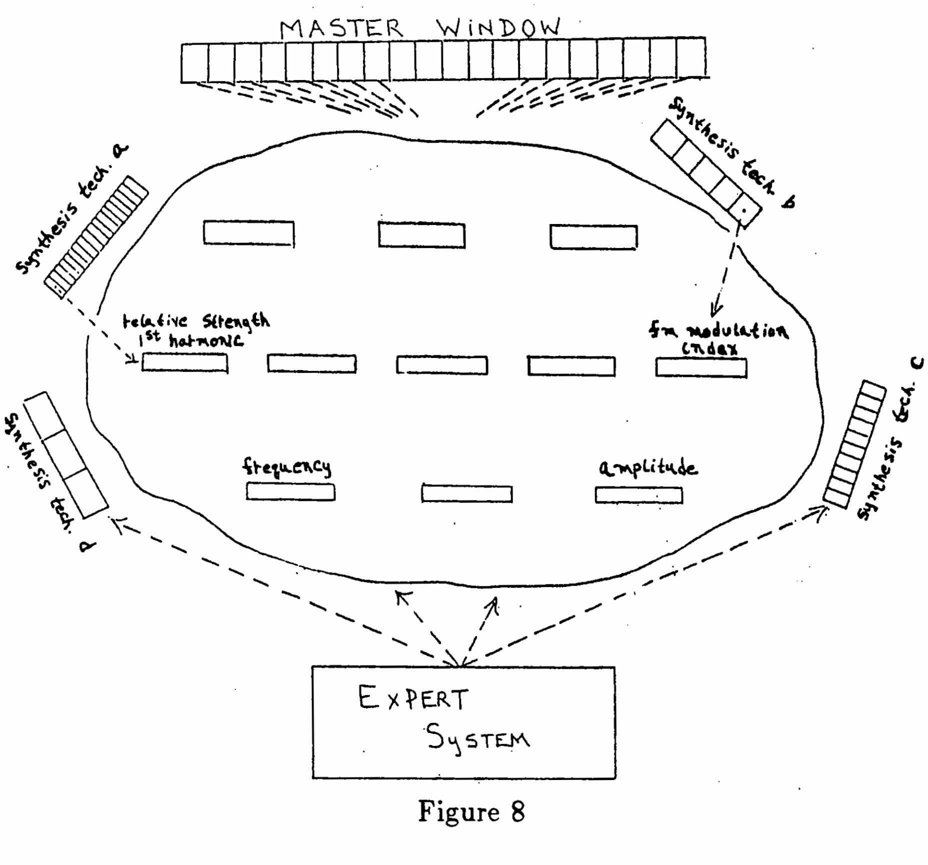

Figure 8

To complete the picture, we have a special window on the space - the *master window*. It contains slot values for all of the attribute nodes in the space. Thus, this provides direct control of the synthesis system by simply copying-over all changes in this master node to the actual synthesis system input.

Figure 9 provides an overview of the entire system discussed.

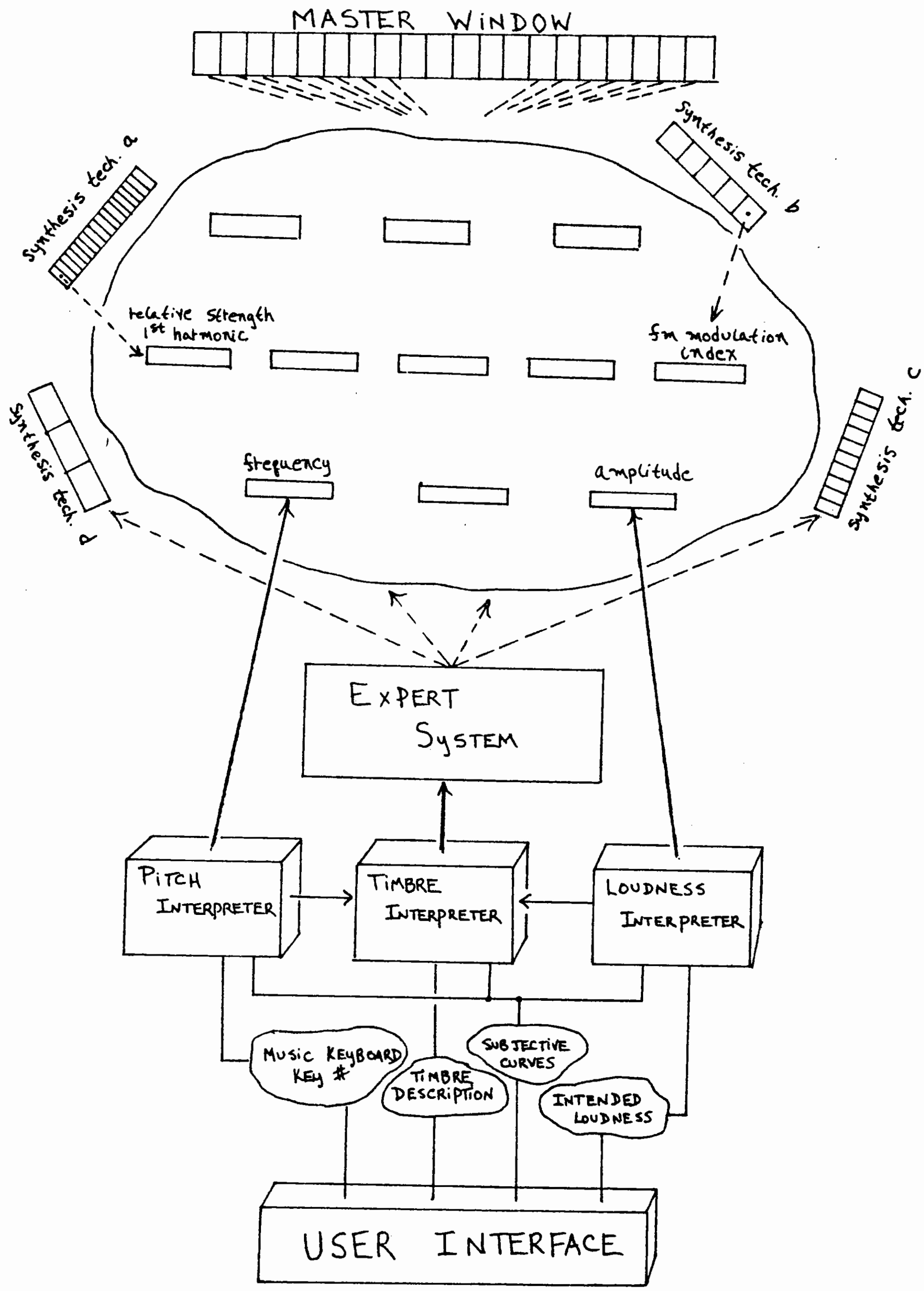

Figure 9

SUMMARY

There are many details to be resolved before such a proposed system can be realized. The concept of interpolating through the timbre space seems a good one. However, all of the underlying physical aspects of sound and their subjective interpretation is not well enough understood to indicate an obvious way to provide *continuous* interpolation. The control of the inharmonic aspect of sound is one specific problem. Also, even when continuous interpolation is achieved, Grey points out that the points interpolated through depend on the specific interpolation scheme used. Thus, interpolations may be *continuous,* but not necessarily *complete.*

The system needs to be usable in "interactive-time" because of its reliance on a feed-back/modification strategy. Since the play-back is of only one sound at a time, though, the time issue is not as crucial. Once these instrument definitions are arrived at, they can be used in a non-interactive-time sound processing environment. Ideally, the system should be fast enough to allow quick passages on the music keyboard for the testing of timbral quality in melodic context.

The flexibility offered by the knowledge engineering scheme allows a modular approach to system modification. For instance, any desire to add new subjective controls, such as location-perception or reverberation, is merely a matter of adding the appropriate subjective interpreter. The replacement or addition of synthesis hardware or synthesis technique would call for a change only in the knowledge base of the expert system.

An afterthought in reanalyzing the benefits of this approach has lead me to consider the value of a secondary mode of interaction. In providing control to the user by means of subjective descriptions, the intent is to offer means of control more natural than physical specifications. However, it is conceivable that at times the user would not even himself know how to describe, in subjective terms, what effect he is seeking. In this case, it might be of use to have the system "take over" by presenting pairs of sounds to the user, requesting responses as to which of each pair is most close to what the user has in mind. In this way, the system could determine what point the user is intending to approach in the timbre space, without even requiring subjective descriptions of the user.

The ideas presented here are intended to be a catalyst for the stimulation of thought in the joining of AI techniques with sound synthesis control. The goal of building such an expert system as described is obviously the most ambitious of the required system components. However, such a system could provide a start in the direction of universal control of any digital synthesis system. The subjective control offered the user could remain the same from one system implementation to another, but the expert system would simply make the best choice of those available to satisfy the user's request. And the biggest benefit, which has been the underlying motivation for the whole approach taken, is that timbral exploration would be available to all, not just the technologically-oriented.

ACKNOWLEDGEMENTS

I would like to thank Daniel Chester and Stephen Gannon for critical reading of final drafts of this paper.

Special thanks to Karen Zabel for many insightful discussions and the general sharing of the academic spirit.

REFERENCES

Davis, Randall (1982). "Expert systems: Where are we? And where do we go from here?" MIT Artificial Intelligence Laboratory report, A.I. Memo No. 665.

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

Michie, Donald (1980). "Knowledge-based Systems", Department of Computer Science Report UIUCDCS-R-80-1001, University of Illinois at Urbana-Champaign (Jan. 80).

Roederer, Juan G. (1979). *Introduction to the Physics and Psychophysics of Music* (Second edition), Springer-Verlag, New York.

Stefik Mark, et. al, *Building Expert Systems,* Teknowledge Series in Knowledge Engineering, Vol. 1, Chap 4, Edited by Hayes-Roth, Waterman and Lenat, Addison-Wesley, Reading, Mass.

Webster, Robin and Miner, Leslie (1982). "Expert Systems, Programming Problem Solving" Technology Jan/Feb 1982.

Wessel, David L. (1979). "Timbre Space as a Musical Control Structure" Computer Music Journal Vol. 3 no. 2, pp. 45-52 (June 79).