Chapter 2

Prior Work

2.1 Timbre

While timbre is a term in common use within the music community, its precise definition has remained elusive. In his thesis on tracking timbre, Shiraishi notes this is in part because:

> The term 'timbre' itself is often confused with two different distinctions. On one hand, we recognize that different acoustic instruments have timbre. The piano has timbre of the piano...for instance. On the other hand, one instrument can be played with a performer's delicate timbre manipulation' (2006, p. 4).

In other words, one can associate a single timbre to a single sound-producing body or to a single sonic "character" possibly produceable by many different sound-producing bodies. Nicol writes that 'ecological listening states that humans are more likely to describe sounds with reference to the apparent source of that sound as opposed to certain characteristics of that sound' implying that the former association with timbre may be more often valid (2005, p. 23). However, for this research the latter association is more relevant and will be adopted going forward.

Johnson and Gounaropoulos make a separate distinction between two different types of timbre that are also often confused, differing in time scale (2006). They write, in their study on associating words with timbral features extracted from an audio signal, that "by timbre we will mean the micro-level spectral characteristics of sound as discussed by Wishart, as opposed to the gross timbral distinctions used e.g. in the MPEG-7 standard" (2006, p. 1). This refers to a confusion between local versus global timbral representations. We find the former of these two most relevant to this research and also adopt this going forward.

It is clear from the above examples that the many different interpretations of the word "timbre" make it a relatively opaque concept. Stowell and Plumbley note that "evidence from perceptual studies tells us that timbre is multidimensional and probably non-linear" (2008, p. 1). Otherwise, the results of this research often conflict, most likely due to the ambiguity associated with the term (Fiebrink, 2005, p. 4), (Ciglar, 2009, p. 11).

There are two approaches to defining timbre (see Figure 1): an additive definition and a subtractive one. Additive definitions explain timbre by associating it with things that it *is*, while subtractive definitions attempt to explain timbre by defining what it *is not*. The subtractive definition, most common in the literature, explains timbre as being the characteristics of a sound left over after removing pitch and loudness (Wessel, 1979, p. 45);

(Toivianen et al., 1998, p. 225). Ciglar points out that this definition assumes that a sound has a definitive pitch in the first place, or equivalently, that unpitched sounds do not have timbre (2009, p. 7). A possible addendum to this definition may be that timbre is what is left after removing loudness and pitch, assuming a pitch even exists to be removed. However, such a definition is still unsatisfactory, and this is why a good deal of research has been carried out to define timbre using an additive definition.

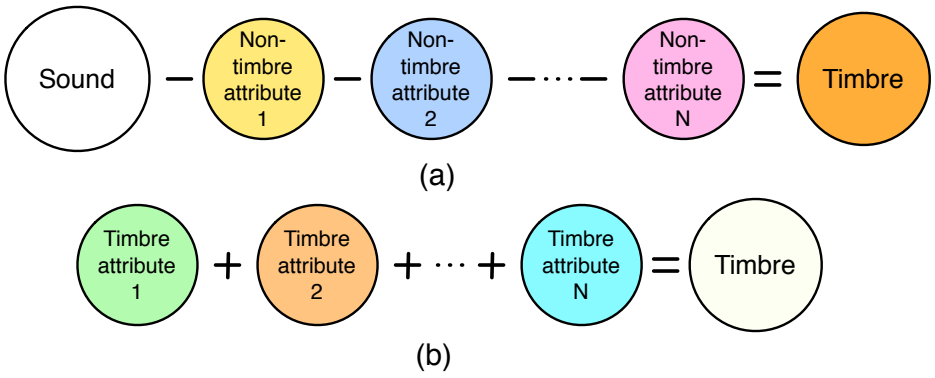

Figure 1: Visual representations of (a) subtractive and (b) additive definitions of timbre.

In order to build an additive definition of timbre, one typically performs controlled perceptual similarity tests, uses the results to build a subjective timbre space where distances are semantically meaningful, and then associates the axes of this space with acoustic features (Pampalk, Herrera, & Goto, 2008, p. 1). Some of the earliest studies of subjective timbre spaces were performed

separately by Grey and Wessel (Grey, 1975); (Wessel, 1979). Both use multidimensional scaling algorithms (MDS) to develop a timbre space from pairwise subjective similarity tests. MDS allows one to generate a space, typically two- or three-dimensional (though by no means limited to these), given a set of distance measurements between points, where the cumulative error in the distances provided between the points is minimized. In other words, MDS provides a way to generate a Euclidean space given a set of subjective measurements. However, in order to assign acoustic features to the axes resulting from MDS analysis on subjective measurements, one is left to find correlations between these axes with tested features. For example, using a two-dimensional timbre space, Wessel finds that the axes correlate with the spectral envelope and the nature of the onset transient (1979, p. 48). However, definitively assigning features to these axes based simply on correlation is misguided. As Caclin, McAdams, Smith, and Winsberg point out in their investigation into the properties of timbre, "Given the multiplicity of acoustical parameters that could be proposed to explain perceptual dimensions, one can never be sure that the selected parameters do not merely covary with the true underlying parameters" (2005, p. 2).

Another problem with this approach is that MDS forces the user to choose the dimensionality of timbre space a priori, which will result in a feature set to explain timbre that may either contain too much or too little information.

Generating a perceptual timbre space using MDS also can be handicapped if not given a wide variety of timbral material that sufficiently covers the perceptual space. As Seago, Holland, and Mulholland note, if not using a wide enough variety of timbral material "a sound in MDS space may have perceptually important features that no other sounds in the same space have—and, by the same token, two sounds could occupy the same location in a given MDS perceptual space, and nevertheless be audibly different" (2008, p. 3).

Yet another drawback of subjective testing is discussed by Prandoni (1994). He notes that many of the pairwise similarity tests for timbre space construction are performed using common instrument sounds. Prandoni writes that this is a major problem in that "the subjective ratings thus obtained are often affected by the listener's high-level notions of the structural features of the playing instruments rather than being a pure combination of sounds" (1994, p. 2). He further points out that these studies find that features related to the attack and steady-state spectral envelope are often found to be highly correlated with the resultant timbre space axes because these are the features that "remain constant among different nuances of color in recognizing an instrument" (p. 8). Therefore, it is not clear whether other features beyond those that allow discrimination between instruments and instrument classes are perceptually relevant, but just not useful in the discrimination task. This is

further complicated by the fact that one of the most often cited timbral features, the shape of the attack, assumes that for every "timbre" point there must be an attack. Nevertheless, acoustic features related to the spectral envelope and attack are the most widely used to generate objective timbre spaces for purposes of measuring timbre similarity.

From Prandoni's comments, one wonders if the features derived from these tests may best describe the intuition of timbre that is tied to the sound-producing body that created it, rather than that which relies solely on the sonic characteristics of the sound. This is supported by the fact that these tests ask a subject to note the similarity between whole instrument tones, rather than between attacks or steady-state portions separately. The result is that a single note played by an instrument will be placed as a single point in timbre space, rather than trace out a trajectory, which would seem more appropriate if the sonic characteristic of that note changes over its duration. In his treatment of both subjective and objective timbre spaces, Nicol writes that the single point vs trajectory delineation is one of the main differentiating characteristics between these two types of spaces (2005, p. 56). In an objective timbre space, each moment in time is represented by a point and as time evolves a trajectory is traced out through timbre space. This representation falls more in line with Johnson and Gounaropoulos' local timbre distinction (2006). It also allows one to incorporate the temporal evolution of timbre into a similarity calculation as

will be necessary when trying to find optimal matches to time-varying timbral content.

In this research, we are interested in such objective timbre spaces as our concept of timbre falls in line with the local timbre distinction and with that which represents the sonic character of the sound, regardless of the sound-production mechanism. Given this distinction, there is still ambiguity regarding "whose" percept of sonic character we refer to. For example, in a large auditorium the sound emanating from a violin will most certainly be perceived by the violinist as having a different character than that perceived by an audience member a hundred feet away due to the addition of reverberation. Since this research focuses on generating a system that is able to understand the sonic characteristics of a given audio file, the *who* can only be the computer, since no input will be given to the system regarding the acoustic space within which the audio file was recorded nor will it be given any information regarding the acoustic space within which the resultant synthesis algorithm will be used to generate sound. Thus, any reverberation or other distortions applied to a sound before it reaches a microphone for recording will be considered as part of the sonic characteristic of the desired sound, inseparable from the sound wave generated at the origin of the sound. Likewise, the computer will assume the synthesis algorithm it generates will be played in an aneachoic chamber in order to match the target sound.

2.2 Sound Synthesis

Timbre is a fundamental parameter of music, yet total and precise timbre control cannot be realized—or, at the very least, is severely limited—by acoustic means. Even the most skilled players are constrained by the physics of their instruments' sound production mechanisms (Wessel & Wright, 2002, p. 11). However, with the advent of digital technology, the sound generator can be separated from its control interface and therefore no longer reliant on its physical properties (Malloch, Birnbaum, Sinyor, & Wanderley, 2006, p. 1). The resulting timbral freedom introduced by digital technology augmented the already increasingly important role timbre had taken in Western compositional practice starting from the turn of the twentieth century (Klingbeil, 2009, p. 1). In investigating the increasingly common compositional use of timbre, Nicola Bernardini and Jran Rudi (2001) write that by providing a different, deeper and total control over timbre [computer music composition has placed a] stronger focus on the timbral aspects of composition - on the micro-level. (p. 3). This increased focus on timbre has spawned the design of numerous sound synthesis algorithms over the past fifty years in hopes to provide the composer with the level of control over timbre that they have enjoyed for years with pitch and loudness.

The different approaches used in designing synthesis algorithms have

varied widely throughout the history of sound synthesis. A number of sources provide surveys of the most popular techniques (Roads, 1996); (Miranda, 1998); (Cook, 2002). Miller Puckette's *The Theory and Technique of Electronic Music* (2007) stands out among the rest in that he not only describes the underlying mathematics behind a number of different synthesis algorithms, but he also provides the user with a blueprint for design and timbral exploration. This book in combination with Puckette's earlier paper, *Combining Event and Signal Processing in the MAX Graphical Programming Environment* (1991), provides a comprehensive look into designing synthesis algorithms using a graphical programming environment. The intent of these writings is not to compare and contrast the efficacy of the algorithms presented, but instead only to illustrate their inner workings. However, when a composer is ready to select a specific synthesis algorithm to work with, the results of such comparisons are paramount to the decision-making process. In order to assess the strengths and weaknesses of each algorithm, one must have a clear understanding of the desirable properties of a timbre producer/manipulator.

These properties have been discussed by a number of respected figures in the field of computer music at various levels of depth (Smith, 1991); (Jaffe, 1995); (Wessel & Wright, 2002). Smith's treatment is relatively superficial in large part because these properties are not his paper's main focus. Jaffe's list of properties presume that a synth's main purpose is to mimic a real-world sound

producing body, which is not always necessarily the case. Wessel and Wright mix the desirable properties of synthesis algorithms with those of control interfaces, often blurring the lines between the two within the same listed property. Most recently, in his treatment of composing with timbre, Nicol (2005, p. 40) provides four desirable properties for an ideal synthesizer: "fast synthesis" (i.e. computational efficiency); a "wide timbral range" (i.e. the ability to produce any desirable timbre); "easy parameterization" (i.e. an intuitive, low-dimensional mapping between parameters and sound); and "low data requirements" (i.e. low memory storage requirements). In other words, Nicol believes that an ideal synthesizer will provide the composer with an intuitive, low-dimensional parameter set that they may use to achieve (and manipulate) any desired timbre. Additionally, the underlying synthesis algorithm will ideally work in real-time and use a small amount of memory.

One of the most comprehensive comparative studies of synthesis algorithms is presented by Tolonen, Välimåki, and Karjalainen (1998). In this report, the authors categorize each synthesis algorithm under evaluation into one of four groups: abstract algorithms; sampling and processed recordings; spectral methods; and physical models (as originally proposed in (Smith, 1991)). Among the many algorithms treated, Tolonen et al. choose the most popular (also known as "classical" synthesis algorithms) to evaluate, and find that variants within each category perform similarly to their classical

counterpart in regards to the proposed criteria (1998, p. 101). The classical synthesis methods investigated in their study were sampling synthesis (sampling and processed recordings), additive synthesis (spectral methods), FM synthesis (abstract algorithms), and digital waveguide synthesis (physical models). A more complete list of algorithms can be found in Table 1.

Processed Recording	Spectral Model	Physical Model	Abstract Algorithm
Concrète	Wavetable F	Ruiz Strings	VCO,VCA,VCF
Wavetable T	Additive	Karplus-Strong Ext.	Some Music V
Sampling	Phase Vocoder	Waveguide	Original FM
Vector	PARSHL	Modal	Feedback FM
Granular	Sines+Noise (Serra)	Cordis-Anima	Waveshaping
Prin. Comp. T	Prin. Comp. F	Mosaic	Phase Distortion
Wavelet T	Chant		Karplus-Strong
	VOSIM		
	Risset FM Brass		
	Chowning FM Voice		
	Subtractive		
	LPC		
	Inverse FFT		
	Xenakis Line Clusters		

Table 1: Synthesis Technique Taxonomy (Smith, 1991)

Sampling synthesis (see Figure 2) “is a method in which recordings of relatively short sounds are played back” (Tolonen et al., 1998, p.10). Using sampling, one can turn any audio recording into a musical instrument by simply assigning the playback of the recording to a switch (Heise, Hlatky, & Loviscach, 2009, p.1). This type of synthesis dates back to the 1920’s and certainly became prominent in 1950 when Pierre Schaeffer founded the Studio

de Musique Concrete in Paris (Tolonen et al., 1998, p. 3). Like all sampling and processed recordings methods, sampling synthesis flexibility is dependent on the size of the database containing the pre-recorded audio segments. Thus, to obtain a truly flexible sampling synthesizer "the required amount of memory storage is huge" (Tolonen et al., 1998, p. 11). As the flexibility of the system grows and one has enough material to generate small variations in timbre, doing so becomes more difficult as the number of samples to choose from during any single discrete movement in timbre space becomes unwieldy. Thus, while sampling synthesis is computationally efficient and requires a low-dimensional parameter set, there is a fundamental tradeoff between its ability to generate a wide variety of timbral material and the amount of memory it requires. It is also nontrivial to transform a given timbre in a smooth way, which, in his more recent survey of synthesis techniques, Smith writes is the true "fundamental problem with sampling synthesis" (2006, p. 22).

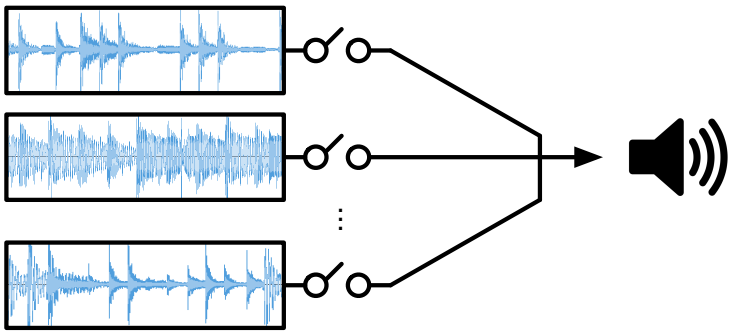

Figure 2: Sampling Synthesis

Granular synthesis (see Figure 3) ‘is a set of techniques that share a common paradigm of representing sound signals by “sound atoms” or grains...the synthetic sound signal is composed by adding these elementary units in the time domain’ (Tolonen, Välimåki, & Karjalainen, 1998, p.13). In his dissertation on analysis/synthesis techniques (which we be covered later), Klingbeil adds that ’granular synthesis may be viewed as a particular specialization of sampling synthesis...[It] offers the possibility to decouple the time and frequency evolution of a sound, as well as impart particular characteristics modulating between rough, smooth, or stochastic textures’ (2009, p. 6). Thus, unlike sampling synthesis, granular synthesis allows one to transform smoothly between timbres. However, this comes at the price of a higher dimensional parameter space, forcing the user to specify ‘the shape of the overall “cloud”, the fundamental frequency, the way in which the individual grains are generated, the structure of the individual grains used, etc’ (Johnson, 1998, p. 5). In Nicol’s dissertation investigating mappings between synthesis parameter spaces and timbre spaces, he notes that the non-intuitive mapping between this high dimensional parameter space and timbre space makes ‘emulation of target timbres a non-trivial process’ (Nicol, 2005, p. 49). Granular synthesis requires less memory than sampling synthesis, but is also less efficient (although real-time algorithms do exist). Vercoe, Gardner, and Scheirer point out that granular synthesis is ‘best suited to the generation of

noisy or textural sounds like water, wind, applause, and fire' (1998, p. 6).

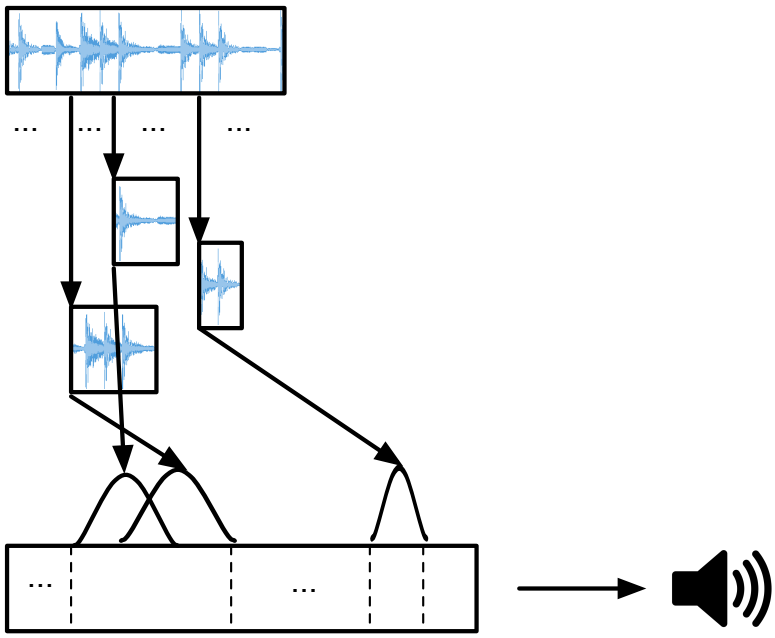

Figure 3: Granular Synthesis

Additive synthesis (see Figure 4) "is a method in which a composite waveform is formed by summing sinusoidal components to produce a sound" (Tolonen et al., 1998, p. 17). Based on Fourier analysis, "additive synthesis can in theory synthesize arbitrary sounds if an unlimited number of oscillators is available" (Tolonen et al., 1998, p. 94). However, as the numbers of oscillators increase, so does the number of controllable parameters and, therefore, what adding oscillators gains in flexibility, it loses in controllability (Klingbeil, 2009, p. 7). It is because of this that additive synthesis is best utilized for

generating harmonic or quasi-harmonic signals where little noise is present (Vercoe et al., 1998, p. 5). Additive synthesis also requires a number of parameters to be changed simultaneously in order to move a small distance in timbre space, which is unattractive. The desirable properties that additive synthesis satisfies are low storage (although the control data can require a lot of memory) and the ability to generate, in theory, any timbre, but its parameter space is high-dimensional and therefore difficult to control.

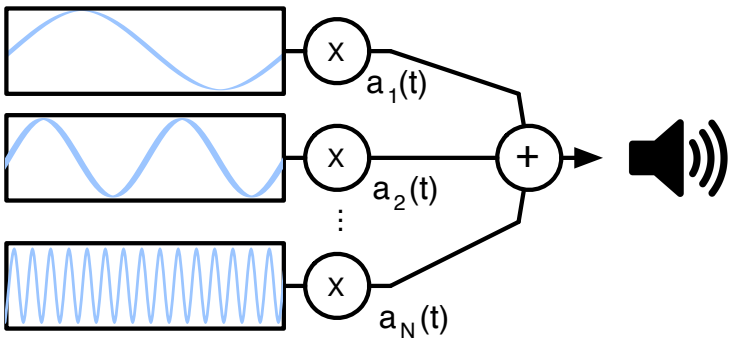

Figure 4: Additive Synthesis

Subtractive synthesis (see Figure 5) is similarly based on Fourier analysis, but works by subtracting sinusoids from a spectrally rich source to generate material rather than building up material by adding sinusoids together, as in additive synthesis. The 'subtraction' is performed by a time-varying filter whose coefficients are supplied by the user. In order to achieve complex and temporally evolving sounds, the parameter space can grow to the size of

additive synthesis and therefore will suffer from the same controllability issues (Tolonen, Välimåki, & Karjalainen, 1998, p. 48). If one uses a simple network of filters to try to reduce the size of the parameter space, 'the resulting tones have a distinctive "analog synthesizer" character that, while sometimes desirable, is difficult to avoid' (Vercoe, Gardner, & Scheirer, 1998, p. 5). Thus, one must trade timbral flexibility with the controllability of the algorithm. Similar to additive synthesis, the control data can require a lot of memory during sound production. Also, the efficiency of the algorithm and the facility to move around timbre space varies with the complexity of the time-varying filter or network of filters.

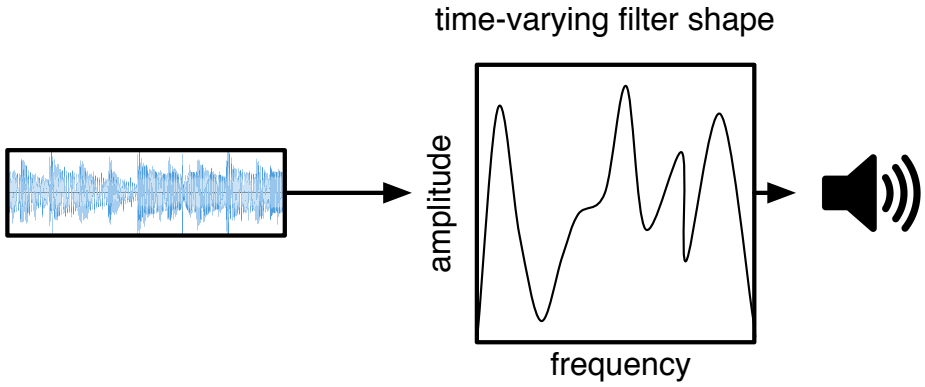

Figure 5: Subtractive Synthesis

FM synthesis (see Figure 6), in its most basic form, contains two oscillators, which are connected so that one oscillators waveform modulates

the frequency of the other. There are only a few parameters with which to generate a large range of timbral material, which is desirable, but due to an inherent nonlinear mapping between parameter space and control space, “FM has become widely regarded as a difficult synthesis type to control” (Mitchell & Creasey, 2007). This is for two reasons. First, the nonlinearity provides a non-intuitive relationship between a given set of parameters and the sound it produces (Nicol, 2005, p. 45). Second, a nonlinear mapping means that small changes in the input parameters can map to large changes in the timbre and thus fine timbral manipulation can be difficult (Jaffe, 1995, p. 2). Another undesirable quality of FM synthesis is that the characteristic FM sound is “fairly metallic, so the FM output is often filtered in order to produce a more natural sound” (Nicol, 2005, p. 45). The benefits of FM synthesis are that is “is very cheap to implement, uses little memory, and [as previously mentioned] the control stream is sparse” (Tolonen et al., 1998, p. 92).

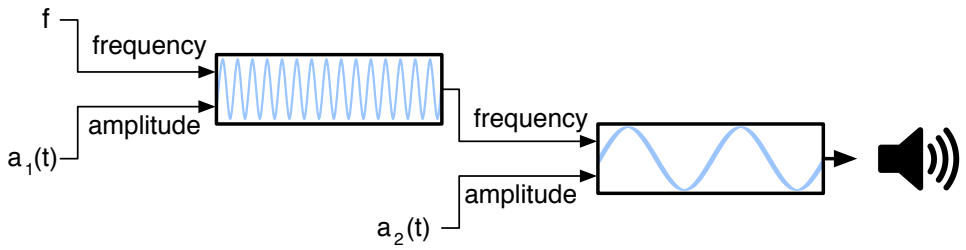

Figure 6: FM Synthesis

Digital waveguides (see Figure 7) are “based on a general solution of the wave equation in a one-dimensional homogeneous medium” (Tolonen et al., 1998, p. 63). They are basic linear time-invariant (LTI) structures that allow one to develop physical models of various instruments. Their parameters are intuitive and small in number, aiding in timbral manipulation. However, specific waveguide networks are designed to imitate a single sound producing body and therefore are timbrally restricted in comparison to additive synthesis. Also, as Smith (1992) comments in his seminal paper on digital waveguides, “new models must be developed for each new kind of instrument, and for many instruments, no sufficiently concise algorithm is known” (p. 86). In fact, as Tolonen et al. point out, any sound producing object that requires a two- or three-dimensional waveguide mesh to represent its governing physics will be computationally expensive to model and, therefore, only simple physical models based on waveguides will be able to run in real-time (1998, p. 99-100). Thus, digital waveguides provide a low-dimensional, intuitive parameter set and low storage requirements, but can be computationally expensive for complex sounds and require separate topologies for each region of timbre space (Nicol, 2005, p. 50).

The above discussion illustrates a fundamental tradeoff between versatility and control that all synthesis topologies face. The in-depth comparative study of various sound synthesis techniques, carried out by

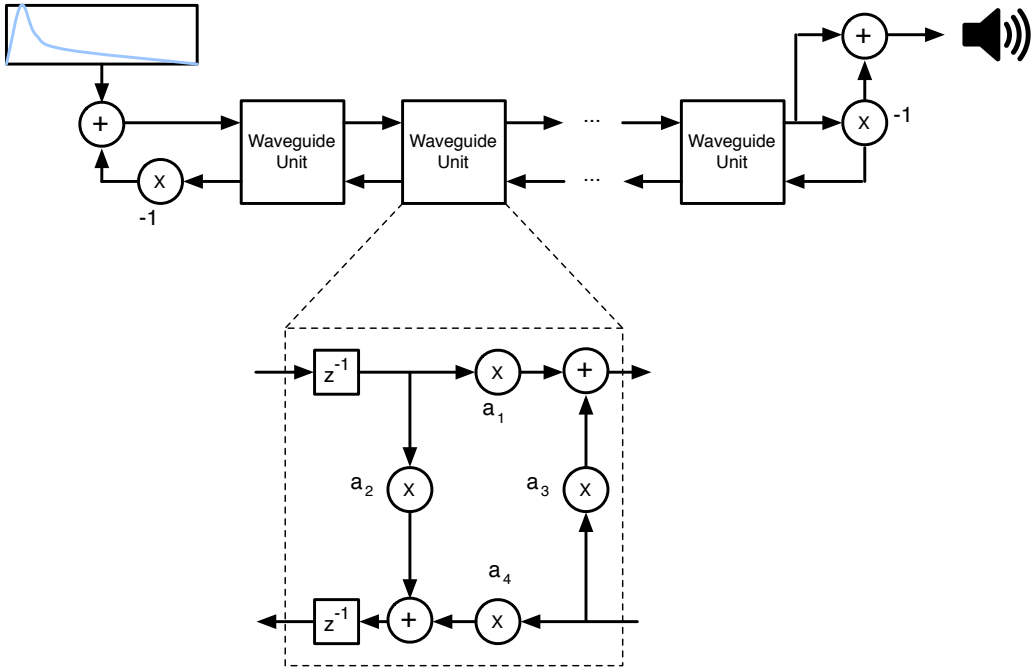

Figure 7: Digital Waveguide Synthesis

Tolonen et al., elucidates the strengths and weaknesses of each algorithm in regards to this tradeoff as well as practical issues related to implementation. However, one practical issue of utmost importance that is not covered in this study is whether each algorithm provides a mechanism by which to realize a specific target timbre. If such a mechanism does not exist, the resultant algorithm would be, at best, as useful as an instrument capable of producing any pitch and offering fine pitch control, but lacking deterministic means to produce a specific desired pitch. Chafe, in presenting a historical perspective on the many ways synthesis has influenced and benefitted composition over the

last fifty years, writes that, historically, imitation has been one of the main uses of sound synthesis since its inception (1999, p. 2). Thus, the importance of being able to achieve a specific desired timbre cannot be overstated.

For sample synthesis, one only needs to record the desired sound and store it for later retrieval. However, as previously stated, as the number of desired timbres increases the storage requirements become difficult to manage and therefore more sophisticated methods are required. For the other synthesis algorithms discussed, there are unfortunately no obvious ways to extract the precise parameter set for a desired timbre. A body of research has emerged specifically to develop methods for this task. These methods are generally referred to as 'parameter estimation' or 're-synthesis' techniques.

2.3 Parameter Estimation

Estimating synthesis parameters for a target timbre is a difficult problem (see Figure 8). Johnson and Gounaroloulos (2006) write that that it is not possible to find an appropriate mapping from the input parameters to a desired result unless "[users] have a very strong understanding of the underlying mechanisms that produce the sound, or a large amount of 'trial-and-error' experience with generating timbral changes within a system" (p. 1). However, even with both strong understanding of a sound synthesis algorithm's sound producing mechanisms and extensive experience using the algorithm to explore

timbre, the ability to realize a particular point or region in timbre space can be time-consuming at best, and is often infeasible, as discussed in Heise et al.'s paper on parameter estimation (2009, p. 1). The requirement that a composer be versed in signal processing theory and, quite often, computer programming in order to even begin to search efficiently in timbre space is not ideal. Mitchell and Creasey (2007) note that such a requirement often leaves the composer "more concerned with the scientific process [driving the algorithm] than artistic creativity" (p. 1).

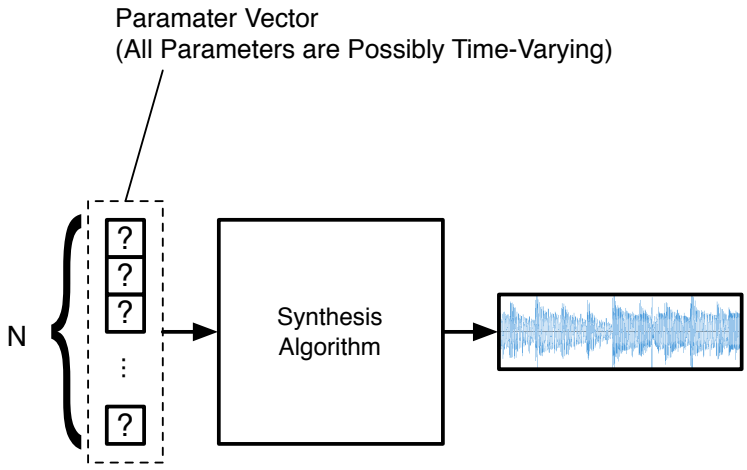

Figure 8: The Task of Producing a Target Timbre

As opposed to leaving the task of target timbre discovery to the composer, re-synthesis and parameter estimation techniques attempt to automatically fit the parameters of a sound synthesis algorithm to imitate a target sound (see Figure 9). Specifically, re-synthesis uses the results of a target

signal's analysis to fit parameters to an underlying synthesis algorithm. Once a particular parameter fitting has been found, the user is left to explore the space around the target using the underlying synthesizer. This means that a re-synthesis technique paired with a specific synthesis algorithm will combine to produce a system that suffers from the same pitfalls associated with the algorithm. It is therefore useful to categorize re-synthesis techniques based on their underlying synthesis algorithms, so that one may compare these techniques within the proper context.

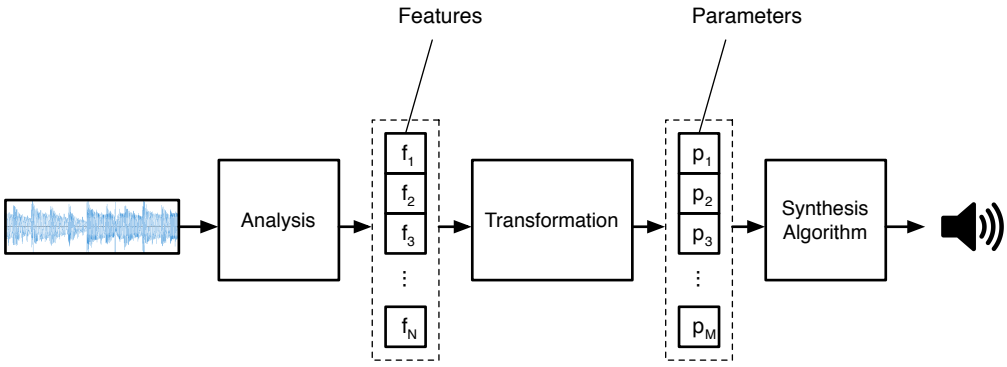

Figure 9: Re-synthesis Approach

An in-depth comparison of some of the most ubiquitous re-synthesis techniques can be found in Klingbeil's dissertation (2009). However, one must look beyond this source in order to evaluate promising state-of-the-art techniques from each synthesis category proposed by Smith (1991).

One of the most popular re-synthesis techniques utilizing a sampling

synthesis method is concatenative synthesis (see Figure 10). Concatenative synthesis, as described by Diemo Schwarz, "use[s] a large database of source sounds, segmented into units that match best the sound or musical phrase to be synthesized, called the target" (2006, p.1). In principle, concatenative synthesis can be used to match any type of feature (e.g. pitch or loudness). To differentiate timbral matching from other objectives, the term "musical mosaicing" is used, a process first developed by Zils and Pachet (2001).

In order to determine the best timbral match in the database to a given target one must rely on a perceptual distance measure. Deriving such a measure is complicated (as will be discussed later) (Schwarz, 2006, p. 13). In addition to searching for the best matching units, musical mosaicing also typically constrains the sequencing of these units to ensure a smooth timbral development (Zils & Pachet, 2001, p. 1). In order to achieve accurate re-synthesis, a large database of units is required to meet both local and global constraints while providing a high-fidelity solution. As the size of this database increases an efficient search becomes difficult. The length of this search may not be problematic for a composer interested in matching one particular target, but if timbral exploration around the target is desired, the search for a new sequence for every slight timbral variation becomes a prohibitive bottleneck (Schwarz, 2006, p. 11). If discontinuities between units and imprecise unit matching is acceptable however, such a search can usually be run in realtime.

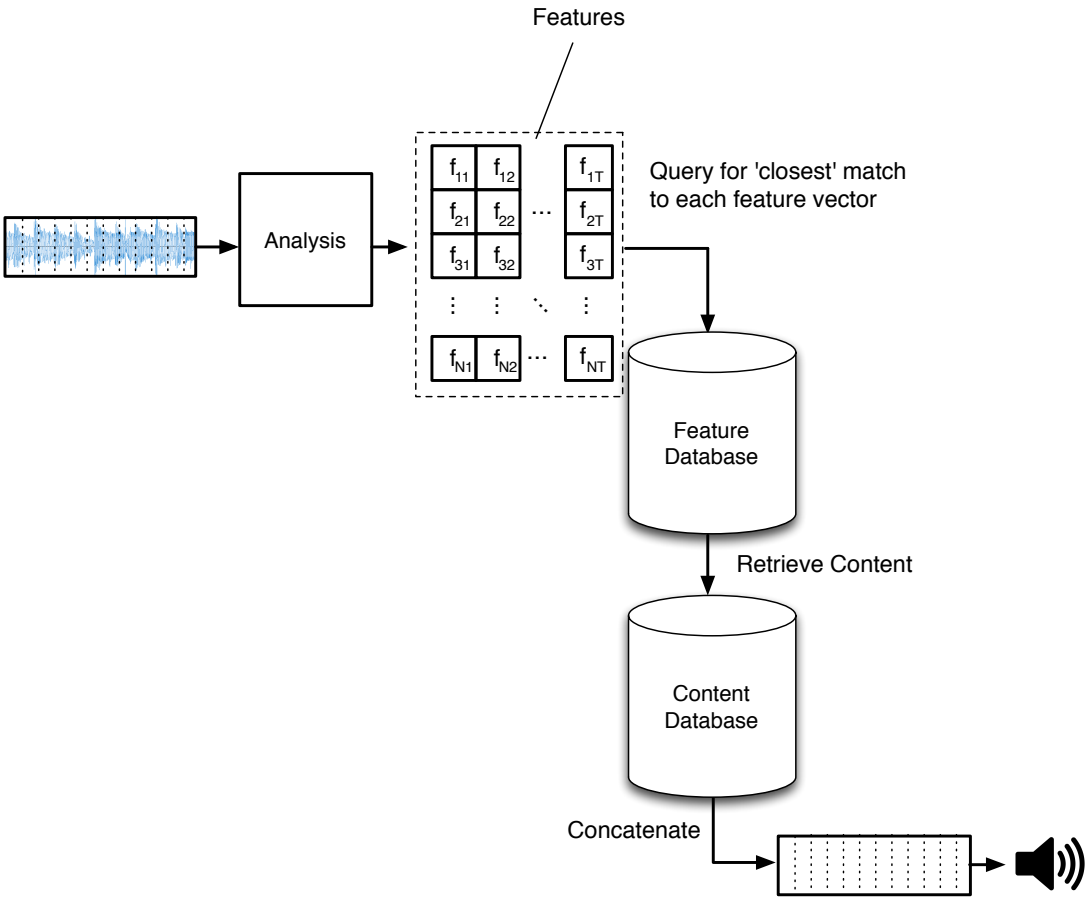

Figure 10: Concatenative Synthesis

For example, Bonada and Serra (2007) have developed a system where they pre-compute timbral features over an extremely large database of vocal sound units that sufficiently cover their vocal feature space, and allow a user to map performance trajectories to that space, retrieving the closest units to that trajectory, and concatenating them to produce an output signal. They call this ‘performance sampling’ (p. 67).Their system is able to generate a wide variety

of vocal timbral evolutions. However, as noted by the authors, the storage requirements are large in order to model only a small region of timbre space and they have yet to determine a way to limit undesirable discontinuities between 'breathy to nonbreath connections' (Bonada & Serra, 2007, p. 78).

An interesting granular synthesis re-synthesis techniques was proposed by Johnson (1998). He provides a system to the user that allows them to score a number of randomly generated output sounds based on their similarity to a user-defined target. These scores are used to 'push' the parameter set corresponding to each generated output towards the target sound, and the user scores the results again. This process continues until the target is reached (p. 2). The system works well as long as the user commits to the process, but because it is not completely automated, it is not an ideal solution.

There are several popular re-synthesis methods that employ spectral models, which fit the parameters of additive synthesizers. Of these, the most popular are the Phase Vocoder and Spectral Modeling Synthesis

The phase vocoder (see Figure 11) was developed at Bell laboratories in 1966 and first introduced to the music community as a re-synthesis tool a decade later by Moorer (1978). It makes use of the spectral representation of a target sound provided by a short-time Fourier transform (STFT). This technique uses the phase spectra returned by the STFT to provide better estimates for the frequency values associated with each bin. These values along

with the magnitudes and phases corresponding to each bin can be used to set the parameters of an additive synthesizer. If the sound is harmonic or quasi-harmonic (i.e. contains little noise), then one can convincing re-synthesize the sound using a small number of peaks for a more efficient implementation.

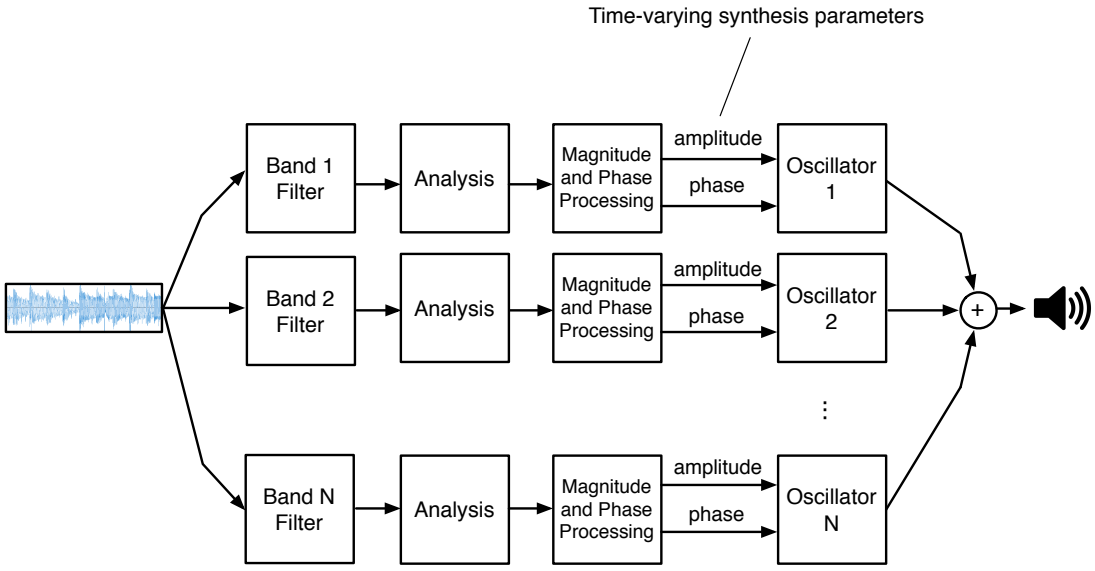

Figure 11: Filter-bank Implementation of the Phase Vocoder

There are two main problems with the phase vocoder. First, transients are often very difficult to model accurately. Robel has been working on this problem for a number of years (2003, 2010) and made good progress, but notes that it is still an unsolved problem (2010, p. 1). Second, the phase vocoder performs poorly on “inharmonic sounds with deep vibrato” due to its inability

to track frequency components across bins and its difficulty in efficiently modeling noise (Serra & Smith, 1990, p. 13). This second problem was the impetus to the development of Spectral Modeling Synthesis (SMS) by Serra and Smith (1990).

Spectral Modeling Synthesis (SMS) (see Figure 12) "models time-varying spectra as a collection of sinusoids controlled through time by piecewise linear amplitude and frequency envelopes [the deterministic part] and a time-varying filtered noise component [the stochastic part]" (Serra & Smith, III, 1990, p. 12). The deterministic portion contains sinusoids that are able to change in frequency, via partial tracking methods, but as stated, these changes are represented by piecewise linear functions, which "affects the generality of the model" (Tolonen et al., 1998, p. 31). Transients are also difficult to model. This has led to the development of an extended technique called Transient Modeling Synthesis (TMS), which "provides a parametric representation of the transient components" (Tolonen et al., 1998, 33). However, TMS requires the accurate segmentation of transients from a given signal, which is a difficult problem in its own, as discussed in (Ciglar, 2009, p. 15). Additionally, Serra and Smith note that:

> the characterization of a single sound by two different representations may cause problems. When different transformations are applied to each representation [which is

> common in order to transform the target sound], it is easy to create a sound in which the two components, deterministic and stochastic, do not fuse into a single entity (1990, p. 23).

Klingbeil adds that "partial tracking becomes particularly difficult in the presence of noise, reverberation, or dense polyphony" and also that SMS "requires a number of input parameters [that] can have a significant effect on the quality of the analysis" (2009, p. 42).

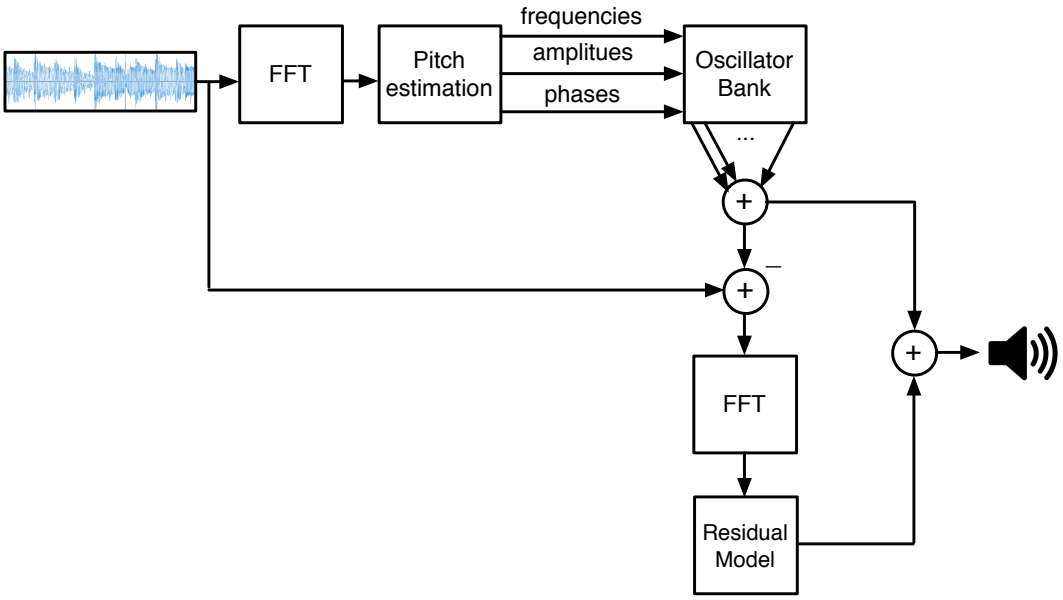

Figure 12: Spectral Modeling Synthesis (SMS)

While there has been less work produced on re-synthesis methods that overlie physical models, and abstract algorithms like FM Synthesis, some interesting research has developed in the last decade.

Vercoe, Gardner and Scheirer point out that physical modeling (e.g. digital waveguide) parameter estimation 'has a particular advantage over equivalent estimation for additive, FM, or other abstract synthesis models in that the resulting parameter set has a clearly understandable interpretation, which aids in further signal manipulation' (1998, p. 11). This is due to physical modeling's low-dimensional and intuitive parameter set. Bensa, Gipouloux, and Kronland-Martinet estimate the parameters for a piano hammer-string model by analyzing a time-frequency representation of a recorded piano tone. Because the mapping of the time-frequency representation to the parameter values is complex, the authors used nonlinear optimization techniques—specifically simulated annealing, which will be discussed later—to search through the parameter space for the appropriate parameter set (2005, p. 499). A similar study was performed by Riionheimo and Vallimaki, who used nonlinear optimization techniques to estimate the parameters of a plucked string physical model (2003). Parameter estimation for physical models is a promising area of research for recreating the sounds that the models are designed for, but due to the high specificity of each model, these techniques will not be applicable outside of a small region of timbre space.

Techniques developed for re-synthesis based on frequency modulation are often referred to as "adaptive-FM" or "FM-matching" techniques (see Figure 13). The first researchers to successfully re-synthesize sounds using FM

Synthesis were Horner, Beauchamp, and Haken (1993). Similar to physical modeling, the relationship between FM synthesis' parameter space and its output's time-frequency representation is unclear. Thus, Horner et al. also made use of a nonlinear optimization technique, a genetic algorithm (GA), to search for an optimal parameter set given a target. In these initial experiments, the parameters were not allowed to vary over time, limiting the applicability of the system (Horner et al., 1993, p. 22). This system has been extended via a number of studies outlined in Horner (2003). One of the more successful FM-matching systems, developed by Mitchell and Sullivan, matches time-varying FM synthesis parameters to various complex sounds using GAs (2005). The FM topographies used were allowed to be more complex than one with a single modulator and carrier signal (e.g. the *double modulator* design uses the sum of two modulators to vary the carrier signal's frequency).

2.4 Meta-synthesis

The previous discussion of the most popular re-synthesis methods had a common theme: each re-synthesis technique discussed is used to fit parameters to one specific synthesis algorithm. As Garcia points out (and as reiterated in (Tolonen et al., 1998, p. 103)), "it is known that different sound synthesis techniques perform better with different types of sounds" (Garcia, 2000, p. 1). Therefore, depending on the target sound, "some parameter matching

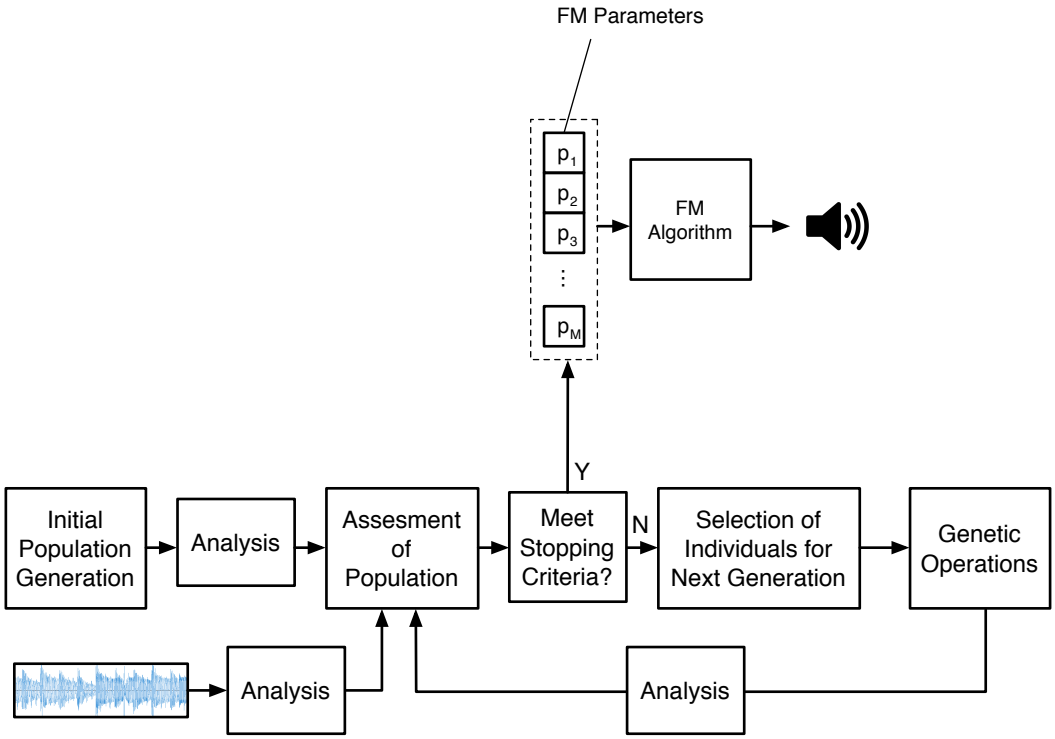

Figure 13: Adaptive FM using Genetic Algorithms

techniques can give poor results when using a fixed topology" (Garcia, 2000, p. 1). In other words, a re-synthesis technique that rests on top of a specific synthesis algorithm will inherit its undesirable features and therefore will be optimal for only certain types of sounds. One solution, as proposed by Misra and Cook (2009), is to use a different synthesis algorithm (and therefore a different re-synthesis technique) for different kinds of sounds, so that each sound is generated by an algorithm that best suits it (see Figure 14) (p. 1). However, knowing which algorithm to choose is often not obvious and,

therefore, placing the burden of this choice on the composer is not ideal. In their concluding comments, directly related to this issue, Misra and Cook write that a way to relieve this burden would be to "present the entire range of [synthesis] techniques to the machine and let it decide which to use on-the-fly" (2009, p. 5). Such a system would require the machine to 'learn' which synthesis algorithms (and corresponding re-synthesis techniques) are best suited for which kinds of sounds. However designed, the learning machine used would have to be trained on enough timbral data to sufficiently blanket timbre space. Also, one must develop a metric by which to measure how 'suited' each algorithm is for each sound, so that one may be assigned the 'winner.' If such a process were possible, the winners of each point in timbre space would aggregate their points into 'regions' within which they lay claim as the appropriate synthesis technique to use. However, the issues of generating sufficient training data and measuring suitability of not just the synthesis algorithm, but also its paired re-synthesis technique (for a given point in timbre space) are not trivial.

Instead of having to generate and provide enough timbral data to blanket timbre space, Puckette (2004) proposes an alternative method for designing such a system (see Figure 15). First, the author defines an 11-dimensional objective timbre space by segmenting a spectrum's magnitude into 11 bands, calculating the loudness in each, and then transforming the result so that these

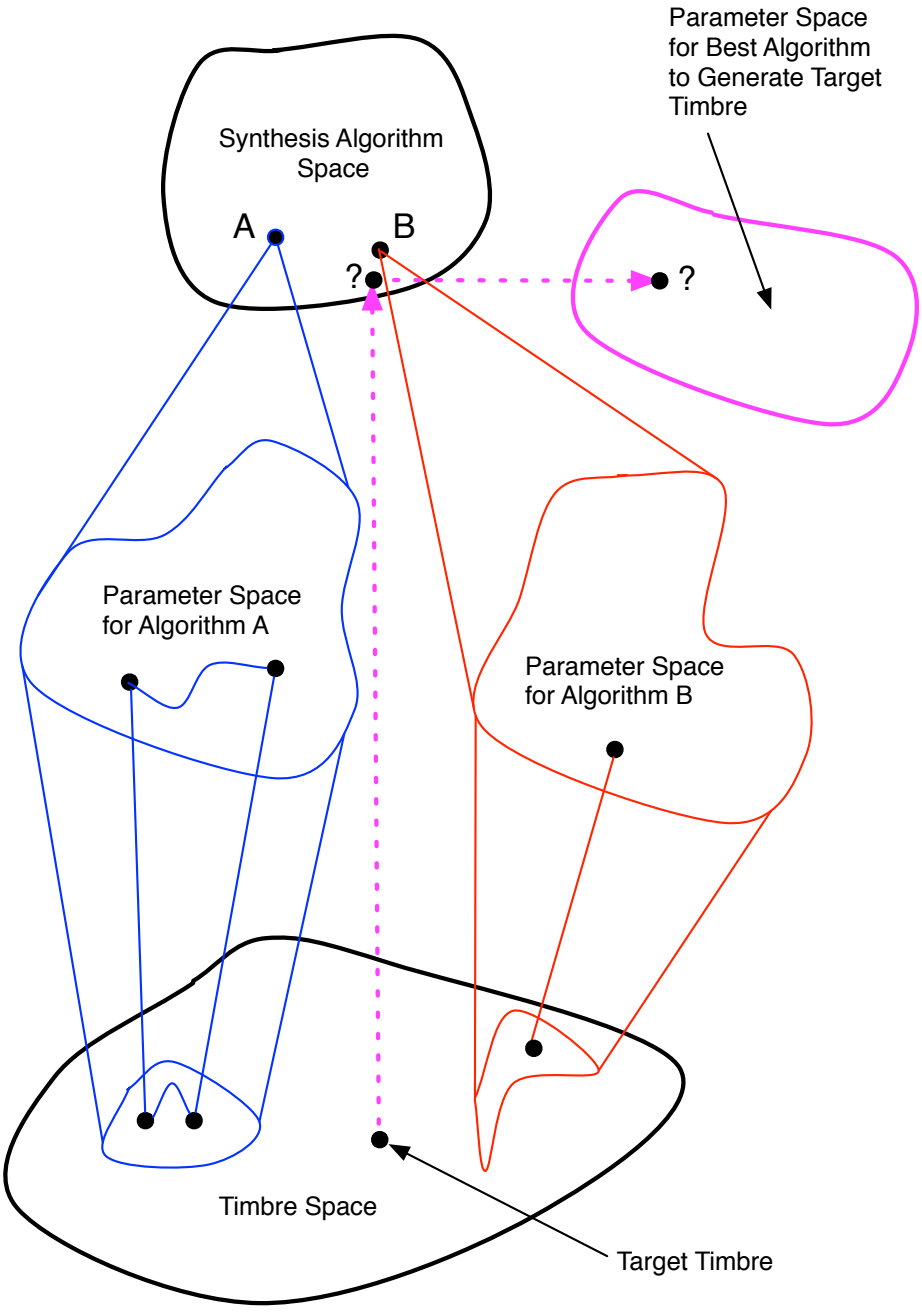

Figure 14: The Meta-Synthesis problem

values are decorrelated (p. 1-2). He then uses a specific synthesizer to generate points in this timbre space by sufficiently blanketing the input parameter space. When a target sound - corresponding to a path in the timbre space - is provided to his system, it re-synthesizes the sound by finding the synthesizer's nearest neighbors (using Euclidean distance) to each point in the target path, determining the 'smoothest' trajectory through these nearest neighbors, and mapping this trajectory back to parameter space to determine how the synthesis algorithm can best produce the target (p. 3). Puckette notes that since his system was initially produced to force a specific synthesizer to produce the target, there may be no nearby synthesis points to the target curve (p. 3). However, one could easily extend Puckette's system by using a number of different synthesizers to produce points in timbre space and then, for a given target, constrain the match trajectory to pass only through points of one synthesizer, so that the resultant parameter set would correspond to the synthesizer which 'best fits' the target. Note that this extended version of Puckette's system would also separate the suitability of the synthesis algorithm from its specific re-synthesis technique, because the re-synthesis process would be the same for all synthesis methods. Therefore, the proposed extended version of Puckette's system would solve many of the issues previously posed when developing a learning machine able to intelligently choose between synthesis algorithms. However, it is not without its own set of problems. First,

it is not clear that euclidean distances in Puckette's objective timbre space are semantically meaningful. Second, there is no indication for when one has sufficiently blanketed the input parameter space of a given synthesizer. Third, a smooth trajectory of points in timbre space does not necessarily lead to a smooth trajectory to points in the parameter space. Since trajectories are drawn through a finite set of points in the timbre space, one must determine how best to interpolate between parameter sets in the input space, which, depending on the mapping, could produce wild fluctuations between points back in timbre space. Fourth, no matter how many synthesizers are provided to the system, it is not guaranteed that all points that are mapped from the various parameter spaces to this timbre space will have close neighbors. Lastly, it is not necessarily the case that the match trajectory will correspond to the 'best suited' synthesis algorithm, because many algorithms have high-dimensional and non-intuitive parameter spaces, and therefore, while they may provide a best fit to the target trajectory, they may provide more a complicated exploration than another algorithm with the next best fit.

Loviscach's work (2009) on making synthesizer control feel more natural for synthesizers with large parameter spaces provides a solution to this last problem. In his work, Loviscach studies correlations between parameter sets for a given synthesizer based on a database of preset data. If two different parameters are highly correlated over a broad range of parameter presets, then

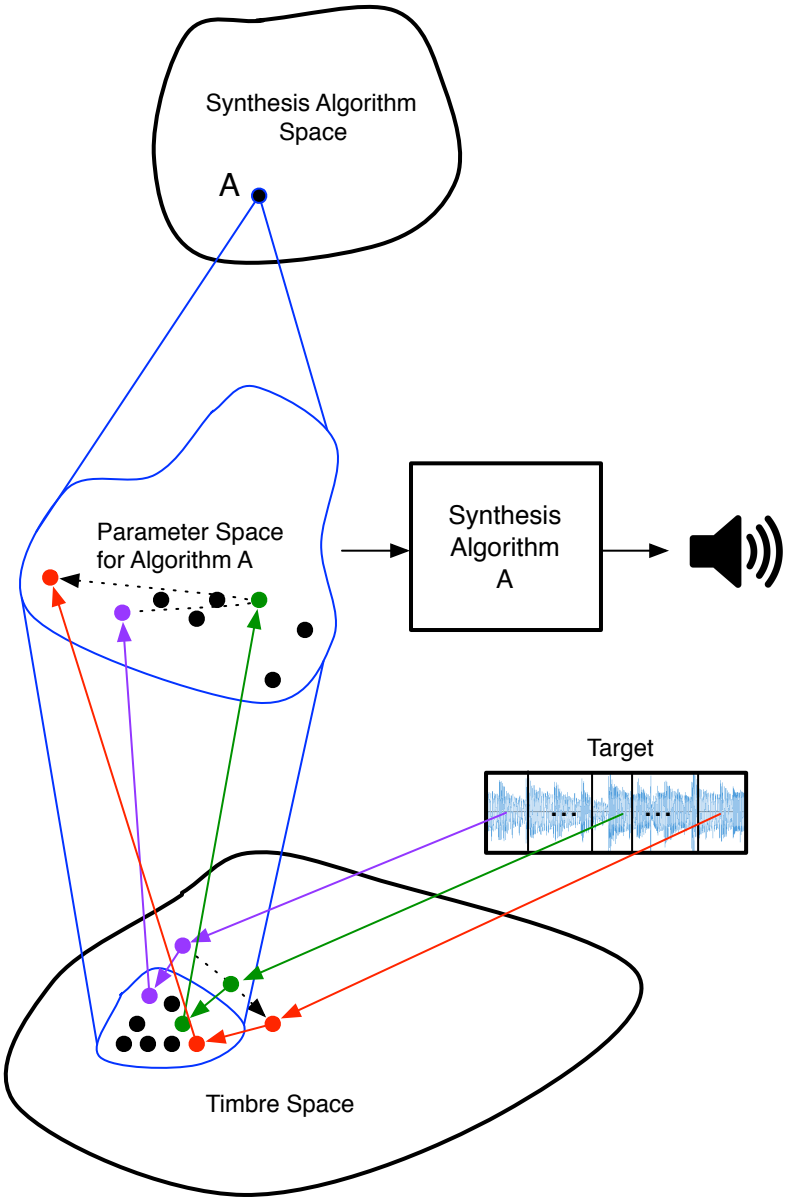

Figure 15: Finding the closest parameter mappings in a partially canvassed objective timbre space.

adjusting one will adjust the other (p. 1). Correlations are found between all pairs of parameters and placed in a two-dimensional field such that highly correlated parameters are placed next to each other. The manipulation of any one parameter will affect all others according to their distance and joint statistics (p. 1). Therefore, assuming that the provided presets for a given synthesizer correspond to a synthesizers natural usage, this system is able to replace a synthesizer's high-dimensional non-intuitive parameter space with a more natural low-dimensional one. Of course, this requires that enough presets exist for each synthesizer so that their parameter correlations can be considered statistically significant. In this case, however, a hybrid system employing techniques from both Puckette's and Loviscach's research could be quite interesting. The major hurdle would be having the enormous amount of training data necessary for both techniques to work well in harmony, making such a system impracticle. In theory, however, this hybrid system would allow a composer to provide a target timbre and, in return, would be given a specific algorithm and parameter set that they could then use to explore the surrounding space in a natural way. If further constraints are placed on storage and efficiency, this system would meet the basic requirements of an ideal synthesis tool. However, other authors have suggested that one could do even better than this.

A limitation of the synthesis learning machine proposed by Misra and

Cook (2009) and carried over into the previous discussion of extending Puckette's and Loviscach's systems is that the machine is only able to choose from a finite set of synthesis topologies when selecting a best-fit topology and corresponding parameter set. Suggested as early as 1998 by Vercoe, Gardner and Scheirer and supported by Garcia (2000, p. 2) and Moreno (2005, p. 1), hybrid synthesis methods - those allowing any combination of the 'classical methods' - can provide better solutions to imitative sound synthesis both perceptually and in matters involving controllability (p. 9).

The research of Carpentier, Tardieu, Harvey, Assayag, and Saint-James (2010) exploits this same well-known fact in the domain of acoustic orchestration. They use acoustic instruments as their "synthesis methods" and allow any realistic linear combination of these instruments to generate a given target (see Figure 16). The re-synthesis problem becomes more difficult than simple parameter fitting for fixed topologies, because there is a combinatorial explosion in the number of topologies available. In order to determine an appropriate combination of instruments and associated playing techniques, they map individual instrument features (obtained from steady-state time-frequency representations) to an objective timbre space, make assumptions about how these features interact in the presence of polyphony, and perform a search over all possible combinations for that which best achieves the target (p. 2). The authors limit their system to steady-state sounds and suggest that a way to

transition between timbres is to do so incrementally, with one instrument dropping out or modifying its state at a time. Such transitions would be limited in how "smooth" one perceives them to be as well as how quickly they may occur. However, a system utilizing this idea and based on linear combinations of synthesis algorithm's may be able to avoid this problem.

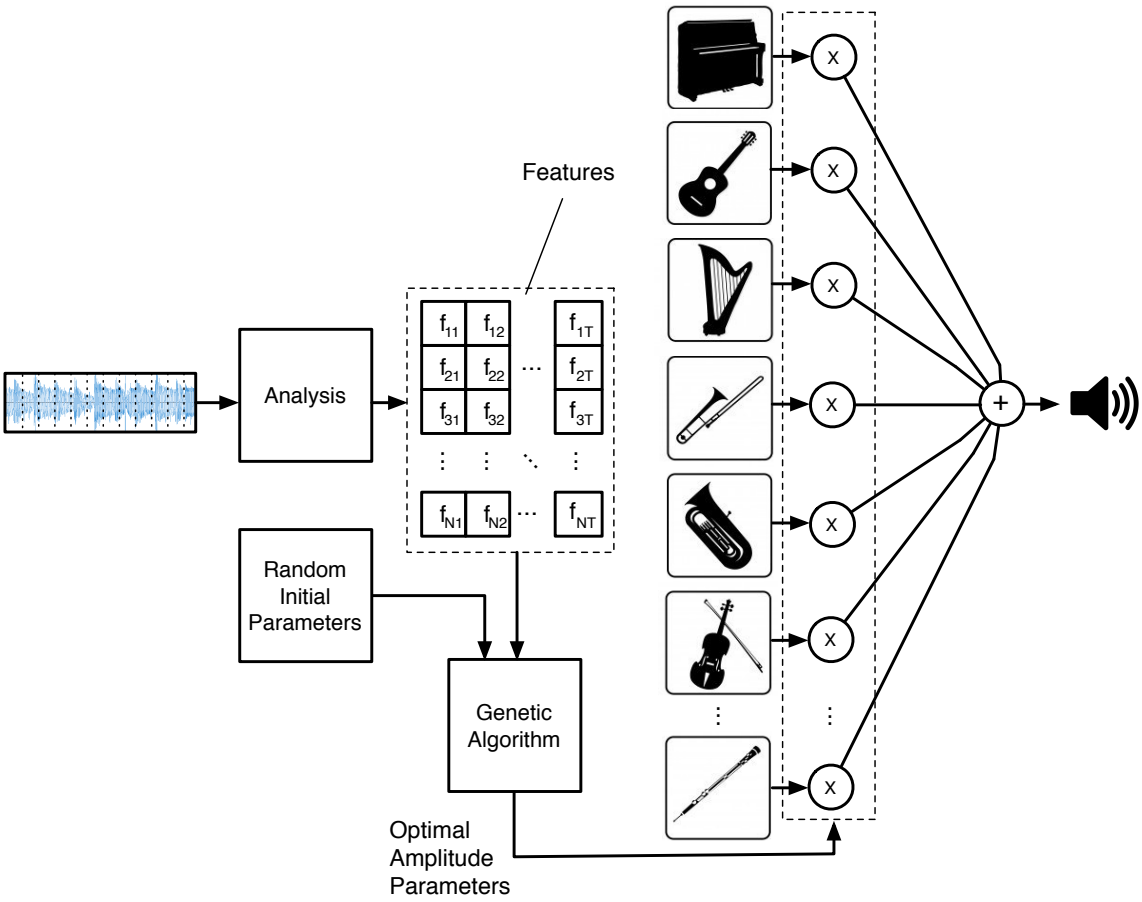

Figure 16: Approximating a complex timbre using a linear combination of orchestral instrumentation.

Carpentier et al. (2010) designed their system specifically as an auto-orchestration mechanism for given target sounds, constraining their search to a realizable orchestral combination. This constraint need not be applied when combining sound synthesis algorithms. It is also not necessary to consider typical sound synthesis algorithms as "atomic" units, as it is in the acoustic instrument case. By removing this limitation, one may be able to generate better-suited synthesis algorithms for a given target using building blocks at a different level of abstraction. Replacing high-level blocks with the lower-level components that make them up, one would be able to generate solutions at least as fit as any of those generated using the higher-level set, and possibly better (Garcia, 2000, p. 2). However, the search space becomes larger as the atoms become lower-level and, therefore, the search itself becomes more difficult. Thus, one must balance the theoretical ability of the system to produce possibly better solutions with the practicality of being able to find one of these solutions. A few studies have attempted to build a system capable of constructing synthesis algorithms, well-suited to a given target, via an intelligent, directed search through synthesis algorithm space (Wehn, 1998); (Garcia, 2000, 2002). Before being able to discuss this research, one must understand the search strategy used in both studies: genetic programming (GP).

2.5 Artificially Intelligent Music Systems

An exhaustive search over synthesis space is not possible, no matter how large the atomic topology unit. Even when this search is restricted over one single topology, the parameter space will often be enormous. Therefore, it is necessary to investigate intelligent ways of searching such a high-dimensional and complex space. In order to develop a directed search technique, one must consider a way to measure how well each point in the input space performs at the given task. The resultant measure is typically represented by an "objective function" that, given a point in the input space, produces a value (often between 0.0 and 1.0) that represents how well the point meets the target objective. The optimal solution to the problem will exist as a global maximum along the objective function's surface (see Figure 17). The closed-form representation of this function is often unknown, making search difficult. However, in some cases, one can make reasonable claims about the shape of the objective surface.

The artificial intelligence community has developed various intelligent search algorithms to address such problems, which direct the search based on hypotheses about the objective surface's shape (see Figure 18), and/or restrict the region of search based on known characteristics of the desired solution (Russell & Norvig, 2009, p. 64-108).

If one can assume that the objective surface is unimodal and smooth,

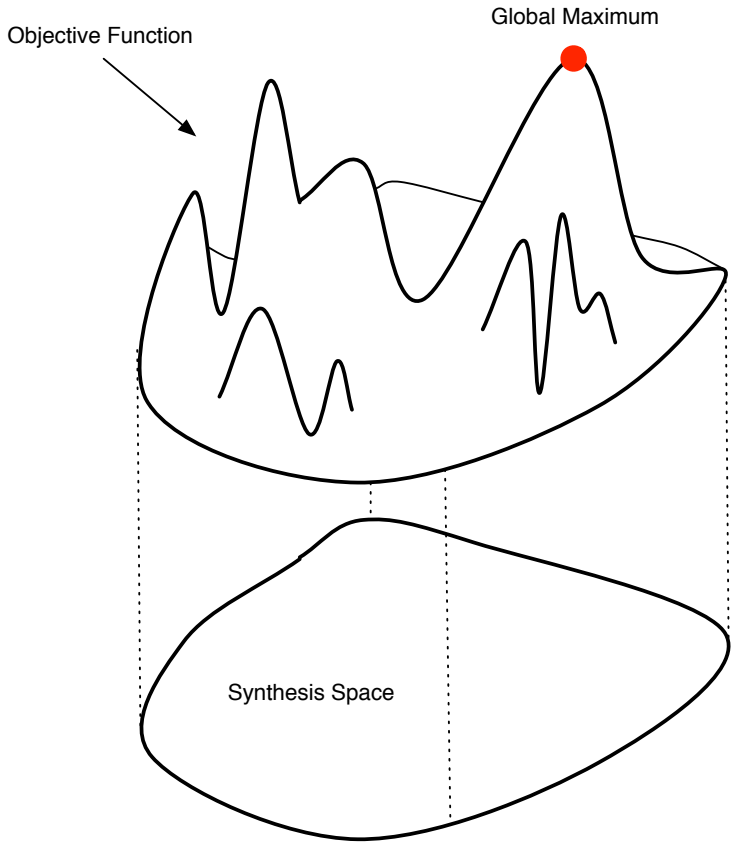

Figure 17: An objective function/surface sitting above synthesis space with its global maximum shown.

then the most efficient search algorithm is called hill-climbing, which simply looks at all neighbors of the current position in the search and moves in the direction of greatest increase in the objective function. However, if the surface is not unimodal, then this algorithm has a chance of converging to a local maximum, which is undesirable. Teller writes that hill-climbing is a fully-exploitative search algorithm, meaning it "focuses the search in the areas

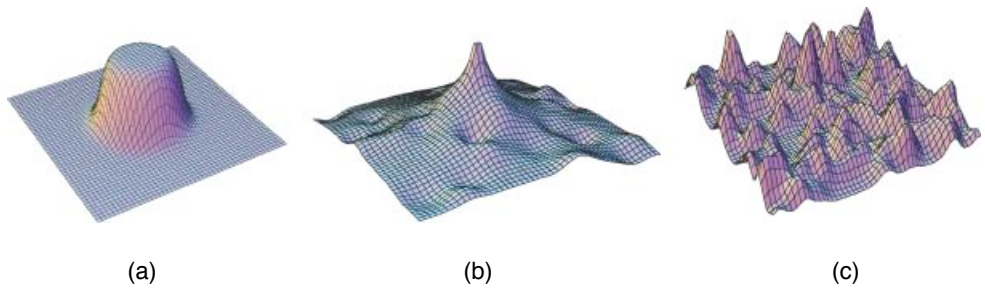

Figure 18: Three different fitness surfaces: (a) unimodal and smooth, (b) multimodal, but relatively smooth and with a clear global maximum, and (c) multimodal and rough with no clear global maximum.

of the search space that have the highest known [objective] values" (1998, p. 23). He explains that "in search there is sometimes a trade-off between exploration and exploitation" whereas opposed to exploitation, "exploration means trying out options that are believed to be locally sub-optimal (in hopes that globally these options will lead to an improved solution)" (1998, p. 23). If no assumptions about the shape of the objective surface can be made, then one must find a good balance between the exploitation and exploration in search so that premature convergence is unlikely to occur. Search strategies with this balance are often employed—and are known to be most productive—in cases where the problem domain is not well understood, and one does not know a priori about the structure of the solution, as is the case with the synthesis problem that we are facing (Vanneschi, 2004, p. 42). Thus, it is not a

coincidence that variants of genetic algorithms (GA)—a strategy that provides mechanisms to directly control the amounts of exploitation and exploration—have been successfully utilized in many of the paremeter estimation studies listed above (Horner et al., 1993); (Johnson, 1998); (Riionheimo & Valimaki, 2003); (Mitchell & Sullivan, 2005). The degree to which GAs have helped solve the FM matching problem has led Horner, one of the initial pioneers of FM matching, to proclaim that it has helped push FM matching "into something of a renaissance period" (2003, p. 28).

As described in his dissertation on making variants of GAs more efficient, Vanneschi writes that GAs "can be imagined as operating via a population of explorers initially scattered at random across the landscape. Those explorers that have found relatively high fitness points are rewarded with a high probability of surviving and reproducing" (2004, p. 70). (Note that when discussing GAs, the objective surface is often referred to as the "fitness" surface and the process of search is called "evolution".) The "pull" strength of explorers from regions with low fitness values to regions with high fitness values is determined by the GA's parameters, allowing one to specify in what ways the algorithm exploits and explores. The ability to search many different regions of the input space in parallel increases the probability of finding many local optima on a complex multi-modal fitness surface and then selecting the best option between all of them (Garcia, 2001, p. 37). For an visual example of

how a GA might obtain a near-optimal parameter set, see Figure 19.

Since it is known that for at least several fixed topology parameter estimation techniques (e.g. FM matching, physical model parameter estimation, granular re-synthesis), the objective surfaces are best searched via genetic algorithms, it is very likely that the objective surface for the more complex problem of simultaneous topology and parameter estimation will also be best searched using genetic algorithms. However, classical genetic algorithms can only be applied when the size and shape of the solution is known beforehand (Vanneschi, 2004, p. 42). In the search for an appropriate synthesis algorithm topology, the size and shape of the ultimate solution is a large part of the problem. Research into more sophisticated search methods, which are better able to search over algorithm space is the subject matter of Automatic Programming.

"Automatic Programming is one of the central goals of computer science...[the goal being to make] computers perform required tasks without being told explicitly how to accomplish these tasks" (Koza, Bennet III, Andre, Keane, & Dunlap, 1997, p.3). Ideally, in an Automatic Programming system, requirements provided by the user only specify what the intended behavior of the program is and not how it should produce that behavior. In a paper overviewing the many different approaches to Automatic Programming, Rich and Waters (1992) state that Automatic Programming has three goals: to make

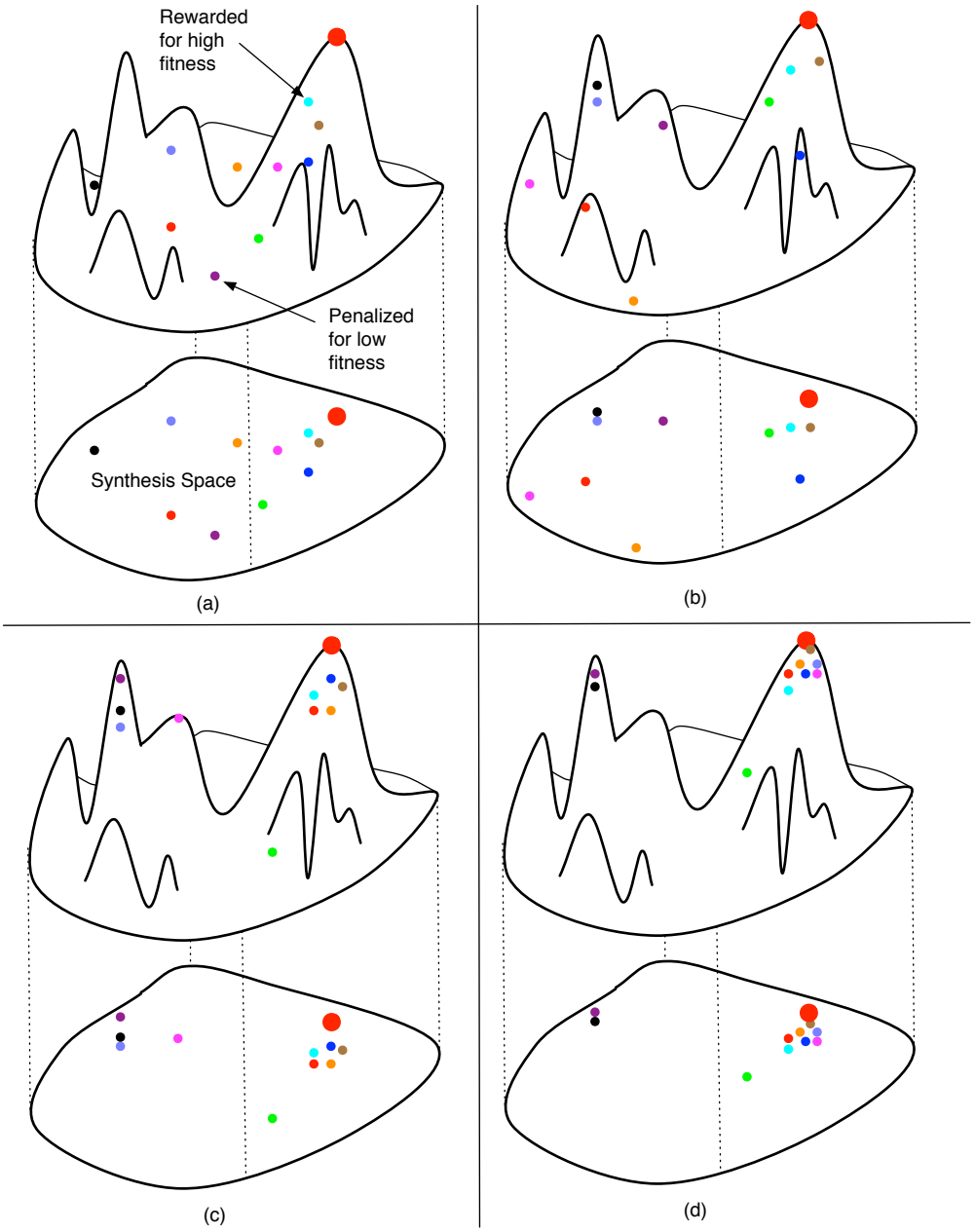

Figure 19: A GA population after: (a) 1, (b) 10 (c) 100, and (d) 1000 iterations (and optimal solution notated by large red dot).

a system end-user oriented, general purpose, and fully automated (p. 4). All approaches in existence at that time focused on two of these goals at the expense of the third. Rich and Waters split these approaches into three categories: bottom-up, narrow-domain, and assistant (p. 3). Bottom-up approaches sacrifice the user-oriented goal and result in high-level programming languages that are general purpose and fully automated (i.e. they automatically generate machine code), with a goal of becoming "very high level in the future" (p. 3). Narrow-domain approaches focuses on a narrow domain, but is end-user oriented and fully automated, with the goal of becoming wider-domain in the future (p. 4). The assistant approaches lead to systems that are user-oriented and general purpose, but not fully automated (e.g. integrated development environments (IDEs)) (p. 4).

The meta-synth problem suggests the need for a narrow-domain system (only designed for generating synthesis algorithms) that is end-user oriented (where the end user only needs to specify a target output) and fully automated (the synthesis topology and optimal parameter set are found and returned). These approaches typically frame the Automatic Programming problem as one of intelligently searching through algorithm space in order to find an algorithm able to produce the desired output, which falls in line with how this paper interprets the problem. The most common intelligent search strategy though algorithm space is, unsurprisingly, based on genetic algorithms and is called

genetic programming (GP) (Koza, 1992). In fact, genetic programming is so ubiquitous in Automatic Programming research that da Silva—in his dissertation on GP—mistakenly defines GP *as* 'the automated learning of computer programs' even though it is actually a subset of Automatic Programming (2008, p. ix).

GP can be thought of as "variable length, tree-based genetic algorithms" (Teller, 1998, p. 29). The search mechanics of GP are analogous to those of GAs, meaning GP also performs a parallel search through the input space for points (re: algorithms and paired parameter values) that meet the specified fitness (re: algorithm output), balancing exploration and exploitation in a similar manner. The individual points in algorithm space are typically represented by tree data structures whose terminal nodes (or leaves) represent input parameters to the other functional node elements (see Figure 20). This representation was first proposed by Koza as a natural way to structure Lisp programs (1992). It has been successfully applied to a number of different programming languages since. However, there are a number of different ways to translate code to this representation as well as a number of operator variants used to perform the search, and as pointed out by Vanneschi, "the art of choosing an appropriate representation and an appropriate set of operators is often a matter of experience and intuition" (2004, p. 6). Beyond choosing the representation and operators, there are also a number of GP parameters that

must be set, and “much of what GP researchers know about these parameters is [also] empirical and based on experience” (p. 32). Progress on systematic ways of selecting specific orientations of these variables has been made recently and will be discussed below.

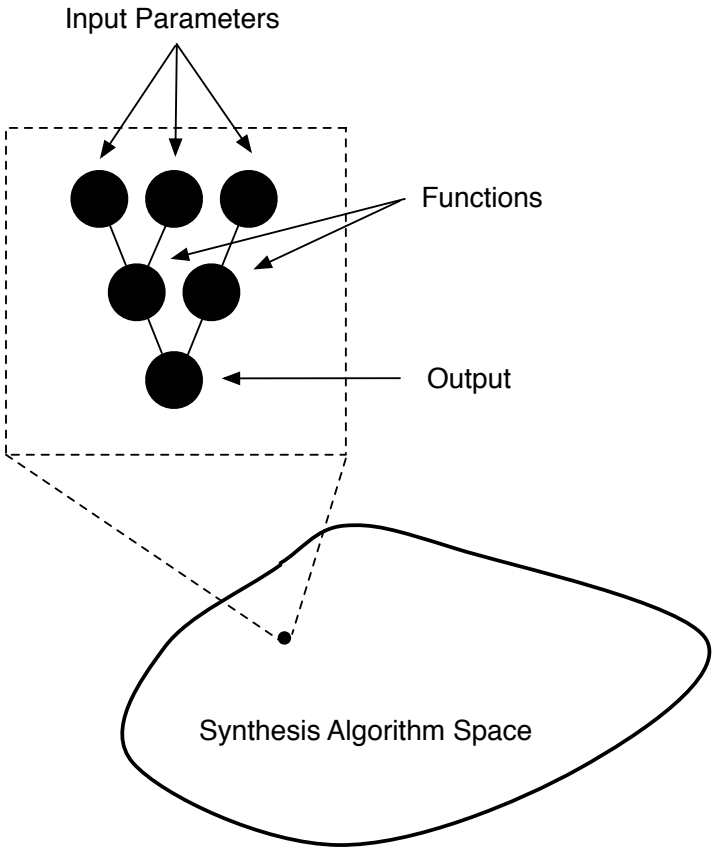

Figure 20: The common Genetic Programming (GP) setup, where a point in algorithm space is represented by a tree.

There has been relatively little research into genetic programming in the arts compared to the hard sciences (e.g. robotics, circuit design) (Hollady &

Robbins, 2007, p. 1). However, there is no inherent reason why this should be the case. In their paper on human-machine interaction, Moroni, von Zuben, and Manzolli note that "human-machine interaction in the context of artistic domains [can be] a framework for exploring creativity and producing results that could not be obtained without such interaction" (2002, p. 185). It is certainly true that a meta-synthesizer would be able to provide a composer with the means to experiment with timbre in ways they would not be able to without it, thus leading them to places of creativity that they would not have been able to explore on their own. However, the first uses of GP in the artistic domain that predated any work on evolving synthesis algorithms were for evolving music-making systems—see (Rowe, 1993); (Spector & Alpern, 1995); (Polito, Daida, & Bersano-Begey, 1997); (Johanson & Poli); (Todd & Werner, 1998); (Costelloe & Ryan, 2004).

As work in GP was being carried out in the music world, there were also advancements in GP for use in signal processing applications not related to music—see (Sharman, Alcazar, & Li, 1994); (Sharman et al. 1997); (Koza et al., 1997); (Miller, 1998); (Uesaka & Kawamata, 2000); (Holladay & Robbins, 2007).

A signal processing application more related to music that has received a lot of attention by GP researchers is automatic feature extraction for various classification tasks. This was predated by GP work on video feature

extraction—see (Harris ,1997); (Teller, 1998).

The first GP system to evolve audio feature extractors was proposed by Conrads, Nordin, and Banzhaf (1998). They used machine-code level GP to extract features for vowel and consonant detection.

Pachet and Roy also used GP to evolve feature extractors for classification tasks (2007). However, their function set contained much more complicated functions than the work of Conrads et al. Their function set includes a Mel-Frequency Cepstral Coefficient (MFCC) calculation, a Fast Fourier Transform (FFT), and a high-pass filter among others. Their work was recently extended by Kobayashi (2009) and Vatolkin, Theimer, and Rudolph (2009).

Taking into account the Automatic Programming research carried out in both digital signal processing and in music composition over the last fifteen years, one may assume that a good amount of research has also been performed in evolving sound synthesis algorithms. However, this is not the case.

There have been a few studies involving sound synthesis evolution using Interactive-GP (where fitness measurements are provided by test subjects) for timbre exploration—see (Dahlstedt, 2001); (Mandelis & Husbands); (McDermott, Griffith, & O'Neill, 2006). However, while each study provides a system that is applicable to any kind of synthesis topology, a fixed topology is chosen for each run of the system. As previously noted, making the user

choose a fixed synthesis topology at the beginning of a run can severely limit the regions of timbre space that are even possible to explore. Additionally, using subjective judgements to steer the search requires that the user stay actively involved during the length of the search. A more desirable solution would replace the need for human fitness evaluation with an objective measure that correlates well. In other words, an objective timbre space would have to be developed where distances are semantically meaningful, so that pairwise similarity tests can be performed by the computer.

Wehn was the first researcher to investigate such a system (1998). In his work, he uses a basic function set comprised of noise, steady-state sinusoids, triangle and square waves, ramp functions, addition, multiplication, and high, low, and bandpass filters from which to generate more complex algorithms (p. 2). The target sounds that Wehn tests his system with do not vary in time and he does not provide a mechanism by which time-varying sounds can be generated, but as a first step towards our goal, his work is important.

In order to assess fitness, Wehn calculates the Euclidean distance between the amplitude spectra of a target sound and those produced by the synthesizer output (p. 2). Thus, Wehn implicitly assumes a timbre space where each point is represented by the elements of a magnitude spectrum. This is a problematic assumption for many reasons, but perhaps the most important being that such a space is extremely high dimensional and therefore suffers

from the “curse of dimensionality” (Powell, 2007). Briefly, this states that as the dimensionality of a space increases, the usefulness of distance measurements decrease. In other words, Euclidean distance in a given timbre space will become less meaningful as the space grows in size. Another problem with using a simple Euclidean distance measure, as noted by Vercoe et al., is that “humans do not measure ‘noise’ or ‘reduction in quality’ of sound in this way” and so distances will not be semantically meaningful even if the magnitude spectrum were low-dimensional (1998, p. 2).

Wehn places a limit on the size of the resultant synthesis topology, so that its parameter set will be low-dimensional and hopefully more intuitive. However, Wehn’s system does not take the other desirable synthesis algorithm features into account (e.g. efficiency, low storage requirements, other aspects of controllability). For example, he does not discuss how a user would gain access nor control the parameters generated for a given target.

Another difficulty with Wehn’s system is that his atomic units are extremely low level. Even basic classical synthesis algorithms would require complex combinations of these units and would therefore require significant evolution. Thus, Wehn’s input space is needlessly high-dimensional, making search in that space more difficult and time-consuming.

Garcia’s research improved on Wehn’s both in fitness representation and atomic function set (2002). By combining Wehn’s function set with more

complex atoms (e.g. variable sine and wavetable oscillators, delays, controlled gain filters, time varying filters) Garcia was able to successfully evolve more complex algorithms, capable of generating time-varying timbres. In his results, Garcia demonstrates his system's ability to independently evolve an FM synthesizer (p. 6). Garcia's system borrows a technique first used by Sharman et al. where topologies are searched for using GP and the optimal parameter set for each topology is separately searched for using simulated annealing (SA) (1996). As described by Bensa et al., simulated annealing "exploits an analogy between the way that metal cools and freezes into a minimum energy crystalline structure and the search for a minimum in a more general system" (2005, p. 501). Basically, what this means is that simulated annealing allows the search for a global optimum to move into areas of lower fitness (in hope of escaping local optima) with a certain probability that decreases over time (i.e. as the search particle 'cools'). Sharman et al. note that simulated annealing, like GP, has proven useful for finding global optima of multimodal functions (1996, p. 1). However, SA is better suited for smoother and less complex multimodal landscapes and provides a more efficient means of search than GP. Thus, by breaking the synthesis space into a topology space (which is not well understood, but likely rough and multimodal) and a parameter space (which is considered to be smooth and multimodal) one is able to search using separate suitable algorithms for each space. The alternative is to have to choose one

search algorithm that may not be optimal for the multi-faceted complexity of the problem at hand. It should be noted that the parameter space will have a different structure for each given topology. Thus, the "breaking up" of the synthesis space into two spaces is really more analogous to searching independently over points in a quotient space (re: topology space) and each point's equivalence class (re: parameter space) in that space.

Like Wehn, Garcia also places a limit on the size of the evolved synthesis topologies, thereby reducing the size of the parameter space associated with the optimal topology found (2001, p. 87). However, also like Wehn, he does not incorporate any of the other desirable properties into the search requirements.

The fitness function proposed by Garcia is more advanced than Wehn's. Garcia reduces the dimensionality of the magnitude spectrum-representation that Wehn uses by incorporating perceptually motivated thresholding to allow only certain bin values to be incorporated in the fitness calculation, based on frequency masking (2002, p. 6). While the reduced dimensionality will make Euclidean distances more relevant in the resultant timbre space, it is not clear that these distances will be semantically meaningful, especially if accumulated over time to measure timbre similarity on time scales greater than the FFT frame size used. For example, this representation will consider two time-shifted or slightly time-scaled versions of the same sound to be timbrally dissimilar. Slight nonlinear time warpings or differences in length are also not

handled well by this representation. Additionally, the specific choice of perceptual thresholding may be better modeled as a soft threshold as opposed to the hard one used. More advanced perceptually motivated distance metrics can be found in (Riionheimo & Valimaki, 2003); (Jehan, 2005). Developing an appropriate fitness function is absolutely crucial in designing an efficient GP system (McDermott et al., 2006, p. 3). Without one, the search process can be led in directions that are not relevant to the problem at hand, thus complicating the search process and often ultimately causing its downfall.

Another potential reason why Garcia's system may not be suited for more complex sounds is that his function set is still quite low-level. For example, it would most likely take a long time to evolve something as common as a reverberation algorithm using his function set, but such an algorithm would be quite useful for generating a number of real-world timbres. Holladay and Robbins show that including such domain specific synthesis modules directly into the function set can make the GP system much more powerful (2007, p. 4). The idea of incorporating such modules into the function set along with low-level topological atoms was first proposed by Koza and has been used widely in GP research (1992).

The above review of using GP to automatically generate synthesis algorithms given a target timbre shows that there is still much research to be done. Specifically, there are three areas of investigation that will make these

systems more powerful, which happen to be the three areas that are most responsible for the successfulness of all GP systems: an appropriate primitive/function set, a well-designed fitness function, and heuristics to refine the search space and thus speed-up the search. In this problem-domain, this translates to specifying an atomic set of topologies that allow for efficient search (by not being too low-level) without preventing optimal solutions (by not being too high-level), an appropriate measure of timbral similarity (re: a better fitness function), and a more thorough treatment of ways in which to restrict the search based on the desirable features of an optimal synthesis algorithm.

2.6 Audio-Implementation Systems

Audio-implementation systems, such as Max, CSound, Supercollider, and ChucK, provide a more technically-oriented composer with the ability to experiment with sound synthesis and audio effect design. Many of the goals these software systems try to meet are in line with the goals of the desired meta-synthesis system we have previously described (Moreno, 2005). In fact, these software solutions can be viewed as meta-synthesis systems that place the onus on the user to search for a specific timbre by combining the atomic building blocks available to them.

A main goal of such systems is to abstract away the low-level audio

programming that would not be beneficial to a composer interested in timbral exploration (Moreno, 2005, p. 1). This has to be balanced however with the goal of providing a user with functional elements that are low-level enough so that they may be combined into topologies able to produce any timbre in a reasonably efficient manner (Moreno, 2005, p. 1). Thus, these systems typically provide functions that are both low-level and useful in many signal processing applications (e.g. sample addition) and high-level modules that would require complex design using only the lower-level functions, like the FFT or reverberation. The balance of the above goals is the exact same balance we face when choosing an appropriate function set for GP to evolve synthesis algorithms. Therefore, we will rely on the many years of development, user-feedback, and re-development of these systems to determine the appropriate level of abstraction necessary for an efficient GP search. If fact, Johnson picked up on this years ago in discussing a possible system by which synthesis algorithms may be evolved by GAs. In noting the suitability of Max as such a system he writes, it would be easy to fit the evolution of Max patches into this framework (1998, p. 6).

2.7 Timbral Similarity

In order to define a better objective fitness function to measure the subjective nature of timbral similarity, we can look to a large body of work

dedicated to the subject.

The reliance of timbre similarity on timbral evolution trajectories is well-supported and fits in line with an objective timbre space representation we have adopted. Toiviainen et al. note that a major component of the perception of timbre and measuring timbral similarity is the time-varying spectral envelope of sound (1998, p.225). This is further supported by Caclin et al. and also separately noted by Ciglar in more recent studies (Caclin et al., 2005, p. 1); (Ciglar, 2009, p. 4).

Therefore, appropriate timbre similarity models will measure similarities (re: sematically relevant distances) between trajectories of perceptually relevant timbre features. Using the results from the subjective timbre space literature, a number of researchers interested in Music Information Retrieval (MIR) have developed such models.

The MIR community has investigated music similarity on a number of different levels. Some research has been aimed at rhythmic similarity (Paulus & Klapuri, 2002), other at harmonic similarity (de Haas, Veltkamp, & Wiering, 2008), and, more recently, structural similarity (Bello, 2009). However, timbre similarity has been a main focal point in the community for a number of years. The quest for an appropriate timbral similarity measure has also seen the development of a number of different timbre features for use within similarity models.

Early work in timbre similarity was aimed at classifying instruments, which is quite appropriate given that work on subjective timbre spaces could alternatively be seen as generating spaces with good discrimination properties for instruments (Loureiro, de Paula, & Yehia, 2000); (Park, 2004); (Timoney, Lysaght, Mac Manus, & Schwarzbacher, 2004); (Zhang & Ras, 2006). This research takes the viewpoint that timbre is related to its sound-production mechanism and therefore develops models that will assign similar timbre to all sounds generated by the same instrument. Many of the features used in the literature are designed specifically for monophonic instrument sounds and do not apply to polyphonic mixtures or more complex timbral evolution (Ciglar, 2009, p. 6). For example, these include features that measure the temporal progression of an instrument's harmonic partials, features related specifically to the attack, sustain, and release portions of a note, as well as strength relations between odd/even partials (Timoney et al., 2004, p. 182). In order to develop a more general timbre feature representation, applicable to both monophonic and polyphonic sounds, the community focused on a more complex problem: genre recognition.

It is important to note that genre recognition systems are primarily interested with timbre on a global time scale (i.e. over an entire song). It is because of this large body of research that Johnson and Gounaropoulos make a distinction between local and global timbre (2006). However, even though

genre recognition systems are not necessarily interested in local timbral evolution (as we are), genre recognition research requires a representation of timbre for complex polyphonic signals, and thus a lot can be learned from such work.

Aucouturier and Pachet were two of the first researchers to investigate a system for genre classification (2002). The basic idea in genre recognition is that each genre has unique timbral properties on the global scale and therefore by measuring these properties given an unlabeled song, one is able to automatically assign it a genre. Aucouturier and Pachet borrow a feature set from the speech community known as Mel-Frequency Cepstral Coefficients (MFCCs) that have now become ubiquitous as timbre features (2002, p. 1). In speech processing research, "MFCCs were developed to model the spectral envelope while suppressing the fundamental frequency" (Jensen, Christense, Ellis, & Jensen, 2009, p. 1). Thus, they are often considered to be a compact representation of the spectral envelope (a typical feature included in the additive definition of timbre) divorced from pitch (a property from the negative definition of timbre). Aucouturier and Pachet model the probability density of all of the MFCCs from a genre using a Gaussian Mixture Model (GMM), which is simply a multimodal (and often multivariate) distribution composed of Gaussian components (see Figure 21). In order to determine the genre of a set of test samples, MFCCs are extracted and their likelihood under each genre

model is calculated. The most likely genre is assigned to the set of test samples (2002, p. 2).

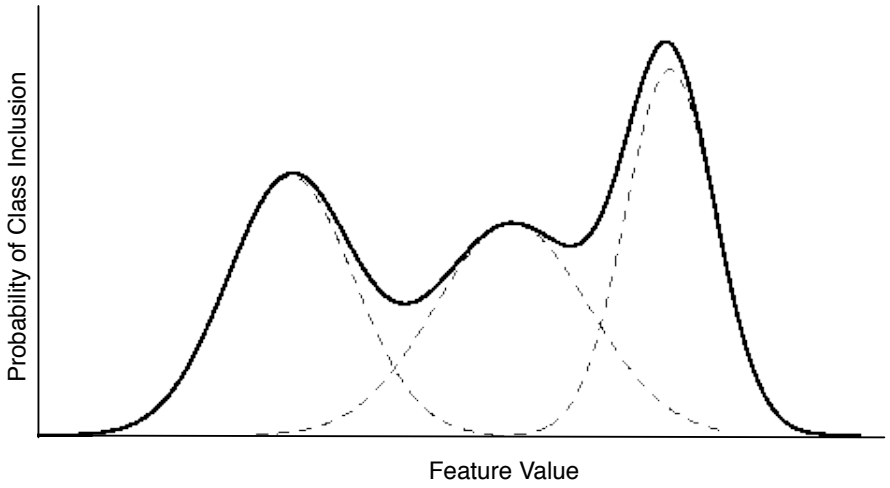

Figure 21: An example of a univariate Gaussian Mixture Model (GMM) taken from (Schwardt, 2005).

Following the work of Aucouturier and Pachet, Pampalk also modeled the timbral content of music with GMMs of MFCCs in his dissertation on sound similarity (2006). However, instead of classifying songs by genre, Pampalk was primarily interested in calculating pairwise similarity between songs. He separately modeled each song with a GMM and calculated the similarity between GMMs using two common distance metrics for probability distributions: Kullback-Leibler (KL) Divergence and Earth Mover's Distance (EMD) (Pampalk, 2006, p. 26). Roughly speaking, KL Divergence (also

known as relative entropy in Information Theory literature) is a measure of how well one distribution approximates another (and is therefore not symmetric) while the EMD is a measure of how "much" one would have to change the shape of one distribution in order for it to look like the other. Determining whether such models are appropriate for pairwise similarity is a difficult task. In order to properly measure the accuracy of a given model, a sound similarity ground truth must be designated.

As noted by Seyerlehner and Widmer,"ideally, a content-based audio similarity metric should approximate the ill-defined 'sounds-like' related for songs" (2008, p. 1). However, Logan, Ellis, and Berenzweig point out that similarity ratings can vary "not only across users, but across time, according to mood and according to context" (2003, p. 1). This is verfied by Pampalk et al. (2008, p. 8). By testing enough users, one may be able to form a meaningful consensus, but user-testing is expensive (both in time and money) and alternate ground truth data is often desired. Therefore, some other labeling is usually accepted (e.g. genre), whether it is provided by experts or by clustering user-provided tags (p. 2). The extent to which this data approximates "sounds-like" is unclear. An even more important question, however, is whether "sounds-like" on the global timescale is truly a measure of similarity between two songs. Many other factors could play a role (e.g. lyrical content, popularity). In Jensen's dissertation, he writes that:

> The inherent problem, that genres are a cultural as much as a musical phenomenon, persists...in our opinion, genre classification as a research topic in signal processing should be abandoned in favor of specialized tests that directly evaluate the improvements of proposed algorithms. The short time features that only capture timbre similarity, or methods using source separation, could e.g. be tested in a polyphonic identification setup that much better shows the capability of the algorithms (2009, p. 11).

Aucouturier and Pachet do note in their 2003 paper that in studying genre recognition, "the problem of the actual perception of timbre is not addressed by current methods" (p. 16). However, this facet of MIR continued to be synonymous with timbre similarity for a number of years and was used to measure the accuracy of systems like Pampalk's (2006). The fact that genre classification results did not see any marked improvements over that time was actually a good thing for timbre research, because it stimulated a wide variety of approaches towards finding better timbre features and incorporating them into better similarity models.

For example, Meng, Ahrendt, and Larsen attempt to improve genre classification results by integrating short-time timbre features, on the order of 30ms, into medium (on the order of 740ms) and long-term (on the order of 9.62s) timbre features in order to better model the temporal evolution of timbre

on longer time scales (2004, p.498). The authors suggest calculating simple statistics, using dynamic PCA (where features are stacked over the desired time horizon and PCA is used to reduce the dimensionality of the resultant feature vector), and modeling the time-varying properties of a sequence of MFCCs by calculating the power spectrum of each (p. 498). They found best results using these latter features, which they called filterbank coefficient (FC) features. These features are able to capture MFCC modulations over a number of modulation rates (with a resolution determined by the length of the time horizon).

Heise et al. use similar features to Meng et al.s FCs (2009). Instead of calculating the power spectrum of each MFCC over a specified time window, they compute a discrete cosine transform (DCT), which decorrelates and compresses the MFCCs spectral data in the first few bins while retaining the characteristics of its shape. The result is a MFCC spectral matrix where each row represents a different MFCC. The authors reduce the dimensionality of this feature matrix by either throwing away the high MFCCs (retaining good MFCC spectral resolution, but smoothing out the signals spectral envelope), or by throwing out the last columns (retaining good signal-spectral-envelope resolution, but smoothing out the MFCC spectral resolution) (2009, p. 3).

Pampalk, Flexer, and Widmer develop fluctuation pattern (FP) features that are derived from a perceptually transformed spectrogram. Again, like

Meng et al. and Heise et al., these authors incorporate the temporal evolution of the features by taking the FFT over each band in their perceptual spectrogram (2005, p. 4). They find that incorporating these features do not improve genre classification performance, but again, this does not necessarily mean that they are not modeling the timbre more accurately (p. 8).

Incorporating the temporal evolution of timbral features over some time horizon using the FFT of each feature over that horizon is an interesting idea that requires more research outside of the domain of genre recognition so that it may be fairly evaluated.

A number of other features have been used alongside MFCCs (on the same time scale) in hopes that additional timbre information will be contained in some or all of them. Most often, due to the curse of dimensionality, feature selection methods are used to find those features that boost performance the most, while filtering out those that provide marginal contribution.

Allamanche, Herre, Hellmuth, Kastner and Ertel start with a set of features including normalized loudness, delta log-loudness, spectral flatness measure, spectral crest factor, real cepstral coefficients, spectral tilt, spectral sharpness, and zero crossing rate along with MFCCs (2002). For feature selection, they perform a greedy search by first finding the single feature that provides best classification, then adding to it the feature that helps it perform best, etc.

McDermott et al. investigate 40 different features commonly used in MIR research and eliminate features based on their redundancy in the presence of others (2005, p. 1). The authors split these features into six groups, based on the type of feature—time domain, Fourier-transform domain, partial domain, trajectory, periodic, or statistical—and find that features within the same group tend to be redundant in the presence of others in that group, an un-alarming result (p. 6).

Kobayashi uses evolutionary algorithms to breed linear combinations of features that provide the best classification performance (2009). His system is based around instrument recognition, but the principle would be valid for any type of classification. As opposed to Pachet and Roy's work (2007) on using GP to breed single discriminatory feature, Kobayashi starts with a general feature set, able to extract information from scalars, vectors, and/or matrices, and evolves the best linear combination of discriminatory features.

Evolving features or proper combinations of features is a natural direction for feature selection. Instead of relying on hand-crafted feature sets and information-theoretic assumptions about what makes a feature "useful", these methods evolve functions that well-suited for their problem-domain. A downside to such methods, however, is that one must evolve a new feature or set of features for each problem.

Another option as opposed to feature selection for timbral similarity is

intelligent feature combination, as proposed by Fu, Lu, Ting, and Zhang (2009). As opposed to selecting a subset of features that improves classification over the entire feature set, feature combination attempts to combine all features in a way that improves performance over any single feature vector.

Seyerlehner et al. take a subtractive approach to feature selection by searching for feature vectors that are not perceptually relevant (e.g. silence) and filtering them out (2009). They note that in modeling a feature vector sequence containing some silent frames using a GMM, the distribution will contain a peak at the silent location in space. Thus, if another feature vector sequences likelihood is computed and it also has silent frames, then a non-negligible likelihood will result no matter how different the non-silent frames are from one another between the two sequences. The authors find that these low energy frames thus contribute greatly to the measure of similarity, which is undesirable (2009, p. 3).

While feature combination/selection methods expand the space of feature possibilities by looking not only at hand-crafted features, but also any linear combinations of them, there is a strong reliance on the assumption that all important timbal information is contained in this hand-crafted feature space and that a combination of hand-crafted features exists that not only contains all timbal information, but also that contains little information about any other musical dimensions (i.e. the feature combination has little noise). A more

direct approach to learning low-noise timbal features is to not limit the feature space to only contain linear combinations of hand-crafted features, but instead to allow more fundamental building blocks from which to generate timbal features.

Humphrey, Glennon, and Bello's (2011) work directly learns timbre features using machine learning architectures known as Convolutional Neural Networks (CNNs). Their approach learns a nonlinear projection from basic high-dimensional spectral representations of audio to a low-dimensional space where distances are semantically meaningful in the sense that two points that are close together will be considered timbrally similar and two points that are far apart, dissimilar. Their work focuses on the instrument recognition task and uses an iterative learning process to ensure that two sounds that are transformed by the same projection that come from the same instrument resolve down to points close in timbre space. The nonlinear projection imposed by the resultant CNN can be viewed as feature extraction, where projected points reside in a feature space that is representative of timbre.

Determining the most appropriate objective timbre representation is still an unsolved problem. While MFCCs are ubiquitous as timbre features, BOF approaches (requiring feature selection or combination) and early fusion techniques have potential to improve upon this representation. However, one still may find the best timbre representation comes from its negative definition

(by removing pitch and loudness) rather than its additive definition, which is equally flawed.

Humphrey, Glennon, and Bello's (2011) work relies on the assumption that, given a set of instruments, more often than not timbrally similar sounds produced within that set are likely to come from the same instrument. While there may be cases where this is not true, given a wide range of content, we believe this assumption to be more valid than that which underlies genre recognition and therefore will utilize this feature learning method in our system.

As opposed to trying to improve the timbre *representation* for a boost in genre classification performance, there has also been a push in the literature to improve the metrics used in timbre similarity engines, rather than relying on the typical calculation of KL-Divergence between GMMs of timbre features, as used by Pampalk (2006). The primary weakness of this often-used metric is that it ignores information about the temporal evolution of timbre. This information is thrown away when modeling the observed series of timbre feature vectors with a GMM probability distribution. While previously discussed research has attempted to incorporate temporal information into the features themselves (this is known as early fusion), other research has focused on incorporating temporal information into the classifier (known as late fusion) (Meng et al. 2004, p. 500).

For example, Flexer, Pampalk, and Widmer model similarity using a Hidden Markov Model (HMM) instead of a GMM (2005). In describing their decision, they note that aspects like spectral fluctuation, attack or decay of an event cannot be modeled without respecting the temporal order of the audio signals, which the GMM does not do (2005, p. 1).HMMs allow one to model time-varying timbral data using probability density functions representing locally stationary timbre and transition probabilities between those stable states (2005, p. 2). Thus, the temporal variation in timbre is directly incorporated into the model. However, in most implementations, first-order HMMs are used (due to their computational efficiency in comparison to higher order models), which retain only the temporal information regarding how neighboring feature vectors are ordered. This combined with the fact that this is a probabilistic mode, means that some information about the temporal evolution is lost. However, one would expect this to result in a much more adept model at calculating timbral similarity. Flexer et al. found, however, that this model did not achieve significant gains in genre classification performance (p. 5). As stated, this does not necessarily mean that HMMs do not provide any significant gains in generating semantically meaningful timbre distance measurements.

Jehan, in his dissertation, calculates similarity between perceptually processed spectral data using dynamic time warping (DTW) (2005, p. 70). By aligning the feature data between two feature vector sequences, one is able to

account for slight time warpings and/or shifts between the sequences, allowing for a simple Euclidean distance calculation between aligned sequence vectors in order to calculate similarity. Jehan also incorporates the importance of the attack towards timbre perception directly into the DTW algorithm by dynamically weighing the path with a half-raised cosine function (p. 71). While DTW retains virtually all of the temporal evolution information of a timbre feature sequence, it has an undesirable time-complexity. However, recently an O(N) variant of DTW, called FastDTW, that restricts the warp path based on reasonable assumptions has been developed (Salvador & Chan, 2004).

Late fusion similarity models like the HMM or DTW are also a step in the right direction away from GMMs. However, other models should be investigated as well. As previously noted, low-order HMMs only retain short-time temporal information, and, while the DTW retains virtually all of the temporal evolution information in a sequence of timbre features, there may be examples where slight time warping or shifting is not enough to align two sequences that are semantically similar. For example, if two sound files contain the same distinct, repetitive timbral gesture, but one file contains a larger number of repetitions than the other, DTW will be unable to globally align them and therefore will consider them timbrally dissimilar. A more appropriate comparison in this case may be analogous to a continuous space sequence similarity measure.

Our approach (as will be laid out in detail later) follows this train of thought by modeling timbre similarity over a curve in an objective timbre space as determining the similarity between sequences of atomic timbre subsequences. Our approach provides similar results to direct Euclidean distance between feature vectors if the sounds compared are time aligned and not time warped and similar results to DTW if time warped and/or not time aligned, but also provides an appropriate similarity measure for when either a different number of repetitions are present in the sounds being compared, when timbre subsequences are re-arranged, or when two sounds' timbre curves only partially overlap - after some alignment and local warping - but diverge otherwise.

Recent research into timbre similarity already presents a wealth of information that will can improve upon the fitness calculations used thusfar in research involving the evolution of synthesis algorithms. However, we believe our approaches to both timbre feature extraction and timbre similarity measurement represent state-of-the-art techniques, which provide the best chance of evolving synthesis algorithms that accurately model any target sound with time-varying timbre.

2.8 State-of-the-Art Genetic Programming

The main deterrent to using any GA variant is that, in comparison to other search techniques, it is a very slow search process. Todd and Werner note that "the main reason for this sometimes-glacial pace is that...evolution builds systems through the gradual accrual of beneficial bits and pieces, rather than through systematic design or rapid learning from the environment" (1998, p. 5). Another reason for GAs inefficiency is the fitness calculation bottleneck, as discussed by Riionheimo and Valimaki (2003, p.10). Typically most of the search time is spent calculating the fitness of each individual. If a fitness calculation is not needed (i.e. if there is a known mathematical relationship between the input parameter space and the output space) then the search would be sped up tremendously, however, in such cases, analytical solutions are often available.

Due to its inefficiency, GAs are often used only as a last resort in search problems. However, for problems that are not well understood or which are known to have complex multimodal fitness landscapes (e.g. Automatic Programming), they often provide the only solution. Much research has been carried out to find ways in which to increase the efficiency of search. Most often, these methods are aimed at restricting the search space using a set of heuristics that include both problem-domain dependent and independent

knowledge. The problem-domain independent methods designed to restrict the search space and/or improve the search "quality" of the GP system influence every part of its structural design. We apply the term heuristics both to architecture decisions as well as limitations imposed on the architecture because, as pointed out by Vanneschi, the art of choosing an appropriate representation and an appropriate set of operators is often a matter of experience and intuition and therefore all design choices are either based on personal or historical trial-and-error (2004, p. 6).

A number of GP enhancements have already been described. For example, the GP/SA architecture used both by Alcazar and Sharman (1996) as well as Garcia (2002) is a way to improve the parameter search using simulated annealing, which is likely better suited for its fitness landscape. Also, incorporating higher-level modules into the function set along with low-level building blocks has been shown to make GP more powerful in various problems (Koza et al., 1997).

Another enhancement to the proposed system is the ability for basic function elements to handle vectors as a native data type. The inability to, for example multiply a stream of numbers by a scalar, is cited by Holladay and Robbins as a major deterrent to Automatic Programming systems that require such operations (2007, p. 1). If such simple operations on vectors are not possible, then the loop structures underlying their execution would have to be

evolved separately.

When applying methods to restrict the search space, one must be careful not to restrict the space within a region that does *not* include the optimal solution. A popular way to naturally restrict the search space (without the possibility of omitting the optimal solution) is to use strongly-typed GP (STGP), which enforces data type constraints, so that only syntactically correct structures are searched (Harris, 1997); (Vanneschi, 2004); (Pachet & Roy, 2007). Basic GP does not enforce such rules and so it is possible that topologies will be searched that are invalid. Enforcing data type constraints during search is a simple way to disregard such topologies and make the search much more efficient. An example of a syntactically-correct Max patch and a syntactically-incorrect Max patch are shown in figure 22.

One of the major problems that GP systems face, efficiency-wise, is code bloat. Bloat, as defined in da Silva's dissertation, is 'an excess code growth without a corresponding improvement in fitness' (p. 2008, p. 2). There are a number of theories as to why bloat occurs in GP systems, but da Silva notes that strong evidence supports that bloat is a side-effect of the search for higher fitness and therefore intricately linked to GP search (p. 9). The problem is analogous to the overfitting problem in machine learning. da Silva writes that 'programs tend to grow until they fit the fitness data perfectly' (p. 9).

da Silva separates bloat control methods into four categories based on

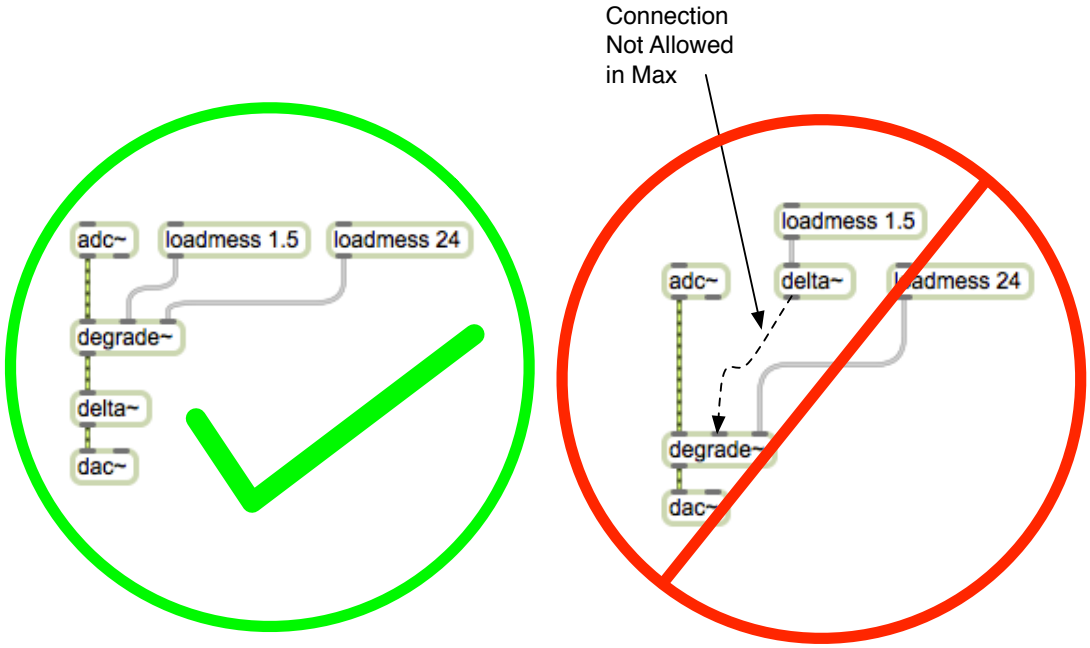

Figure 22: A syntactically correct Max patch that would be included in a STGP search space and a syntactically incorrect one that would not be (due to the labeled disallowed connection).

where in the search process control is applied: evaluation, selection, breeding, and survival (2008, p. 11). In evaluation methods, the desire for parsimony is directly incorporated into the fitness function (p. 11). Selection methods apply rules that favor parsimonious solutions when selecting individuals for breeding (p. 12). Breeding methods introduce genetic operators that are designed specifically to restrict code growth (p. 12). Finally, survival methods only allow candidates from one generation to move on to the next if they meet

certain size requirements (p. 12).

In da Silva's dissertation, he rigorously compares the most popular methods in each category using a number of test problems, differing in complexity, representation, and domain. da Silva finds that a combination of Dynamic Maximum Tree Depth (DMTD) and dynamic Resource-Limited GP (dRLGP) outperformed all other methods consistently (2008, p. 86).

DMTD is an extension of the static-depth limits (SMTDs) imposed by both Wehn (1998) and Garcia (2002) in their research to restrict the size of the evolved synthesis topology and, more importantly from a user perspective, the parameter set. The difference between DMTD and SMTD, as the names imply, is that DMTD imposes a depth limit that changes throughout the search process. Specifically, DMTD works by first setting an initial depth limit when generating the first population. A new individual that breaks this limit will be replaced by its parent, unless it is the most fit individual so far in the run. In this case, the dynamic limit is adjusted to match this length and breeding continues with the new limit until another most-fit-individual breaks it (da Silva, 2008, p. 17).

dRLGP is a variant of Resource-Limited GP (RLGP). In RLGP, a single resource limit is imposed on the entire GP population (da Silva, 2008, p. 21). Every topology node and parameter node in a population is considered a resource. The limiting of such "resources" models the limiting of natural

resources available to a given biological population, considered a main component of evolution (p. 21). The process works as follows: offspring of a population are sorted by fitness, followed by their parents; programs in this list are given resources going from most fit to least fit; if an individual needs more resources than are available, it is passed over and not allowed into the next population. dRLGP adjusts the size of the resource limit if the mean population fitness is greater than the mean fitness of the best population in the run thusfar (p. 22).

Imposing such limits will pressure the system into developing parsimonious solutions (which is good for synthesis algorithm controllability and efficiency) while severely restricting the search space (which makes the search process itself more efficient). A further benefit of using these limits to control code bloat as opposed to evaluation, selection, or breeding methods, is that they are parameterless and can be used with any selection method (da Silva, 2008, p. 97)

Another solution that is commonly used to improve the efficiency of the search (which also has added benefits related to the quality of the evolved algorithms) is called the "Parallel and Distributed Genetic Programming" (PADGP) technique based on parallelizing the GP search process (Vanneschi, 2004, p. 13). In PADGP, the search divides the initial population into a number of subpopulations that evolve independently, with frequent exchange of

individuals between subpopulations (p. 174). The only communication between subpopulations occurs when the exchange of individuals is necessary. The ability to evolve subpopulations independently allows one to parallelize the search process, speeding up the search tremendously (p. 173). However, the true benefit of the island model is that it prevents another major problem of GP, premature convergence (p. 15).

Vanneschi writes that "premature convergence, both from the view of the variety of the fitness values and of syntactic structures, is an experimental evidence in almost all the GP applications" (2004, p. 15). Convergence is "used to describe the state of affairs when the population becomes largely homogenous, consisting largely of variations on a single style of solution" (Vanneschi, 2004, p. 26). When this happens prematurely, the system will output a sub-optimal solution. Crossover, one of the two basic genetic operators used in GP (along with Mutation) allows propagation of small blocks of code that were useful in previous populations (or, at the very least, part of useful algorithms) into future populations. While this is the main mechanism for code re-use during the GP search, it "tends to encourage a uniformity in building solutions and can contribute to convergence problems" (p. 30). Vanneschi shows, via rigorous testing, that the PADGP is a natural way of maintaining diversity, and thus preventing premature convergence, inside GP populations (p. 213). He compares the effects of this model to a number of

other methods designed specifically to limit premature convergence and finds that it works as well or better than these other methods (p. 221). He notes that additionally "one advantage of multiple populations as means for maintaining diversity is that, in contrast to the clever methods above, diversity is maintained 'for free', so to speak, without any particular algorithmic device beyond the simple communication among islands" (p. 213). Vanneschi also suggests that using the island model has a better chance at finding global optima in multimodal landscapes due to the division of the space into a number of mini-parallel searches, as opposed to a singular parallel search over the entire space (p. 192). He writes that this has the effect of speeding up the search as well and that parallel GP would provide for a more efficient search even if implemented on a single processor (p. 230).

In order to further restrict the search space, one can also develop heuristics appropriate for the specific problem-domain. A common example is restricting the possible parameter values by discretizing a finite range, as suggested in (Riionheimo & Valimaki, 2003, p. 6). This process provides the possibility of omitting the optimal solution from search and so restriction on parameter values should not be too severe. Riionheimo and Valimaki suggest attempting to limit the range of parameters so that they are able to just cover all possible musical "tones/' and discretize with a resolution that is below the discrimination threshold (2003, p. 6) However, determining such ranges and

resolutions is a complex process. Martens and Marui found that, at least for vibrato flange and stereo chorus, the useful ranges of parameters were roughly the same (p. 4). However, a more detailed study by McDermott et al. show that, in general, "useful" parameter ranges vary for different synthesis topologies and therefore must be found for each new topology one is presented with (2005, p. 5). Therefore, it is important to err on the side of caution when specifying such limits. One can also place limits on the functions used in each run. However, in general, "the function and terminal sets should be chosen so as to verify the requirements of closure and sufficiency", so restricting the function set can be as dangerous as limiting parameter ranges (Vanneschi, 2004, p. 22).

A method used to improve the ability of GP to find optimal solutions that is directed at the fitness function definition is presented by McDermott et al. (2006). By defining a set of "increasingly discriminating fitness functions (IFFs)" one is able to reward minor progress at each stage of the search in a way not possible using a statically defined fitness function (p. 15). The basic idea is to reshape the fitness landscape throughout the search so that the optimal solution becomes more and more difficult to obtain as the search progresses. Early on in the search when individuals in the population perform poorly, the grading mechanism is more lenient and differences between extremely poor solutions and only slightly better solutions is magnified. As the

mean fitness over the population increases, the fitness function can grow to become more and more difficult to satisfy. In effect, one is manually evolving the fitness function along with the population.

It is our belief that the research discussed above along with improved measures of timbral similarity can combine to push synthesis topology evolution past a point of only being able to generate simple, "toy-example" topologies. By utilizing a meta-synthesis software environment like Max, that was specifically designed to provide synthesis building blocks with the exact level of abstraction required by an efficient GP system, we expect our results to contribute greatly to this area of research.