INDEPENDENT MANIPULATION OF HIGH-LEVEL SPECTRAL ENVELOPE SHAPE FEATURES FOR SOUND MORPHING BY MEANS OF EVOLUTIONARY COMPUTATION

Marcelo Caetano, Xavier Rodet
Analysis/Synthesis Team
IRCAM
{caetano,rodet}@ircam.fr

ABSTRACT

The aim of sound morphing is to obtain a sound that falls *perceptually* between two (or more) sounds. Ideally, we want to morph perceptually relevant features of sounds and be able to independently manipulate them. In this work we present a method to obtain perceptually intermediate spectral envelopes guided by high-level spectral shape descriptors and a technique that employs evolutionary computation to independently manipulate the timbral features captured by the descriptors. High-level descriptors are measures of the acoustic correlates of salient timbre dimensions derived from perceptual studies, such that the manipulation of the descriptors corresponds to potentially interesting timbral variations.

1. INTRODUCTION

There is a burgeoning interest in the search for computational techniques that allow the user to obtain perceptually relevant sound transformations and seamless transitions because computer sound manipulations are widespread in audio applications. Feature-based synthesis is a promising candidate to attain perceptually relevant sound transformations when the features closely capture salient perceptual dimensions [1]. Among the many different possible transformations [2], we will focus specifically on morphing acoustic musical instrument sounds. There seems to be no consensus on what sound morphing is. Most authors seem to agree that morphing involves the hybridization of two (or more) sounds by blending auditory features. One frequent requirement is that the result should fuse into a single percept, somewhat ruling out simply mixing the sources [3], [4], because the ear is still usually capable of distinguishing them due to a number of cues and auditory processes. Although many different methods are described as morphing [5], the result is usually associated with what many authors describe as *timbre interpolation* [4], [6], [7]. The goal of timbre interpolation is to obtain a hybrid sound that is perceived to come from more than one source at the same time. For instance, we could seek to get a hybrid between a violin and a trumpet sound. We should bear in mind that this definition of morphing supposes that timbre is the perceptual phenomenon responsible for sound source identification [8], somewhat ignoring that the same sound can present timbral variations related to the perceptual dimensions of timbre unveiled by psychoacoustic experiments [9]. A *crescendo* trumpet note, for example, becomes perceptually brighter as it gets louder.

Most morphing techniques described in the literature consist in interpolating the parameters of a model used to represent both sounds we wish to morph between, regardless of features [3], [6] [7], [10], [13], [11]. These techniques usually aim at obtaining a sound with an intermediate timbre [3], [6]. The basic idea behind the interpolation principle is that if we can represent different sounds by simply adjusting the parameters of a model, we should obtain a somewhat smooth transition between two (or more) sounds by interpolating between these parameters. Interpolation of sinusoidal modeling is amongst the most common approaches [3], [4], [6], [10]. The sinusoidal parameters can be directly interpolated [3], [6], or by means of another technique [10]. A few authors have proposed to detach the spectral envelope from the pitch information and interpolate them separately [7], [11]. Even more interesting seems to be the approach of designing the spectral envelope separately [12], [13] and imposing the result later for synthesis [14].

Our main motivation is to find a morphed sound that would not only be perceived as a hybridization of the sources, but would also be perceptually intermediate with respect to known salient timbre dimensions, such as brightness [9]. In other words, instead of simply obtaining morphed sounds, we want to control the morphing process perceptually. We want to be able to decide how much of each timbrally related feature we will include from each source. So, for example, when morphing between a bright trumpet sound and a duller clarinet sound, we want to be able to control the perceived brightness of the trumpet-clarinet hybrid. For this, we need to be able to independently manipulate individual features. There have been different proposals in the literature to use features to guide synthesis and transformations. Yee-King [15] uses a genetic algorithm to tune the parameters of an FM synthesis model according to target MFCC values. Hoffman [16] presents preliminary results on an MFCC-based synthesis module that uses some descriptors as guides, while Le Groux [17] uses a support vector machine approach to map an additive synthesis PCA-reduced model to descriptors such as fundamental frequency and loudness. In turn, Park [18] proposes ways of modulating various descriptor-based features, although not independently, such that varying one feature also changes the others in unexpected ways. Verfaille [1] details a general framework to manipulate low-level features to obtain sound transformations that control certain perceptually related features.

However, for morphing, only recently did we start to take perceptual aspects into consideration [4], [5], [12], and the result is the addition of one more step in the process, feature calculation. In most models proposed, linear variation of interpolation parameters does not produce perceptually linear morphs [5]. For the moment, we are interested in being able to control perceptual features of sounds related to the spectral shape, so we will study spectral envelope morphing [13] and manipulation techniques, such as that proposed by Caetano [12]. In this work, we aim at spectral envelope design guided by high-level features, such that in general terms our approach consists of first obtaining an envelope with the general desired number of formant peaks and then shaping it so as to manipulate the high-level features. More spe-

cifically, we will study a method that produces more perceptually relevant morphed envelopes with the desired intermediate number of formant peaks guided by high-level spectral shape descriptors. High-level descriptors are acoustic correlates of timbral dimensions, such that manipulation of the descriptors corresponds to potentially interesting timbral variations. So we define the most suitable representation of the spectral envelope that allows manipulating the spectral envelope shape while retaining the number of peaks. Finally, we describe a technique that uses a genetic algorithm (GA) [19] to independently manipulate the timbral dimensions guided by the descriptors.

The next section presents the basic notion behind the spectral envelope and sound source identification. Then we describe how slight variations of a spectral envelope can be perceived as presenting slightly different timbral features (different neighboring points in timbre space), while still being associated with the same instrument. Next, we introduce the high-level spectral shape descriptors, which capture the acoustic correlates of timbre dimensions. We proceed with the description of the methods to obtain morphed spectral envelopes and study their perceptual impact. The next step is to introduce a technique that uses evolutionary computation to independently manipulate timbral features associated with an envelope. Finally, we describe the experiment we devised to validate our proposal, followed by an evaluation of the results and the conclusions and future perspectives.

2. SOURCE-FILTER MODEL AND SOUND SOURCE IDENTIFICATION

Listeners use many acoustical properties to identify sonic events, such as the spectral shape, formant frequencies, attack and/or onset and decay and/or offset, noise, among others [8]. The cues to identification and timbre vary across notes, durations, intensities and tempos. One model of sound production is based on two possibly interactive components, the source and the filter. The basic notion is that the source applies excitation energy to generate a vibration pattern composed of several vibration modes (modeled as sinusoidal components). This pattern is imposed on the filter, which acts to modify the relative amplitudes of the components of the source input. Resonators, by their nature, tend to amplify certain frequencies louder than others. These resonant frequency regions, or formant peaks, are uniquely related to the size and shape of the instrument and its resonator. We obtain estimates of the excitation and the filter by calculating the spectral envelope, which is a smooth curve that approximately matches the peaks of the spectrum. The peaks of the spectral envelope (also called formants in voice research) correspond roughly to the vibration modes of the source-filter model. The number and absolute position of spectral peaks in frequency is important for musical instrument (sound source) identification. However, we cannot underestimate the perceptual impact of slight variations of spectral shape. The relationship between fundamental frequency and timbre, for example, is readily apparent in some acoustic instruments. The clarinet, for instance, has three distinct registers, that is, three distinct pitch ranges with three different timbral characteristics. It is remarkable that a single instrument can have such a variety of timbres, but the example of the clarinet proves the impact of variations of spectral shape on an instrument's timbre and even temporal evolution. The relationship between applied energy and timbre is relatively clear. As more energy is input to the instrument, higher modes of vibration are achieved such that more partials are present in the frequency spectrum. This is why a note played *forte* is not just louder than *piano*, but also brighter in timbre. A classical example is Risset's discovery that brassy trumpet sounds, usually described as bright, present a broader spectrum resulting from the appearance of higher partials. The spectral centroid, defined later in Section 4, was found [9], [23] to be highly correlated with the dimension of timbre usually verbally labeled as brightness. Therefore, brassy trumpet sounds that are perceived as bright can be characterized at the signal level as presenting a high spectral centroid value. Notably, a *crescendo* trumpet note would exhibit an increasing spectral centroid, while roughly preserving the general position of the first formant peaks (because other peaks appear at high frequency regions).

The vast majority of research in sound perception has focused either on the acoustic properties of musical instruments [20] or on the perception of sounds as unveiled by psychoacoustic experiments [9]. The challenge we face today is to find the link between the two in order to be able to manipulate the sounds in a more perceptually meaningful way.

3. ACOUSTIC CORRELATES OF TIMBRE SPACES

In this section we briefly present timbre perception, timbre spaces and the most relevant acoustic correlates of timbral dimensions obtained in the literature of timbre perception. The concept of timbre is related to the subjective response to the perceptual qualities of sound objects and events [8]. We know that source identification is not reduced to waveform memorization because the intrinsic dynamic nature of the sources produces variations [8]. Timbre perception is inherently multidimensional, involving features such as the attack, spectral shape, and harmonic content.

Historically, Helmholtz was the first to propose an acoustic model of musical instrument sounds. Helmholtz characterized what he called musical tone as a waveform that follows an amplitude envelope that consists of the attack, the steady state and the decay, as shown in Figure 1. During the attack, the amplitude increases from zero to its peak value. In the steady state portion the amplitude is constant and finally decreases back to zero during the decay. Helmholtz concluded that sounds that evoke the sensation of pitch possess fixed waveforms that do not change in the course of the tone, apart from the amplitude envelope, whose temporal evolution has great impact on the perception of the tone, according to him. The classical Helmholtz model breaks down when we examine musical instrument sounds on a small scale. When the harmonic content of sound is examined with the STFT over small time periods, we discover that, contrary to the Helmholtz model, a sound's spectrum changes profoundly over time. During the attack portion of a sound, harmonic content may change rapidly and unpredictably. This phenomenon is called the initial transient. During the release, upper partials tend to disappear more quickly before the entire sounds fades away. While the sustain portion of the sound, when it exists, is certainly more stable than the attack or decay, it is hardly as static as Helmholtz would suggest. Clearly, the basic premise of the classical Helmholtz model - a static spectral envelope with a fixed amplitude envelope temporal evolution – is by no means an accurate and robust characterization of a wide range of acoustic musical instrument sounds.

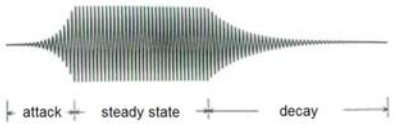

Figure 1: Classical acoustic model of musical instrument sounds. Original figure from [22].

Since the pioneering work of Helmholtz, multidimensional scaling techniques figure among the most prominent when trying to quantitatively describe timbre. Handel [8] gives a comprehensive review of the early timbre space studies. Grey [23] investigated the multidimensional nature of the perception of musical instrument timbre and constructed a three-dimensional timbre space and proposed acoustic correlates for each dimension. He concluded that the first dimension corresponded to spectral energy distribution (measured by the spectral centroid), the second and third dimensions were related to the temporal variation of the notes (onset synchronicity). Krumhansl [24] conducted a similar study using synthesized sounds and also found three dimensions related to attack, synchronicity and brightness (spectral energy distribution). Krimphoff [25] studied acoustic correlates of timbre dimensions and concluded that brightness is correlated with the spectral centroid and rapidity of attack with rise time in a logarithmic scale. McAdams [9] conducted similar experiments with synthesized musical instrument timbres and concluded that the most salient dimensions were log rise time, spectral centroid and degree of spectral variation. More recently, Caclin [26] studied the perceptual relevance of a number of acoustic correlates of timbre-space dimensions with MDS techniques and concluded that listeners use attack time, spectral centroid and spectrum fine structure in dissimilarity rating experiments.

These results suggest that slight changes in the spectral shape produce perceptual changes in timbre that can be measured with high-level spectral shape descriptors. Notably, a morphed sound/spectral envelope with intermediate descriptors should be perceived not only as a hybrid of the source envelopes, but especially as perceptually intermediate.

4. HIGH-LEVEL DESCRIPTORS

In this section we present the general scheme used to calculate the descriptors used in this work, depicted in Figure 2. The sound signal is highlighted with a dark background, all the purely signal processing stages have white background and the steps where we calculate the descriptors present a light background. Peeters [27] describes exhaustively how to calculate all the descriptors we use in this work and proposes to use them in audio classification tasks instead of traditional MFCCs. We are going to present every step of the descriptor extraction scheme with emphasis on the descriptor calculation procedures. The basic signal processing step is the STFT, which refers to the blocks that read "signal frame" and "FFT".

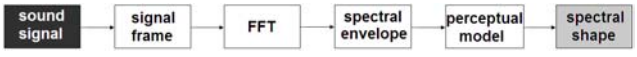

Figure 2. Simplified scheme to calculate the descriptors.

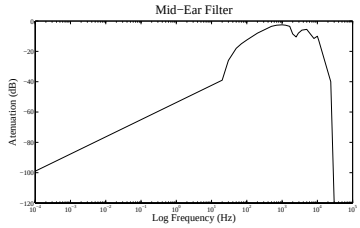

Figure 3: Mid-ear filter applied to extracted spectral envelopes

4.1. Spectral Shape

The calculation of the spectral shape descriptors consists of three steps, spectral envelope estimation, application of the perceptual model, and finally calculation of the spectral shape descriptors, namely, spectral centroid, spread, skewness, kurtosis and slope [27]. For every frame, we calculate the spectral envelope using a cepstral smoothing technique, called *true envelope* [28]. Next, we apply the perceptual model, which consists of the mid-ear filter shown in Figure 3 evaluated on the mel frequency scale. We should notice that this calculation is like the MFCC-based spectral envelope used in [7] without critical band smoothing, such that we do not lose information. Finally we calculate the spectral shape descriptors with the mid-ear attenuated, mel-warped spectral envelope. The spectral shape descriptors considered are calculated as if the magnitude spectrum were a probability distribution. So we associate the frequency bins i of the DFT with the sample space and the probabilities to observe them with the magnitude of the normalized spectral envelope, given by

$$p(k) = \frac{|S(k)|}{\sum_k |S(k)|} \tag{1}$$

such that the spectral shape descriptors are defined as the moments of $p(k)$, where k is the frequency index. The spectral centroid is measured as the mean of $p(k)$ and the spectral spread as the standard deviation, shown in equations (2) and (3) respectively.

$$\mu = \sum_k k p(k) \tag{2}$$

$$\sigma^2 = \sum_k k^2 p(k) \tag{3}$$

The third and forth standardized moments are respectively skewness and kurtosis, shown in equations (4) and (5).

$$\gamma_3 = \frac{\sum_k (k-\mu)^3 p(k)}{\sigma^3} \tag{4}$$

$$\gamma_4 = \frac{\sum_k (k-\mu)^4 p(k)}{\sigma^4} \tag{5}$$

And finally the spectral slope is given by equation (6), where i is the FFT frequency bin index

$$\lambda = \frac{1}{\sum_i p(i)} \frac{N\sum_i p(i)k(i) - \sum_i p(i)\sum_i k(i)}{N\sum_i k^2(i) - \left(\sum_i k(i)\right)^2} \tag{6}$$

5. SPECTRAL ENVELOPE MORPHING BY HIGH-LEVEL DESCRIPTORS

In this section we explain our motivation for using spectral shape descriptors as guides in obtaining perceptually relevant morphed spectral envelopes. The aim of morphing spectral envelopes is to obtain a result that is perceived not only as a hybrid between the original sounds, but especially perceptually intermediate between them. Slaney [7] explains the concept by analogy with image morphing, where the aim is to gradually change from one image to the other, producing perceptually convincing intermediates (or hybrids) along the way. Other authors have proposed the same analogy [3]. Figure 4 shows such an example of image morphing with faces. Clearly, it is not enough to blindly interpolate parameters (pixels, for instance, for the images) since there are a number of important features in the faces that we must take into account. Finding those features is an important task, and developing techniques to obtain intermediate (hybrid) images that use those features as cues is the key to a successful morph. The analogy with spectral envelope morphing is immediate. Each frame of the STFT is interpreted as a snapshot of the spectrum of the sounds seen through a time window. So the task of morphing spectral envelopes becomes similar to image morphing, each hybrid envelope must present intermediate features to be perceptually convincing. Here we argue that high-level descriptors capture salient timbre dimensions of sounds, so we use them as a guide to morph spectral shapes. An important concept that can be inferred from Figure 4 is the fact that there are many possible intermediate steps between the two images shifting from the first (S_1) to the second (S_2) spectral envelope. So, if we consider each intermediate hybrid spectral envelope as the result of a different combination of S_1 and S_2, this convex combination can be mathematically expressed as equation (7)

$$M(\alpha,t) = \alpha(t)S_1 + [1-\alpha(t)]S_2 \tag{7}$$

and each step is characterized by one value of a single parameter (α), called morphing factor, as shown at the bottom of Figure 4. The morphing factor should vary between 0 and 1, such that $\alpha = 1$ and $\alpha = 0$ produces S_1 and S_2 respectively.

Usually, morphing techniques propose to interpolate the parameters of a model without making sure that the results will (perceptually) correspond in the feature space. Moreover, depending on the parameters we interpolate, the morphed spectral envelope will not present the desired number and position of peaks. Ideally, we want the morphed spectral envelope to have an intermediate number and position of peaks and to match as closely as possible the values of target perceptually related features, in this case, spectral shape descriptors. So we present in Figure 5 an example of the perceptual impact of morphing spectral envelopes by three different methods proposed in the literature, namely, linear predictive coefficients (LPC) [12], [29] line spectral frequencies (LSF) [5], [30], [32], and dynamic frequency warping (DFW) [13], [32].

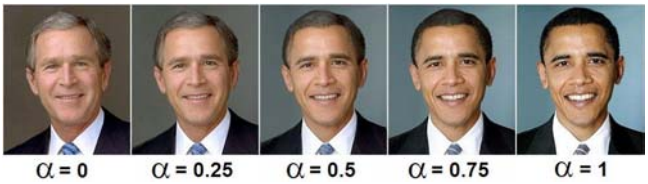

Figure 4: Depiction of image morphing to exemplify the aim of sound morphing

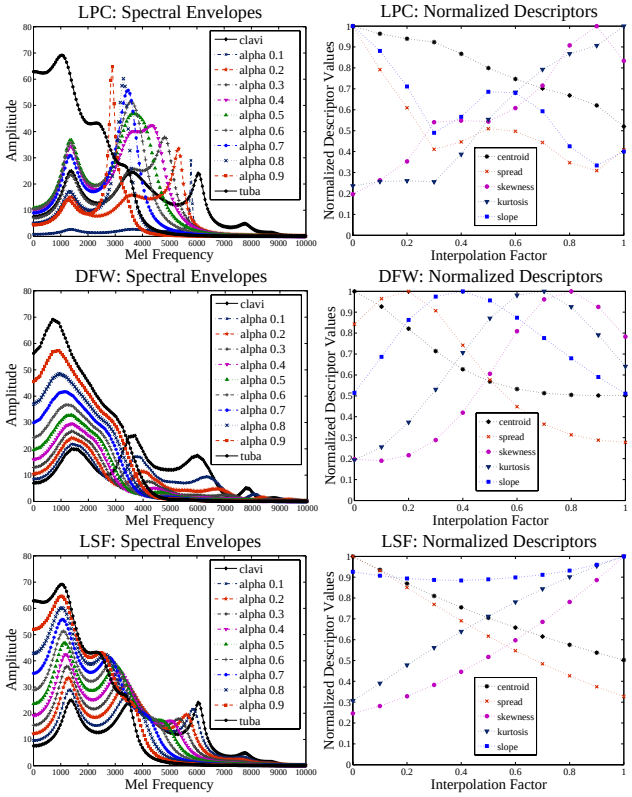

Figure 5. Spectral envelopes (left) and spectral shape descriptor values (right) corresponding to linearly varying the interpolation factor α from 0 to 1 for LPC (top), LSF (middle) and DFW (bottom) spectral envelope interpolation.

Figure 5 compares the result of varying the interpolation factor from 0 to 1 for LPC, LSF and DFW both in number of peaks and in shape for two very different spectral envelopes, labeled clavi and tuba. Notice that, for this example, LPC (top) does not interpolate well either the peaks or the shape. Neither the spectral envelopes nor the descriptors vary smoothly from one extreme to the other as desired. Although DFW (middle) visually seems to render a satisfactory shift, the descriptors vary in unexpected ways. LSFs (bottom) interpolate better both the peaks and the shape. Ideally, we expect the hybrid spectral envelope to smoothly change from source to target, with peaks shifting, appearing or splitting and disappearing or merging. The change is also very satisfactory in the descriptor domain, varying almost as straight lines. Ideally, we want a spectral envelope morphing method that generates hybrid envelopes with not only an intermediate number of peaks, but also with intermediate values of features, as measured by the descriptors. When we set $\alpha = 0.5$, we seek for a morphed envelope whose descriptors are halfway between those of S_1 and S_2. Table 1 shows an example of the accuracy of the three methods for $\alpha = 0.5$. We show the target value for each descriptor considered, calculated applying equation (7) using the descriptor values as S_1 and S_2, and the value measured for the spectral envelope produced by each morphing method. We consider a method that generates morphed envelopes with descriptor values closer to the target values to render more perceptually relevant morphed spectral envelopes [5]. This is an objective measure of the perceptual impact of the morphed envelopes obtained with each method.

Descriptor	**Target**	**LPC**	**DFW**	**LSF**
Centroid ($x10^{+3}$)	3.20	3.57	3.34	3.27
Spread ($x10^{+6}$)	2.70	2.01	2.62	2.70
Skewness	1.27	1.21	0.94	1.27
Kurtosis	12.4	12.5	7.46	9.15
Slope ($-1x10^{-16}$)	3.14	1.12	1.54	3.12

Table 1. Target (α = 0.5) and measured spectral shape descriptors.

6. INDEPENDENT MANIPULATION OF FEATURES BY EVOLUTIONARY COMPUTATION

In this section we explain the technique we use to manipulate perceptually relevant timbral features of the morphed envelopes to produce interesting timbral variants. In other words, we are still looking for a morphed envelope that is perceived halfway (α = 0.5) between a clarinet and a trombone, but now we dot not want all the timbral features to be also automatically in the middle. We want to be able to control them independently and obtain a hybrid clarinet-trombone morphed envelope that sounds as bright as the original trombone, for example. This would correspond to obtaining a hybrid image morph of the Bush-Obama faces where all the features are halfway, except the nose, which still resembles Obama's. The technique consists in generating a prototype morphed envelope with intermediate features, and then manipulating the features independently to match new target values set for each descriptor. The variants are obtained by setting the morphing factor α independently for each descriptor, representing each feature we want to control. The prototype morphed envelope can potentially be obtained with any method because we can convert the representation of a spectral envelope. This means that we can obtain the LSF representation of an envelope estimated with true envelope [28], linear prediction [29], or any other method. Particularly, we can calculate the LSF representation of an envelope generated with DFW [32]. Therefore, we will verify if there is one prototyping method that outperforms the other between prototype envelopes generated with LSF and DFW. We clearly need a suitable representation that allows local manipulation of the spectral shape without completely changing the overall prototype envelope (i.e., the number and location of peaks).

6.1. Line Spectral Frequency Pairs

Line spectral frequency pairs (LSFs) are an alternative parameterization of LPC [29] with a one to one correspondence. The two LSF polynomials are given by

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \quad (8)$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \quad (9)$$

where $A(z)$ is the linear prediction polynomial of order p [29]. The roots of the polynomials in equations (8) and (9) determine the LSFs. If $A(z)$ is minimum phase, the roots of $P(z)$ and $Q(z)$ are on the unit circle, are real, interleaved with each other, and always lead to stable envelopes when arranged in ascending order [30]. LSFs also present the useful tendency to be located where the peaks of the envelope they represent are. Figure 6 shows that each pair tends to be close together when near a peak of the spectral envelope and far apart when not, depicting another useful property of LSFs. The closer the line spectrum pair is, the narrower the peak.

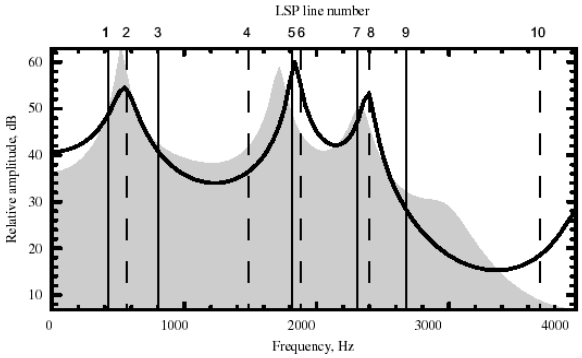

Figure 6: Depiction of LSFs and corresponding spectral envelope. Original figure from [30].

Based on these properties of LSFs, McLoughlin [30] exemplifies how we can manipulate the LSFs to produce small changes in the shape of the spectral envelope and Morris [31] presents a method for modifying formant peak locations and bandwidths in the line spectrum domain. Figure 6 shows the original spectral envelope in grey and a modified envelope (solid line) with its corresponding LSFs. Since there are LSF pairs that correspond roughly to specific spectral peaks, we generally can make changes to a specific peak without changing much the overall spectral envelope. With this in mind, for a prototype morphed envelope, there must be a variant with slightly different LSFs whose shape matches more closely the independently set target spectral shape descriptors. Because fine tuning the LSFs to match all descriptors at once is a difficult task (highly nonlinear mapping), we use a genetic algorithm (GA) to perform the search. The GA will manipulate the LSF representation of prototype morphed envelopes obtained with both LSF and DFW.

6.2. Genetic Algorithms

Genetic algorithms (GAs) are the most commonly used paradigm of evolutionary computation due to the robustness with which they explore complex search spaces. They codify the parameters of a model into a chromosome-like structure so that each individual corresponds to a point in the parameter space, depending on the values of the parameters. The resulting search space contains the candidate solutions, and the evolutionary operators will implement exploration and exploitation of the search space aiming at finding quasi-global optima. The GA iteratively manipulates populations of individuals at a given generation by means of the simple genetic operations of selection, crossover and mutation. The evolutionary process combines survival of the fittest with the exchange of information in a structured yet random way. The standard genetic algorithm [19] consists of the steps shown in Table 2.

```
(*Initialize Population*)
   (*Main Cycle*) generations
  repeat
(*Competition Cycle*)
     Crossover
     Mutation
     Fitness Evaluation
     Selection
until termination criterion met
```

Table 2. Standard Genetic Algorithm.

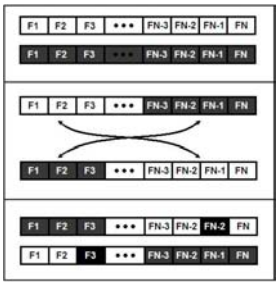

Figure 7. Depiction of the chromosome (top), crossover (middle) and mutation operations (bottom).

6.3. Codification and Evolutionary Operators

In this section we explain how we apply the steps presented in Table 2 to the population of candidate prototype spectral envelopes in the search for the variant envelope model that best matches the independent target envelope shape descriptor values.

6.3.1. Initialization

We initialize the population of N candidate solutions by producing variants of a prototype spectral envelope. We obtain the prototype morphed envelope with either LSF or DFW. The variants are obtained by adding a perturbation vector (normal distribution $N(0,10^{-3})$) to the LSFs of the prototype morphed envelope, so that each individual is a variant of the prototype with a slightly different spectral shape. Each individual in the population is codified as a chromosome that lists LSF pairs in ascending order, as represented on the top of Figure 7. This initialization process is intended to sample the search space (spectral shape descriptors) while restricting the positions of the peaks of the envelopes.

6.3.2. Crossover

Crossover is responsible for the exploitation of regions of interest of the search space by means of the exchange of information between individuals of a population. Crossover consists of selecting two parent individuals, the crossover points, and swapping the chromosome segments (represented by different shades in Figure 7) between them, thus generating two offspring. We mate each individual of the current population with one randomly chosen partner (uniform distribution) using a one-point crossover operator with a uniform distribution [19]. Both offspring are inserted in the population and the parents are also kept. We use a one-point crossover operator, which consists of selecting one mating partner for each individual in the population (those are the parent chromosomes), randomly (uniform distribution) choosing a crossover point and swapping the segments between the parents, thus generating the offspring chromosomes shown in the middle of Figure 7, where each segment is represented by a different shade. The result of the crossover operation is two offspring individuals comprised of LSFs that come from both parents. Thus the offspring consist of variant spectral envelopes whose spectral shape inherits characteristics of both parents. It is important to notice that both offspring generated are inserted in the population and the parents are also kept, thus increasing the number of candidate solutions by 2. That is, now we have three times the initial number of individuals in the population, or $3N$, because for each individual we mate, they produce two offspring that are inserted in the current generation and we keep the current individuals in the population.

6.3.3. Mutation

Mutation is responsible for the exploration of the search space by randomly replacing the value of one randomly selected pair of LSFs, thus allowing different regions of the search space to be investigated. The mutation operation, applied to all individuals in this increased population, is depicted at the bottom of Figure 7 and consists of randomly (uniform distribution) choosing a mutation point, represented in black in Figure 7, and adding a perturbation to it (normal distribution $N(0,10^{-3})$). Thus mutation is a kind of perturbation of the LSF pair selected, resulting in a slightly different shape.

6.3.4. Fitness Evaluation

Next we measure the fitness of all individuals in the current population using the fitness function (*ff*) in equation (10). The fitness function operates on the feature space (the spectral shape descriptors) and here it is a very simple error (or distance) measure between the target and the calculated descriptor values. Equation 10 below shows the computation as the absolute value of the difference between target descriptors (T) and the descriptor values (c_i) calculated for each individual in the current generation, weighted (ω_i) and normalized by the target value T for each spectral shape descriptor used in the method so that they are all dimensionless and therefore can be compared.

$$ff = \frac{\sum_i \left| \omega_i \frac{(T_i - c_i)}{T_i} \right|}{\sum_i \omega_i} \tag{10}$$

6.3.5. Selection

Finally, the selection operator discards individuals with low fitness values, only keeping individuals that correspond to promising regions of the search space for the next generation. Selection is done by sorting the individuals of the population of the current generation by increasing values of fitness and selecting the first N as the population for the next generation. We use a strategy called elitism that consists of keeping the best individual found so far in all generations even if it is lost in the current population due to crossover and mutation. The termination criterion is met when either a minimum fitness threshold or the maximum number of generations is reached.

7. EXPERIMENT AND RESULTS

We aim to show that the GA allows us to obtain hybrid morphed envelopes that closely match the target perceptual descriptor values even if we use independent morphing factors for the descriptors. We also want to verify which method used to obtain the prototype envelope, DFW or LSF, renders resultant envelopes that retain the desired overall number and position of peaks. So we set a variable morphing factor (α = [0.1, 0.3, 0.5, 0.7, 0.9]) for the spectral centroid and kept the others constant at α = 0.5. We expect the results to show hybrid spectral envelopes whose centroid varies as desired while keeping the other descriptors considered unchanged. We measure the quality of the results as how closely they match the values of target perceptual spectral shape descriptors while retaining the desired overall number and position of peaks of the prototype spectral envelope. Table 3 shows the target and measured descriptor values for each morphing factor α indicated, applied only to the perceptual centroid. Figure 8 depicts the resultant envelopes for LSF and DFW proto-

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			$\alpha = 0.7$			$\alpha = 0.9$		
Descriptor	**Targ**	**DFW**	**LSF**	**Targ**	**DFW**	**LSF**	**Targ**	**DFW**	**LSF**	**Targ**	**DFW**	**LSF**	**Targ**	**DFW**	**LSF**
Centr ($x10^{+3}$)	2.35	3.53	2.35	2.78	2.78	2.78	3.20	2.69	3.20	3.63	2.43	3.63	4.05	2.42	4.05
Spread ($x10^{+6}$)	2.70	4.52	1.66	2.70	3.91	2.08	2.70	2.71	2.57	2.70	1.94	2.99	2.70	1.44	3.49
Skew	1.27	1.07	1.82	1.27	1.86	1.47	1.27	2.48	1.17	1.27	2.69	0.93	1.27	1.93	0.68
Kurt	12.4	5.35	16.6	12.4	8.64	12.6	12.4	14.9	9.52	12.4	21.2	7.49	12.4	20.7	5.84
Slope ($-1x10^{-16}$)	3.14	5.63	1.89	3.14	5.84	2.64	3.14	5.11	3.39	3.14	3.54	3.97	3.14	1.85	4.39

Table 3. Target and measured perceptual spectral shape descriptor values for DFW and LSF after the application of the GA with morphing factor independently set.

type spectral envelopes. If we compare the number and position of peaks of the results presented in Figure 8 with the corresponding morphed prototypes in Figure 5, we readily see that the desired intermediate number and position of peaks was retained for LSF. Although DFW generates morphed prototypes with smoothly varying intermediate number and position of peaks, since the corresponding descriptor values are farther from the target (see Table 1), the GA compromises the original number and position of peaks in favor of shape.

Table 3, on the other hand, shows that, for LSFs, as the spectral centroid shifts as expected, the other values vary slightly. In general, LSFs outperformed DFW in matching the target descriptors. We do not control the individual accuracy of descriptors; therefore, because they all have different ranges, the precision of matching the spectral centroid differs from the other descriptors. Without the weights, descriptors with smaller ranges tend to be matched with greater precision. The weights allow us to tip the scales and focus on the descriptors of interest. If we compare the corresponding columns in Table 1 with the column labelled α = 0.5 in Table 3, we see that DFW performed poorly in matching the target descriptors even after the application of the GA. On the other hand, for LSFs, in this case and all others presented, even though the GA unmatched the other descriptors a little, the descriptor of interest (centroid) is always a perfect match. We do not directly compare the results quantitatively with the manipulation of LPCs by a GA [12] because qualitatively they are not equivalent. The method presented in [12] does not manipulate LSFs and therefore would probably not render results with the desired number and position of peaks for, as we showed in section 5, the interpolation of LPCs is highly unstable.

We did not find any studies on how the accuracy of the spectral shape descriptors affects the perception of the features they are related to, such that, it is not possible as of now to decide how perceptually relevant the values are and how accurate the matching should be. We would need to study whether there is a sort of just noticeable difference (JND) for the descriptors in order to infer how their individual accuracies affect perception. The most important aspect of the results lies in the independent control of the descriptors given that the relevance of the descriptors values is only relative and there is no scale at present with which to perform a deep quantitative analysis. Controlling the descriptors individually is a first step toward the study of perceptually motivated feature-based sound transformations such as morphing. All the results presented in this article are available online http://recherche.ircam.fr/anasyn/caetano/morph.html.

8. CONCLUSION AND FUTURE PERSPECTIVES

The aim of sound morphing is to find a morphed sound that would not only be perceived as a hybridization of the original sounds, but would also be perceptually intermediate with respect to known salient timbre dimensions. We studied methods to obtain a morphed spectral envelope that contains both the desired intermediate number and position of resonant peaks, but also corresponding intermediate perceptual features, measured by high-level spectral shape descriptors. High-level descriptors are acoustic correlates of timbral dimensions, such that manipulation of the descriptors corresponds to potentially interesting timbral variations. We compared the morphed spectral envelopes obtained with LPC [29], DFW [32] and LSFs [30] and found that LSFs generate more smoothly varying hybrids according to two criteria, number and position of spectral peaks and values of spectral shape descriptors.

Then, we defined LSFs as the most suitable representation of the spectral envelope that allows manipulating the spectral envelope shape while roughly retaining the desired number and position of peaks of the prototype morphed envelope we wish to transform. Finally, we described a technique that uses a genetic algorithm (GA) to independently manipulate perceptually relevant timbral features of the morphed spectral envelopes guided by the descriptors to produce interesting timbral variants.

We verified that we have finer control over sound transformations perceptually if we are able to control how individual features behave under each transformation. Using our proposed technique, we can perform transformations where only one descriptor changes, or they all change in different ways, so there would be many different possible transformations between two different sounds regarding the perceptual shape, instead of only one. For example, we could obtain a morphed clarinet-trumpet sound and impose the brightness of the trumpet on the morphed result.

However, we would need to perform listening tests to verify how these differences manifest perceptually, since the perceptual impact of manipulating such nuances of timbral perception is still unclear.

Future perspectives of this work include investigating the impact of the timbral variants on the spectral shape of the origi-

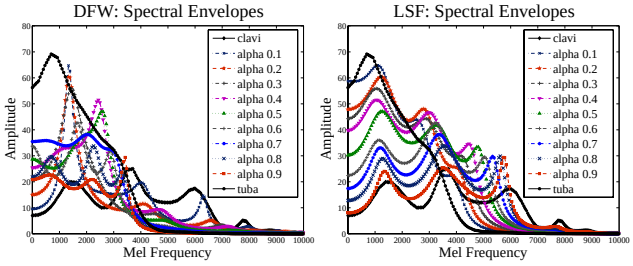

Figure 8. Spectral envelopes resulting from the application of the GA with a varying morphing factor independently controlling the perceptual spectral shape descriptors for LSF and DFW prototype envelopes.

nal sounds alone. We could investigate if the timbral variation-technique presented here could render a trumpet sound brighter, like it was played louder. Also, we could verify whether improvements on the perceptual model used to calculate the descriptors would render more perceptually relevant results. Finally, there are many possible variations of the application of the GA that should be investigated, such as a different crossover operator, mutation, selection, fitness function, among others. One could even test other optimization methods or even different mapping strategies to obtain the variant spectral envelopes.

9. ACKNOWLEDGEMENTS

This work is supported by the Brazilian Governmental Research Agency CAPES (process 4082-05-2).

10. REFERENCES

[1] V. Verfaille, U. Zolzer, D. Arfib. Adaptive Digital Audio Effects (a-DAFx): A New Class of Sound Transformations. IEEE Trans. Audio, Speech, and Language Processing, 14(5), pp. 1817-1831, 2006.

[2] X. Amatriain, J. Bonada, A. Loscos, J.L.Arcos, V. Verfaille. Content-Based Transformations. Journal of New Music Research, 32(1), March 2003, pp. 95-114.

[3] Fitz, K., Haken, L., Lefvert, S., Champion, C., O'Donnell, M. "Cell-Utes and Flutter-Tongued Cats: Sound Morphing Using Loris and the Reassigned Bandwidth-Enhanced Model" Computer Music Journal, 27 (3), pp. 44-65, 2003.

[4] Williams, D., Brookes, T. "Perceptually-Motivated Audio Morphing: Softness", *AES 126th Convention*, 2009.

[5] M. Caetano, X. Rodet. "Automatic Timbral Morphing of Musical Instrument Sounds by High-Level Descriptors, to appear in Proc. ICMC 2010.

[6] Tellman, E., Haken, L., Holloway, B. "Timbre Morphing of Sounds with Unequal Numbers of Features," J. Audio Eng. Soc. vol. 43, no. 9, pp 678-689, September 1995.

[7] Slaney, M., Covell, M., Lassiter, B. "Automatic Audio Morphing". *Proc. ICASSP*, IEEE, 1996.

[8] Handel, S. "Timbre perception and auditory object identification." In B.C.J. Moore (ed.), Hearing (pp. 425-461). New York: Academic Press, 1995.

[9] McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., Krimphoff, J. "Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specifities and latent subject Classes". Psychol. Res., 58, pp. 177-192, 1995.

[10] Boccardi, F., Drioli, C. "Sound Morphing with Gaussian Mixture Models" *Proc. DAFx*, pp. 44-48, 2001.

[11] Ahmad, M., Hacihabiboglu, H., Kondoz, A. M. "Morphing of transient sounds based on shift-invariant discrete wavelet transform and singular value decomposition" *Proc. ICASSP*, 2009.

[12] Caetano, M., Rodet, X. Evolutionary Spectral Envelope Morphing by Spectral Shape Descriptors, Proc. ICMC 2009.

[13] Ezzat, T., Meyers, E., Glass, J., Poggio, T. "Morphing Spectral Envelopes using Audio Flow" *Proc. ICASSP*, 2005.

[14] D.Arfib, F. Keiler, U. Zoelzer (2002): "Source-Filter processing", in DAFx digital audio effects, pp 293-361, ed U. Zoelzer, publisher Wiley and sons

[15] Yee-King, M., Roth, Synthbot: An Unsupervised Software Synthesizer Programmer, proc ICMC, 2008.

[16] Hoffman, M., Cook, P. "Feature-based Synthesis: Mapping from Acoustic and Perceptual Features to Synthesis Parameters" Proc. ICMC, 2006.

[17] Le Groux, S.; Verschure, P. Perceptsynth: Mapping Perceptual Musical Features to Sound Synthesis Parameters, Proc. ICASSP, pp. 125-128, 2008.

[18] Park, T., Biguenet, J., Li, Z., Conner, R., Travis, S. Feature Modulation Synthesis, Proc. ICMC, 2007.

[19] Davis, L. "Handbook of Genetic Algorithms". New York:Van Nostraud Reinhold,1991.

[20] Risset, J. C. Computer Study of Trumpet Tones. Murray Hill, N.J.: Bell Telephone Laboratories, 1966.

[21] Von Helmholtz, H. On the Sensations of Tone. London, Longman, 1885.

[22] Dodge, C. Jerse, T. A. Computer Music: Synthesis, composition and performance. Schirmer Books, Macmillan, New York, 1985. ISBN 0-02-873100-X.

[23] Grey, J. M., and Moorer, J. A., "Perceptual Evaluations of Synthesized Musical Instrument Tones". Journ. Ac. Soc. Am., 62, 2, pp 454-462, 1977.

[24] Krumhansl, C. L. 1989. "Why is Musical Timbre So Hard to Understand?" In S. Nielzén and O. Olsson, eds. Structure and Perception of Electroacoustic Sound and Music. Amsterdam: Excerpta Medica.

[25] Krimphoff, J., S. McAdams, and S. Winsberg. "Caractérisation du Timbre des sons Complexes. II: Analyses Acoustiques et Quantification Psychophysique" Journal de Physique 4(C5), pp. 625-628, 1994.

[26] Caclin, A., McAdams, S., Smith, B. K., Winsberg, S. "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones". J. Acoust. Soc. Am. 118 (1), pp. 471-482, 2005.

[27] Peeters, G. "A large set of audio features for sound description (similarity and classification) in the CUIDADO project". Project Report, 2004.

[28] A. Röbel, X. Rodet, "Efficient Spectral Envelope Estimation and its Application To Pitch Shifting And Envelope Preservation," Proc. DAFx, 2005.

[29] Makhoul, J. "Linear prediction: A tutorial review" Proc. IEEE, vol. 63, pp. 561-580, Apr. 1975.

[30] McLoughlin, I. V. Line spectral pairs. Signal Processing 88 (3), pp. 448-467, 2008.

[31] R. W. Morris and M. A. Clements. "Modification of Formants in the Line Spectrum Domain". IEEE Sig. Proc. Letters, 9(1), Jan. 2002.

[32] Pfitzinger, H., R. DFW-based Spectral Smoothing for Concatenative Speech Synthesis. Proc. ICSLP 2004, vol. 2, pp. 1397-1400. Korea. Oct, 2004.