AUTOMATIC TIMBRAL MORPHING OF MUSICAL INSTRUMENT SOUNDS BY HIGH-LEVEL DESCRIPTORS

Marcelo Freitas Caetano, Xavier Rodet

AUTOMATIC TIMBRAL MORPHING OF MUSICAL INSTRUMENT SOUNDS BY HIGH-LEVEL DESCRIPTORS

Marcelo Caetano, Xavier Rodet

Ircam
Analysis/Synthesis Team
{caetano,rodet}@ircam.fr

ABSTRACT

The aim of sound morphing is to obtain a result that falls *perceptually* between two (or more) sounds. In order to do this, we should be able to morph perceptually relevant features of sounds instead of blindly interpolating the parameters of a model. In this work we present automatic timbral morphing techniques applied to musical instrument sounds using high-level descriptors as features. High-level descriptors are measures of the acoustic correlates of salient timbre dimensions derived from perceptual studies, so that matching the descriptors becomes the goal itself to render the results more perceptually meaningful.

1. INTRODUCTION

The 20th century witnessed a compositional paradigm shift from pitch and duration to timbre [32]. The advent of the digital computer revolutionized the representation and manipulation of sounds, opening up new avenues of exploration. Timbre manipulation led to the development of transformational techniques usually referred to as morphing. Among the several possible applications of morphing [27], the exploration of the sonic continuum in composition [32] stands out as the most exciting to date. Jonathan Harvey's 'Mortuos plango, vivos voco' morphs seamlessly from a vowel sung by a boy to the complex bell spectrum consisting of many partials. Another example is Trevor Wishart's Red bird where the word 'listen' gradually morphs into birdsong [32]. Wishart himself mentions Michael McNabb's 'Dreamsong' and its particularly striking opening and closing morphs [33]. These authors did morphing by hand mainly using studio techniques. This work investigates techniques to automatically achieve similar results by simply choosing what sounds we want to morph between and how we want to do the transformation, especially because many different transformations fall under the umbrella of morphing, as we will explain in more detail in Section 2. 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 [6], [27], because the ear is still usually capable of distinguishing them due to a number of cues and auditory processes. Sill, many different transformations are described as morphing, such as interpolated timbres [27], smooth, seamless transitions between sounds [1] or cyclostationary morphs [26], each of which will be thoroughly reviewed in Section 2. Most authors propose to interpolate the parameters of a model [1], [2], [7], [13], [21], [24], [26], [27] without worrying about the perceptual impact of the process. These authors often conclude that the linear interpolation of the parameters do not correspond to linearly varying the corresponding features [1], [12], [26]. Some authors proposed timbre spaces [8], [3], where each dimension is correlated to a perceptual feature. Caetano [4] figures among the first to make a distinction between interpolation of parameters and morphing of features. Our motivation is the hybridization of perceptual features of musical instrument sounds that are related to salient timbral dimensions unveiled in psychoacoustic experiments [3], [15], [18]. In other words, instead of simply obtaining hybrid sounds, we want to control the hybridization process perceptually. In this work, we describe techniques to automatically obtain *perceptually* intermediate quasi-harmonic musical instrument sounds using high-level descriptors as guides. High level descriptors are measures of acoustic correlates of timbre dimensions obtained by perceptual studies, such that sounds whose features are intermediate between two would be placed between them in the underlying timbre space used as guide.

The next section contains a comprehensive review of the terminology and processes usually called morphing, followed by the techniques proposed to achieve the desired results. Next, we briefly review timbre perception and timbre spaces, and introduce high-level descriptors. Then, we propose a timbral morphing technique that consists of extracting the features, interpolating between them in the descriptor domain, thought to capture perceptual timbral features, and resynthesizing the morphed sound with parameter values that correspond to the morphed features. We emphasize methods to obtain a morphed spectral envelope with hybrid descriptor values. Finally, we present the conclusions and future perspectives of the morphing technique.

2. WHAT IS SOUND MORPHING?

After a thorough review of the literature on the hybridization of sounds, we realized there is much confusion in terminology. One of the aims of this article is to clarify a little bit the techniques referred to as morphing and the terminology itself used in the literature. Apart from sound morphing, some authors refer to this hybridization process as audio morphing [26], while others prefer timbre morphing [27] or even timbre interpolation [12] to refer to similar goals, and some choose to use these terms interchangeably. The result has been called hybrid [9] [7], intermediate [4], interpolated [12] or even mongrel sound [13]. In this work, we reserve the term sound for the auditory impression or the sensation perceived by the sense of hearing, whereas audio refers more specifically to the signal. Moreover, we make a distinction between interpolation and morphing. Interpolation acts on the parameters of a model, being restricted to the signal level, whereas we reserve morphing for the hybridization of perceptual qualities. So we propose sound morphing as the most appropriate term to our goals, and we talk about hybrid or intermediate sounds. We focus on timbral features independent from loudness and pitch (LP-timbre, as defined by Letowsi [17]), especially those related to the spectral envelope shape [4], so we will make an additional important distinction between timbre morphing and the term we chose to use here, timbral morphing, while attempting to find a good definition for sound morphing. There seems to be no widely accepted definition of morphing in the literature. Instead, most authors either attempt to provide a definition of their own or simply explain what the aim of their work was. Some definitions are too system dependent to be useful, Fitz [6] defines morphing as "the process of combining two or more Lemur files to create a new Lemur file with an intermediate timbre", others are too general, such as Boccardi's [2] "modifying the time-varying spectrum of a source sound to match the time-varying spectrum of a given number of target sounds". Definitions based on the concept of timbre are common [12], [27], [20], [7]. Usually, these authors define timbre morphing as "the process of combining two or more sounds to create a new sound with intermediate timbre" [27] or "to achieve a smooth transition from one timbre to another" [12]. We should notice that these refer to different goals. All in all, we prefer to avoid any definition that relies heavily on a concept as loosely defined and misunderstood as timbre, that can encompass many different perceptual dimensions of sounds [17]. Although nobody defines what they mean by timbre, most authors seem to refer to timbre as the set of attributes that allow sound source identification. In musical instrument contexts, this usually means that timbre becomes a synonym of musical instrument and thus timbre morphing reduces to hybrid musical instrument sounds. It is possible, though, to morph between sounds from the same instrument (different loudness or even different temporal features) [27], [26]. Instead, we prefer to define the aim of morphing as obtaining a sound that is perceptually intermediate between two (or more), such that our goal becomes to hybridize perceptually salient features of sounds related to timbre dimensions, which we term timbral morphing.

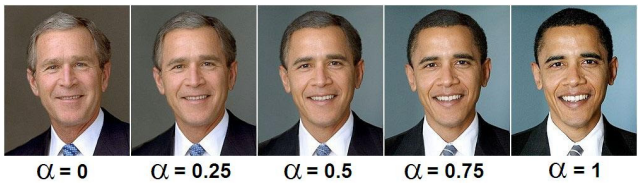

Figure 1. Depiction of image morphing to exemplify the aim of sound morphing.

Slaney [26], on the other hand, prefers to avoid a direct definition altogether and explains the concept by analogy with image morphing instead, where the aim is to gradually change from one image (the source) to the other (the target) producing convincing intermediates (or hybrids) along the way. Other authors have proposed the same analogy [7]. Nonetheless, they rely on the concept of sound object especially because they do not restrict their goal to musical instrument sounds. Figure 1 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. Here we argue that high-level descriptors capture salient timbre dimensions of sounds, so we use them to align temporal features and to morph spectral shapes. An important concept that can be inferred from Figure 1 is the fact that there are many possible intermediate steps between the two images shifting from the source to the target. The original images/sounds from now on shall arbitrarily be called source and target for formalization purposes only because the morph should not be different if they change positions. So, if we consider each intermediate image/sound as the result of a different combination of source and target, this convex combination can be mathematically expressed as equation (1) and each step is characterized by one value of a single parameter (α), called interpolation or morphing factor, as shown at the bottom of Figure 1. The morphing factor should vary between 0 and 1, such that $\alpha = 1$ and $\alpha = 0$ produce source and target respectively. Convex combinations of more than two objects (images, sounds) are also possible, as well as using a time varying morphing factor, giving rise to dynamic transformations.

$$M(\alpha,t) = \alpha(t)\hat{S}_1 + [1-\alpha(t)]\hat{S}_2 \quad (1)$$

Due to the intrinsic temporal nature of sounds, a better analogy would be that of movie morphing [26], where the aim must be reviewed to better fit the dynamic nature of the media, depicted in Figure 2. Now our sound morphing analogy has closer correspondences. For example, each movie frame could correspond to an STFT frame resulting from the analysis of the sounds we intend to morph between. Also, we can imagine that each frame's visual features have a corresponding set of sonic features that also evolve in time and that this evolution in time itself carries important information about how we perceive the movie (sound). Notice that Figure 2 depicts movies (sounds) with different numbers of frames, therefore, different lengths (supposing the same frame rate). This is a somewhat trickier problem than image morphing because of the added temporal dimension. Now we need to choose what kind of transformation we intend to do. We could simply make a movie that contains an intermediate number of frames, but we need to account for important temporal information to make it more convincing. If the first movie shows an explosion at the beginning (similarly to the abrupt attack of a plucked string or a percussive sound) and the other a butterfly gently flapping its wings and then flying away, we might need to align relevant temporal cues to produce an interesting morph. Moreover, there are a number of possible transitions between the two. Do we want an intermediate movie that contains morphed images of each frame (here called *static* or *stationary morphing* because α is constant), or are we going for a movie that starts as the first and dynamically changes into the other (here called *dynamic morphing* because α varies in time)? We could choose to run the first frames of the first movie until we stop at a selected frame, gradually morph it into another selected frame of the second, and then proceed by showing the rest of it (*warped dynamic morphing*), choosing to somehow warp the length of the result in order to achieve a given effect. Finally, another possibility would be to produce several hybrid sounds in different intermediate points (i.e., different values of α) of the path between source and target (*cyclostationary morphing*).

With these considerations in mind, a world of possibilities opens up, from the trajectory followed by the morph determined by α to the choice of source and target sounds to be morphed between. We just need to bear in mind that all these choices affect the quality of the results and might even be somewhat intertwined. For instance, it might be easier to morph between a trumpet note and the singing voice than drums.

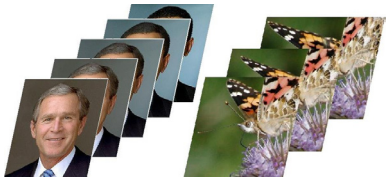

Figure 2. Depiction of two movies shown frame by frame.

3. MORPHING SOUNDS

The aim of this section is to review the morphing techniques and highlight the aim of using descriptors to guide the transformation. Most morphing techniques proposed in the literature consist in describing a model used to analyse the sounds and interpolating the parameters of the model regardless of features [7], [12], [27], [6], [2], [20], [21], [1]. 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 modelling is amongst the most common approaches [7], [2], [11], [20], [21], [27], [30], [33]. Tellman [27] offers us one of the earliest descriptions of a morphing technique, which is based on a sinusoidal representation [6]. The morphing scheme consists of interpolating the result of the Lemur [6] analysis and involves time-scale modification to morph between different attack and vibrato rates. More recently, Fitz [7] presented a morphing technique also using a sinusoidal representation, and morphing is achieved again by simply interpolating the parameters of the model. Hope [13], [14] prefers to interpolate the parameters of a Wiegner Distribution analysis. Boccardi [2], in turn, uses GMM to interpolate between additive parameters (SMS) [25]. Röbel [24] proposes to model sounds as dynamical systems with neural networks and to morph them by interpolating the attractors corresponding to those dynamical systems. Ahmad [1] applies a discrete wavelet transform and singular value decomposition to morph between transient sounds. They interpolate linearly between the parameters and state that other interpolation strategies with a better perceptual correlation should be studied.

A few authors have proposed to detach the spectral envelope from the frequency information and interpolate them separately [1], [5], [4], [26]. Slaney [26] proposes to represent the sounds to be morphed between in a multidimensional space that encodes spectral shape and pitch in orthogonal axes and warp the dimensions of this representation to obtain the desired result. However, they represent spectral shape by MFCCs and pitch information by a residual spectrogram calculation, which are then interpolated using dynamic time warping and harmonic alignment as guides. They conclude by stating that the method should be improved with perceptually optimal interpolation functions. Ezzat [5] uses a spectral smoothing technique to morph spectral envelopes. They analyse soberly the problem of interpolating spectral envelopes and argue that this approach accounts for proper formant shifting between source and target. We shall verify this claim in Section 6, and also verify if it accounts for the morphing of timbral features as a perceptually motivated morphing algorithm should. Finally, only recently did we start to take perceptual aspects into consideration [4], [30],

[31], [11], and the result is the addition of one more step in the process, feature calculation. In most models proposed, linear variation of interpolation parameter does not produce perceptually linear morphs [12], so recently authors have started to study the perceptual impact of their models and how to interpolate the parameters so that the results vary roughly linearly on the perceptual sphere. Williams [30], [31] studies an additive-based perceptually-motivated technique to morph sounds guided by the spectral centroid. They selectively amplify or attenuate harmonics of sawtooth or square waves to tilt the centroid towards that of the target sound. Hikichi [12] uses MDS spaces [18] constructed from the sources and morphed sounds to figure out how to warp the interpolation factor in the parameter space so that it will linearly morph in the perceptual domain. Hatch [11] poses the problem of feature interpolation very clearly but it remains unclear how he matches target values of spectral centroid, for example. Caetano [4] proposes to morph spectral envelopes guided by descriptors controlling the spectral shape by changing the parameters of the spectral envelope model with the aid of a genetic algorithm. In this work, we are going to present strategies to achieve perceptually relevant morphing of quasi-harmonic musical instrument sounds taking most temporal and spectral timbral aspects of sounds into account. Particularly in the spectral domain, we are looking for a spectral envelope representation that best approximates linear interpolation on the perceptual timbre space when we linearly interpolate the parameters. We use quasi-harmonic musical instrument sounds so that the partials have a simple correspondence, although the techniques herein described could easily be extrapolated to vocal sounds (singing voice) or inharmonic sounds.

4. 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 [10]. We know that source identification is not reduced to waveform memorization because the intrinsic dynamic nature of the sources produces variations [10]. Timbre perception is inherently multidimensional, involving features such as the attack, spectral shape, and harmonic content. Since the pioneering work of Helmholtz [29], multidimensional scaling techniques figure among the most prominent when trying to quantitatively describe timbre. McAdams [18] gives a comprehensive review of the early timbre space studies. Grey [8] investigated the multidimensional nature of the perception of musical instrument timbre, constructed a three-dimensional timbre space, and proposed acoustic correlates for each dimension. He concluded that the first dimension corresponded to spectral energy distribution (spectral

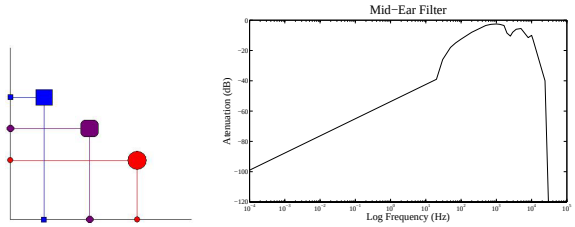

Figure 3. Left: Illustration of two-dimensional timbre space with two sound objects depicted as the circle and the square and one intermediate sound object depicted as the square with rounded corners. Right: Mid-ear filter applied to the spectral envelopes.

centroid), the second and third dimensions were related to the temporal variation of the notes (onset synchronicity). Krumhansl [16] conducted a similar study using synthesized sounds and also found three dimensions related to attack, synchronicity and brightness. Krimphoff [15] 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 [18] 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 [3] 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.

Listeners use many acoustical properties to identify events, such as the spectral shape, formant frequencies, attack (onset) and decay (offset), noise, among others [10]. The cues to identification and timbre vary across notes, durations, intensities and tempos [10]. One model of sound production is based on two possibly interactive components, the source and the filter [10]. The basic notion is that the source is excited by energy to generate a vibration pattern composed of several vibration modes (modelled as sinusoidal components). This pattern is imposed on the filter, which acts to modify the relative amplitudes of the components of the source input [10]. We obtain estimates of 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 and here we refer to it as spectral form to distinguish from the spectral shape, which is correlated with dimensions of timbre spaces obtained from perceptual studies. We note that envelope form and shape complement each other, since there are several possible spectral envelopes with different forms and the

same shape, i.e., values of descriptors. So we say that to obtain perceptually intermediate spectral envelopes we need not only to take spectral form but also spectral shape into account. In other words, we need to obtain a spectral envelope with an intermediate number and absolute position of formant peaks and also intermediate brightness (centroid), roughness (spread), etc. Obtaining an intermediate spectral shape corresponds to placing the sounds between two (or more) in the corresponding underlying timbre space that generated the dimensions. Supposing that timbre space is orthogonal (like in MDS studies), then intermediate points in high-dimensional space have intermediate values for each dimension (that is, intermediate descriptors), as illustrated on the left-hand side of Figure 3. We see a two-dimensional orthogonal abstraction of timbre space where each dimension corresponds to a feature captured by a descriptor. We also see two sound objects represented by the circle and the square and their corresponding features reflected as the values of the descriptors on each axis. The intermediate sound object represented by the square with rounded corners must have intermediate features, and therefore intermediate values of descriptors.

5. HIGH-LEVEL DESCRIPTORS

We measure timbral features with high-level descriptors, such that a sound with intermediate descriptors should be perceived as intermediate. We adopted temporal and spectral features in our study to account for prominent timbre dimensions. The temporal features are log attack and decay times, energy (temporal) envelope, and temporal evolution of harmonic contents, usually referred to as shimmer and jitter. The spectral features are form (formant peaks) and shape (centroid, spread, skewness, kurtosis and slope). Notice that the spectral features are extracted from both the sinusoidal and noise components of the analysis. In this section we present the general scheme used to calculate all he descriptors used in this work, depicted in Figure 4. 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 [23] describes exhaustively how to calculate all the descriptors we use in this work and proposes to use them in audio classification tasks. 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 ("signal frame" and "FFT").

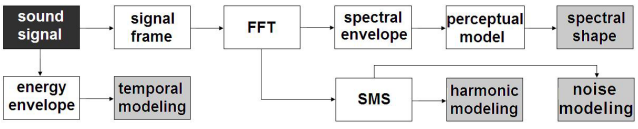

Figure 4. Simplified scheme to calculate the descriptors.

5.1. Temporal Modeling

This step accounts for the estimation of the attack and release times as described in [23]. Firstly we calculate the amplitude (or temporal) envelope, which is a smooth curve that outlines the waveform. We estimate the attack and release times from here. It is important to note that the energy envelope itself must be interpolated in the morphing process. The next steps of the descriptor calculation scheme are repeated for every signal frame, such that variations naturally arising from the (presumably) acoustical nature of the sound source will give rise to shimmer and jitter.

5.2. 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 [23]. For every frame, we calculate the spectral envelope using a cepstral smoothing technique (true envelope [28]). Next, we apply the perceptual model, which consists of the mid-ear filter shown on the right of Figure 3 evaluated on the mel scale. We should notice that the result is similar to the MFCC-based spectral envelope used in [26] without critical band smoothing. Finally we calculate the spectral shape descriptors for the mid-ear attenuated, mel-warped spectral envelope.

5.3. Harmonic Modeling

Here, we need to finally extract the remaining pitch information, i.e., the instantaneous values of the frequencies of the partials. There are many possible ways to do this, but for the sake of fidelity, we chose to perform an SMS-based sinusoidal plus residual analysis [25] (again on every signal frame) and keep only the frequency values of the sinusoidal part. The amplitudes of the partials are already accounted for by the spectral envelope estimation step. Temporal variations on the frequencies of the partials guarantee the naturalness of the tone.

5.4. Noise Modeling

The result of the SMS analysis is a sinusoidal component and a residual that models the noise part of the sound signal. In order to account for this perceptually important feature, we extract the spectral envelope and repeat the 'spectral shape' analysis here. The residual is modeled as pink noise modulated by the envelopes frame by frame.

6. MORPHING BY DESCRIPTORS

The final step of the morphing process consists of morphing between the descriptors with a desired morphing

factor α and then resynthesizing a sound with parameter values that correspond to the morphed features. Some temporal features are somewhat independent from the spectral ones (attack and release times), while others are intrinsically intertwined with them (jitter, shimmer), such that we can manipulate attack an release times by time stretch/compress completely independently from other features, but jitter and shimmer are intrinsically contained in the time-varying nature of the analysis and will naturally morph as we interpolate the parameters. Our approach relies on the alignment of temporal features such as attack and release time, and a spectral envelope morphing technique that produces intermediate envelopes with the desired form (number of peaks) and intermediate spectral shape features. The tricky part is exactly the mapping between spectral shape descriptors and spectral envelope parameters. As other authors noted earlier [1], [4], [12], linear variation of most spectral envelope parameters does not guarantee that the perceptual features will also change linearly, so we will present a study about which spectral envelope representations closely approximate linear interpolations in the descriptor space when linearly interpolated. Ezzat [5] briefly reviews techniques to morph spectral envelopes. First they acknowledge that simply interpolating the envelope curve does not account for proper formant shifting. We should mention that this is exactly what most techniques do when they directly interpolate the amplitudes of a sinusoidal model. Then, they state that interpolating alternative representations of the envelopes, such as linear prediction or cepstral parameters, also poses problems and propose to use dynamic frequency warping (DFW) instead. So, the main motivation of this section is to verify this claim by investigating the perceptual impact of several spectral envelope interpolation schemes [22], namely, the envelope curve (ENV), linear prediction coefficients (LPC), reflection coefficients (RC), line spectral frequencies (LSF), cepstral coefficients (CC) and dynamic frequency warping (DFW). The rest of this section explains each step in our morphing technique.

6.1. Temporal Alignment

First, using the end of attack and beginning of release times estimated [23], we time stretch or compress the attack, sustain and release portions of both sounds to align them temporally [27], [10]. For the attack and release times we use logarithmic interpolation.

6.2. Spectral Envelope Shape

We represent morphing by descriptors as weighted interpolation in the feature space representation, much like in [1], [26]. The fundamental difference is that our space corresponds to perceptual dimensions so there is no direct inversion for resynthesis. Instead, we are trying to find the

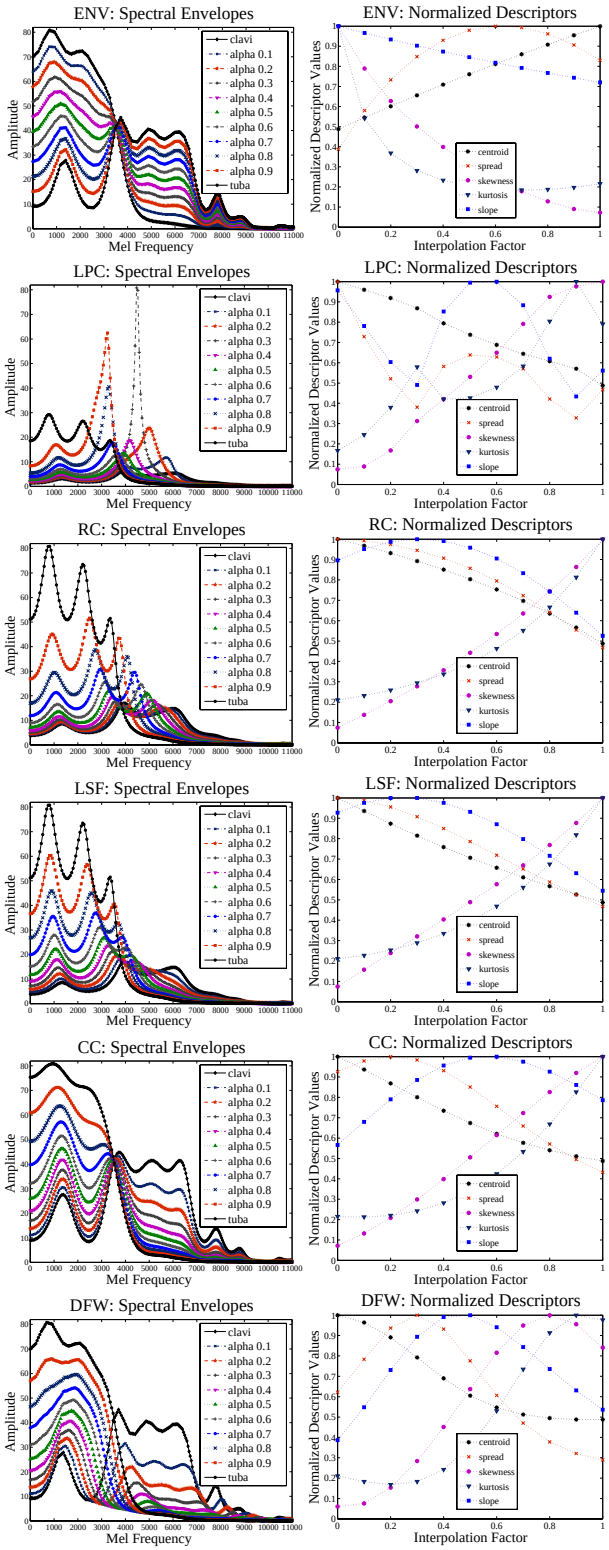

Figure 5. Perceptual impact of interpolating between the parameters of several spectral envelope models. The curves are shown on the left and the corresponding descriptor variation on the right.

spectral envelope model whose associated descriptors interpolate the closest possible to linearly when its parameters are linearly interpolated. Figure 5 illustrates the impact on the spectral shape descriptor domain of interpolating cepstral, linear prediction, and dynamic frequency warping based spectral envelope model parameters for two very challenging envelopes. On the left, Figure 5 shows the source and target envelopes in solid lines and nine intermediate envelopes corresponding to linearly varying the interpolation factor by 0.1 steps in dashed and dotted lines; on the right, we see the associated values of the spectral shape descriptors for each step. When evaluating Figure 5 we have to take into account spectral from and shape, that is, we want the envelope model that accounts properly for formant shifting and whose spectral shape descriptors vary as a straight line. The apparent difference in shape of the source and target for linear prediction based envelopes (LPC, RC and LSF) is due to the conversion from cepstral estimation. The conversion from cepstral to linear prediction based spectral envelope introduces artifacts, but we still consider that the result is better than extracting the envelope directly with linear prediction [28]. Figure 5 confirms for this case (we will extrapolate the conclusions) that interpolating envelope curves does not account for formant shifting and most spectral shape descriptors do not vary in a straight line. Moorer [19] states that LPCs do not interpolate well because they are derived from impulse responses, and therefore too sensitive to changes, and Figure 5 seems to confirm that. Figure 5 also shows that the linear interpolation of cepstral based envelope representations like Slaney [26] proposes neither shifts the formants nor results in linear variation of descriptors. The same applies for the DFW based spectral envelope morphing proposed by Ezzat [5]. On the other hand, RC and LSF behave fairly well under both constraints in this case just like Paliwal [22] states for LSFs. The only inconvenient could be the initial distortion caused by the conversion from using a cepstral smoothing envelope estimation technique.

6.3. Harmonic structure

Here we propose to morph quasi-harmonic musical instrument sounds with the same pitch, so that the partials have a one to one correspondence and no pitch shift is required. Since the spectral shape and form are morphed separately with the spectral envelope, we simply interpolate the partials frequency values to account for frequency fluctuations (jitter, shimmer), inharmonicity and other temporal features that are encoded in the frequency variation with time.

6.4. Amplitude Envelope

Here we simply interpolate the amplitude envelope curve and modulate the amplitude of the morphed sinusoidal component with it.

6.5. Stochastic Residual

We morph the spectral envelopes of the residual noise signal and synthesize a morphed residual by filtering pink noise with it and mixing it into the morphed sinusoidal component.

7. CONCLUSIONS AND FUTURE PERSPECTIVES

In this work, we describe techniques to automatically morph salient timbral dimensions of quasi-harmonic musical instrument sounds guided by high-level descriptors. High-level descriptors are acoustic correlates of timbre dimensions obtained in psychoacoustic studies, such that sounds whose features are intermediate between two would be placed between them in the underlying timbre space. So, interpolating the descriptor values becomes the goal itself to render the results more perceptually meaningful. We also reviewed the definitions and goals of sound morphing in the literature to try and establish common grounds for future research. Moreover, we reviewed the morphing techniques proposed so far and whether they took the perceptual impact into account. Finally, we evaluated the perceptual impact of interpolating the parameters of several spectral envelope models aiming to find which models correspond the closest to morphing in the underlying timbre space, that is, in the perceptual domain as measured by the descriptors. We investigated direct interpolation of the envelope curve, LPC, RC, LSF, CC, and DFW. We concluded that RC and LSF correspond the closest to morphing the descriptors linearly when linearly interpolated. Examples avaliable on http://recherche.ircam.fr/equipes/analyse-synthese/caetano/icmc2010.html.

Future perspectives of this work include experimenting with different trajectories in timbre space determined by different time-varying morphing factors. It is also interesting to explore techniques to independently morph each timbre dimension by manipulating the descriptors with different morphing factors. Some technical aspects could be improved, such as extracting the temporal envelope for each partial and estimating the attacks independently to simulate onset asynchrony, include vibrato modeling and treatment, extending the technique to inharmonic sounds (would need different interpolation of harmonic structure), improving the estimation of attack time for percussive or plucked sounds. Also tremolo could be dealt with by developing a better energy envelope morphing than simply interpolate the curves. Finally, we could possibly extend the model to any sound object to finally be able to obtain a ‘barking trumpet’, for example.

8. ACKNOWDGEMENTS

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

9. REFERENCES

[1] 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.

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

[3] 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.

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

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

[6] Fitz, K., Haken, L. *Sinusoidal Modeling and Manipulation Using Lemur*. Computer Music Journal, 20 (4), pp. 44-59, 1996.

[7] 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.

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

[9] Haken, L., Fitz, K.,Christensen, P. "Beyond Traditional Sampling Synthesis: Real-Time Timbre Morphing Using Additive Synthesis" in Beauchamp, J. W., ed *Sound of Music: Analysis, Synthesis, and Perception*. Berlin: Springer-Verlag, 2006.

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

[11] Hatch, W. "High-Level Audio Morphing Strategies" MA Thesis, Music Technology Dep., McGill University, 2004.

[12] Hikichi, T., Osaka, N. *Sound Timbre Interpolation Based on Physical Modeling*. Acoustical Science and Technology, 22 (2), pp. 101-111, 2001.

[13] Hope, C. J., Furlong, D. J. "Endemic Problems in Timbre Morphing Processes: Causes and Cures". *Proc. ISSC*, 1998.

[14] Hope, C. J., Furlong, D. J. "Time-frequency Distributions for Timbre Morphing: the Wigner distribution versus the STFT". *Proc. SBCM*, pp. 99-110, 1997.

[15] 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.

[16] 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.

[17] Letowski, T. *Timbre, Tone Color, and Sound Quality: Concepts and Definitions*. Archives of Acoustics, 17 (1), pp. 17-30, 1992.

[18] 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.

[19] Moorer, J. A., "The Use of Linear Prediction of Speech in Computer Music Applications" J. Audio Eng. Soc., 27 (3), pp. 134-140, 1979.

[20] Osaka, N. "Concatenation and Stretch/Squeeze of Musical Instrumental Sound Using Morphing" *Proc. ICMC*, 1995.

[21] Osaka, N. "Timbre Interpolation of Sounds Using a Sinusoidal Model" *Proc. ICMC*, 1995.

[22] Paliwal, K. "Interpolation Properties of Linear Prediction Parametric Representations" *Proc. Eurospeech*, 1029-1032, 1995.

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

[24] Roebel, A. "Morphing Dynamical Sound Models" Proc. IEEE Workshop Neural Net Sig. Proc, 1998.

[25] Serra, X. "Musical Sound Modeling with Sinusoids Plus Noise" in *Musical Signal Processing*, Swets & Zeitlinger, 1997.

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

[27] 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.

[28] Villavicencio, F., Robel, A., Rodet, X. "Improving LPC Spectral Envelope Extraction of Voiced Speech by True Envelope Estimation". *Proc. ICASSP*, 2006.

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

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

[31] Williams, D., Brookes, T. "Perceptually-Motivated Audio Morphing: Brightness", *AES 122nd Convention*, 2007.

[32] Wishart, T. *On Sonic Art*. Simon Emerson: Harwood Academic Publishers, ISBN 3-7186-5847-X, 1998.

[33] Wishart, T. *SoundHack*. Computer Music Journal, 21 (1), pp. 10-11, 1997.