Musical Instrument Sound Morphing Guided by Perceptually Motivated Features

Marcelo Caetano, *Member, IEEE*, and Xavier Rodet

***Abstract*—Sound morphing is a transformation that gradually blurs the distinction between the source and target sounds. For musical instrument sounds, the morph must operate across timbre dimensions to create the auditory illusion of hybrid musical instruments. The ultimate goal of sound morphing is to perform perceptually linear transitions, which requires an appropriate model to represent the sounds being morphed and an interpolation function to obtain intermediate sounds. Typically, morphing techniques directly interpolate the parameters of the sound model without considering the perceptual impact or evaluating the results. Perceptual evaluations are cumbersome and not always conclusive. In this work, we seek parameters of a sound model that favor linear variation of perceptually motivated temporal and spectral features used to guide the morph towards more perceptually linear results. The requirement of linear variation of feature values gives rise to objective evaluation criteria for sound morphing. We investigate several spectral envelope morphing techniques to determine which spectral representation renders the most linear transformation in the spectral shape feature domain. We found that interpolation of line spectral frequencies gives the most linear spectral envelope morphs. Analogously, we study temporal envelope morphing techniques and we concluded that interpolation of cepstral coefficients results in the most linear temporal envelope morph.**

***Index Terms*—Musical instrument sounds, sound morphing, source-filter model.**

I. Introduction

SOUND morphing figures prominently among the sound transformation techniques studied in the literature due to its great creative potential and myriad possible outcomes. Sound morphing has been used in music compositions [1]–[3], sound synthesizers [4], and even in psychoacoustic experiments, notably to study timbre spaces [5]. However, there seems to be no consensus in the literature on which transformations fall into the category of sound morphing and there certainly is no widely accepted definition of the morphing process for sounds. Most authors seem to agree that sound 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, ruling out simply mixing or crossfading the sounds [4], [6] because the ear is capable of distinguishing them due to a number of cues and auditory processes. Still, many different sound transformations are described as morphing, such as interpolated timbres [4], smooth or seamless transitions between sounds [7] or cyclostationary morphs [8]. In a previous work [9], we thoroughly reviewed the different types of sound transformation that are usually termed morphing and evaluated how the temporal nature of the morphing transformation (stationary, dynamic, etc) directly interferes in the requirements of the process.

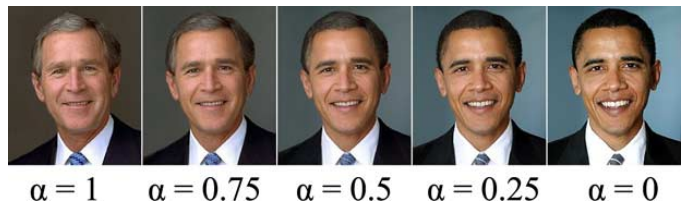

Fig. 1. Depiction of image morphing to exemplify the aim of sound morphing. Found online at http://tinypic.com/images/404.gif, currently publicly available at http://paulbakaus.com/wp-content/uploads/2009/10/bush-obama-morphing.jpg.

When morphing musical instrument sounds, we usually want to transform across timbre dimensions to create the auditory illusion of hybrid musical instruments, gradually blurring the categorical distinction between the source and target sounds. Fig. 1 illustrates this effect for images. A challenging aspect of such transformations is to control the morph on the *algorithmic* and *perceptual* levels with a single coefficient α, called morphing or interpolation factor [9]. Ideally, we would like to obtain a morphed sound *perceptually* halfway between source and target when $\alpha = 1/2$, and be able to recursively repeat the process for $\alpha = 1/2^n$. Equivalently, linear variation of α should lead to a perceptually linear transformation. The concept of *perceptual linearity* in sound morphing lies at the core of this work, where we use *perceptually motivated features* to guide the transformation and evaluate *linearity* in the feature domain. We assume that linear variation in the feature domain indicates perceptual linearity when the features capture perceptually relevant information.

Most morphing techniques proposed in the literature directly apply the interpolation principle without taking perceptual aspects into consideration [4], [6], [7], [10]–[15]. In this work, *parameter* refers to coefficients from which we can resynthesize sounds (e.g., spectral peaks), while *feature* refers to coefficients used to describe or identify a particular aspect of a sound (e.g.,

Manuscript received August 21, 2012; revised December 22, 2012; accepted March 26, 2013. Date of publication April 25, 2013; date of current version May 08, 2013. This work was supported by the Brazilian government under a CAPES grant (process 4082-05-2) while M. Caetano was pursuing the Ph.D. degree at the Analysis/Synthesis team, IRCAM. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. James Johnston.

M. Caetano is with the Signal Processing Laboratory, Foundation for Research and Technology—Hellas, Institute of Computer Science (FORTH-ICS), GR 700133, Heraklion, Crete, Greece (e-mail: caetano@ics.forth.gr).

X. Rodet is with the Analysis/Synthesis Team, Institut de Recherche at Coordination Acoustique/Musique (IRCAM), 75004, Paris, France.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2260154

1558-7916/$31.00 © 2013 IEEE

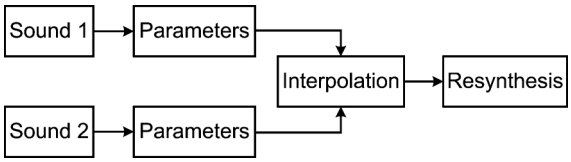

Fig. 2. Depiction of the classic morphing scheme using the interpolation principle, which assumes that perceptually intermediate representations possess intermediate parameter values.

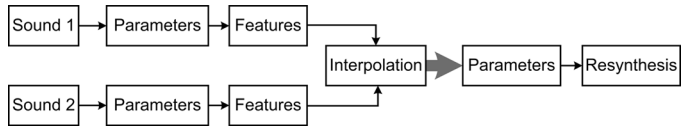

Fig. 3. Depiction of the morphing by feature interpolation principle adopted in this work, which advocates that perceptually intermediate representations present intermediate feature values rather than intermediate parameter values. Notice that the step represented by the grey arrow implies retrieving parameters from features.

spectral centroid). Features are commonly related to sound perception, so it is usually not possible to resynthesize sounds directly from feature values. The interpolation principle, depicted in Fig. 2, supposes that if we can represent different sounds by simply adjusting the parameters of a model, we should obtain a somewhat smooth transition between two sounds by interpolating the parameter values of their representations.

Interpolation of sinusoidal models [16], [17] is among the most common approaches to sound morphing [4], [6], [11], [13]–[15], [18], [19]. In what is perhaps the first major work devoted specifically to morphing, Tellman *et al.* [4] proposed to interpolate the amplitude and frequency values resulting from the sinusoidal model dubbed Lemur [6]. Their focus is synthesizers and how to produce sounds with intermediate features such as loudness and vibrato from pre-recorded sounds. Their morphing scheme involves time-scale modification to morph between different attack and vibrato rates.

More recently, Fitz *et al.* [10] presented a morphing technique using the enhanced-bandwidth sinusoidal modeling called Loris [10], and morphing is achieved again by simply interpolating the parameters of the model. They recognize the need to temporally align the sounds to be morphed. However, they do not have an automatic procedure to do so, rather, they annotate by hand what they consider to be the perceptually relevant temporal cues, such as start and end of attack.

Hope and Furlong [20], [21] prefer to interpolate the parameters of a Wigner distribution analysis. Boccardi and Drioli [11], in turn, used Gaussian mixture models to interpolate between sinusoidal modeling parameters [16], [17]. Röbel [22] proposed to model sounds as dynamical systems with artificial neural networks and to morph them by interpolating the corresponding attractors. Ahmad *et al.* [7] applied 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 amplitude from the frequency of the partials with spectral envelopes and morph them separately [7], [8], [23]–[27]. Slaney *et al.* [8] proposed to morph spectral envelopes by cross-fading (time-varying interpolation) between the mel-frequency cepstral coefficients [28] that represent each spectral envelope, focusing on dynamically varying sounds such as words. First of all, they use the widely known dynamic time warping (DTW) algorithm to align temporal events in the sounds. Their conclusion is that the method should be improved with more perceptually optimal interpolation functions. Pfitzinger [23] used dynamic frequency warping (DFW), a frequency domain counterpart of DTW, in a spectral smoothing approach applied to concatenative speech synthesis.

Ezzat *et al.* [24] studied the use of DFW to morph spectral envelopes in the context of musical sounds, analyzing soberly the problem of interpolating spectral envelopes and arguing that the spectral envelope morphing technique should shift the peaks of the spectral envelope (also called formant peaks) between source and target. They acknowledge that simply interpolating the envelope curve does not account for proper formant shifting, which is where direct interpolation of the amplitudes of a sinusoidal model commonly fails to render more perceptually linear results. Then, they state that interpolating alternative representations of the envelopes, such as linear prediction or cepstral coefficients, also poses problems and propose to use DFW instead. However, formant shifting alone does not guarantee perceptual linearity.

In most proposed models, linear variation of interpolation parameters 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 and Brookes [14], [15] studied a perceptually-motivated technique to morph simple synthetic sounds guided by the spectral centroid. Hikichi [12] used multidimensional scaling (MDS) spaces [29], [30] 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. In [26], [27], we proposed to morph spectral envelopes guided by features controlling the spectral shape by changing the parameters of the spectral envelope model with the aid of a genetic algorithm.

Sound transformations that use features to control perceptually related aspects such as *pitch*, *loudness*, or *brightness* are called *content-based transformations* [31] or *adaptive sound effects* [32] in the literature. The aim of such transformations is to use the feature values to control the result perceptually. For instance, doubling the value of the *spectral centroid* to obtain a sound that is twice as *bright*. In [9], [25], we introduced the concept of *sound morphing by feature interpolation*, illustrated in Fig. 3, as an alternative to directly interpolating parameters of a sound representation. In this article, we present an in-depth study of musical instrument sound morphing using the feature values as objective measure of linearity, followed by a listening test to cross-evaluate the results perceptually.

Morphing by feature interpolation advances that sounds with intermediate values of features are perceptually intermediate when the features capture perceptually relevant information. Therefore, we should extract features from these parameters, interpolate the feature values, and retrieve the set of parameter values that correspond to the interpolated feature values. However, the step represented by the grey arrow in Fig. 3 would

require retrieving parameter values from feature values, a notoriously difficult problem [32]. Instead, in this work, we seek to interpolate parameters of a sound model to obtain morphed sounds whose values of features are as close as possible to the interpolated feature values. We propose to use the feature values as objective measure of the perceptual impact of the morphing transformation, requiring the features to vary in a straight line when α changes in equal steps (linearly).

Section II introduces the features used in this work along with their psychoacoustic background from MDS studies. Then, we present an overview of the proposed musical instrument morphing procedure and the temporal and spectral models used. Next, we discuss morphing the spectral and temporal envelopes guided by the features, followed by the evaluation of linearity in the feature domain. Finally, we conclude and discuss future perspectives.

II. The Features Used as Guides

The features used in this work are derived from the acoustic correlates of timbre spaces from multidimensional scaling (MDS) studies of timbre perception. We include temporal and spectral features to capture the most perceptually salient dimensions of timbre perception, namely, the attack time and the distribution of spectral energy. The temporal features we use are the *log attack time* and the *temporal centroid*. The spectral shape features *spectral centroid*, *spectral spread*, *spectral skewness*, and *spectral kurtosis* we use are a measure of the balance of spectral energy.

A. Acoustic Correlates of Timbre Spaces

MDS techniques figure among the most prominent when trying to quantitatively describe timbre. The MDS algorithm maps subjective distances (perceptual dissimilarity between musical instrument sounds) into an orthogonal metric space which has the number of dimensions specified by the investigator. McAdams [29] and Handel [33] independently propose comprehensive reviews of the early timbre space studies. Grey [30] 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 centroid*), the second and third dimensions were related to the temporal variation of the partials (*onset synchronicity* and *spectral fluctuation*). Krumhansl [34] conducted a similar study using synthesized sounds and also found three dimensions related to *attack*, *synchronicity* and *brightness*. Krimphoff [35] studied acoustic correlates of timbre spaces and concluded that *brightness* is correlated with the *spectral centroid* and *rapidity of attack* with *rise time* in a logarithmic scale. McAdams [29] 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 [36] 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.

In timbre spaces obtained with MDS, the distances between pairs of instruments represent the perceptual dissimilarity between them. Timbre space representations are essentially sparse in nature. The space is mostly void and the musical instruments occupy non-overlapping areas. When the morphed sound is perceptually intermediate between two musical instrument sounds, it would be placed between them in the underlying timbre space, "filling" the voids and allowing exploration of the sonic continuum. When the features guiding the morph are acoustic correlates of timbre dimensions, intermediate feature values would correspond to intermediate positions in the timbre space.

B. Temporal Features

The attack is the beginning of the acoustic stimulus, present in all sounds. Psychoacoustic (dis)similarity studies [29], [30], [33]–[37] discovered that the attack is among the most perceptually salient features of musical instrument sounds. These studies have shown that the attack time is perceived roughly on a logarithmic scale.

The *log attack time* Λ is calculated as shown in (1), where λ_1 stands for the beginning of the attack A and λ_2 for the end (see Fig. 5). The *temporal centroid* is the measure of the balance of energy distribution along the course of a sound and is calculated as in (2), where τ represents the temporal centroid, t is time, and $a(t)$ is the value of the temporal envelope for time t. The temporal centroid has been shown [38] to be especially important when comparing percussive and sustained sounds because that is when it varies more significantly, allowing us to distinguish between the two classes. Still, in the context of strictly sustained sounds, the attack times and temporal centroids vary significantly enough to be relevant.

$$\Lambda = \log(\lambda_2 - \lambda_1) \tag{1}$$

$$\tau = \frac{\sum_t t a(t)}{\sum_t a(t)} \tag{2}$$

C. Spectral Shape Features

The spectral shape features are calculated on every frame, which permits to follow their temporal variation. The spectral centroid is one of the most salient features in psychoacoustic studies [29], [30], [33]–[37] correlated with the verbal attribute "brightness." Spectral spread is a measure of the bandwidth of the spectrum. Spectral skewness and spectral kurtosis were shown to be significantly correlated with 2 out of 27 dimensions of 10 timbre spaces tested in a study [37] of acoustic correlates of timbre dimensions.

The spectral shape features δ_i are the first four standardized moments of the normalized magnitude spectrum $p(k)$ viewed as a probability distribution defined in (3), where $|X(k)|$ is the magnitude spectrum, the frequencies k are the possible outcomes, and the probabilities to observe them are $p(k)$.

$$p(k) = \frac{|X(k)|}{\sum_k |X(k)|} \tag{3}$$

Following this definition, the spectral shape features δ_i are

$$\delta_1 = \mu = \sum_k k p(k) \tag{4}$$

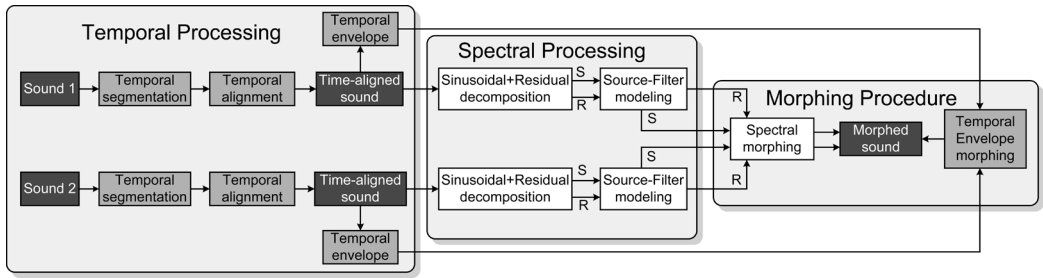

Fig. 4. Depiction of the general steps of the musical instrument sound morphing procedure. There are three distinct parts, temporal processing, spectral processing, and morphing procedure. Blocks with dark grey background represent waveforms, blocks with light grey background represent temporal feature extraction and processing, and blocks with white background represent spectral feature extraction and processing.

$$\delta_2 = \sigma^2 = \sum_k (k-\mu)^2 \, p(k) \tag{5}$$

$$\delta_3 = \frac{\sum_k (k-\mu)^3 \, p(k)}{\sigma^3} \tag{6}$$

$$\delta_4 = \frac{\sum_k (k-\mu)^4 \, p(k)}{\sigma^4}. \tag{7}$$

The spectral centroid δ_1 is the mean of $p(k)$ and the spectral spread δ_2 is the variance around the mean, shown respectively in (4) and (5). The spectral skewness δ_3, shown in (6), measures the asymmetry of $p(k)$ around the spectral centroid, while spectral kurtosis, shown in (7), is a measure of the peakedness of $p(k)$ relative to the normal distribution.

Notice that the spectral shape features δ_i have different units. The spectral centroid δ_1 is measured in Hertz, the spectral spread δ_2 in Hertz squared, and both the spectral skewness δ_3 and spectral kurtosis δ_4 are dimensionless. Furthermore, we can use different frequency and amplitude scales when calculating the spectral shape features to better approximate the spectral information that reaches the ear. In this work, we use the mel scale [39] to warp the frequency axis and logarithmic amplitude to better represent *loudness* perception.

III. Modeling Musical Instrument Sounds

The morphing technique we developed comprises three steps, temporal processing, spectral processing, and the morphing procedure. Fig. 4 illustrates each step. The blocks represent modeling and processing operations, and the arrows indicate the order in which they are applied. Blocks with dark gray background represent waveforms, light gray background represents temporal modeling or processing, and white background is spectral modeling or processing. The temporal processing step consists of temporal segmentation, temporal alignment, and temporal envelope estimation. The spectral processing step comprises sinusoidal plus residual decomposition followed by source-filter modeling of both the sinusoidal and residual components, which are morphed separately and then mixed back together.

A. Temporal Segmentation

Temporal segmentation consists in estimating the boundaries of four perceptually important regions, namely, *attack*, *transition*, *sustain*, and *release*. Naturally, the sounds can be annotated by hand [4], [10], but in this work we want to automatically segment the sounds. Ideally, morphing algorithms should take two (or more) sounds as input and automatically output the morphed sound according to the value of α.

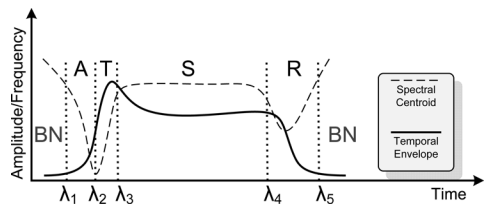

Fig. 5. Amplitude/Centroid Trajectory (ACT) model used in the automatic temporal segmentation of musical instrument sounds. The solid line represents the temporal envelope and the dashed line is the spectral centroid. The numbers stand for the boundaries of the perceptually salient regions, represented by the letters.

The automatic segmentation technique we proposed elsewhere [40] uses the Amplitude/Centroid Trajectory (ACT) model [41] depicted in Fig. 5, where A stands for attack, T is transition, S is sustain, R is release, and BN is background noise. The ACT uses the temporal envelope and the spectral centroid to estimate the boundaries (numbered lambdas) of the regions (letters). From these estimations, we calculate the length of each region.

B. Temporal Alignment

The attack is characterized by fast transients, and the sustain part is much more stable. Therefore, if we combine a sound that has a long attack with another sound with a short one without prior temporal alignment, the region where attack transients are combined with more stable partials would not sound natural. To achieve a more perceptually seamless morph, we need to temporally align these regions so that their boundaries coincide before combining them.

The temporal alignment procedure makes sure that the different regions A, T, S, and R are synchronized for the sounds being morphed. Algorithmically, temporal alignment means aligning the numbered lambdas λ_i from the ACT model as follows. For each sound, we measure the length of each region (labeled with letters) by computing the time difference using the markers (numbered lambdas). Then, we interpolate between the lengths of the regions according to (8) to obtain their corresponding lengths in the morphed sound. For the attack A we interpolate Λ from (1) instead. The interpolated lengths are represented by a letter that stands for the region and subscripts indicating both sounds, e.g., S_{12} for the sustain as shown in (8)

$$S_{12} = \alpha S_1 + [1-\alpha] \, S_2 \tag{8}$$

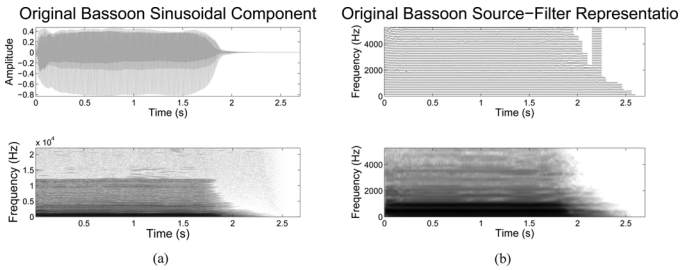

Fig. 6. Comparison between the spectro-temporal view of the sinusoidal and SF representations. Part a) shows the waveform (top) and spectrogram (bottom). Part b) shows the source (top) and filter (bottom). The source is represented as the temporal variation of the frequencies of the partials, while the filter is a time-varying spectral envelope. (a) Waveform and spectrogram and (b) source and filter

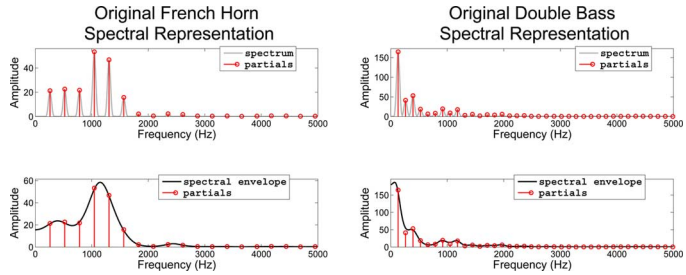

Fig. 7. Spectral view of the source-filter model. Each subfigure shows the traditional sinusoidal representation at the top and the source-filter representation at the bottom for one analysis frame.

where S_1 represents the length of the sustain of the first sound and S_2 of the second. The stretch/compress factors ν_{S1} and ν_{S2} applied on the first and second sounds respectively are calculated as in (9)

$$\nu_{S_1} = \frac{S_1}{S_{12}}, \quad \nu_{S_2} = \frac{S_2}{S_{12}}. \tag{9}$$

Finally, we simply time stretch/compress each region by the corresponding ratio, for instance, S_1 by ν_{S1}, etc. After temporal alignment, both musical instrument sounds are ready to be morphed in the spectral domain.

C. Temporal Envelope Estimation

The amplitude modulations of musical instrument sounds and speech are important perceptual cues. Accurate estimation of the temporal envelope of a complex waveform (such as music or speech) is not a trivial task. Ideally, the amplitude envelope should outline the waveform connecting the main peaks and avoiding over fitting. In this work, temporal envelope estimation is performed with the *true amplitude envelope* (TAE) technique we developed [42], based on cepstral smoothing. TAE gives a reliable estimation that follows closely sudden variations in amplitude and avoids ripples in more stable regions with near optimal order selection depending on the fundamental frequency of the signal.

D. Sinusoidal Plus Residual Decomposition

The aligned musical instrument sounds are decomposed into a sinusoidal and a residual parts, which are modeled independently as source and filter. For musical instrument sounds, the sinusoidal component contains most of the acoustic energy present in the signal because musical instruments are designed to have very steady and clear modes of vibration. The residual component, obtained by subtraction of the sinusoidal component from the original recording, contains mostly noisy modulations.

E. The Source-Filter Model

The source-filter (SF) model we developed [43] represents source and filter independently, as shown in Fig. 6. The sinusoidal component comprises a time-varying spectral envelope (filter) and the time-varying frequency values for the partials (source). The residual component is modeled as white noise (source) driving a time-varying spectral envelope (filter).

The selection of the spectral envelope estimation method for the sinusoidal and residual components is very important. The estimation of the spectral envelope is intimately linked to the SF model because it corresponds to the identification of the parameters of the filter. The main goal of this deconvolution between source and filter by means of spectral envelope estimation is to eliminate the harmonic structure of the spectrum, which is associated with the source. Ideally, for the sinusoidal component, the spectral envelope should be a smooth curve that approximately matches the peaks of the spectrum. Wen and Sandler [44] propose to use the channel vocoder to model the filter part. However, Röbel [45] showed that "true envelope" (TE) outperformed the spectral envelope estimation methods tested, minimizing the estimation error for the peaks of the spectrum. Thus we use TE to estimate the spectral envelope curve of the sinusoidal component.

Fig. 7 presents a comparison of the sinusoidal and the SF representation of the amplitudes of partials. The top part of each subfigure shows the original spectrum (solid line) and the partials (vertical spikes), i.e., the spectral peaks selected by the peak-picking algorithm. At the bottom part, we see the partials from sinusoidal analysis (vertical spikes) and the spectral envelope curve estimated with "true envelope" (solid curve) representing the amplitude of the partials. Both representations retain essentially the same information (amplitude and frequency of partials) in different ways. The frequencies of the partials are now the values at which we "sample" the spectral envelope curve. The sinusoidal model has a more accurate representation of the amplitudes of the partials, while the SF representation is much more flexible to perform sound transformations [43]. We use linear prediction to estimate the spectral envelope of the residual component because the envelope curve follows the average energy of the magnitude spectrum rather than fit the amplitudes of the spectral peaks.

IV. Morphing Musical Instrument Sounds

The morphing steps comprise spectral envelope morphing, interpolation of frequencies of partials, and temporal envelope morphing. Each frame is morphed separately in the spectral domain. The morphed temporal envelope modulates the morphed spectral frames upon resynthesis. For each frame, the morphed spectral envelope gives the amplitude of each partial at the value of the interpolated frequencies. Sound examples can be found

on http://recherche.ircam.fr/anasyn/caetano/overview.html. This section explains thoroughly each morphing step, paying particular attention to how the selected features guide spectral and temporal envelope morphs. Linearity in the feature domain will be used in Section V as objective measure to select a representation for the spectral and temporal envelope morphing techniques. Finally, a listening test was performed to validate the results of the objective evaluation. Section V presents a systematic evaluation using twenty-six (26) pairs of sounds.

A. Spectral Envelope Morphing

The peaks of the spectral envelope are the frequency regions where spectral energy is concentrated. For musical instruments, these perceptually important spectral regions are associated with timbre perception [33]. The spectral envelope morphing technique must shift in frequency the peaks of the spectral envelope [24]. Moreover, the amplitudes of these peaks must also change accordingly to ensure that the transition will be perceived as smoothly as possible. In other words, the balance of spectral energy should gradually shift from source to target when the spectral envelope morph is perceived linearly [25]. Therefore, in this work, the spectral envelope morphing technique must satisfy both requirements, namely, spectral envelope peak shifting, and variation of spectral shape features as close as possible to a straight line when its parameters are interpolated in equal steps (i.e., linearly).

Fig. 8 shows an example to illustrate the morph using several representations proposed in the literature: the envelope curve (ENV) [4], [6], cepstral coefficients (CC) [8], dynamic frequency warping (DFW) [23], [24], linear prediction coefficients (LPC) [46], reflection coefficients (RC) [46], and line spectral frequencies (LSF) [47]–[50]. Fig. 8(a) shows the source and target envelopes in solid lines and nine intermediate envelopes in dashed and dotted lines corresponding to linearly varying the interpolation factor α by 0.1 steps. Fig. 8(b) shows the associated values of the spectral shape descriptors for each step. We want the technique that properly accounts for peak shifting and exhibits linearity in the spectral shape feature domain.

Fig. 8 suggests that ENV does not account for peak shifting. In this case, visually, most spectral shape descriptors change fairly close to a straight line. Interestingly, Fig. 8 reveals that the interpolation of CC does not shift the peaks of the spectral envelope in frequency. In fact, the figure suggests that the result of interpolating CC is very similar to ENV. The variation of spectral shape features reveals that these are different transformations. Fig. 8 shows that DFW results in peak shifting. However, the spectral shape features do not vary close to a straight line. Moorer [46] states that LPC do not interpolate well because they are derived from impulse responses, and therefore too sensitive to changes. Fig. 8 seems to confirm that. In the literature [46], RC are a more robust alternative representation of LPC. Fig. 8 reveals that the transformation is smooth. However, the spectral shape features do not change linearly under interpolation of RC. Itakura [51] proposed LSFs as an attractive alternative representation for LPC because of several properties [50], including peak shifting [47]–[49]. The example in Fig. 8 shows that LSF are indeed suitable parameters to represent and interpolate the spectral envelopes because the peaks shift properly and the spectral shape features change rather linearly.

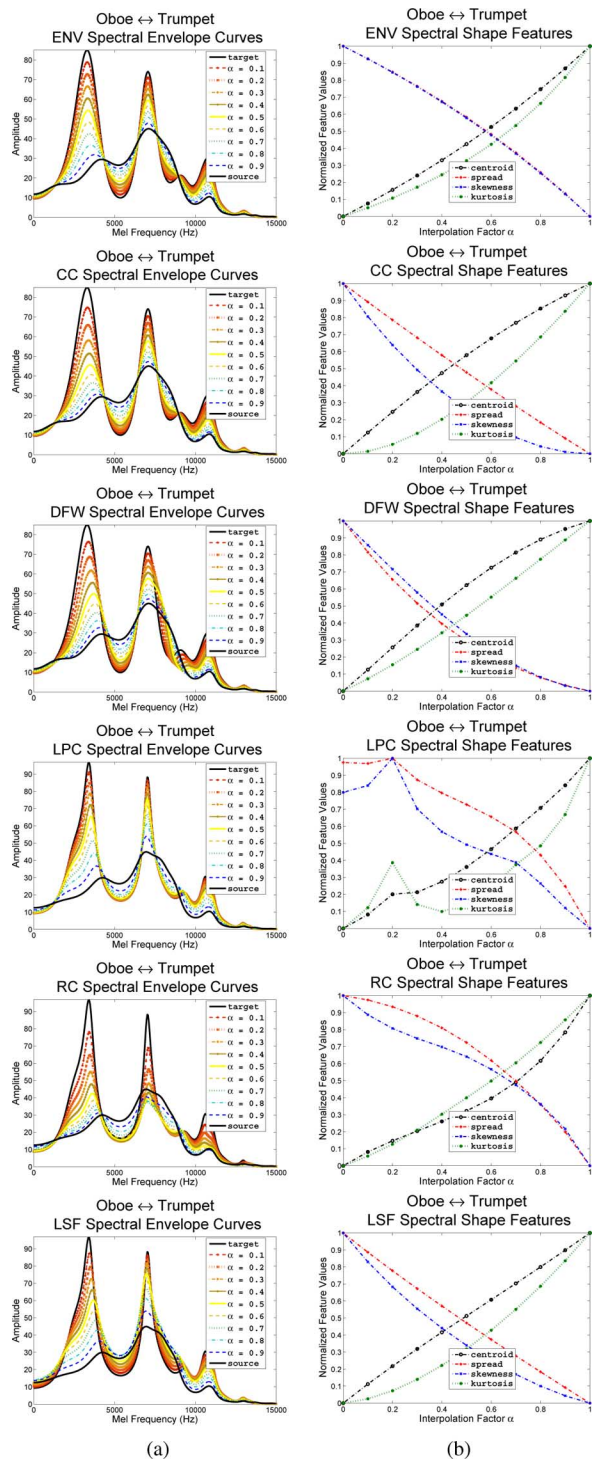

Fig. 8. Spectral envelope morphing guided by spectral shape features. The figure shows the variation of the values of spectral shape features when morphing spectral envelopes using the main approaches proposed in the literature. Part a) shows the spectral envelope curves and part b) shows the corresponding variation of feature values. We want the spectral envelope morphing algorithm that leads to linear variation of spectral shape features. (a) Spectral envelopes. (b) Spectral shape features.

B. Interpolation of Partials Frequencies

The frequencies of the partials, the source in the SF model, carry perceptually important information in the form of temporal frequency modulations. For example, when one of the sounds to be morphed presents vibrato, the transformation should gradually change between more stable partials and vibrato modulations. This can be achieved by interpolating the interval ς in cents between frequency f_{n1} and frequency f_{n2} as expressed in (10), where f_{n1} represents the frequency value of the nth partial of the first sound, and f_{n2} the frequency value of the nth partial of the second sound. Equation (11) presents how to interpolate the interval ς in cents rather than the frequency values f_{n1} and f_{n2} directly. We define the frequency f_α with the aid of (10) as a fraction α of the interval ς in cents and use the value of $f_{n\alpha}$ as the frequency of the nth interpolated partial.

$$\varsigma = 1200 \log_2 \left(\frac{f_{n1}}{f_{n2}} \right) \tag{10}$$

$$f_{n\alpha} = f_{n1} 2^{(\alpha\varsigma/1200)} = f_{n1} 2^{\alpha \log_2 (f_{n1}/f_{n2})} \tag{11}$$

The correspondence between the partials should be carefully considered. Osaka [19] proposed an algorithm to find the optimal solution to the problem of correspondence between two sets of partials derived from sinusoidal analysis by minimizing the distance between the frequency intervals for all possible matches of partials (one-by-one). For near harmonic musical instrument sounds, simply matching the partial number might be enough. However, one sound might have more partials than the other, in which case we could simply discard the unmatched partials. Another possibility is to include a harmonic estimate of the unmatched partial f_n based on the fundamental frequency f_1 and the harmonic number n as $f_n \simeq n f_1$. However, this substitution can only be used when both sounds are nearly harmonic. When there is a slight harmonic deviation (such as piano sounds, whose upper partials deviate farther and farther from perfectly harmonic), we must only interpolate the intervals in cents between pairs of partials that were detected. Alternatively, we can use a model of the inharmonicity to predict the frequencies of upper partials that were not detected and therefore do not have a match. In this work, we empirically verified that discarding unmatched partials gives better results for the musical instrument sounds used than including the harmonic estimation.

C. Temporal Envelope Morphing

In this work, morphing the temporal envelope guided by the temporal centroid τ is analogous to morphing the spectral envelope because the same estimation and representation techniques can be applied [25], leading to similar morphing transformations. Also, the temporal centroid τ is the time-domain counterpart of the spectral centroid δ_1, and as such, its values behave in the same fashion under the same transformations. Analogously to Fig. 8, Fig. 9(a) shows the source and target temporal envelope curves as solid lines with nine intermediate temporal envelope curves corresponding to linearly varying α by 0.1 steps. Fig. 9(b) shows the corresponding variation of the temporal centroid. The temporal envelope morphing techniques considered are interpolation of the envelope curve (ENV) directly and interpolation of the cepstral coefficients (CC) used to represent it [25]. We discard morphing techniques that shift peaks of the envelope because this behavior is undesirable for the temporal envelope.

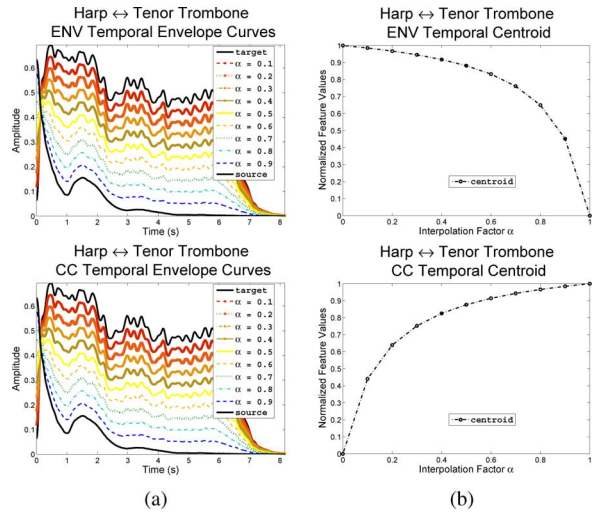

Fig. 9. Temporal envelope morphing guided by perceptually motivated features. The figure shows the temporal envelope curves on the left-hand side and the corresponding variation of the temporal centroid on the right-hand side. (a) Temporal envelope. (b) Temporal centroid.

V. Evaluation

In total, eighteen (18) musical instrument sounds from the *Vienna Symphonic Library* covering the *woodwind*, *brass*, and (plucked and bowed) *string* musical instrument families were used in the evaluation. All the sounds used have the same pitch (C4), duration (2s), and dynamics (*forte*) and present different attacks (slow, normal, and *staccato*). The bowed strings also have *vibrato*. The results presented below include a total of twenty six (26) pairs of sounds from different musical instruments. We chose not to morph between the same musical instrument with different attacks under the hybrid musical instrument constraint.

The requirement of linearity in the feature domain can be formulated as a minimum squared error and applied both in the spectral and temporal feature domains. Therefore, the variation of spectral shape features and the temporal centroid are evaluated. The interpolation of the frequency of the partials, on the other hand, cannot be formulated or evaluated similarly. The representation that renders the minimum quadratic error is the most linear in the feature domain under consideration, and selected as the most appropriate for the morphing scheme.

A. Objective Evaluation

The requirement of linearity of features led to a simple objective error measure to investigate which spectral envelope representation renders the smallest error when interpolated for all pairs of sounds morphed. Fig. 10 illustrates the error calculation as the deviations ε between the calculated feature values "o" and the interpolated feature values "x" for each normalized spectral shape feature $\delta_i(\alpha)$. The interpolated feature values "x" are obtained with a straight line connecting the calculated feature values for the source $\hat{\delta}(\alpha = 0)$ and target $\hat{\delta}(\alpha = 1)$. The features are all normalized between 0 and 1, so in practice

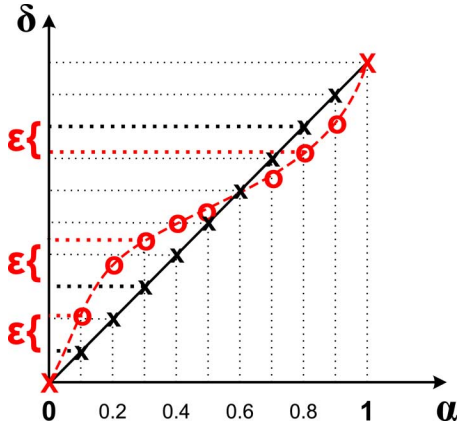

Fig. 10. Error calculation. The figure depicts the calculation of the feature interpolation error. The interpolated feature values α_m obtained as a linear regression are represented as "x", while the calculated feature values $\hat{\delta}(\alpha_m)$ are represented as "o".

$\delta_i(\alpha) = \alpha$ holds for the interpolated features and the calculated features are represented as $\hat{\delta}(\alpha_m)$.

The error function $\epsilon(\delta_i)$ in (12) is the square root of the sum of the quadratic deviations ε_m^2 between the calculated feature values $\hat{\delta}(\alpha_m)$ and the interpolated feature values α_m for each normalized spectral shape feature δ_i, where M is the number of linear steps between $\alpha_1 = 0$ and $\alpha_M = 1$ and the subscript i represents each spectral shape feature.

$$\epsilon(\delta_i) = \sqrt{\sum_{m=1}^{M} \varepsilon_m^2} = \sqrt{\sum_{m=1}^{M} \left(\hat{\delta}(\alpha_m) - \alpha_m\right)^2} \qquad (12)$$

For each pair of sounds, the error $\epsilon(\delta_i)$ is evaluated for each feature i for all spectral envelope representations considered, and then averaged across features to obtain an error estimation $\tilde{\epsilon}(\delta)$ for each spectral envelope morphing method for a given pair of sounds. Finally, a global error value ϵ_T for each method is obtained as the average across all pairs of sounds.

B. Linearity of Spectral Envelope Morphing

Fig. 11 shows the error between the interpolated values (α_m) and the calculated values $(\hat{\delta}(\alpha_m))$ of the spectral shape features for each spectral envelope representation. Part a) shows the error for each feature individually for one pair of sounds, and part b) shows the average error across all twenty-six (26) pairs of sounds used. Part a) shows $\epsilon(\delta_i)$ for each feature, and $\tilde{\epsilon}(\delta)$ on the right-hand side (marked "Total"), representing the average performance of each method for each pair of sounds. Notice that, in practice, $\epsilon(\delta_i)$ is averaged over N frames, so Fig. 11 also shows the 95% confidence interval across frames.

Part b) shows the total error ϵ_T for all twenty-six (26) pairs of sounds tested. The lowest error bar in this plot gives the most linear spectral envelope morphing method in the spectral shape feature domain for the musical instrument sounds tested. Fig. 11 reveals that interpolation of LSFs presented the smallest quadratic deviation when morphing the spectral envelope.

C. Linearity of Temporal Envelope Morphing

The evaluation of the linearity of temporal envelope morphing representations uses the error function $\epsilon(\delta_i)$ analogously

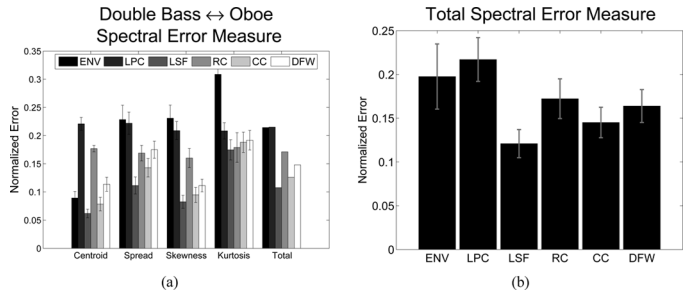

Fig. 11. Error analysis for spectral envelope morphing. The figure shows the error between the interpolated values (α_m) and the calculated values $(\hat{\delta}(\alpha_m))$ for the spectral shape features. Part a) shows the error for each feature individually for one pair of sounds, and part b) shows the average error across all twenty-six (26) pairs of sounds used and the lowest error bar in this plot gives the most linear spectral envelope morphing method in the spectral shape feature domain. (a) Single error. (b) Total error.

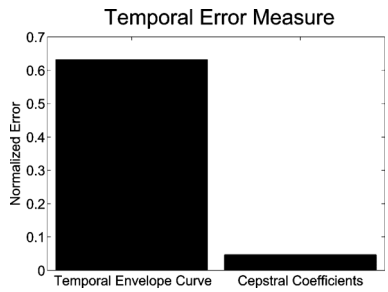

Fig. 12. Error analysis for temporal envelope morphing. The figure shows the error between the interpolated values (α_m) and the calculated values $(\hat{\delta}(\alpha_m))$ for the temporal centroid for both temporal envelope morphing methods.

to the spectral counterpart. The feature used is the temporal centroid τ, and the temporal envelope morphing methods compared are interpolation of curves (ENV) and cepstral coefficients (CC) for the same twenty-six (26) pairs of musical instruments. Fig. 12 shows the comparison of the error values, indicating that the interpolation of the cepstral coefficient representation (CC) of the temporal envelope leads to the smallest error. Interestingly, simple visual inspection of Fig. 9(a) is not enough to choose between ENV or CC, showing the need to adopt the smallest quadratic deviation criterion.

D. Perceptual Comparison of Linearity in Sound Morphing

Finally, we performed a listening test to compare the linearity of morphing transformations using musical instrument sounds between the SF model developed and traditional sinusoidal analysis. The SF model used LSFs to morph the spectral envelopes, while the sinusoidal morphing used the standard interpolation of frequency and amplitude values. The temporal alignment step is the same for both methods, only the spectral morphing procedure changes. The listening test presented 11 pairs of cyclostationary morphs (with 9 intermediate versions each) and asked the participants which was "smoother." The test is available online (http://recherche.ircam.fr/anasyn/caetano/survey/smoothness.html). Participants could either choose a method, or have no preference. In total, the results of 58 participants aged between 22 and 53 were used.

The listening test revealed that, in general, linearity depends on the pair of sounds used. There is no clearly predominant morphing technique. For some pairs, many participants manifested

no preference. The task was considered very difficult because it required participants to compare the intervals between the nine steps of the morph, judging several characteristics of sounds at the same time and remembering them for comparison across steps. The big cognitive load of the task compromised the evaluations in some cases. The number of steps used was considered inappropriate because the task relies on memory to perform the comparison.

VI. Conclusion and Future Perspectives

In this work, we describe techniques to automatically morph musical instrument sounds across timbre dimensions guided by perceptually motivated spectral and temporal features that capture the most salient dimensions of timbre perception. The temporal features are *log attack time* and *temporal centroid*, and the spectral shape features (a measure of the balance of spectral energy) are *spectral centroid*, *spectral spread*, *spectral skewness*, and *spectral kurtosis*. The concept of morphing by feature interpolation adopted in this work considers that sounds with intermediate values of features are perceptually intermediate when the features capture perceptually relevant information. Therefore, the values of the features are considered an objective measure of the perceptual impact of the morphing transformation and the objective evaluation we adopted requires that the features vary in a straight line when the morphing factor α used to control the transformation changes linearly.

We describe the temporal and spectral steps of our morphing algorithm, along with the models used, which include temporal segmentation and alignment of perceptually salient regions, temporal envelope estimation, and source-filter modeling. We investigate several spectral envelope morphing techniques previously proposed in the literature, namely the envelope curve (ENV), cepstral coefficients (CC), dynamic frequency warping (DFW), linear prediction coefficients (LPC), reflection coefficients (RC), and line spectral frequencies (LSF), to determine which representation renders the most linear transformation in the spectral shape feature space. We adopted a minimum quadratic deviation approach to evaluate the linearity of the transformations in the feature domain. We found that interpolation of line spectral frequencies (LSF) gives the most linear spectral envelope morphs and properly shifts the peaks of the spectral envelope in frequency. We also investigated temporal envelope morphing techniques guided by feature values analogously to the spectral envelope. For the temporal envelope, we found that cepstral coefficients (CC) give the most linear transformation without shifting the peaks of the temporal envelope, which is considered undesirable behavior.

The innovative aspect of the work described here lies in the adoption of an objective evaluation criterion (linearity in the feature domain), which resulted in an error measure that allows comparison across different morphing techniques. Future perspectives of this work include investigating the perceptual linearity of the results. This challenging task requires human evaluation in listening tests. However, the investigator would need to develop a procedure to efficiently evaluate or compare the perceptual linearity of morphing techniques. Nonlinearities play an important role in musical instruments and their sound production mechanism, such as attack transients or a *brighter* sound when played louder. Perceptual aspects of nonlinear, nonstationary, and inharmonic characteristics of musical instrument sounds certainly constitute an interesting direction to follow the work towards more gradual morphing transformations.

References

[1] T. Wishart, *On Sonic Art*, ser. On Sonic Art, S. Emmerson, Ed. New York, NY, USA: Imagineering Press, 1996.

[2] M. McNabb, "Dreamsong: The composition," *Comput. Music J.*, vol. 5, no. 4, pp. 36–53, 1981.

[3] J. Harvey, "Mortuos plango, vivos voco: A realization at ircam," *Comput. Music J.*, vol. 5, no. 4, pp. 22–24, 1981.

[4] E. Tellman, L. Haken, and B. Holloway, "Morphing between timbres with different numbers of features," *J. Audio Eng. Soc.*, vol. 43, no. 9, pp. 678–689, 1995.

[5] J. M. Grey and J. A. Moorer, "Perceptual evaluations of synthesized musical instrument tones," *J. Acoust. Soc. Amer.*, vol. 62, no. 2, pp. 454–462, 1977.

[6] K. Fitz and L. Haken, "Sinusoidal modeling and manipulation using Lemur," *Comput. Music J.*, vol. 20, no. 4, pp. 44–59, 1996.

[7] M. Ahmad, H. Hacihabiboglu, and A. Kondoz, "Morphing of transient sounds based on shift-invariant discrete wavelet transform and singular value decomposition," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, 2009, pp. 297–300.

[8] M. Slaney, M. Covell, and B. Lassiter, "Automatic audio morphing," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, 1996, pp. 1001–1004.

[9] M. Caetano and X. Rodet, "Automatic timbral morphing of musical instrument sounds by high-level descriptors," in *Proc. Int. Comput. Music Conf.*, 2010.

[10] K. Fitz, L. Haken, S. Lefvert, C. Champion, and M. O'Donnel, "Cellutes and flutter-tongued cats: Sound morphing using Loris and the reassigned bandwidth-enhanced model," *Comput. Music J.*, vol. 27, no. 3, pp. 44–65, 2003.

[11] F. Boccardi and C. Drioli, "Sound morphing with Gaussian mixture models," in *Proc. Int. Conf. Digital Audio Effects*, 2001, pp. 44–48.

[12] T. Hikichi, "Sound timbre interpolation based on physical modelling," *Acoust. Science Technol.*, vol. 22, no. 2, pp. 101–111, 2001.

[13] N. Osaka, "Timbre interpolation of sounds using a sinusoidal model," in *Proc. Int. Comput. Music Conf.*, 1995.

[14] D. Williams and T. Brookes, "Perceptually motivated audio morphing: Brightness," in *Proc. Audio Eng. Soc., 122nd Conv.*, 2007.

[15] D. Williams and T. Brookes, "Perceptually motivated audio morphing: Softness," in *Audio Eng. Soc., 126nd Conv.*, 2009.

[16] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[17] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, no. 4, pp. 49–56, 1990.

[18] W. Hatch, "High-level audio morphing strategies," M.S. thesis, Music Technol. Dept., McGill Univ., Montreal, QC, Canada, 2004.

[19] N. Osaka, "Concatenation and stretch/squeeze of musical instrumental sound using morphing," in *Proc. Int. Comput. Music Conf.*, 1995.

[20] C. J. Hope and D. J. Furlong, "Endemic problems in timbre morphing processes: Causes and cures," in *Proc. Irish Signals Syst. Conf.*, 1998.

[21] C. J. Hope and D. J. Furlong, "Time-frequency distributions for timbre morphing: The Wigner distribution versus the STFT," in *Proc. Brazilian Symp. Comput. Music*, 1997.

[22] A. Röbel, "Morphing dynamical sound models," in *Proc. IEEE Workshop Neural Netw. Signal Process.*, 1998.

[23] H. Pfitzinger, "Dfw-based spectral smoothing for concatenative speech synthesis," in *Proc. Int. Conf. Spoken Lang. Process.*, 2004, vol. 2, pp. 1397–1400.

[24] T. Ezzat, E. Meyers, J. Glass, and T. Poggio, "Morphing spectral envelopes using audio flow," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, 2005.

[25] M. Caetano and X. Rodet, "Sound morphing by feature interpolation," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, 2011, pp. 161–164.

[26] M. Caetano and X. Rodet, "Evolutionary spectral envelope morphing by spectral shape descriptors," in *Proc. Int. Comput. Music Conf.*, 2009.

[27] M. Caetano and X. Rodet, "Independent manipulation of high-level spectral envelope shape features for sound morphing by means of evolutionary computation," in *Proc. Int. Conf. Digital Audio Effects*, 2010.

[28] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.
[29] S. McAdams, S. Winsberg, S. Donnadieu, G. de Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specifities and latent subject classes," *Psychological Res.*, vol. 58, no. 3, pp. 177–192, 2005.
[30] J. M. Grey and J. W. Gordon, "Multidimensional perceptual scaling of musical timbre," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1270–1277, 1977.
[31] X. Amatriain, J. Bonada, Á. Loscos, J. L. Arcos, and V. Verfaille, "Content-based transformations," *J. New Music Res.*, vol. 32, no. 1, pp. 95–114, 2003.
[32] V. Verfaille, U. Zölzer, and D. Arfib, "Adaptive Digital Audio Effects (a-DAFx): A new class of sound transformations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1817–1831, Sep. 2006.
[33] S. Handel, "Timbre perception and auditory object identification," in *Hearing*, B. Moore, Ed. New York, NY, USA: Academic, 1995, pp. 425–461.
[34] C. L. Krumhansl, "Why is musical timbre so hard to understand?," in *Structure and perception of electroacoustic sound and music*, S. Nielzén and O. Olsson, Eds. New York, NY, USA: Excerpta Medica, 1989, pp. 43–54.
[35] J. Krimphoff, S. McAdams, and S. Winsberg, "Charactérisation du timbre des sons complexes. II: Analyses acoustiques et quantification psychophysique," *J. Physique IV*, vol. 4, no. 1, pp. C5.625–C5.628, 1994.
[36] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," *J. Acoust. Soc. Amer.*, vol. 118, no. 1, pp. 471–482, 2005.
[37] S. McAdams, G. Bruno, P. Susini, G. Peeters, and V. Rioux, "A meta-analysis of acoustic correlates of timbre dimensions (a)," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 3275–3275, 2006.
[38] J. Skowronek and M. McKinney, "Features for audio classification: Percussiveness of sounds," in *The language of electroacoustic music*, W. Verhaegh, E. Aarts, and J. Korst, Eds. Dordrecht, The Netherlands: Springer Netherlands, 2006, pp. 103–118.
[39] S. S. Stevens, J. Volkman, and E. Newman, "A scale for the measurement of the psychological magnitude of pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.
[40] M. Caetano, J. J. Burred, and X. Rodet, "Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues," in *Proc. Int. Conf. Digital Audio Effects*, 2010.
[41] J. Hajda, "A new model for segmenting the envelope of musical signals: The relative salience of steady state versus attack, revisited," in *Proc. Audio Eng. Soc. Conv. 101*, 11, 1996.
[42] M. Caetano and X. Rodet, "Improved estimation of the amplitude envelope of time-domain signals using true envelope cepstral smoothing," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, 2011, pp. 4424–4427.
[43] M. Caetano and X. Rodet, "A source-filter model for musical instrument sound transformation," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, 2012, pp. 137–140.
[44] X. Wen and M. Sandler, "Source-filter modeling in the sinusoidal domain," *J. Audio Eng. Soc.*, vol. 58, no. 10, pp. 795–808, 2010.
[45] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. Int. Conf. Digital Audio Effects*, 2005, pp. 30–35.
[46] J. A. Moorer, "The use of linear prediction of speech in computer music applications," *J. Audio Eng. Soc.*, vol. 27, no. 3, pp. 134–140, 1979.
[47] K. Paliwal, "Interpolation properties of linear prediction parametric representations," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1995, pp. 1029–1032.
[48] R. Morris and M. Clements, "Modification of formants in the line spectrum domain," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 19–21, Jan. 2002.
[49] I. V. McLoughlin, "Review: Line spectral pairs," *Signal Process.*, vol. 88, no. 3, pp. 448–467, 2008.
[50] T. Backström and C. Magi, "Properties of line spectrum pair polynomials—A review," *Signal Process.*, vol. 86, pp. 3286–3298, 2006.
[51] F. Itakura, "Line spectrum representation of linear prediction coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, pp. S35–S35, 1975.

Marcelo Caetano received the Ph.D. degree in signal processing from UPMC Paris 6 University in 2011 under the supervision of Xavier Rodet. He is presently a Marie Curie postdoctoral fellow with the Signal Processing Laboratory at FORTH. Dr. Caetano's research interests range from musical instrument sounds to music modeling, including analysis/synthesis models for sound transformation and music information retrieval.

Xavier Rodet is currently emeritus researcher at IRCAM on the Analysis/Synthesis team. He has been working on digital signal processing for speech, singing voice synthesis, and automatic speech recognition. Computer music is his other main domain of interest. Dr. Rodet has been working on understanding spectro-temporal patterns of musical sounds and on synthesis-by-rules.