Audio Morphing for Percussive Hybrid Sound Generation

Andrea Primavera[1], Francesco Piazza[1] and Joshua D. Reiss[2]

[1]*A3Lab - DII - Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy*

[2]*Centre for Digital Music, Queen Mary University of London, London E1 4NS, UK*

Correspondence should be addressed to Andrea Primavera (a.primavera@univpm.it)

ABSTRACT

The aim of audio morphing algorithms is to combine two or more sounds to create a new sound with intermediate timbre and duration. During the last two decades several efforts have been made to improve morphing algorithms in order to obtain more realistic and perceptually relevant sounds. In this paper we present an automatic audio morphing technique applied to percussive musical instruments. Based on preprocessing of the sound references in frequency domain and linear interpolation in time domain, the presented approach allows one to generate high quality hybrid sounds at a low computational cost. Several results are reported in order to show the effectiveness of the proposed approach in terms of audio quality and acoustic perception of the generated hybrid sounds, taking into consideration different percussive samples. mean opinion score and multidimensional scaling were used to compare the presented approach with existing state of the art techniques.

1. INTRODUCTION

Morphing is a general term used to describe a set of techniques widely used in image, video and audio processing. In audio processing, morphing is typically employed to accomplish two different tasks: the former is to obtain a smooth transition between two sounds while the latter is to generate a new intermediate sound which has the characteristics of the two given samples (e.g., timbre and duration). This paper focuses on the second of these two tasks. Commonly applied in cinema and videogames to design innovative and original sounds, audio morphing is also used in music production to create new artificial musical instruments. For instance, a flute tone could be morphed with a trumpet to create a half-flute and half-trumpet effect. Morphing algorithms are employed in a lot of commercial synthesizers to create and reproduce new hybrid sounds.

During the last two decades several algorithms have been presented to morph and to interpolate different audio signals. Linear interpolation has been one of the first techniques proposed to perform this task. Although computationally convenient, this approach is not widely used because it is unable to create high quality hybrid sounds when the timbre of the input signals is significantly different; both the references are individually perceivable in the final sound, creating an undesirable effect. To cope with this problem, different algorithms [1] [2] [3] [4], respectively based on the sinusoidal model [5] and the spectral envelope exploiting the MFFCs, have been presented in the literature. The sinusoidal model represents the input audio samples as a summation of partials which are interpolated to synthesize new hybrid sounds. Fundamental frequency and harmonic components are linearly interpolated in order to obtain an intermediate timbre between the reference signals. An approach based on sinusoids plus residual [6] was introduced by Serra in 1990, and used in several different audio morphing algorithms [7] [8]. However, although widely employed in many contexts, as voice and wind instruments morphing, two main problems afflict harmonic models: the former is related to the limitations of the model to reproduce sounds characterized by noisy spectrum as percussive samples, the latter is due to the fact that linear interpolation of the parameters of the harmonic model does not correspond to a linear variation of the perception of sound.

To cope with these problems an automatic audio morphing technique applied to percussive musical instruments is proposed here. There are two key points of the presented approach: preprocessing of the audio references performed in the frequency domain and time domain

linear interpolation to execute the morphing. Starting from the consideration that only percussive sounds are used, in the preprocessing phase the release portions, obtained exploiting the Amplitude Centroid Trajectory (ACT) model, are time stretched in order to align them temporally.

Several tests have been carried out to evaluate the effectiveness of the proposed approach, comparing it with the previous state of the art through subjective and objective comparisons. Real recordings of drum kits have been employed using the BFD2 sound library [9]. Two different algorithms, linear interpolation (LI) and sinusoidal plus residual (SMS) have been compared with the proposed approach, linear interpolation with automatic preprocessing (LIP). Objective measures such as spectral centroid evolution and release time have been used in order to analyze the relationship of temporal and spectral features to the interpolation factor. Moreover, two different listening tests, based on Mean Opinion Score (MOS) and MultiDimensional Scaling (MDS), have been performed in order to evaluate the audio perception of the generated hybrid sounds as a function of the interpolation factor.

The paper is organized as follows. Section 2 describes the state of the art relative to audio morphing techniques. Then, the proposed approach is presented in Section 3. Section 4 describes the different tests performed in order to prove the effectiveness of the presented approach, also comparing it with the current state of the art. Finally, conclusion are drawn and directions for future works are discussed in Section 5.

2. BRIEF OVERVIEW OF AUDIO MORPHING TECHNIQUES

In most approaches, the morphing is achieved by performing an interpolation of the main sound behaviors obtained from analysis/synthesis techniques such as the Short Time Fourier Transform (STFT) or sinusoidal modeling. In the following subsections the most common algorithms used to perform audio morphing are briefly described.

2.1. Linear Interpolation

Linear interpolation and crossfading have been one of the first approaches used to implement the audio morphing tasks. Given two audio references, s_1, s_2, and the interpolation factor α, the generated hybrid sound y is described by the following relationship:

$$y(t) = \alpha s_1(t) + (1-\alpha)s_2(t) \tag{1}$$

In particular, crossfading is employed to perform a gradual transformation from one sound into another one (time varying value of α are required to perform this operation) while linear interpolation is used to create new hybrid sounds with intermediate characteristics. Although computationally convenient, these approaches are not widely used because they are unable to create high quality hybrid sounds when the timbre of the input signals is significantly different. Both the reference signals could be individually perceivable in the final sound.

2.2. Sinusoidal Plus Residual

The most common morphing algorithms are based on the sinusoidal model. In this analysis/synthesis technique the time varying spectral behaviors are modeled as sums of sinusoids called partials. On the basis of the foregoing consideration, the signal $s(t)$ can be expressed as:

$$s(t) = \sum_{n=1}^{N} A_n(t)\cos[\theta_n(t)] \tag{2}$$

where A_n and θ_n are respectively the amplitude and the phase of the nth sinusoid, while N is the number of partials taken into account. A more general representation is obtained using the sinusoidal plus residual model for which $s(t)$ is equal to:

$$s(t) = \sum_{n=1}^{N} A_n(t)\cos[\theta_n(t)] + e(t) \tag{3}$$

with $e(t)$ the residual component. Exploiting this model to perform the morphing, three different interpolation factors are provided to independently control the generated hybrid sound: the fundamental frequency, the harmonics timbre and the residual envelope. For the sake of brevity further details are not presented in this paper but a full description of the sinusoidal plus residual model and its use to perform audio morphing can be found in literature [10].

3. PROPOSED ALGORITHM

An automatic audio morphing algorithm employed to generate new hybrid percussive sounds is presented here. The main features of the proposed approach are preprocessing of the audio references performed in the frequency domain and time domain linear interpolation to execute the morphing. An overall scheme of the presented algorithm is shown in Figure 1.

3.1. Preprocessing

The preprocessing phase is performed on both the signal references before the execution of the linear interpolation

in order to obtain hybrid sounds that change perceptually linearly when the corresponding interpolation factor varies linearly. The basic idea is to time scale the release portions of both sounds in order to align them temporally. Attack portions are not modified in order to avoid introducing distortion and unnatural artifacts in percussive sounds.

Preprocessing consists of three different phases; all of them performed in frequency domain:

- **Time Align**: Since different drum samples can be used, it is necessary to time align the samples before to execute the morphing. Cross-correlation is used to evaluate the delay and to perform the automatic temporal alignment.
- **AR Model Analysis**: Starting from the consideration that only percussive sounds are used, the Attack/Release (AR) model is employed instead of the more complex Attack/Decay/Sustain/Release (ADSR). Attack and release portions are respectively determined using the Amplitude Centroid Trajectory (ACT) [11] [12] procedure and analyzing the reverberation time T60 [13] of the audio samples. The main step of the ACT procedure is the spectral centroid evaluation; the spectral centroid represents the barycenter of the spectrum and it is defined as follows:

$$Spectral\ centroid(t) = \frac{\sum_{b=1}^{M} f_b(t) a_b(t)}{\sum_{b=1}^{M} a_b(t)} \quad (4)$$

  where, f is the frequency (in Hz) and a is the amplitude of the b up to M bands, evaluated exploiting the FFT operation.

  On the basis of the foregoing consideration, the attack ends when the spectral centroid slope changes direction (i.e., local minimum) while the release ends when the sound decays by 60dB. An example of the attack and release partitioning is depicted in Figure 2, where the normalized spectral centroid evolution as a function of time for a percussive audio sample is reported.

- **Time Stretching**: Release portions are time stretched or compressed as a function of the interpolation factor value.

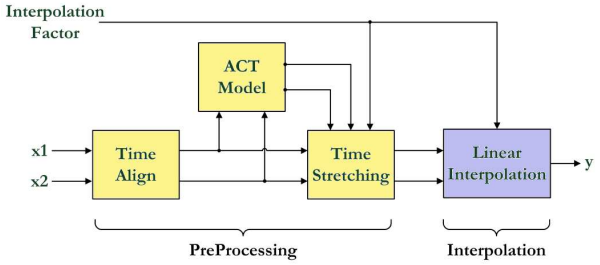


Fig. 1: Block diagram of the proposed morphing algorithm: yellow indicates the frequency domain while blue represents the time domain.

3.2. Linear Interpolation

Linear interpolation is typically not used in audio morphing algorithms, however since drum elements (e.g., tom and snare or hi-hat and cymbal) are usually characterized by similar pitch and an extremely noise spectrum, this approach is preferable to the widely used techniques based on additive synthesis.

4. RESULTS

Several tests have been carried out to evaluate the effectiveness of the proposed approach through subjective and objective comparisons. Two different algorithms, typically used to perform audio morphing, have been compared with the proposed approach:

- Linear interpolation (LI);
- Proposed approach based on Linear interpolation and preprocessing (LIP);
- Sinusoidal plus residual (SMS).

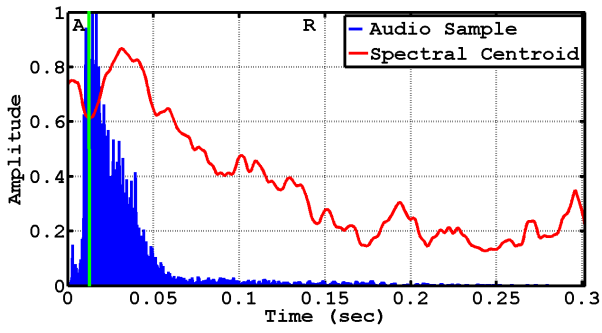


Fig. 2: Time evolution of the spectral centroid for a percussive sample.

Many tests have been done using multiple real percussive samples extracted from the BFD2 database [9], an acoustic drum production environment composed of 55GB of rare, vintage, boutique and classic drumkits. However, for the sake of brevity only three morphing example have been reported in order to show the effectiveness of the proposed approach compared with the previous state of the art:

- Morphing between different kind of snare drums (i.e, electronic and classic snare);
- Morphing between toms of different size and brand;
- Morphing between hi-hats of different brand.

Similar results have been obtained for all the other analyzed percussive sounds in the many executed tests.

4.1. Analysis of Time and Spectral Features

The release time and the spectral centroid evolution as a function of the interpolation factor are shown in Figures 2-4. By a first analysis of the Figures 3(a)-5(a), it is easy to see that for the proposed approach (LIP) release time changes linearly with the change of the interpolation factor. This does not occur for LI and SMS. The spectral centroid does not seem affected by the preprocessing phase since the trend of LI and LIP is quite similar. In conclusion, the preprocessing phase allows one to obtain a linear evolution of the release time without changing the frequency contents of the generated hybrid sound.

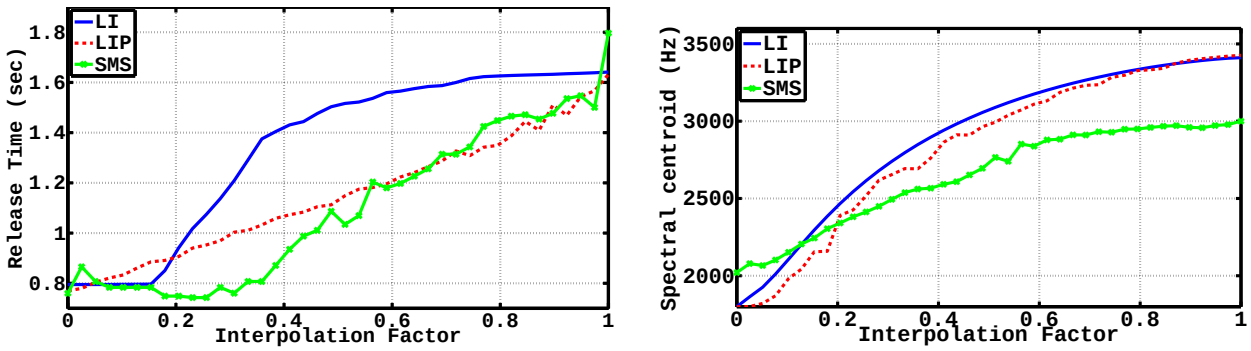
It is also interesting to observe that the spectral centroid behavior, obtained with the SMS approach, looks quite different from that obtained with LI or LIP, especially in Figure 5(b). This is due to the limitations of the SMS applied to percussive sounds; indeed, since percussive sounds are characterized by noisy spectrum, harmonic models are not able to synthesize well this kind of sound. For this reason even the values of the two extrema differ greatly from the reference signals.

4.2. Listening Test

In order to evaluate the audio perception of the produced hybrid sounds two different listening tests have been performed. The first aimed to evaluate the estimated interpolation factor and the audio quality of the generated sounds. The second, based on the MultiDimensional Scaling (MDS) analysis [14] [15], attempted to create the perceived audio space. For the sake of brevity, only the results obtained from the morphing of a classic and an electronic snare are reported. Eight hybrid sounds have been produced using various percentages of morphing (0, 0.15, 0.29, 0.43, 0.58, 0.72, 0.86, 1.00) and the aforementioned drum samples. Moreover, in order to have a more confident result, all the audio files throughout the experiment have been normalized according to the experimenter's ear. The sound tracks stored in wave file, at 16 Bit 44.1 kHz and mono format, were reproduced through a PC with professional headphone (i.e., AKG); the number of involved subjects was 20 (17 males and 3 females) with ages ranging from 21 to 35 and with some experience of critical listening and sound recording, as suggested in [16]. Before executing the main tests, listeners had to perform a familiarization exercise, in which they could listen to the full range of stimuli in a randomized order for as long as they felt necessary.

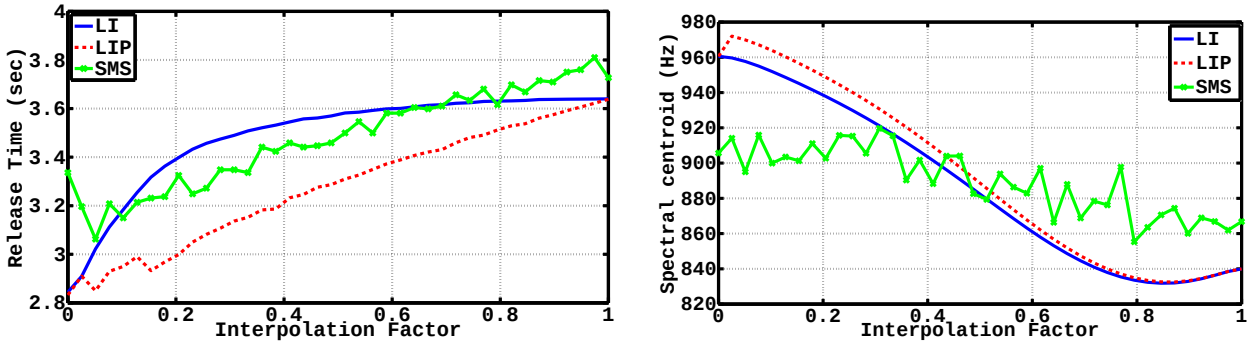
In the first listening test, subjects were asked to grade, on a scale divided into 100 discrete intervals, the estimated interpolation factor and the audio quality of the reproduced hybrid sound. Estimated interpolation factor was obtained by asking the test subject to evaluate the percentage of composition of the two audio references (i.e, classic snare and electronic snare) in the sample. Audio quality was obtained by asking the test subject to assess the perceived sound quality in terms of naturalness or unnaturalness (e.g, presence of some audio artifacts). The results obtained for the 3 tested algorithms are shown in Figure 6 and Table 1, respectively the estimated interpolation factor and the perceived audio quality. By a first analysis of Figure 6, it is easy to notice that for the proposed approach (LIP) the estimated interpolation factor changes roughly linearly with the change of the interpolation factor. This does not occur for classic linear interpolation (LI) where the factor 0.5 was perceived as near 0.8, obtaining a non linear morphing of the perceived hybrid sounds. The SMS approach also changes in a non linear way. This is due to the artifacts and distortions presented in the synthesized sounds, as confirmed by the results reported in Table 1 and Figure 6 where the error bar values of the SMS algorithm are higher than the other approaches. The hybrid sounds created with SMS are highly unnatural, confirming the limitation of additive synthesis in the reproduction of percussive sounds.

In the second listening test, a pairwise dissimilarity experiment was employed to test the perceived dimensionality of the analyzed morphing algorithms. Since the audio quality generated by the SMS approach is not sufficient to justify the use on percussive sounds, only LI



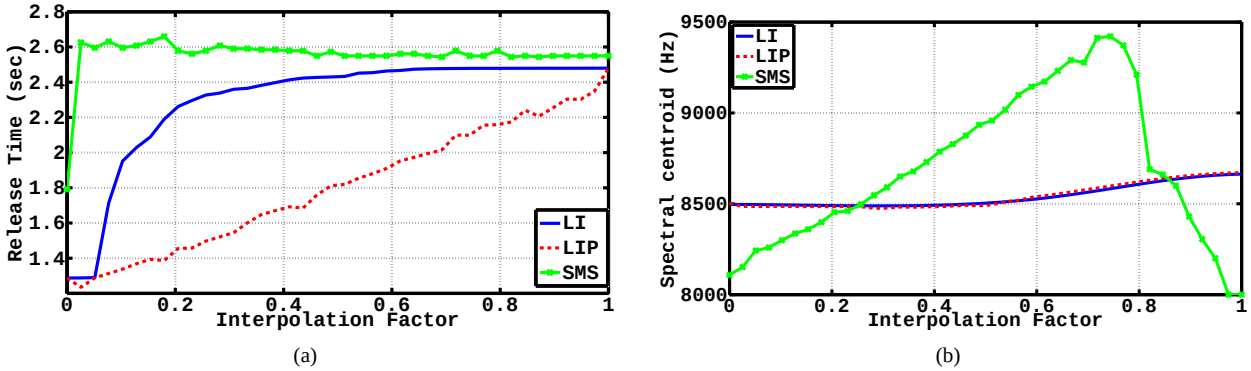
(a) (b)

Fig. 3: Release time (a) and spectral centroid evolution (b) as a function of the interpolation factor (morphing performed between a classic and electronic snare).



(a) (b)

Fig. 4: Release time (a) and spectral centroid evolution (b) as a function of the interpolation factor (morphing performed between toms of different sizes and brands).



(a) (b)

Fig. 5: Release time (a) and spectral centroid evolution (b) as a function of the interpolation factor (morphing performed between hi-hats of different brands).

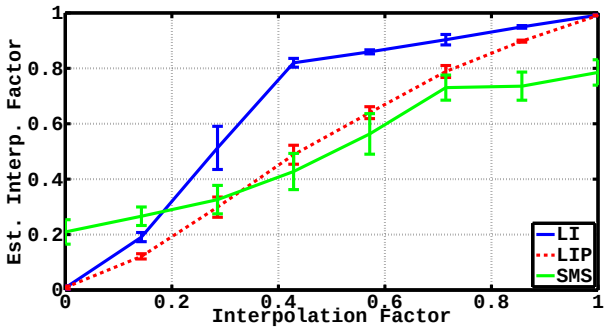


Fig. 6: Estimated interpolation factor.

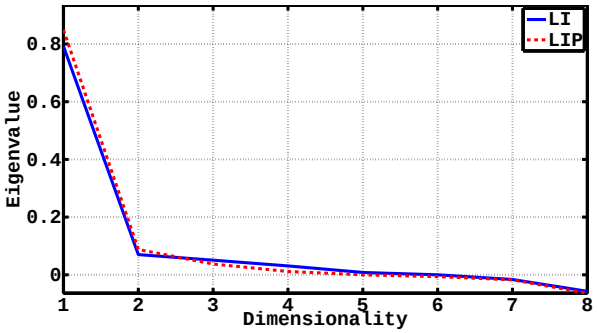


Fig. 7: Eigenvalue analysis.

and LIP have been evaluated. The same 8 hybrid sounds, used in the previous listening test, were employed for a total of 36 pairs of samples to evaluate for each algorithm. During the test, subjects were asked to grade, again on a scale divided into 100 discrete intervals, with end-points labeled as "totally dissimilar" and "the same", the similarity of the reproduced pair.

The listeners' response has been used to create a dissimilarity matrix analyzed by multidimensional scaling in order to reproduce the perceived audio space. Through the analysis of the eigenvalues returned by the MDS procedure it is possible to select the correct number of dimension used in the visualization of the information; in this case the scree plot, shown in Figure 7, indicates that two dimensions are enough to represent the variables. The corresponding two dimensional perceptual spaces obtained for the LI and LIP approaches are respectively shown in Figure 8 and Figure 9.

For both spaces there is a movement of the stimuli along the first dimension and the distribution of the audio samples in LIP space appears more uniformly distributed, confirming the effectiveness of the proposed approach. There is a high correlation between the release time and the movement along the first dimension. This means that the first and more important dimension is strictly related to the release time, and release time is an important factor in the evaluation of percussive audio morphing algorithms.

In Figure 10 the evolution of the first dimension and the release time as a function of the interpolation factor are shown for LI and LIP approaches. In order to illustrate the strong relationship between release time and first dimension the curves have been shifted on the y-axis to enhance the readability. In Table 2 the values of correlation and p-value are reported. Concurrently, the second dimension was compared with the audio quality as a function of the interpolation factor obtained from the first listening test. Table 3 reports the values of correlation and the obtained p-value. Although the correlation between the second dimension and the audio quality is not zero, the values obtained are not sufficient to establish a strong relationship between them. However, this may be due to the measurement noise and, increasing the number of listeners, may increase the correlation factor.

Table 1: Perceived audio quality expressed in a scale from 0 to 1 (0 bad - 1 excellent).

	LI	**LIP**	**SMS**
Quality	0.90	0.88	0.10
STD	0.06	0.07	0.08

Table 2: Correlation between first dimensions and release time for LI and LIP approaches.

	LI	**LIP**
CORR	0.986	0.980
P-VAL	$7.2*10^{-6}$	$1.9*10^{-5}$

Table 3: Correlation between second dimensions and audio quality for LI and LIP approaches.

	LI	**LIP**
CORR	0.66	0.77
P-VAL	0.07	0.02

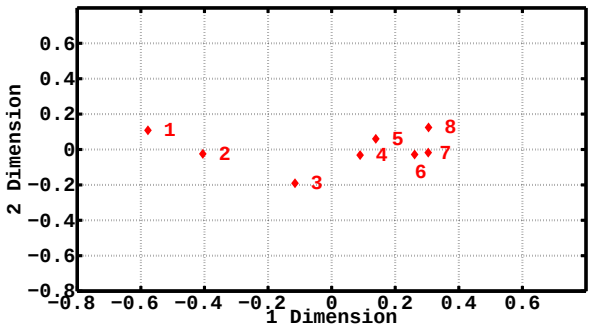


Fig. 8: Multidimensional scaling analysis obtained using linear interpolation.

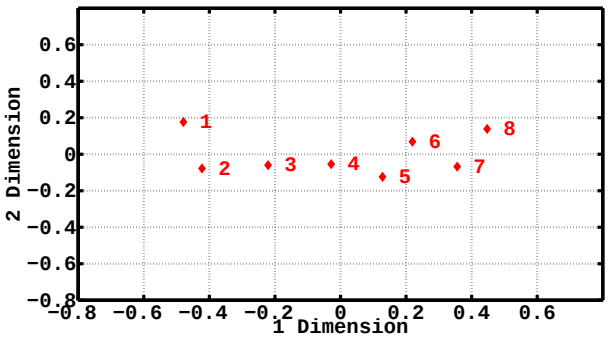


Fig. 9: Multidimensional scaling analysis obtained using linear interpolation with preprocessing.

5. CONCLUSION

An automatic audio morphing technique applied to percussive musical instruments has been presented in this work. Based on preprocessing of the sound references in frequency domain and linear interpolation in time domain, the presented approach allows one to generate high quality hybrid sounds at a low computational cost. The preprocessing phase aligns the release portions, exploiting the ACT model analysis and time stretching, while the linear interpolation guarantees better audio quality with percussive sounds than the widely used additive synthesis.

Different tests have been carried out according to objective and subjective measures comparing the proposed approach with other algorithms presented in the literature (i.e., classic linear interpolation and additive synthesis plus residual noise). Spectral centroid and release time have been evaluated in order to analyze the relationship between time and spectral features with the interpolation factor. This analysis confirms that release time changes linearly with the change of the interpolation factor. The same result have been obtained from the performed listening tests confirming a linear evolution in the perception of the hybrid sounds using the presented algorithm. Moreover, multidimensional scaling suggests that release time is the most important feature in the perception of hybrid percussive sounds.

Future work will be oriented towards a further investigation of the proposed approach analyzing the obtained performance for more dissimilar sounds. Moreover, a real time implementation of the approach considering an embedded platform and the refinement of the preprocessing phase, extending it with the ADSR model, will be taken into account.

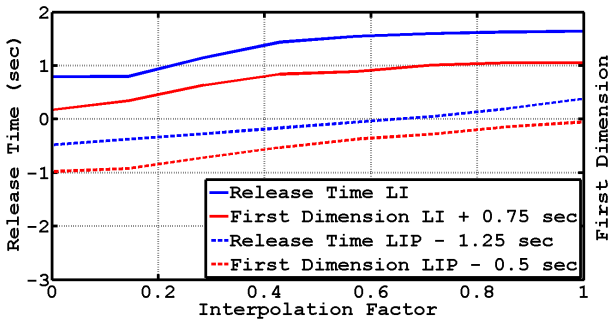


Fig. 10: Release time and first dimension as a function of the interpolation factor for LI and LIP.

6. REFERENCES

[1] M. H. Serra, D Rubine, and R. Dannenberg, "Analysis and synthesis of tones by spectral interpolation," *J. Audio Eng. Soc.*, vol. 38, no. 3, pp. 111–128, March. 1990.

[2] E. Tellman, L Haken, and B. Holloway, "Timbre morphing of sounds with unequal numbers of features," *J. Audio Eng. Soc.*, vol. 43, no. 9, pp. 678–689, September. 1995.

[3] N Osaka, "Timbre interpolation of sounds using a sinusoidal model," in *Proc. Int. Computer Music Conference*, Banff, Canada, Sep 1995, vol. 34.

[4] M. Slaney, M. Covell, and B. Lassiter, "Automatic Audio Morphing," in *Proc. IEEE International*

Conference on Acoustics, Speech and Signal Processing, Atlanta, May. 1996.

[5] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, August 1986, vol. 34, pp. 744–823.

[6] X. Serra and J. Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *J. Computer Music*, vol. 14, no. 4, pp. 12–24, 1990.

[7] F. Boccardi and C. Drioli, "Sound morphing with Gaussian mixture models," in *Proc. International Conference on Digital Audio Effects)*, Limerick, Ireland, December 2001.

[8] P. Cano, A. Loscos, J. Bonada, M de Boer, and X. Serra, "Voice Morphing System for Impersonating in Karaoke Applications," in *Proc. of the ICMC2000*, Berlin, Germany, 2000.

[9] "BFD2 version 2.0.1 Acoustic Drum Production Environment, fxpansion," .

[10] Udo Zoelzer, *DAFX Digital Audio Effects*, J. Wiley & Sons, 2010.

[11] John Hajda, "A New Model for Segmenting the Envelope of Musical Signals: The Relative Salience of Steady State versus Attack, Revisited," in *Proc. 101th Audio Engineering Society Convention (AES'96)*, Los Angeles, USA, Nov. 1996.

[12] J.J. Burred, X. Rodet, and M. Caetano, "Automatic Segmentation of the Temporal Evolution of Isolated Acoustic Musical Instrument Sounds Using Spectro-Temporal Cues," in *Proc. International Conference on Digital Audio Effects (DAFX'10)*, Graz, AU, Sep. 2010.

[13] Sabine, W.C., "Collected Papers on Acoustics," 1964, reprinted by Dover, New York.

[14] D. Williams and T. Brookes, "Perceptually motivated audio morphing: warmth," in *Proc. of 128th Audio Engineering Society Convention (AES'10)*, London, UK, May. 2010.

[15] A. Zacharakis and J. Reiss, "An additive synthesis technique for independent modification of the auditory perceptions of brightness and warmth," in *Proc. of 130th Audio Engineering Society Convention (AES'11)*, London, UK, May. 2011.

[16] ITU-R BS. 1534, "Method for subjective listening tests of intermediate audio quality," 2001.